Data Segmentation and High Dimensional Time Series Analysis

Estimation and Computation

DOMINIC OWENS

School of Mathematics
UNIVERSITY OF BRISTOL

A dissertation submitted to the University of Bristol in accordance with the requirements of the degree of DOCTOR OF PHILOSOPHY in the Faculty of Science.

SEPTEMBER 2023

Word count: 35338

Abstract

Time series analysis, the study of time-ordered data, is an historic and accomplished subfield of statistics. A great number of methods have been proposed for the analysis of time series data, and these have found use in many application areas. Often, however, these methods correspond to models that do not account for two properties commonly found in the data. These are non-stationarity, wherein the joint distribution of the underlying process is not constant, and high dimensionality, where the number of concurrent series is possibly larger than the sample size. This thesis proposes new methods for the statistical analysis of data with either or both of these properties. We preface the thesis with a review of relevant literature.

First, in Chapter 3, we describe a data segmentation procedure for high dimensional data which follow a regression model, where the parameters are piecewise-constant with respect to time. In two steps, the method first compares parameter estimates over a moving window to detect changes, then uses local refinements minimising a comparative loss for the final location estimates. We prove that the method consistently detects and locates all changes at the minimax-optimal rate (up to logarithmic factors) under Gaussianity, and is consistent under dependence and heavy tails, and when changes are multiscale. The computational cost is small relative to competing algorithms.

In Chapter 4, we describe a segmentation procedure for multiple time series which follow a piecewise-stationary vector autoregression (VAR) model. The method uses a moving sum detector to look for changes in the expectation of estimating functions. We prove this is consistent for the number and locations of changes when the dimension is fixed. A series of methodological extensions are proposed so that the algorithm may be used in less idealised settings, for example with changes which are multiscale or only detectable with a local inspection parameter.

Chapter 5 describes an extension of the model and method from Chapter 4, where we treat high-dimensional time series as observations from a dynamic factor model, allowing the latent factors to follow a piecewise-stationary VAR. By applying the proposed method to estimated principal components, we show that we consistently segment the data. We give a similar method for segmenting a piecewise stationary factor-augmented regression, and combine this with methods for forecasting under structural breaks, giving a comprehensive method for diffusion index forecasting for non-stationary data.

Finally in Chapter 6 we discuss the factor-adjusted VAR model, designed for high dimensional time series which exhibit strong serial and cross-sectional correlations, as well as sparse idiosyncratic structure. We give an overview of the model, particularly estimation procedures for the common and idiosyncratic components and of the implicit network structures, as well as forecasting methods. A software package is described, with visualisation methods and tools for the data-driven selection of tuning parameters, and we extensively study the computational properties of the method.

We end the thesis with a discussion and directions for future work.

Dedication and acknowledgements

"A lot's gonna change
in your lifetime"

Natalie Mering

While this Dedication gives me the opportunity to give thanks to everyone to whom it is owed, I owe it foremost to my supervisor, Dr. Haeran Cho. For giving me her time, her patience, and particularly for encouraging my attention to detail, I give thanks, as without these I would not be the researcher I am today. My thanks go also to my coauthor Prof. Matteo Barigozzi for collaborating with me on two papers, during which I learned so much, and to Ed, David, and Ralph from CheckRisk for getting the idea for Chapter 5 off the ground.

To everyone in the Institute for Statistical Science I am indebted for the excellent and stimulating environment, especially to Dr. Skevi Michael, Prof. Anthony Lee, and Prof. Oliver Johnson for keeping me on track, and to Dr. Matteo Fasiolo and Dr. Henry Reeve while I learned to teach. Enough thanks cannot go to my friends and colleagues involved with the Compass CDT, where I spent four valuable years of my life.

Sometimes I take for granted the number of people I may call my friend. I feel very fortunate that each of you have been here with me (you know who you are, and if you're reading this that's a strong hint), and I have much to thank you all for, not least for giving me all the moments when I could forget about the writing of this thesis.

Inevitably, I would not be where I am if not for my family - Mum, Dad, and Isobel - who have supported me the longest of all, and cultivated my nascent interest in mathematics. Thanks everyone.

Author's declaration

I declare that the work in this dissertation was carried out in accordance with the requirements of the University's Regulations and Code of Practice for Research Degree Programmes and that it has not been submitted for any other academic award. Except where indicated by specific reference in the text, the work is the candidate's own work. Work done in collaboration with, or with the assistance of, others, is indicated as such. Any views expressed in the dissertation are those of the author.

SIGNED: ..... DATE: .....

TABLE OF CONTENTS

List of Tables

List of Figures

Figure **Page**

CHAPTER 1

INTRODUCTION

In many fields of statistical practice we observe time series data with non-stationary behaviour, such that the joint distribution of the underlying process changes over time. This is a pertinent problem given that many models for time series data rely on (at least) second-order stationarity; without taking this into account, our inferences and predictions may be highly inaccurate. One modelling approach, perhaps the simplest, is to treat the data as piecewise stationary, so that the behaviour is stationary between multiple unknown change points. *Data segmentation*, or *change point analysis*, is the problem of estimating the number and locations of these changes.

In recent years, improvements in data collection and storage have made high-dimensional datasets readily available. In fields including macroeconomics, finance, and climate sciences, we have access to time series datasets in which the number of concurrent series may be large relative to number of observations. This setting poses a challenge for the many classical statistical methods designed for a small, fixed number of series, calling for the development of theory and methods which allow the number of series to diverge with the sample size.

Indeed, these two characteristics are often concurrent, as illustrated by the number of applications throughout this thesis. In four chapters of this thesis, we propose statistical methods which address one or both of these problems. The contributions of each chapter are described below.

Chapter 2: Literature review We describe a number of the regression models commonly used for time series analysis, namely linear regression, vector autoregression, and factor models. We also pose the data segmentation problem, describing methods for estimating structural changes in the canonical mean change problem and under the regression models of interest.

Chapter 3: High-dimensional data segmentation in regression settings permitting heavy tails and temporal dependence We propose a data segmentation methodology for the high-dimensional linear regression problem, where the regression parameters are allowed to undergo multiple changes. The proposed methodology, MOSEG, proceeds in two stages. The data are first scanned for multiple change points using a moving window-based procedure, which is followed by a location refinement stage. MOSEG is computationally efficient as the first stage takes place on a coarse grid, and MOSEG is theoretically consistent in estimating both the total number and the locations of the change points without requiring independence or sub-Gaussianity. In particular, it nearly matches minimax optimal rates when Gaussianity is assumed. We also propose MOSEG.MS, a multiscale extension of MOSEG which, while comparable to MOSEG in terms of computational complexity, achieves theoretical consistency for a broader parameter space that permits multiscale change points. We demonstrate good performance of the proposed methods in comparative simulation studies and for an economic dataset. The R software implementing MOSEG and MOSEG.MS is available from `https://github.com/Dom-Owens-UoB/moseg`.

A version of Chapter 3 has been submitted for publication as Cho and Owens (2022), High-dimensional data segmentation in regression settings permitting heavy tails and temporal dependence.

Chapter 4: Moving sum data segmentation for vector autoregressive time series We propose methods to segment multiple time series which follow a vector autoregressive model. We use moving sum statistics to detect and locate multiple change points, giving asymptotic guarantees for size, power, and the consistent estimation of the number and locations of change points. A series of methodological extensions are proposed: (i) By evaluating the detector over a coarse grid, we significantly reduce computational complexity while still achieving consistent estimation. (ii) Dimension reduction methods, combined with a parametric bootstrap, allow the analysis of larger panels. (iii) A recursive segmentation procedure is proposed for improved detection sensitivity. (iv) A multiple-bandwidth procedure allows for the presence of multiscale changes. The methods are validated by simulation studies and two applications to real datasets. An efficient implementation is available in the R package `mosumvar` at `github.com/Dom-Owens-UoB/mosumvar`.

Chapter 5: Segmenting and forecasting macroeconomic data with factor-augmented regression models We propose a methodology for diffusion index forecasting under non-stationarity. Firstly, we give a data segmentation methodology for high-dimensional time series which follow a factor model. The factors are assumed to follow a vector autoregressive model, the parameters of which are allowed to undergo multiple changes. We give a similar method for the segmentation of factor-augmented regression models. Based on the low-dimensional method proposed in Chapter 4, we show these consistently estimate change points, are computationally efficient, and can be used with methodological extensions such as a multiscale algorithm. We finally use the model to produce forecasts, employing weighted estimation methods which account

for the estimated change points. We test the performance of the methods in simulations and with real macroeconomic data, where ours show favourable forecasting performance when compared with rolling window estimation. An implementation in the R language is available via `https://github.com/Dom-Owens-UoB/mosumfvar`.

Chapter 6: Factor-adjusted network estimation and forecasting for high-dimensional time series A suite of methodologies is proposed for the network estimation and forecasting of high-dimensional time series under a factor-adjusted vector autoregressive model, which permits strong spatial and temporal correlations in the data. These are implemented in the package **fnets** for the R language. Additionally, we provide tools for visualising the networks underlying the time series data after adjusting for the presence of factors. The package also offers data-driven methods for selecting tuning parameters including the number of factors, vector autoregressive order and thresholds for estimating the edge sets of the networks of interest in time series analysis. We demonstrate various features of **fnets** on simulated datasets as well as real data on electricity prices and asset volatilities. An efficient implementation is available in the R package `fnets` on CRAN, and at `https://github.com/Dom-Owens-UoB/fnets`.

Content from Chapter 6 has been published in the Journal of Business and Economic Statistics as Barigozzi et al. (2023), FNETS: Factor-adjusted network estimation and forecasting for high-dimensional time series, and published in The R Journal as Owens et al. (2023), fnets: An R Package for Network Estimation and Forecasting via Factor-Adjusted VAR Modelling.

Chapter 7: Discussion We end the thesis with a summary and reflection on the contributions, highlighting directions for future work.

This thesis is structured as a series of independent works, each intended as contributions in their own right. As such, some content may be repeated or reiterated between chapters.

CHAPTER 2

LITERATURE REVIEW

We review the literature relevant to the two topics of this thesis, namely high-dimensional time series regression models (Section 2.1), then data segmentation (Section 2.2).

2.1 High-dimensional time series regression models

The models considered in this thesis are all cases of linear regression models, wherein the conditional mean of a continuous response is modelled as a sum of paired products of regressors and coefficients. These may be multivariate, such that there are possibly multiple regressors and responses, and for time series data, where the regressors indexed at t may be formed from any observations available up until t.

2.1.1 Linear regression

We observe pairs $(Y_t, \boldsymbol{X}_t)$, $t = 1, \ldots, n$, with $\boldsymbol{X}_t = (X_{1t}, \ldots, X_{pt})^\top \in \mathbb{R}^p$. These are modelled so that

$$Y_t = \boldsymbol{X}_t^\top \boldsymbol{\beta} + \varepsilon_t, \tag{2.1}$$

where $\boldsymbol{\beta} \in \mathbb{R}^p$ is a fixed parameter vector, and ε_t is additive noise such that $\mathbb{E}(\varepsilon_t) = 0, \mathsf{Var}(\varepsilon_t) = \sigma^2 \in (0, \infty)$. With fixed dimension p, this is perhaps the canonical statistical model, and is widely studied due to its simplicity, interpretability, and ubiquity in applications (Dobson and Barnett, 2018). Here, the least squares estimator is

$$\widehat{\boldsymbol{\beta}}^{OLS} = \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \sum_{t=1}^{n} (Y_t - \boldsymbol{X}_t^\top \boldsymbol{\beta})^2 = (\sum_{t=1}^{n} \boldsymbol{X}_t \boldsymbol{X}_t^\top)^{-1} (\sum_{t=1}^{n} \boldsymbol{X}_t Y_t).$$

When ε_t is Gaussian (and under some mild conditions), this is the maximum likelihood estimator. Allowing p to diverge with n, this estimator quickly becomes infeasible as the sample covariance will not be invertible.

One solution is the sparsity assumption, where only a handful of variables contribute to modelling the response. Formally, this is a restriction on $\mathfrak{s} = |\boldsymbol{\beta}|_0 = \sum_{i=1}^{p} \mathbb{I}\{\beta_i \neq 0\}$, the ℓ_0-norm of $\boldsymbol{\beta}$, i.e. the number of non-zero elements. Here we discuss estimation approaches designed for this setting, see Bühlmann and van de Geer (2011) and Tibshirani (2011) for an overview.

The Lasso estimator In their seminal work, Tibshirani (1996) introduce the Lasso estimator for $\boldsymbol{\beta}$, which solves

$$\widehat{\boldsymbol{\beta}}^{\text{Lasso}}(\lambda) = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \frac{1}{n}\sum_{t=1}^{n}(Y_t - \boldsymbol{X}_t^\top \boldsymbol{\beta})^2 + \lambda|\boldsymbol{\beta}|_1.$$

This objective penalises the estimator's ℓ_1-norm, i.e. $|\boldsymbol{\beta}|_1 = \sum_{i=1}^{p}|\beta_i|$. This introduces some bias into the estimator as a trade-off for reducing the variance. Moreover, very small estimated coefficients will be pushed towards zero, effectively performing variable selection. This can be improved by thresholding the estimator at some value. In practice, this means we can include potentially many irrelevant predictors in our regression without affecting the predictive power.

Denote the sample covariance matrix as $\widehat{\boldsymbol{\Sigma}} = n^{-1}\sum_{t=1}^{n}\boldsymbol{X}_t\boldsymbol{X}_t^\top$. For a set $S \subset \{1,\dots,p\}$ with $|S| = \mathfrak{s}$, let $\boldsymbol{\beta}_S \in \mathbb{R}^{\mathfrak{s}}$ be the subvector of $\boldsymbol{\beta}$ with components in S. Let $\mathscr{C}_S = \{\boldsymbol{\beta}\in\mathbb{R}^p : |\boldsymbol{\beta}_{S^C}|_1 \leq 3|\boldsymbol{\beta}_S|_1\}$ be a cone.

van de Geer and Bühlmann (2009) discuss conditions under which the Lasso estimator is consistent. *Compatibility* holds for S if for some $\omega > 0$ and for all $\boldsymbol{\beta}\in\mathscr{C}_S$, it holds that

$$|\boldsymbol{\beta}_S|_1^2 \leq \boldsymbol{\beta}^\top\widehat{\boldsymbol{\Sigma}}\boldsymbol{\beta}\cdot\mathfrak{s}/\omega^2.$$

This cannot be verified in general, as S is unknown. If the cardinality $\mathfrak{s}$ is known, we may instead consider the *restricted eigenvalue* (variously, *restricted strong convexity*) condition, that is, for all $\boldsymbol{\beta}\in\mathscr{C}_S$ with $|\boldsymbol{\beta}|_0 \leq \mathfrak{s}$, there exists a constant $C > 0$ such that

$$\omega|\boldsymbol{\beta}|_2^2 - C\mathfrak{s}|\boldsymbol{\beta}|_1^2 \leq \boldsymbol{\beta}^\top\widehat{\boldsymbol{\Sigma}}\boldsymbol{\beta},$$

which implies compatibility. Under either of these conditions, by Bühlmann and van de Geer (2011) Theorem 6.1, when $\lambda \geq 4\sigma\sqrt{\frac{\log p}{n}}$, with high probability we have a bound for the prediction error

$$|n^{-1}\sum_{t=1}^{n}\boldsymbol{X}_t^\top(\widehat{\boldsymbol{\beta}}^{\text{Lasso}} - \boldsymbol{\beta})|_2^2 \leq 4\lambda^2\mathfrak{s}/\phi^2,$$

and for the ℓ_1-error

$$|\widehat{\boldsymbol{\beta}}^{\text{Lasso}} - \boldsymbol{\beta}|_1 \leq 4\lambda\mathfrak{s}/\phi^2,$$

where ϕ lower bounds the eigenvalues of $\widehat{\boldsymbol{\Sigma}}$.

Due to the selection property of the Lasso estimator, we can consider the set of non-zero elements $\widehat{S} = \{\widehat{\beta}_i : \widehat{\beta}_i \neq 0\}$ as an estimator of S. The goal is typically to show that $\mathrm{P}[\widehat{S} = \widehat{S}] \to 1$ as $n \to \infty$. Meinshausen and Bühlmann (2006) show this when *neighborhood stability* (or *irrepresentability*) holds which, roughly speaking, requires submatrices of the sample covariance to be far from linearly dependent, and when the smallest non-zero coefficient is sufficiently large, i.e. $\min_{i\in S}|\beta_i|$ grows faster than $\sqrt{\mathfrak{s}\log(p)/n}$.

The Dantzig selector Candes and Tao (2007) propose the Dantzig selector, which solves the ℓ_1-constrained problem

$$\widehat{\boldsymbol{\beta}}^{\mathrm{DS}}(\lambda) = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} |\boldsymbol{\beta}|_1, \quad \text{such that} \quad |\sum_{t=1}^{n} \boldsymbol{X}_t(Y_t - \boldsymbol{X}_t^\top \boldsymbol{\beta})|_\infty \leq \lambda,$$

where $|\boldsymbol{\beta}|_\infty = \max_{i=1,\ldots,p} |\beta_i|$. This is the dual problem of the Lasso objective, and can be solved using linear programming.

Let $(\boldsymbol{X}_t)_S$ be the subvector of $\boldsymbol{X}_t$ with components in S. The *isometry constant* $c_{\mathfrak{s}}$ is the smallest such that

$$(1-c_{\mathfrak{s}})|\boldsymbol{\beta}_S|_2^2 \leq |\sum_{t=1}^{n} (\boldsymbol{X}_t)_S^\top \boldsymbol{\beta}_S|_2^2 \leq (1+c_{\mathfrak{s}})|\boldsymbol{\beta}_s|_2^2$$

for all S with $|S| \leq \mathfrak{s}$. Consider also S' with $|S'| \leq \mathfrak{s}'$. The *orthogonality constant* $d_{\mathfrak{s},\mathfrak{s}'}$ where $\mathfrak{s}+\mathfrak{s}' \leq p$ is the smallest such that

$$\boldsymbol{\beta}_S^\top (\boldsymbol{X}_t)_S (\boldsymbol{X}_t)_{S'}^\top \boldsymbol{\beta}_{S'} \leq d_{\mathfrak{s},\mathfrak{s}'} |\boldsymbol{\beta}_S|_2 |\boldsymbol{\beta}_{S'}|_2.$$

Together these constants define the *restricted isometry property* $c_{2\mathfrak{s}} + d_{\mathfrak{s},2\mathfrak{s}} < 1$. When this holds, and $\lambda \geq \sqrt{2\log p}$, we have the error bound

$$|\widehat{\boldsymbol{\beta}}^{\mathrm{DS}} - \boldsymbol{\beta}|_2^2 \leq C\lambda^2(\sigma^2 + |\boldsymbol{\beta}|_2^2)$$

for some constant C with high probability.

ℓ_2 regularisation Similarly, the ridge estimator of Hoerl and Kennard (1970) solves

$$\widehat{\boldsymbol{\beta}}^{\mathrm{Ridge}}(\lambda) = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \frac{1}{n}\sum_{t=1}^{n} (Y_t - \boldsymbol{X}_t^\top \boldsymbol{\beta})^2 + \lambda|\boldsymbol{\beta}|_2^2.$$

This penalises the ℓ_2-norm, the square root of the sum of the squared values of all β_i, i.e. $|\boldsymbol{\beta}|_2 = \sqrt{\sum_{i=1}^p |\beta_i|^2}$. The estimator achieves variance reduction, even when $\boldsymbol{X}_t$ has non-isotropic covariance, but does not perform variable selection.

The elastic net (Zou and Hastie, 2005) solves

$$\widehat{\boldsymbol{\beta}}^{\mathrm{ElasticNet}}(\lambda,\alpha) = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \frac{1}{2n}\sum_{t=1}^{n} (Y_t - \boldsymbol{X}_t^\top \boldsymbol{\beta})^2 + \lambda\left[\alpha|\boldsymbol{\beta}|_1 + \frac{1}{2}(1-\alpha)|\boldsymbol{\beta}|_2^2\right],$$

where $\alpha \in [0,1]$ controls the trade-off between the ℓ_1- and ℓ_2-penalties. When strong correlations are present between two components of $\boldsymbol{X}_t$, the Lasso will perform poorly at selecting the true non-zero variable, motivating the ℓ_2 penalty. In particular, this avoids some of the computational difficulties which come with solving the Lasso objective. This estimator is termed "naive" as it may possess a large bias. As such, the correction $\frac{1-\alpha}{\alpha}\lambda\widehat{\boldsymbol{\beta}}^{\mathrm{ElasticNet}}(\lambda,\alpha)$ is proposed.

Structured penalties In applications, specific structures may occur in the parameter vector, and we can tailor the penalty to accommodate this. For example, when group structure exists among the variables so that the parameter vector may be partitioned, we may use the group penalty of Yuan and Lin (2006) to encourage some groups' parameters to be set to zero. In this setting, $\{1,\dots,p\} = \bigcup_{i=1}^{m} g_i$ is split into m groups g_i, which possibly form a partition. Letting $\boldsymbol{\beta}_{g_i}$ be the vector with components in g_i, the group Lasso solves

$$\widehat{\boldsymbol{\beta}}^{\text{GroupLasso}}(\lambda) = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \frac{1}{n}\sum_{t=1}^{n}(Y_t - \boldsymbol{X}_t^\top\boldsymbol{\beta})^2 + \sum_{i=1}^{m}\lambda_i|\boldsymbol{\beta}_{g_i}|_2.$$

Huang and Zhang (2010) show that when the groups are sufficiently controlled, this estimator outperforms the standard Lasso estimator in ℓ_2 error.

The adaptive penalty of Zou (2006) adds data-adaptive weights w_i to each coefficient, so that we solve

$$\widehat{\boldsymbol{\beta}}^{\text{AdaptiveLasso}}(\lambda) = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \frac{1}{n}\sum_{t=1}^{n}(Y_t - \boldsymbol{X}_t^\top\boldsymbol{\beta})^2 + \lambda\sum_{i=1}^{p} w_i|\beta_i|.$$

This outperforms the Lasso in basis recovery, particularly when mixtures of large and small coefficients are present.

Tibshirani et al. (2005) introduce the fused Lasso, solving

$$\widehat{\boldsymbol{\beta}}^{\text{FusedLasso}}(\lambda) = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \frac{1}{n}\sum_{t=1}^{n}(Y_t - \boldsymbol{X}_t^\top\boldsymbol{\beta})^2 + \lambda_1\sum_{i=1}^{p}|\beta_i| + \lambda_2\sum_{i=2}^{p}|\beta_i - \beta_{i-1}|. \tag{2.2}$$

Using two regularisation parameters (λ_1,λ_2), this penalises differences between consecutive pairs β_{i-1} and β_i, which is useful when the index ordering has some meaning, for example for time series data or in change point detection problems.

2.1.2 Vector autoregression

Suppose that we observe vectors $\boldsymbol{X}_t, t = 1,\dots,n$, with $\boldsymbol{X}_t = (X_{1t},\dots,X_{pt})^\top \in \mathbb{R}^p$. In the Vector Autoregressive (VAR) model, these are modelled as

$$\boldsymbol{X}_t = \sum_{l=1}^{d}\boldsymbol{A}_l\boldsymbol{X}_{t-l} + \boldsymbol{\varepsilon}_t, \tag{2.3}$$

so that the present $\boldsymbol{X}_t$ is linked to past observations up to d time steps behind, via transition matrices $\boldsymbol{A}_l, l = 1,\dots,d$ (see Figure 2.1). The process is *stationary* (equivalently, *stable*) if and only if

$$\det\left(\boldsymbol{I}_p - \sum_{l=1}^{d}\boldsymbol{A}_l z^l\right) \neq 0 \quad \text{for } |z| \leq 1. \tag{2.4}$$

We assume that $\{\boldsymbol{\varepsilon}_t\}_{t=1}^n$ is a zero-mean, white noise process such that $\mathbb{E}(\boldsymbol{\varepsilon}_t) = \boldsymbol{0}$, $\text{Cov}(\boldsymbol{\varepsilon}_t) = \boldsymbol{S}$ for some positive definite matrix $\boldsymbol{S} \in \mathbb{R}^{p\times p}$, and $\text{Cov}(\boldsymbol{\varepsilon}_t, \boldsymbol{\varepsilon}_{t'}) = \boldsymbol{O}$ for any $t \neq t'$. The best linear predictor for one step ahead is $\boldsymbol{X}_{t+1|t} = \sum_{l=1}^{d}\boldsymbol{A}_l\boldsymbol{X}_{t-l}$, and this can be iterated forwards in time for larger

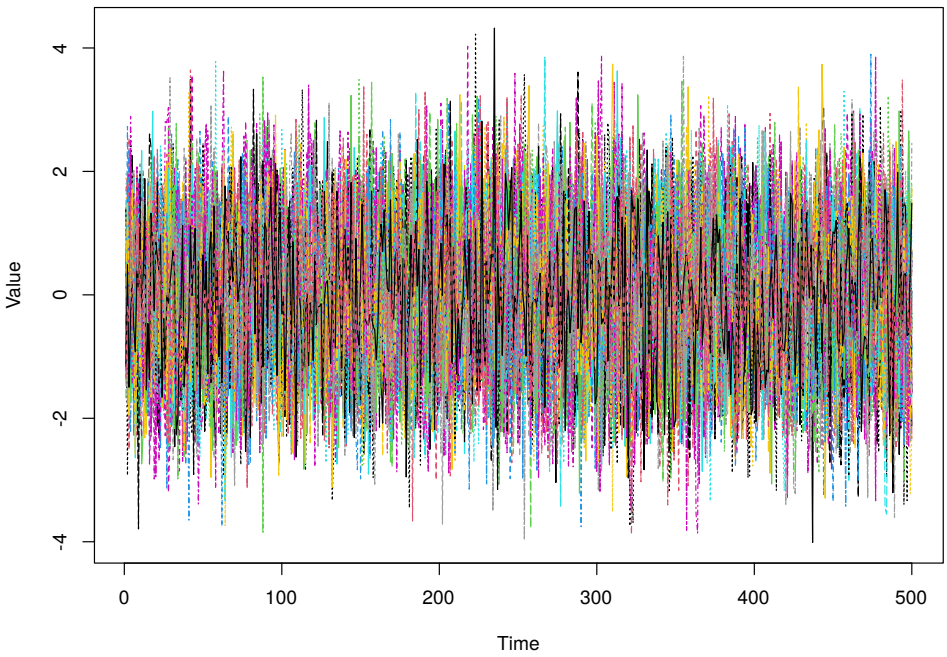


Figure 2.1: Time series data simulated from a sparse, stationary VAR process (2.3) ($n = 500, p = 50, d = 1$), with each colour representing a different series.

forecast horizons. When p is fixed and $n \to \infty$, the transition matrices can be estimated by, for example, the least squares estimator

$$\widehat{\boldsymbol{A}}^{OLS} = \arg\min_{\boldsymbol{A} \in \mathbb{R}^{p \times p}} \sum_{t=2}^{n} |\boldsymbol{X}_t - \boldsymbol{A}\boldsymbol{X}_{t-1}|_2^2 = (\sum_{t=1}^{n-1} \boldsymbol{X}_t \boldsymbol{X}_t^\top)^{-1} (\sum_{t=2}^{n} \boldsymbol{X}_t \boldsymbol{X}_{t-1}^\top).$$

Here we suppose $d = 1$; VARs with $d \geq 2$ may be rewritten as a VAR(1) process and estimated similarly. This is similar to the Yule-Walker estimator, which sums over all available samples in the right-hand side of the above equation.

VARs are popular models for time series data in many disciplines, including economics (Koop, 2013), finance (Barigozzi and Brownlees, 2019), neuroscience (Kirch et al., 2015) and systems biology (Shojaie and Michailidis, 2010). By fitting a VAR model to the data, we can infer dynamic interdependence between the variables, and forecast future values. In particular, estimating the non-zero elements of the VAR parameter matrices recovers directed edges between the components of vector time series in a Granger causal network (see Figure 2.2). By estimating the precision matrix (the inverse of the covariance matrix) of the innovations, we can define a network representing their contemporaneous dependencies via partial correlations. Finally, the inverse of the long-run covariance matrix of the data simultaneously captures lead-lag and contemporaneous co-movements of the variables.

VAR models provide a framework for network inference. Dahlhaus (2000) define graphical models for data with temporal dependence, analysing multivariate series for connections defined by partial correlations. Eichler (2007) define path diagrams for time series, encoding Granger

causal relationships. Billio et al. (2012) use factor-type assumptions to extend these definitions for high dimensional series, and apply this to connectedness in finance and insurance applications. Barigozzi and Brownlees (2019) model large series with a sparse vector autoregression where the precision matrix of the innovations is also sparse, defining a collection of networks summarising temporal and contemporaneous connections. Chen et al. (2023) allow the inferred network to vary with time.

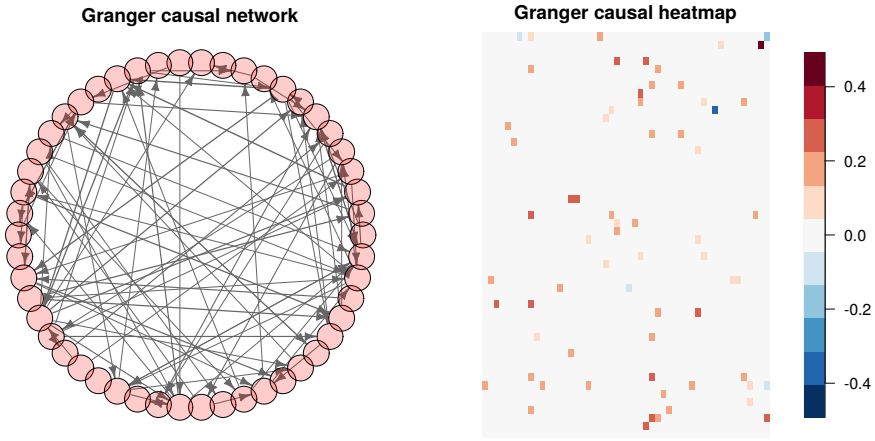


Figure 2.2: Granger causal network (left) and heatmap (right) for parameters estimated from data simulated as in Figure 2.1

Structured estimation Fitting VAR models to the data quickly becomes a high-dimensional problem, as the number of parameters grows quadratically with the dimensionality of the data. There exists a mature literature on regularisation methods for estimating VAR models in high dimensions under suitable structural assumptions on the VAR parameters. Basu and Michailidis (2015) give finite-sample bounds for ℓ_1-penalised transition matrix estimation under sparsity. Similarly, Kock and Callot (2015) give bounds on prediction errors and analyse the adaptive Lasso. Han et al. (2015) adapt the Dantzig selector to the sparse VAR setting. Nicholson et al. (2020) study a variety of penalties, with a particular focus on the case where $d \geq 2$. Basu et al. (2019) allow each transition matrix to be the sum of a low rank matrix and a sparse matrix, which are estimated using a combination of the ℓ_1 penalty and a penalty on the nuclear norm, i.e. the sum of the eigenvalues. The ridge estimator is employed by e.g. Ballarin (2021) and De Mol et al. (2008) in a Bayesian setting. See Kock et al. (2020) for an overview of general penalties.

Consistency of the methods we have discussed is usually derived under the assumption that the spectral density matrix of the data has bounded eigenvalues. This can be overcome by assuming that much of the variance can be explained by the factor-type models studied in Section 2.1.3, and the remaining structure is sparse. See for example Fan et al. (2020, 2021); Krampe and Margaritella (2021).

2.1.3 Factor models

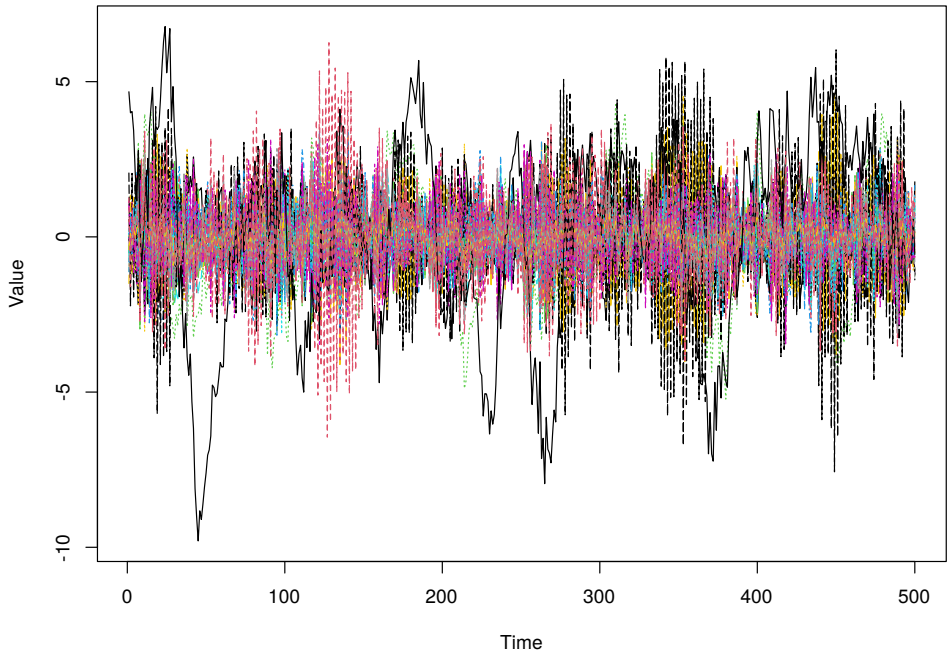


Figure 2.3: Time series data simulated from a GDFM process (6.2) ($n = 500, p = 50, q = 2$)

We give an overview of the factor modelling of high-dimensional time series, where strong cross-sectional or serial correlations are explained by linearly loading onto finite-dimensional vectors of factors. These strong dependencies are often observed in financial and macroeconomic data, and so factor models have found applications including capital asset pricing (Ross, 1976), financial risk management (Campbell et al., 1998), and nowcasting (Giannone et al., 2008). Barhoumi et al. (2014); Hallin (2022); Hallin et al. (2019); Lippi et al. (2023); Stock and Watson (2011) review the literature.

We observe $\boldsymbol{X}_t, t = 1, \ldots, n$, where

$$\boldsymbol{X}_t = \boldsymbol{\chi}_t + \boldsymbol{\varepsilon}_t.$$

The processes $\boldsymbol{\chi}_t$ and $\boldsymbol{\varepsilon}_t$ are the latent *common* and *idiosyncratic* components respectively. In general, $\boldsymbol{\varepsilon}_t$ is treated as a white noise process which is independent of all $\boldsymbol{\chi}_t$.

Generalised dynamic factor model Among many time series factor models, the Generalised Dynamic Factor Model (GDFM, Forni et al. (2000)) provides the most general approach where the p-variate factor-driven component $\boldsymbol{\chi}_t$ admits the representation

$$\boldsymbol{\chi}_t = \mathscr{B}(L)\boldsymbol{u}_t = \sum_{\ell=0}^{\infty} \boldsymbol{B}_\ell \boldsymbol{u}_{t-\ell} \quad \text{with} \quad \boldsymbol{u}_t = (u_{1t}, \ldots, u_{qt})^\top \text{ and } \boldsymbol{B}_\ell \in \mathbb{R}^{p \times q} \tag{2.5}$$

for some fixed q, where L stands for the lag operator. The q-variate random vector $\boldsymbol{u}_t$ contains common factors which are loaded across the variables and time by the filter $\mathscr{B}(L) = \sum_{\ell=0}^{\infty} \boldsymbol{B}_\ell L^\ell$,

and it is assumed that u_{jt} are i.i.d. with $\mathbb{E}(u_{jt}) = 0$ and $\mathrm{Var}(u_{jt}) = 1$ (see Figure 2.3). Forni et al. (2000, 2004, 2005) propose to estimate this in the frequency domain, applying Principal Components Analysis (PCA) to the Fourier transform of the autocovariances of $\boldsymbol{X}_t$. Forni et al. (2015) show that the model (6.2) admits a low-rank VAR representation with $\boldsymbol{u}_t$ as the innovations under mild conditions, and Forni et al. (2017) propose the estimators of $\boldsymbol{B}_\ell$ and $\boldsymbol{u}_t$ based on this representation, using an estimator for the autocovariance of $\boldsymbol{\chi}_t$ derived from frequency-domain estimates of the spectral density.

Static factor model The GDFM (6.2) reduces to a static factor model (Bai, 2003; Fan et al., 2013; Stock and Watson, 2002a) when $\mathscr{B}(L)$ admits a decomposition $\mathscr{B}(L) = \mathscr{M}^{(1)}(L)\mathscr{M}^{(2)}(L)$ with $\mathscr{M}^{(k)}(L) = \sum_{\ell=0}^{m_k} \boldsymbol{M}_\ell^{(k)} L^\ell$ for $k = 1, 2$, where $\boldsymbol{M}^{(1)} \in \mathbb{R}^{p\times q}$ and $\boldsymbol{M}^{(2)} \in \mathbb{R}^{q\times q}$. Then, we can write

$$\boldsymbol{\chi}_t = \sum_{\ell=0}^{m_1} \boldsymbol{M}_\ell^{(1)} \boldsymbol{f}_{t-\ell} = \boldsymbol{\Lambda}\boldsymbol{F}_t \quad \text{where} \quad \boldsymbol{F}_t = (\boldsymbol{f}_t^\top, \ldots, \boldsymbol{f}_{t-m_1}^\top)^\top \quad \text{and} \quad \boldsymbol{f}_t = \sum_{\ell=0}^{m_2} \boldsymbol{M}_\ell^{(2)} \boldsymbol{u}_{t-\ell}, \tag{2.6}$$

with $r = q(m_1 + 1)$ as the dimension of static factors $\boldsymbol{F}_t$. In this case, PCA with the sample covariance will provide a consistent non-parametric estimator for the loadings and factors (Bai and Ng, 2002; Stock and Watson, 1999). For forecasting, we can show that $\boldsymbol{\chi}_{n+a|n} = \boldsymbol{\Gamma}_\chi(-a)\boldsymbol{E}_\chi \mathscr{M}_\chi^{-1} \boldsymbol{E}_\chi^\top \boldsymbol{\chi}_n$, where $\mathscr{M}_\chi \in \mathbb{R}^{r\times r}$ is a diagonal matrix with the r eigenvalues of $\boldsymbol{\Gamma}_\chi(0)$ on its diagonal and $\boldsymbol{E}_\chi \in \mathbb{R}^{p\times r}$ the matrix of the corresponding eigenvectors, and the autocovariance (ACV) matrices of $\boldsymbol{\xi}_t$ are denoted by $\boldsymbol{\Gamma}_\xi(\ell) = \mathbb{E}(\boldsymbol{\xi}_{t-\ell}\boldsymbol{\xi}_t^\top)$ for $\ell \geq 0$ and $\boldsymbol{\Gamma}_\xi(\ell) = (\boldsymbol{\Gamma}_\xi(-\ell))^\top$ for $\ell < 0$.

State-space models When $\boldsymbol{M}_\ell^{(2)}, \ell = 0, \ldots, m_2$ are square-summable (for example, when m_2 is finite), we may approximate $\boldsymbol{f}_t$ in (6.3) by a VAR as per (2.3), where $\boldsymbol{u}_t$ are the innovations. In two papers, Doz et al. (2011, 2012) propose to use the Kalman filter and Quasi-Maximum Likelihood methods to estimate the factors, loadings, and autoregression parameters. The strength of this representation is its simplicity and interpretability while still accounting for dynamics, allowing the use of adapted estimation methods when there are missing or irregularly-spaced data. From this model, we can produce forecasts for the factor series as we would with a VAR for observed data, and apply the estimated loadings to forecast the panel.

Factor-augmented regression models Factor-augmented regression (FAR) models extend the linear regression model (2.1) to include latent regressors, so that Y_t is regressed onto $\boldsymbol{z}_t = \left(\boldsymbol{F}_t^\top, \boldsymbol{w}_t^\top\right)^\top \in \mathbb{R}^{p+r}$, letting $\boldsymbol{w}_t^\top \in \mathbb{R}^p$ be observed covariates. This provides a method for estimating regressions when many possibly-relevant predictors are available, particularly when the predictors are highly correlated. Stock and Watson (1999) and Bai and Ng (2009) propose to estimate the factors with PCA and the regression parameters with least squares.

2.2 Data segmentation

We give a general piecewise-stationary model for time series data. Suppose we observe time-ordered data $\{\boldsymbol{X}_t\}_{t=1}^n$ of vectors $\boldsymbol{X}_t = (X_{1t}, X_{2t}, \ldots, X_{pt})^\top \in \mathbb{R}^p$, possibly accompanied by a series of regressors $\mathbb{X}_t$. Under piecewise stationarity we draw observations from a switching process, i.e. there exist concurrent series $\{\boldsymbol{X}_t^{(j)}\}_{t\in\mathbb{Z}}$ and $\{\mathbb{X}_t^{(j)}\}_{t\in\mathbb{Z}}$, $j = 1, \ldots, q+1$ but we observe $\boldsymbol{X}_t = \boldsymbol{X}_t^{(j)}$ when $k_{j-1} + 1 \leq t \leq k_j$. Each process is distributed such that

$$\boldsymbol{X}_t^{(j)} \sim \mathbb{F}(\mathbb{X}_t^{(j)}, \boldsymbol{\theta}_j),$$

where $\boldsymbol{\theta}_j$ is a vector of regime-specific parameters and each $\boldsymbol{\theta}_{j-1} \neq \boldsymbol{\theta}_j$. Data segmentation, or change point analysis, refers to the estimation of the unknown number q, and if $q \geq 1$, the unknown locations $\{k_j\}_{j=1,\ldots,q}$.

Our focus is estimation, though it may also be of interest to perform inference on the number and locations (Frick et al., 2014; Liu et al., 2022b). We restrict ourselves to the world of parametric models, but we could generalise our model further to include non-parametric distributional changes (McGonigle and Cho, 2023; Padilla et al., 2019). Also, we place ourselves in the *offline* setting, where all data are available to the analyst retrospectively. This is in contrast to the *online* (or *sequential*) setting, where observations are gathered regularly as time continues, and there may be limits on the amount of previous data that can be stored (Xie et al., 2021).

2.2.1 The canonical problem: univariate mean changes

The canonical problem studied in the data segmentation literature is that of mean changes in a univariate series, dating back to Page (1954). Here, we observe a single ($p = 1$) series such that $X_t = \mu_j + \varepsilon_t$ for $k_{j-1} + 1 \leq t \leq k_j$, where each μ_j is a constant and $\{\varepsilon_t\}_{t\in\mathbb{Z}}$ is a mean-zero process, see Figure 2.4 for an example. When $q \geq 1$, the difficulty of the problem is defined by the *signal-to-noise ratio* $\text{SNR} = \delta^2\Delta/\sigma^2$, where $\delta = \min_{j=2,\ldots,q+1} |\mu_j - \mu_{j-1}|$ is the minimum jump size, $\Delta = \min_{j=1,\ldots,q+1}(k_j - k_{j-1})$ is the minimum segment length, and $\sigma^2 = \mathbb{E}(\varepsilon_t^2)$. Intuitively, detection and localisation is easier when jumps are large and spaced far apart, and when the noise is small. Consistency of methods usually follows by placing restrictions directly on the SNR, or implicitly via assumptions on the data generating process.

To solve this problem, many segmentation algorithms are available in the literature. These are broadly divided by Cho and Kirch (2021a) into *local* and *global* methods, and correspondingly *scan-type* and *penalisation* methods by Yu (2020).

Local methods Local methods tend to be variants which scan for changes with the weighted cumulative sum (CUSUM) detector on $t = s, \ldots, e$

$$T_{s,k,e}(X) = \sqrt{\frac{(k-s+1)(e-k)}{e-s+1}}(\bar{X}_{s,k} - \bar{X}_{k+1,e}), \text{ where } \bar{X}_{a,b} = \frac{1}{b-a+1}\sum_{t=a}^{b} X_t, \tag{2.7}$$

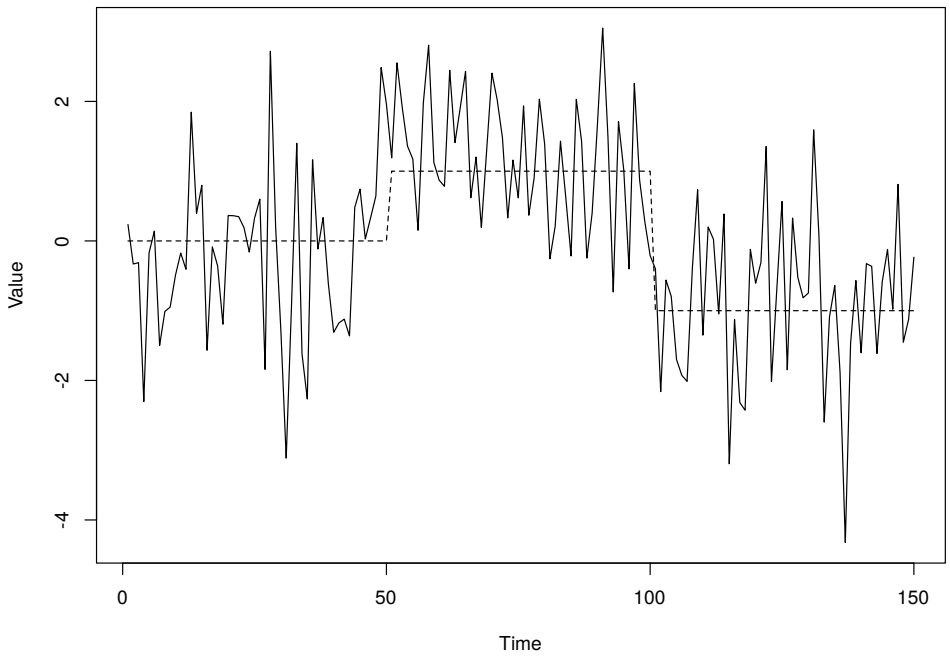


Figure 2.4: Time series data simulated from a process with a piecewise constant mean (dashed line) ($n = 150, k_1 = 50, k_2 = 100$)

where $1 \leq s < k < e \leq n$. When $q = 0$, under our conditions on $\{X_t\}_{t\in\mathbb{Z}}$ we will have $\max_{1<k<n} |T_{1,k,n}(X)| \leq D_n$ with high probability for certain thresholds D_n. Conversely, when $q \geq 1$ we will have that $\max_{1<k<n} |T_{1,k,n}(X)| > D_n$ under conditions on the changes. These facts combined give a framework for testing for the presence of changes. Indeed when $q = 1$, $\widehat{k} = \arg\max_{1<k<n} |T_{1,k,n}(X)|$ consistently estimates k_1, giving a method for identifying possibly one change. The question is then how to detect and locate multiple changes. Any algorithm will ideally meet the minimax-optimal detection rate, such that $\max_{j=1,\ldots,q} |\widehat{k}_j - k_j| = O_P(\delta^{-2})$, which requires that the SNR grows faster than $\log(n)$, as per e.g. Yu (2020).

Binary Segmentation (BS, Scott and Knott (1974); Vostrikova (1981)) was the first attempt in this direction. The algorithm begins by scanning with $s = 1$ and $e = n$. If $\max_{s<k<e} |T_{s,k,e}(X)| \leq D_n$, the algorithm terminates. Otherwise, $\widehat{k} = \arg\max_{s<k<e} |T_{s,k,e}(X)|$ is declared a change point, and the algorithm recurs with $s = 1, e = \widehat{k}$ and with $s = \widehat{k} + 1, e = n$. This continues until each branch of the algorithm terminates (see Figure 2.5). This has computational complexity $O(n \log n)$. Fryzlewicz (2014) show that when $\Delta \geq c_1 n^a, \delta \geq c_2 n^{-b}$ and $D_n \geq c_3 n^{3/4}$ for constants $c_1, c_2, c_3 > 0$ and $a \leq 1, b \geq 0$, with high probability the algorithm detects exactly q changes and localises at the rate $O_P(n^{-2}\delta^{-2}\Delta^{-2}\log n)$.

The BS algorithm is shown to be consistent when $q \geq 2$ in Venkatraman (1993), but under certain configurations the location estimators will perform poorly. To overcome this, Fryzlewicz (2014) propose to draw M random intervals (s, e) over which to scan, allowing (with high probability) the isolation of each change inside an interval in which it will be detected. This requires less

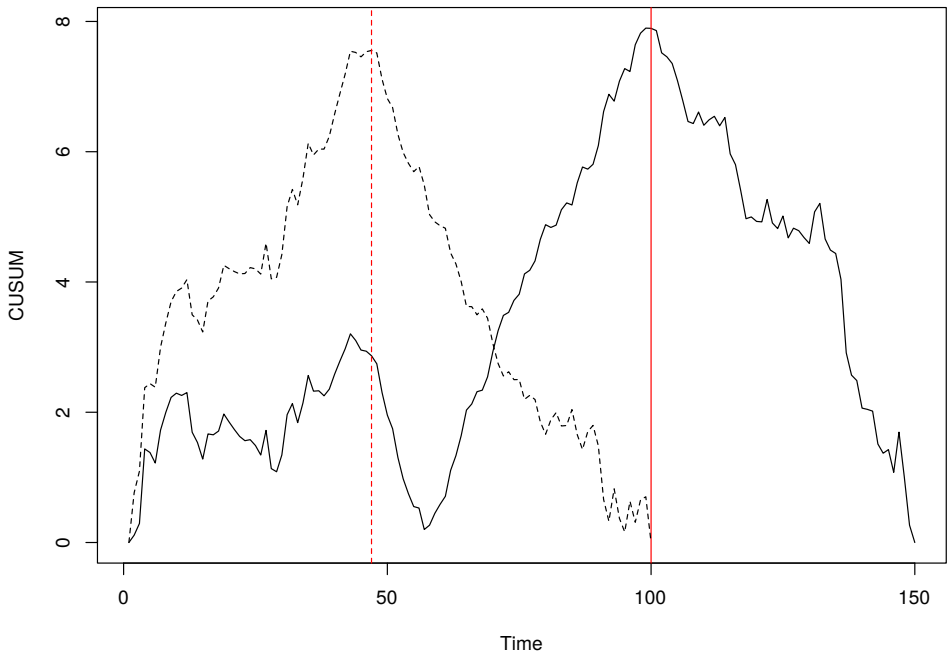


Figure 2.5: CUSUM detectors (2.7) (absolute value) resulting from Binary Segmentation on data in Figure 2.4. Estimated change points are marked in red. Solid and dashed lines indicate the first and second iteration respectively.

stringent assumptions on the spacing between changes, and attains a the minimax-optimal localisation rate, though incurs complexity $O(nM)$. Under the milder condition that $\Delta^{1/2}\delta \geq c\log^{1/2} n$ for some $c > 0$, and $D_n \geq c\log^{1/2} n$, the algorithm localises at the rate $O_P(\delta^{-2}\log n)$, which matches the minimax-optimal rate up to a logarithmic factor.

Alternatively, moving sum (MOSUM) procedures scan over all intervals with $s = k - G + 1$ and $e = k + G$, where $k = G, \ldots, n - G$ and G is a bandwidth. Eichinger and Kirch (2018) propose to locate multiple changes as values $\widehat{k}$ which locally maximise $|T_{s,k,e}(X)|$ over intervals which are large enough relative to G, and such that $|T_{s,k,e}(X)| > D_n$ (see Figure 2.6). This has complexity $O(n)$. Under mild conditions on the errors, and when G is chosen so that $2G \leq \Delta$ and $\delta \geq \log(n/G)/G$, the algorithm localises at the optimal rate.

The multiscale setting, where large frequent and small infrequent changes occur in the same series (see Figure 2.7 for an example), poses a challenge for a fixed-bandwidth procedure. The difficulty of the problem here is defined by $\min_{j=2,\ldots,q+1} |\mu_j - \mu_{j-1}| \min(k_j - k_{j-1}, k_{j-1} - k_{j-2})$, i.e. the minimum over all changes of the product of the jump size and the neighbouring segment length. Messer et al. (2014) and Messer et al. (2018) perform the MOSUM procedure with multiple bandwidths $\mathcal{G} = \{G_1, \ldots, G_H\}$, merging the resulting change points of each call in ascending order of the bandwidth. Cho and Kirch (2021b) allow for asymmetric bandwidth choices, and perform a local pruning step for coherent estimation.

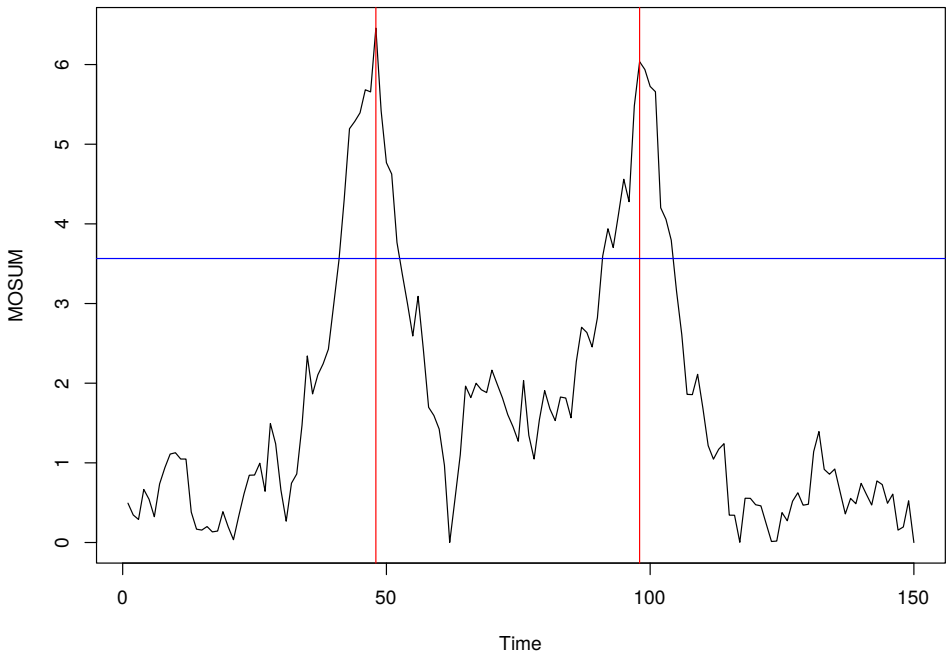


Figure 2.6: MOSUM detector $|T_{k-G+1,k,k+G}(X)|$ with $G = 20$ for data in Figure 2.4. Estimated change points are marked in red, and the threshold in blue. Produced with the MOSUM package (Meier et al., 2021).

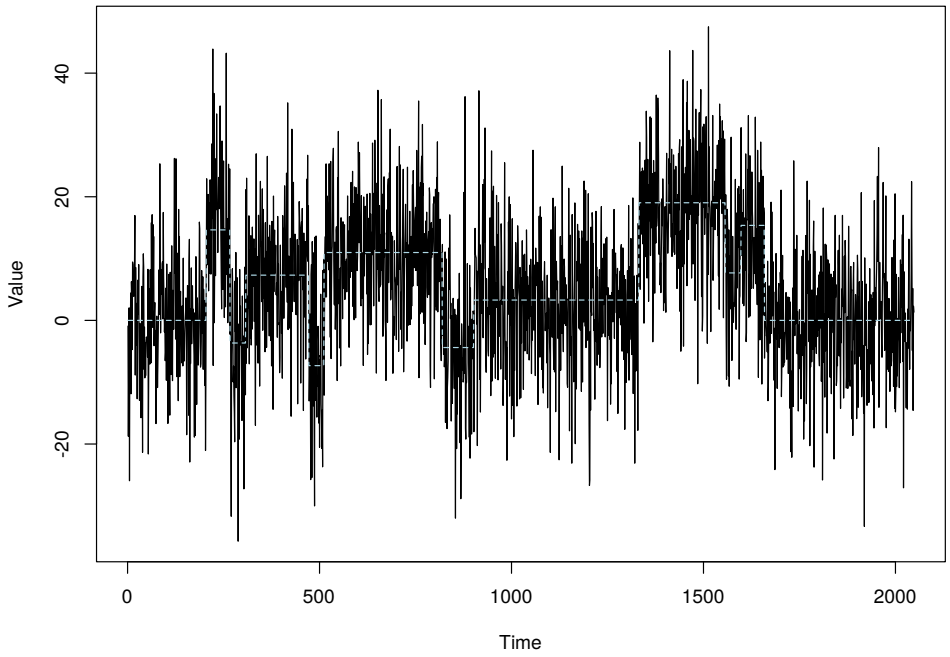


Figure 2.7: Time series data simulated from a process with a piecewise constant mean and multiscale changes (dashed line) (the "blocks" signal of Fryzlewicz (2014))

Global methods These aim to find estimators $\widehat{k}_1,\ldots,\widehat{k}_{\widehat{q}}$ which optimise an objective function over the whole data series. ℓ_0-penalisation methods are formulated as

$$\arg\min_{\substack{1<k_1\leq\cdots\leq k_q<n \\ 0\leq q\leq q_{\max}}} \left\{\text{Cost}(X_1,\ldots,X_n;k_1,\ldots,k_q)+\text{pen}(q,k_1,\ldots,k_q)\right\},$$

where $\text{Cost}(X_1,\ldots,X_n;k_1,\ldots,k_q)$ is a cost function of the data and candidate changes, $\text{pen}(q,k_1,\ldots,k_q)$ is a penalty on the model complexity, and $q_{\max}$ is an upper bound on the number of changes. For the canonical mean change problem, Yao (1988) propose to minimise the Schwarz Criterion (Schwarz, 1978), so that

$$\text{Cost}(X_1,\ldots,X_n;k_1,\ldots,k_q) = -n\log\left(\sum_{j=0}^{q}\sum_{t=k_j+1}^{k_{j+1}+1}\left(X_t-\bar{X}_{k_j+1,k_{j+1}}\right)^2\right),$$

and $\text{pen}(q,k_1,\ldots,k_q)=(2q+1)\log n$. For more general exponential family models, Killick et al. (2012) propose the cost

$$\text{Cost}(X_1,\ldots,X_n;k_1,\ldots,k_q) = -\sum_{j=0}^{q}\sup_{\theta_j}\log\ell(X_{k_j+1},\ldots,X_{k_{j+1}};\theta_j),$$

where $\log\ell(X_1,\ldots,X_n;\theta_j)$ is a parametric likelihood function, and $\text{pen}(q,k_1,\ldots,k_q)=\lambda(q+1)$ is a linear penalty with a constant $\lambda>0$. To solve this with dynamic programming, the worst case complexity is $O(q_{\max}n^2)$, though this can be improved to $O(n)$ expected time for certain problems (Killick et al., 2012; Rigaill, 2010).

To overcome these computational concerns, the problem can be relaxed to one of ℓ_1-penalisation, so that

$$\text{Cost}(X_1,\ldots,X_n;\boldsymbol{g}) = \frac{1}{n}\sum_{t=1}^{n}(X_t-g_t)^2+\lambda\sum_{t=1}^{n-1}|g_{t+1}-g_t|, \quad \boldsymbol{g}=(g_1,\ldots,g_n)^\top. \tag{2.8}$$

Change point locations are identified by jumps in the $\widehat{\boldsymbol{g}}$ minimising (2.8) such that $\widehat{g}_{\widehat{k}}\neq\widehat{g}_{\widehat{k}+1}$. This uses the fused Lasso (2.2), and Harchaoui and Lévy-Leduc (2010) propose an algorithm which solves this with complexity $O(q_{\max}^3+q_{\max}n\log n)$. This is reformulated as a group Lasso problem by Bleakley and Vert (2011), incurring a cost of $O(n\log n)$.

Separately, we mention Bayesian approaches (Adams and MacKay, 2007; Fearnhead and Liu, 2007). Generally, as well as a likelihood on the data, priors are placed on the number and locations of the change points, and inferred using numerical methods. These perform estimation and give uncertainty quantification.

2.2.2 Regression models

Moving beyond the canonical mean change problem, in this thesis we are primarily concerned with the regression models of the types discussed in Section 2.1. We give a discussion here of the literature relevant to each.

2.2.2.1 Linear regression

In the setting of Section 2.1.1, the regression parameter is allowed to vary over time, so that

$$Y_t = \begin{cases} \boldsymbol{X}_t^\top \boldsymbol{\beta}_1 + \varepsilon_t & \text{for } k_0 = 0 < t \leq k_1, \\ \boldsymbol{X}_t^\top \boldsymbol{\beta}_2 + \varepsilon_t & \text{for } k_1 < t \leq k_2, \\ \vdots & \\ \boldsymbol{X}_t^\top \boldsymbol{\beta}_{q+1} + \varepsilon_t & \text{for } k_q < t \leq n = k_{q+1}. \end{cases} \tag{2.9}$$

At each change point k_j, the vector of parameters undergoes a shift such that $\boldsymbol{\beta}_j \neq \boldsymbol{\beta}_{j+1}$ for each $j = 1, \ldots, q$.

The data segmentation problem under (6.4), when the dimension p is fixed, has been approached with dynamic programming procedures (Bai and Perron, 1998; Qu and Perron, 2007) and moving sum procedures (Kirch and Reckrühm, 2022), among others. In high-dimensional settings, when there exists at most one change point, Lee et al. (2016) formulate the Lasso problem to allow the detection of a single change, while Kaul et al. (2019b) aim to locate a single change using plug-in parameter estimates to minimise a prediction loss, giving a computationally cheap method. For the general case with unknown q, several data segmentation methods exist which adopt dynamic programming (Leonardi and Bühlmann, 2016; Rinaldo et al., 2021; Xu et al., 2022), fused Lasso (Bai and Safikhani, 2022; Wang et al., 2021b) or wild binary segmentation (Wang et al., 2021a) algorithms for the detection of multiple change points, and Bayesian approaches also exist (Datta et al., 2019). Qian et al. (2023) propose a method for fast segmentation using a fixed grid of parameters in the search space. Gao and Wang (2022) consider the case where the vector of the change itself is sparse without requiring the sparsity of $\boldsymbol{\beta}_j$. A related yet distinct problem of testing for the presence of a single change point under the regression model has been considered in Wang and Zhao (2022) and Liu et al. (2022a), and results on inference for the locations are proposed in Zhang (2023).

2.2.2.2 Vector autoregression

We observe $\{\boldsymbol{X}_t\}_{t=1}^n$ consisting of time-ordered vectors $\boldsymbol{X}_t = (X_{1t}, X_{2t}, \ldots, X_{pt})^\top \in \mathbb{R}^p$, which follow a piecewise stationary VAR model

$$\boldsymbol{X}_t = \begin{cases} \boldsymbol{X}_t^{(1)}, & k_0 + 1 = 1 \leq t \leq k_1, \\ \boldsymbol{X}_t^{(2)}, & k_1 + 1 \leq t \leq k_2, \\ \vdots & \\ \boldsymbol{X}_t^{(q+1)}, & k_q + 1 \leq t \leq k_{q+1} = n, \end{cases} \tag{2.10}$$

where each $\{\boldsymbol{X}_t^{(j)}\}_{t\in\mathbb{Z}}$ is a stationary VAR(d) process (in the sense of (2.4), or Equation (2.1.9) of Lütkepohl (2005)), i.e.

$$\boldsymbol{X}_t^{(j)} = \boldsymbol{a}_j \mathbb{X}_{t-1}^{(j)} + \boldsymbol{\varepsilon}_t, \text{ where } \boldsymbol{a}_j = \begin{bmatrix} \boldsymbol{a}_j(1)^\top \\ \vdots \\ \boldsymbol{a}_j(p)^\top \end{bmatrix} \in \mathbb{R}^{p\times(dp+1)} \text{ and } \mathbb{X}_{t-1}^{(j)} = \begin{bmatrix} 1 \\ \mathbb{X}_{1,t-1}^{(j)} \\ \vdots \\ \mathbb{X}_{p,t-1}^{(j)} \end{bmatrix} \in \mathbb{R}^{dp+1}$$

for $j = 1,\ldots,q+1$. Here, $\mathbb{X}_{i,t-1}^{(j)} = (X_{i,t-1}^{(j)},\ldots,X_{i,t-d}^{(j)})^\top$ collects the d lagged values of $X_{it}^{(j)}$, the i-th channel of $\boldsymbol{X}_t^{(j)}$, and $\boldsymbol{a}_j(i)$ collects the parameters involved in predicting the i-th channel. Then under (5.3), there are q change points at unknown locations k_j, $1 \le j \le q$, such that $\boldsymbol{a}_j \neq \boldsymbol{a}_{j+1}$ for all j, and our aim is to estimate the total number and the locations of the q change points. The innovation process $\{\boldsymbol{\varepsilon}_t\}_{t=1}^n$ is such that $\mathbb{E}(\boldsymbol{\varepsilon}_t) = \boldsymbol{0}$, $\mathrm{Cov}(\boldsymbol{\varepsilon}_t) = \boldsymbol{S}$ for some positive definite matrix $\boldsymbol{S} \in \mathbb{R}^{p\times p}$, and $\mathrm{Cov}(\boldsymbol{\varepsilon}_t, \boldsymbol{\varepsilon}_{t'}) = \boldsymbol{O}$ for any $t \neq t'$.

Detection of a single change point under a VAR model has been considered in many papers, including Dvořák and Prášková (2013) and Dvořák (2017) who use Wald-type statistics, Hlávka et al. (2017) who use empirical characteristic functions, and Kirch et al. (2015) who use score statistics. Bai (2000) and Bai and Perron (2003) use dynamic programming to estimate multiple changes in the parameters and the error covariance matrix, which is generalised by Qu and Perron (2007) to offer a range of tests. Bayesian inference (Ahelegbey et al., 2021) and group Lasso (Li et al., 2020) approaches have also been studied.

In the high-dimensional setting, the model requires structural assumptions for consistent estimation. Under sparsity, Safikhani and Shojaie (2022) and Safikhani et al. (2022) use a fused Lasso estimator, while Cho et al. (2022) use a moving-window detector based on regularised Yule-Walker estimation. Wang et al. (2019) propose dynamic programming. Under the same model, Maeng et al. (2021) address the related problem of anomaly detection. Under approximate sparsity, Bai et al. (2021) use a fused Lasso estimator for fast detection of changes in Granger-causal networks. Under low-rank assumptions, Bai et al. (2023) compare parameter estimates over moving windows. Bai et al. (2020b) detect changes under a simultaneous low rank and sparse assumption.

2.2.2.3 Factor models

We consider an observation model as in (5.3), and each $\{\boldsymbol{X}_t^{(j)}\}_{t\in\mathbb{Z}}$ is drawn from a stationary factor model of the form (6.2). There is a vast literature on instability in factor models. Broadly, these can be divided according to the part (or parts) of the model in which instability is present. These are the (i) loadings, (ii) factor number, and (iii) second-order structure. For a single break in (i) and (ii), tests and estimators are proposed by e.g. Breitung and Eickmeier (2011), Chen et al. (2014), Han and Inoue (2015), Corradi and Swanson (2014), Bai et al. (2020a), and Koo et al. (2023), and multiple changes are considered in Su and Wang (2017), Ma and Su (2018), and Liu

and Zhang (2021b). Less focus, however, has been given to changes in (iii). Barigozzi et al. (2018) consider a static factor model with changes in both the factor and idiosyncratic components. Cho et al. (2022) do the same assuming a generalised dynamic factor model, while Barigozzi and Trapani (2020) propose a sequential method. Kim et al. (2021) consider testing for a single change under a state-space model.

Considering in particular the factor-augmented regression model, Corradi and Swanson (2014) and Massacci (2019) perform inference for a single break, while Wang et al. (2015) perform estimation. Banerjee et al. (2008) allow for instability in the factor model, while Stock and Watson (2009) allow for instability in the regression relationship.

CHAPTER 3

HIGH-DIMENSIONAL DATA SEGMENTATION IN REGRESSION SETTINGS PERMITTING HEAVY TAILS AND TEMPORAL DEPENDENCE

3.1 Introduction

Regression modelling in high dimensions has received great attention with the development of data collection and storage technologies, and numerous applications are found in natural and social sciences, economics, finance and genomics, to name a few. There is a mature literature on high-dimensional linear regression modelling under the sparsity assumption, see Bühlmann and van de Geer (2011) and Tibshirani (2011) for an overview. When observations are collected over time in highly nonstationary environments, it is natural to allow for shifts in the regression parameters. Permitting the parameters to vary over time in a piecewise constant manner, data segmentation, a.k.a. multiple change point detection, provides a conceptually simple framework for handling nonstationarity in the data.

In this chapter, we consider the problem of multiple change point detection under the following model: We observe $(Y_t, \mathbf{x}_t)$, $t = 1, \ldots, n$, with $\mathbf{x}_t = (X_{1t}, \ldots, X_{pt})^\top \in \mathbb{R}^p$ where

$$Y_t = \begin{cases} \mathbf{x}_t^\top \boldsymbol{\beta}_0 + \varepsilon_t & \text{for } \theta_0 = 0 < t \le \theta_1, \\ \mathbf{x}_t^\top \boldsymbol{\beta}_1 + \varepsilon_t & \text{for } \theta_1 < t \le \theta_2, \\ \vdots & \\ \mathbf{x}_t^\top \boldsymbol{\beta}_q + \varepsilon_t & \text{for } \theta_q < t \le n = \theta_{q+1}. \end{cases} \tag{3.1}$$

Here, $\{\varepsilon_t\}_{t=1}^n$ denotes a sequence of errors satisfying $\mathbb{E}(\varepsilon_t) = 0$ and $\mathrm{Var}(\varepsilon_t) = \sigma_\varepsilon^2 \in (0, \infty)$ for all t, which may be serially correlated. At each change point θ_j, the vector of parameters undergoes a change such that $\boldsymbol{\delta}_j = \boldsymbol{\beta}_j - \boldsymbol{\beta}_{j-1}$ for all $j = 1, \ldots, q$, so that the size of the change is $\delta_j = |\boldsymbol{\beta}_j - \boldsymbol{\beta}_{j-1}|_2$. Then, our aim is to estimate the set of change points $\Theta = \{\theta_j, 1 \le j \le q\}$ by estimating both the

total number q and the locations θ_j of the change points.

The data segmentation problem under (3.1) is considered by Bai and Perron (1998), Qu and Perron (2007), Zhao et al. (2022) and Kirch and Reckrühm (2022), among others, when the dimension p is fixed. In high-dimensional settings, when there exists at most one change point ($q = 1$), Lee et al. (2016) and Kaul et al. (2019b) consider the problem of detecting and locating the change point, respectively. For the general case with unknown q, several data segmentation methods exist which adopt dynamic programming (Leonardi and Bühlmann, 2016; Rinaldo et al., 2021; Xu et al., 2022), fused Lasso (Bai and Safikhani, 2022; Wang et al., 2021b) or wild binary segmentation (Wang et al., 2021a) algorithms for the detection of multiple change points, and Bayesian approaches also exist (Datta et al., 2019). A related yet distinct problem of testing for the presence of a single change point under the regression model has been considered in Wang and Zhao (2022) and Liu et al. (2022a), and Gao and Wang (2022) consider the case where $\boldsymbol{\beta}_j - \boldsymbol{\beta}_{j-1}$ is sparse without requiring the sparsity of $\boldsymbol{\beta}_j, j = 0, \ldots, q$.

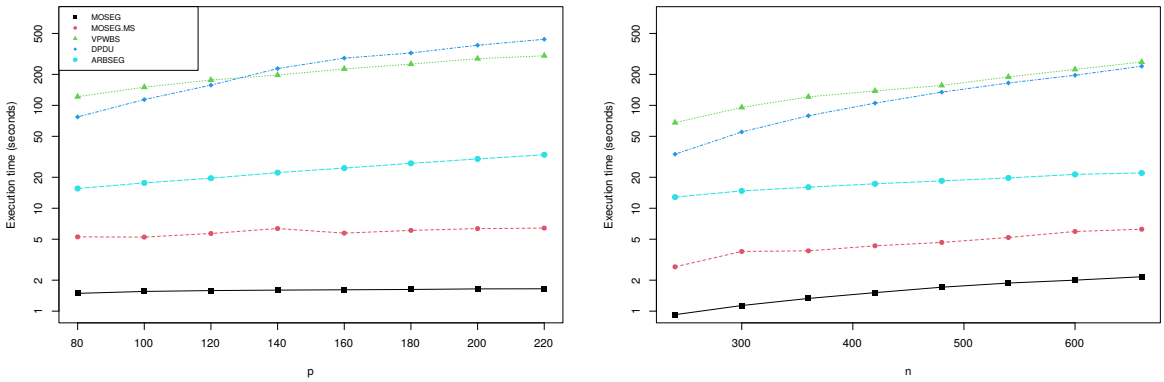


Figure 3.1: Execution time in seconds of MOSEG and MOSEG.MS and competing methodologies on simulated datasets (y-axis is in log scale for ease of comparison). Left: p varies while $n = 450$ is fixed. Right: n varies while $p = 100$ is fixed. For each setting, 100 realisations are generated and the average execution time is reported. See Section 3.4.2 for full details.

Against the above literature background, we list the contributions made in this chapter by proposing computationally and statistically efficient data segmentation methods.

(i) **Computational efficiency.** For the data segmentation problem under (3.1), often the computational bottleneck is the local estimation of the regression parameters via penalised M-estimation such as Lasso. We propose MOSEG, a moving window-based two-stage methodology, and its multiscale extension, which are both highly efficient computationally. In the first stage, MOSEG scans the data for multiple change points using a moving window of length G on a *coarse* grid of size $O(nG^{-1})$, which is followed by a simple location refinement step minimising the local residual sum of squares. The adoption of a coarse grid in the first stage contributes greatly to the reduction of Lasso estimation steps while losing little detection power. Figure 3.1 demonstrates the computational competitiveness

of the proposed MOSEG and MOSEG.MS where they greatly outperform the existing methodologies in their execution time for a range of n and p.

(ii) **Multiscale change point detection.** We propose a multiscale extension of the single-bandwidth methodology MOSEG. Referred to as MOSEG.MS, it is fully adaptive to the difficult scenarios with *multiscale* change points, where large frequent parameter shifts and small changes over long stretches of stationarity are simultaneously present, while still enjoying computational competitiveness. To the best of our knowledge, MOSEG.MS is the only data segmentation methodology under the model (3.1) for which the detection and localisation consistency is derived explicitly for the broad parameter space that permits multiscale change points. Also, while there exist several data segmentation methods that propose to apply moving window-based procedures with multiple bandwidths, MOSEG.MS is the first extension in high dimensions with a guaranteed rate of localisation.

(iii) **Theoretical consistency in general settings.** We show the consistency of MOSEG and MOSEG.MS in estimating the total number and the locations of multiple change points. Under Gaussianity, their separation and localisation rates nearly match the minimax lower bounds up to a logarithmic factor. Moreover, in our theoretical investigation, we permit temporal dependence as well as tail behaviour heavier than sub-Gaussianity. This, compared to the existing literature where independence and (sub-)Gaussianity assumptions are commonly made, shows that the proposed methods work well in situations that are more realistic for empirical applications.

The rest of the chapter is organised as follows. Section 3.2 introduces MOSEG, the single-bandwidth methodology, and establishes its theoretical consistency. Then in Section 3.3, we propose its multiscale extension, MOSEG.MS, and show that it achieves theoretical consistency in a broader parameter space. Numerical experiments in Section 3.4 demonstrate the competitiveness of the proposed methods in comparison with the existing data segmentation algorithms and Section 3.6 provides a real data application to equity premium data. In the Appendix, we present all the proofs. The R software implementing MOSEG and MOSEG.MS is available from `https://github.com/Dom-Owens-UoB/moseg`.

Notation. For a random variable X, we write $\|X\|_\nu = [\mathbb{E}(|X|^\nu)]^{1/\nu}$ for $\nu > 0$. For $\mathbf{a} = (a_1, \ldots, a_p)^\top \in \mathbb{R}^p$, we write $\mathrm{supp}(\mathbf{a}) = \{i, 1 \le i \le p : a_i \neq 0\}$, $|\mathbf{a}|_0 = \sum_{i=1}^p \mathbb{I}_{\{a_i \neq 0\}}$, $|\mathbf{a}|_1 = \sum_{i=1}^p |a_i|$, $|\mathbf{a}|_2 = (\sum_{i=1}^p a_i^2)^{1/2}$ and $|\mathbf{a}|_\infty = \max_{1 \le i \le p} |a_i|$. For a square matrix $\mathbf{A}$, let $\Lambda_{\max}(\mathbf{A})$ and $\Lambda_{\min}(\mathbf{A})$ denote its maximum and minimum eigenvalues, respectively. For a set $\mathscr{A}$, we denote its cardinality by $|\mathscr{A}|$. For sequences of positive numbers $\{a_n\}$ and $\{b_n\}$, we write $a_n \lesssim b_n$ if there exists some constant $C > 0$ such that $a_n / b_n \le C$ as $n \to \infty$. Finally, we write $a \vee b = \max(a, b)$ and $a \wedge b = \min(a, b)$.

3.1.1 Literature review and comparison with the existing methods

Table 3.1 provides an overview of the theoretical properties of MOSEG and MOSEG.MS in comparison with the methods proposed in Wang et al. (2021a), Kaul et al. (2019a) and Xu et al. (2022) for the change point problem in (3.1) under Gaussianity, as well as their computational complexity. For a given methodology, let $\widehat{\mathcal{K}}$ denote the set of estimated change points. When the magnitude of change Δ, measured by either

$$\Delta^{(1)} = \min_{1\le j\le q} \delta_j^2 \cdot \min_{0\le j\le q} (\theta_{j+1}-\theta_j) \text{ or } \Delta^{(2)} = \min_{1\le j\le q} \delta_j^2 \min(\theta_j-\theta_{j-1}, \theta_{j+1}-\theta_j), \tag{3.2}$$

diverges faster than the separation rate $s_{n,p}$ associated with the method, all q changes are detected by $\widehat{\mathcal{K}}$ with asymptotic power one and their locations are consistently estimated with the rate $\ell_{n,p}$, such that $\min_{1\le j\le q} \min_{\widehat{k}\in\widehat{\mathcal{K}}} w_j|\widehat{k}-\theta_j| = O_P(\ell_{n,p})$. Here, w_j refers to the relative difficulty in locating θ_j which is related to the jump size δ_j. Let Lasso(p) denote the cost of solving a Lasso problem with p variables.

Table 3.1: Comparison of data segmentation methods developed for the model (3.1) in their theoretical properties under Gaussianity and computational complexity (for given tuning parameters). Here, $\mathfrak{s} = \max_{0\le j\le q}|\mathcal{S}_j|$ and $\mathfrak{S} = |\cup_{j=0}^{q}\mathcal{S}_j|$, where $\mathcal{S}_j$ is the set of non-zero components of $\boldsymbol{\beta}_j$. The separation rate $s_{n,p}$ is a lower bound for (3.11), while $\ell_{n,p}$ bounds the scaled localisation error as in the text. Let Δ be the magnitude of change as in (3.2) and w_j is the relative difficulty in locating θ_j.

	Separation		Localisation		Computational complexity
	$s_{n,p}$	Δ	$\ell_{n,p}$	w_j	
MOSEG	$\mathfrak{s}\log(p\vee n)$	$\Delta^{(1)}$	$\mathfrak{s}\log(p\vee n)$	δ_j	$O(\frac{n}{rG}\cdot\text{Lasso}(p))$
MOSEG.MS	$\mathfrak{s}\log(p\vee n)$	$\Delta^{(2)}$	$\mathfrak{s}\log(p\vee n)$	δ_j	$O(\frac{n}{rG_1}\cdot\text{Lasso}(p))$
Wang et al. (2021a)	$\mathfrak{s}\log(p\vee n)$	$\Delta^{(1)}$	$\mathfrak{s}\log(n)$	δ_j	$O(n\log^2(n)\cdot\text{GroupLasso}(p))$
Kaul et al. (2019a)	$\mathfrak{S}\log(p\vee n)$	$\Delta^{(1)}$	$\mathfrak{S}\log(p)$	δ	$O(\tilde{q}\cdot\text{Lasso}(p)+\text{SA}(\tilde{q}))$
Xu et al. (2022)	$\mathfrak{s}\log(p\vee n)$	$\Delta^{(1)}$	$\mathfrak{s}\log(p\vee n)$	δ_j	$O(n^2p^2+n^2\cdot\text{Lasso}(p))$

Wang et al. (2021a) propose a method which learns the projection that is well-suited to reveal a change over each local segment and combines it with the wild binary segmentation algorithm (Fryzlewicz, 2014) for multiple change point detection. Kaul et al. (2019a) propose to minimise an ℓ_0-penalised cost function given a set of candidate estimators of size $\tilde{q}$. Their theoretical analysis implicitly assumes that $\min_j(\theta_{j+1}-\theta_j)$ scales linearly in n, and the simulated annealing adopted for minimising the penalised cost, denoted by SA($\tilde{q}$) in Table 3.1, has complexity ranging from $O(\tilde{q}^4)$ on average to being exponential in the worst case. Xu et al. (2022) investigate the dynamic programming algorithm of Rinaldo et al. (2021) for minimising an ℓ_0-penalised cost function in a more general setting. In Table 3.1, we report the separation and localisation rates derived in Xu et al. (2022) for the pre-estimators from the dynamic programming algorithm; in their proposal, the pre-estimators are further refined and their exact minimax optimality is established under a stronger condition on the size of changes, namely that $\Delta^{(1)}/(\mathfrak{s}^2\log^3(pn))\to\infty$.

We also mention Zhang et al. (2015) where the data segmentation problem is treated as a high-dimensional regression problem with a group Lasso penalty, which only provides that the estimation bias is of $o_P(n)$. Leonardi and Bühlmann (2016) consider both dynamic programming and binary segmentation algorithms are considered for change point estimation, and we refer to Rinaldo et al. (2021) for a detailed discussion on their results.

From Table 3.1, we conclude that MOSEG.MS is highly competitive both computationally and statistically. In specifying the properties of Kaul et al. (2019a), the global sparsity $\mathfrak{S} = |\cup_{j=0}^{q} \mathscr{S}_j|$ can be much greater than the segment-wise sparsity $\mathfrak{s}$, particularly when the number of change points q is large. We investigate the theoretical properties of MOSEG.MS in the broadest parameter space possible which is formulated with $\Delta^{(2)}$ instead of $\Delta^{(1)}$ as in all the other papers; recall that from the discussion following (3.18) comparing $\Delta^{(l)}$, $l = 1,2$, we always have $\Delta^{(1)} \leq \Delta^{(2)}$ and the former can be much smaller than the latter when large shifts over short intervals and small changes over long stretches of stationarity are simultaneously present in the signal.

Besides, the theoretical properties of MOSEG.MS reported in Table 3.1 do not require independence unlike other works (with the exception of Xu et al. (2022)), and extend beyond i.i.d. sub-Gaussianity. In the presence of serial dependence and sub-Weibull tails (through having $\gamma > 1$ as in Condition 1 (a)), Xu et al. (2022) require that $\Delta^{(1)} \gtrsim (\mathfrak{s}\log(np))^{4\gamma+2\gamma'-1}$ for the detection of all change points, where a smaller value of $\gamma' \in (0,\infty)$ imposes a faster decay of the serial dependence. This is comparable to the detection boundary of MOSEG, $\Delta^{(1)} \gtrsim (\mathfrak{s}\log(np))^{4\gamma+3}$ which is implied by Assumption 3.5 (a) under Condition 1 (a). We remark that Condition 1 assumes algebraically decaying serial dependence whereas γ' of Xu et al. (2022) governs the rate of exponentially decaying serial dependence. The localisation rate in Corollary 3.3 (i) is also comparable to that attained by the preliminary estimators of Xu et al. (2022) produced by a dynamic programming algorithm; as noted above, under a stronger condition on $\Delta^{(1)}$, they derive a further refined rate.

3.2 Single-bandwidth methodology

We introduce MOSEG, a single-bandwidth two-stage methodology for data segmentation in regression settings. We first describe its two stages in Section 3.2.1, establish its theoretical consistency in Section 3.2.2 and verify meta-assumptions made for the theoretical analysis in Section 3.2.3 for a class of linear processes with serial dependence and heavier tails than that permitted under sub-Gaussianity.

3.2.1 MOSEG

3.2.1.1 Stage 1: Moving window procedure on a coarse grid

Single-bandwidth moving window procedures have successfully been adopted for univariate (Eichinger and Kirch, 2018; Preuss et al., 2015a; Yau and Zhao, 2016), multivariate (Kirch and Reckrühm, 2022) and high-dimensional (Cho et al., 2022) time series segmentation. Often in

a moving window-based data segmentation procedure, the key challenge is to carefully design a detector statistic which, when adopted for scanning the data for changes, has good detection power against the type of changes which is of interest to detect.

For a given bandwidth $G \in \mathbb{N}$ satisfying $G \leq n/2$, our proposed detector statistic is

$$T_k(G) = \sqrt{\frac{G}{2}} \left| \widehat{\boldsymbol{\beta}}_{k,k+G} - \widehat{\boldsymbol{\beta}}_{k-G,k} \right|_2, \quad G \leq k \leq n-G. \tag{3.3}$$

Here, $\widehat{\boldsymbol{\beta}}_{s,e}$ denotes an estimator of the vector of parameters obtained from $(Y_t, \mathbf{x}_t)$, $s+1 \leq t \leq e$, for any $0 \leq s < e \leq n$. The statistic $T_k(G)$ contrasts the local parameter estimators from two adjacent data sections over $\{k-G+1, \ldots, k\}$ and $\{k+1, \ldots, k+G\}$. Then, $T_k(G)$ is expected to form local maxima near the change points where the local parameter estimators differ the most, and thus it is well-suited for detecting and locating the change points under the model (3.1).

We propose to obtain the local estimator $\widehat{\boldsymbol{\beta}}_{s,e}$ via Lasso, as

$$\widehat{\boldsymbol{\beta}}_{s,e}(\lambda) = \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \sum_{t=s+1}^{e} (Y_t - \mathbf{x}_t^\top \boldsymbol{\beta})^2 + \lambda \sqrt{e-s} |\boldsymbol{\beta}|_1 \tag{3.4}$$

for some tuning parameter $\lambda > 0$. In what follows, we suppress the dependence of this estimator on λ when there is no confusion. The estimand of $\widehat{\boldsymbol{\beta}}_{k-G,k}$ is

$$\boldsymbol{\beta}^*_{k-G,k} = \frac{1}{G} \sum_{j=L(k-G+1)}^{L(k)} \left\{ (\theta_{j+1} \wedge k) - ((k-G) \vee \theta_j) \right\} \boldsymbol{\beta}_j, \tag{3.5}$$

where $L(t) = \{j, 0 \leq j \leq q : \theta_j + 1 \leq t\}$ denotes the index of a change point θ_j that is the closest to t while lying strictly to its left. In short, $\boldsymbol{\beta}^*_{k-G,k}$ is a weighted sum of $\boldsymbol{\beta}_j$ with the weights corresponding to the proportion of the intervals $\{k-G+1, \ldots, k\}$ overlapping with $\{\theta_j+1, \ldots, \theta_{j+1}\}$.

Scanning the detector statistic $T_k(G)$ over all $k \in \{G, \ldots, n-G\}$ requires the computation of the Lasso estimator $O(n)$ times. This is far fewer than $O(n^2)$ times required by dynamic programming algorithms for ℓ_0-penalised cost minimisation (Rinaldo et al., 2021; Xu et al., 2022), but it may still pose a computational bottleneck when the data sequence is very long or its dimensionality ultra high. Instead, we propose to evaluate $T_k(G)$ on a coarser grid only for generating *pre-estimators* of the change points. Let $\mathscr{T}$ denote the grid over which we evaluate $T_k(G)$, which is given by

$$\mathscr{T} = \mathscr{T}(r, G) = \left\{ G + \lfloor rG \rfloor m, 0 \leq m \leq \left\lfloor \frac{n-2G}{rG} \right\rfloor \right\} \tag{3.6}$$

with some constant $r \in [G^{-1}, 1)$ that controls the coarseness of the grid. When $r = G^{-1}$, we have the finest grid $\mathscr{T} = \{G, \ldots, n-G\}$ and the grid becomes coarser with increasing r.

Motivated by Eichinger and Kirch (2018), who considered the problem of detecting multiple shifts in the mean of univariate time series using a moving window procedure, we propose to accept all significant local maximisers of $T_k(G)$ over $k \in \mathscr{T}$ as the pre-estimators of the change

points. That is, for some threshold $D > 0$ and a tuning parameter $\eta \in (0,1]$, we accept all $\widetilde{\theta} \in \mathscr{T}$ that simultaneously satisfy

$$T_{\widetilde{\theta}}(G) > D \quad \text{and} \quad \widetilde{\theta} = \arg\max_{k \in \mathscr{T}: |k-\widetilde{\theta}| \leq \eta G} T_k(G). \tag{3.7}$$

That is, at such $\widetilde{\theta}$, the detector $T_{\widetilde{\theta}}(G)$ exceeds the threshold and attains a local maximum over the interval of length ηG. We denote the set collecting all pre-estimators fulfilling (3.7), by $\widetilde{\Theta} = \{\widetilde{\theta}_j, 1 \leq j \leq \widehat{q} : \widetilde{\theta}_1 < \ldots < \widetilde{\theta}_{\widehat{q}}\}$ with $\widehat{q} = |\widetilde{\Theta}|$ as the estimator of the number of change points. This grid-based approach substantially reduces the computational complexity by requiring the Lasso estimators to be computed only $O(n/\lfloor rG \rfloor)$ times. Even so, it is sufficient for detecting the presence of all q change points, provided that r is chosen not too large (see Theorem 3.1 (i) below). We remark that the idea of utilising only a sub-sample of the data for detecting the presence of change points, has been proposed for univariate mean change point detection in Lu et al. (2017). The next section describes the location refinement step applied to the pre-estimators of change point locations.

3.2.1.2 Stage 2: Location refinement

Once the set of pre-estimators $\widetilde{\Theta}$ is generated by the first-stage moving window procedure on a coarse grid, we further refine the location estimators. It involves the local evaluation and minimisation of the following objective function

$$Q\left(k; a, b, \widehat{\boldsymbol{\gamma}}^{\mathrm{L}}, \widehat{\boldsymbol{\gamma}}^{\mathrm{R}}\right) = \sum_{t=a+1}^{k} (Y_t - \mathbf{x}_t^\top \widehat{\boldsymbol{\gamma}}^{\mathrm{L}})^2 + \sum_{t=k+1}^{b} (Y_t - \mathbf{x}_t^\top \widehat{\boldsymbol{\gamma}}^{\mathrm{R}})^2 \text{ for } k = a+1, \ldots, b, \tag{3.8}$$

for suitably chosen a, b, $\widehat{\boldsymbol{\gamma}}^{\mathrm{L}}$ and $\widehat{\boldsymbol{\gamma}}^{\mathrm{R}}$. A similar idea has been considered for location refinement in the change point literature, see e.g. Kaul et al. (2019b) and Xu et al. (2022).

For each $j = 1, \ldots, \widehat{q}$, let $\widetilde{\theta}_j^{\mathrm{L}} = \widetilde{\theta}_j - \lfloor G/2 \rfloor$ and $\widetilde{\theta}_j^{\mathrm{R}} = \widetilde{\theta}_j + \lfloor G/2 \rfloor$, and consider the following local parameter estimators

$$\widehat{\boldsymbol{\beta}}_j^{\mathrm{L}} = \widehat{\boldsymbol{\beta}}_{0 \vee (\widetilde{\theta}_j^{\mathrm{L}} - G), \widetilde{\theta}_j^{\mathrm{L}}} \quad \text{and} \quad \widehat{\boldsymbol{\beta}}_j^{\mathrm{R}} = \widehat{\boldsymbol{\beta}}_{\widetilde{\theta}_j^{\mathrm{R}}, (\widetilde{\theta}_j^{\mathrm{R}} + G) \wedge n}, \tag{3.9}$$

which serve as the estimators of $\boldsymbol{\beta}_{j-1}$ and $\boldsymbol{\beta}_j$, respectively. Then in Stage 2, we propose to obtain a refined location estimator of θ_j from its pre-estimator $\widetilde{\theta}_j$, as

$$\widehat{\theta}_j = \arg\min_{\widetilde{\theta}_j - G + 1 \leq k \leq \widetilde{\theta}_j + G} Q\left(k; \widetilde{\theta}_j - G, \widetilde{\theta}_j + G, \widehat{\boldsymbol{\beta}}_j^{\mathrm{L}}, \widehat{\boldsymbol{\beta}}_j^{\mathrm{R}}\right), \tag{3.10}$$

for all $j = 1, \ldots, \widehat{q}$. Referring to the methodology combining the two stages as MOSEG, we provide its algorithmic description in Algorithm 1. In Figure 3.2, we demonstrate the algorithm on data simulated with a single change located at $\theta_1 = 100$.

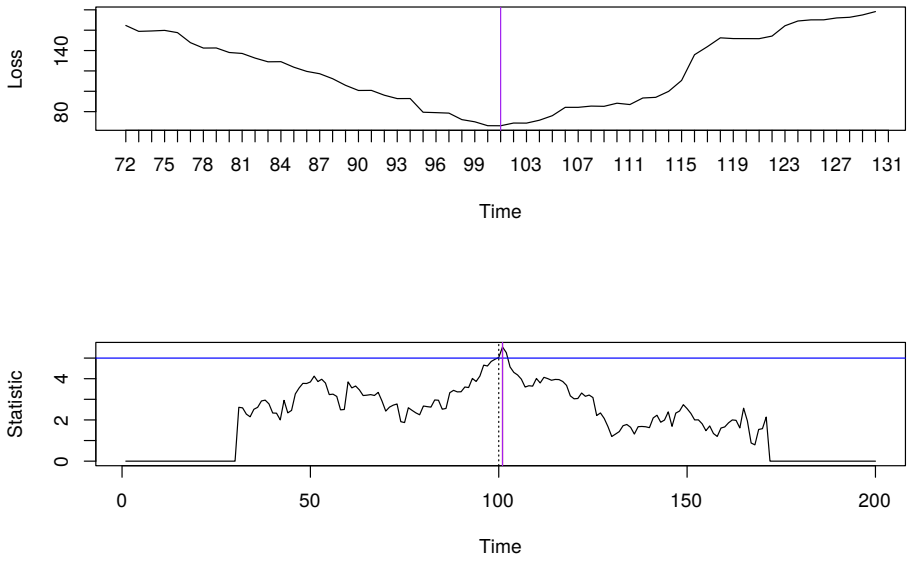


Figure 3.2: Results from Algorithm 1 for data simulated with a single change located at $\theta_1 = 100$. The lower panel plots $T_k(G)$ in black, with the chosen threshold marked horizontally. The stage one estimator $\widetilde{\theta}_1$ is marked vertically in red, and the corresponding stage 2 estimator $\widehat{\theta}_1$ is marked in purple. The upper panel visualises $Q(k;\widetilde{\theta}_1 - G, \widetilde{\theta}_1 + G, \widehat{\boldsymbol{\beta}}_1^{\mathrm{L}}, \widehat{\boldsymbol{\beta}}_1^{\mathrm{R}})$.

Algorithm 1: MOSEG: Single-bandwidth two-stage data segmentation methodology under a regression model.

input : Bandwidth G, grid resolution r, penalty λ, threshold D, $\eta \in (0,1]$

initialise: $\widetilde{\Theta} = \emptyset$, $\widehat{\Theta} = \emptyset$

// Stage 1

Compute $T_k(G)$ in (3.3) for all $k \in \mathscr{T} = \mathscr{T}(r,G)$

Add all $\widetilde{\theta}$ satisfying $T_{\widetilde{\theta}}(G) > D$ and $\widetilde{\theta} = \arg\min_{k \in \{\widetilde{\theta} - \lfloor \eta G \rfloor + 1, \ldots, \widetilde{\theta} + \lfloor \eta G \rfloor\} \cap \mathscr{T}} T_k(G)$ to $\widetilde{\Theta}$, and set $\widetilde{\Theta} = \{\widetilde{\theta}_j, 1 \le j \le \widehat{q}\}$

// Stage 2

for $j = 1, \ldots, \widehat{q}$ **do**

  Identify $\widehat{\theta}_j = \arg\min_{\widetilde{\theta}_j - G + 1 \le j \le \widetilde{\theta}_j + G} Q(k; \widetilde{\theta}_j - G, \widetilde{\theta}_j + G, \widehat{\boldsymbol{\beta}}_j^{\mathrm{L}}, \widehat{\boldsymbol{\beta}}_j^{\mathrm{R}})$ with $\widehat{\boldsymbol{\beta}}_j^{\mathrm{L}}$ and $\widehat{\boldsymbol{\beta}}_j^{\mathrm{R}}$ computed as in (3.9), and add it to $\widehat{\Theta}$

end

return $\widehat{\Theta}$

3.2.2 Consistency of MOSEG

To establish the consistency of MOSEG, we make the following assumptions on $(\mathbf{x}_t, \varepsilon_t)$, $1 \le t \le n$. Assumption 3.1 is commonly made in the literature on high-dimensional regression and change

point problems thereof.

Assumption 3.1. We assume that $\mathbb{E}(\mathbf{x}_t) = \mathbf{0}$, $\mathbb{E}(\varepsilon_t) = 0$ and $\mathsf{Var}(\varepsilon_t) = \sigma_\varepsilon^2$ for all $t = 1, \ldots, n$, and that $\mathsf{Cov}(\mathbf{x}_t) = \mathbf{\Sigma}_x$ has its eigenvalues bounded, i.e. there exist $0 \leq \omega \leq \bar{\omega} < \infty$ such that

$$\omega \leq \Lambda_{\min}(\mathbf{\Sigma}_x) \leq \Lambda_{\max}(\mathbf{\Sigma}_x) \leq \bar{\omega}.$$

Assumptions 3.2 and 3.3 below extend the deviation bound and restricted strong convexity (RSC) conditions required for high-dimensional M-estimation (Loh and Wainwright, 2012; Negahban et al., 2012; van de Geer and Bühlmann, 2009), to change point settings. They are met e.g. by a class of linear processes accommodating serial dependence and non-Gaussian tail behaviour, as verified later in Section 3.2.3. We explicitly state these meta-assumptions to highlight that the consistency of MOSEG derived in this section is not limited to such processes only.

Assumption 3.2 (Deviation bound)**.** There exist fixed constants $C_0, C_{\mathrm{DEV}} > 0$ and some $\rho_{n,p} \to \infty$ as $n, p \to \infty$, such that $\mathsf{P}(\mathscr{D}^{(1)} \cap \mathscr{D}^{(2)}) \to 1$, where

$$\mathscr{D}^{(1)} = \left\{ \max_{0 \leq s < e \leq n,\, e-s \geq C_0 \rho_{n,p}^2} \left| \frac{1}{\sqrt{e-s}} \sum_{t=s+1}^{e} \varepsilon_t \mathbf{x}_t \right|_\infty \leq C_{\mathrm{DEV}} \rho_{n,p} \right\},$$

$$\mathscr{D}^{(2)} = \left\{ \max_{\substack{0 \leq s < e \leq n,\, e-s \geq C_0 \rho_{n,p}^2 \\ |\{s+1,\ldots,e\} \cap \Theta| \leq 1}} \left| \frac{1}{\sqrt{e-s}} \sum_{t=s+1}^{e} (Y_t - \mathbf{x}_t^\top \boldsymbol{\beta}_{s,e}^*) \mathbf{x}_t \right|_\infty \leq C_{\mathrm{DEV}} \rho_{n,p} \right\}.$$

Assumption 3.3 (Restricted strong convexity)**.** There exist fixed constants $C_{\mathrm{RSC}} > 0$ and $\tau \in [0, 1)$ such that $\mathsf{P}(\mathscr{R}^{(1)} \cap \mathscr{R}^{(2)}) \to 1$, where

$$\mathscr{R}^{(1)} = \left\{ \sum_{t=s+1}^{e} \mathbf{a}^\top \mathbf{x}_t \mathbf{x}_t^\top \mathbf{a} \geq (e-s)\omega |\mathbf{a}|_2^2 - C_{\mathrm{RSC}} \log(p)(e-s)^\tau |\mathbf{a}|_1^2 \text{ for all } 0 \leq s < e \leq n \text{ satisfying } e - s \geq C_0 \rho_{n,p}^2 \text{ and } \mathbf{a} \in \mathbb{R}^p \right\},$$

$$\mathscr{R}^{(2)} = \left\{ \sum_{t=s+1}^{e} \mathbf{a}^\top \mathbf{x}_t \mathbf{x}_t^\top \mathbf{a} \leq (e-s)\bar{\omega} |\mathbf{a}|_2^2 + C_{\mathrm{RSC}} \log(p)(e-s)^\tau |\mathbf{a}|_1^2 \text{ for all } 0 \leq s < e \leq n \text{ satisfying } e - s \geq C_0 \rho_{n,p}^2 \text{ and } \mathbf{a} \in \mathbb{R}^p \right\}.$$

For each $j = 0, \ldots, q$, we denote by $\mathscr{S}_j = \mathrm{supp}(\boldsymbol{\beta}_j)$ the support of $\boldsymbol{\beta}_j$, and by $\mathfrak{s} = \max_{0 \leq j \leq q} |\mathscr{S}_j|$ the maximum segment-wise sparsity of the regression parameters. We make the following assumptions on the size of change $\delta_j = |\boldsymbol{\beta}_j - \boldsymbol{\beta}_{j-1}|_2$ and the spacing between the neighbouring change points through imposing conditions on G.

Assumption 3.4. There exists some constant $C_\delta > 0$ such that $\max_{1 \leq j \leq q} \delta_j \leq C_\delta$.

Assumption 3.5. The bandwidth G fulfils the following conditions with τ, $\rho_{n,p}$ and ω introduced in Assumptions 3.1, 3.2 and 3.3.

(a) $2G \leq \min_{1\leq j\leq q+1}(\theta_j - \theta_{j-1})$.

(b) There exists a fixed constant $C_1 > 0$ such that

$$\min_{1\leq j\leq q} \delta_j^2 G \geq C_1 \max\left\{\omega^{-2}\mathfrak{s}\rho_{n,p}^2, \left(\omega^{-1}\mathfrak{s}\log(p)\right)^{1/(1-\tau)}\right\}.$$

Assumption 3.4 is a technical condition under which we focus on the more challenging regime where the size of change is allowed to tend to zero; an analogous condition found in Lee et al. (2016), Kaul et al. (2019b), Wang et al. (2021a) and Xu et al. (2022). In particular, it rules out the case where $\mathrm{Var}(Y_t)$ diverges for some t, since $\mathrm{Var}(Y_t) \geq \omega \sum_{j=0}^{q} |\boldsymbol{\beta}_j|_2^2 \cdot \mathbb{I}_{\{\theta_j+1\leq t\leq \theta_j\}} + \sigma_\varepsilon^2$. Assumption 3.5 (a) relates the choice of bandwidth G to the minimum spacing between the change points. Together, (a) and (b) specify the *separation rate* imposing a lower bound on

$$\Delta^{(1)} = \min_{1\leq j\leq q} \delta_j^2 \cdot \min_{0\leq j\leq q} (\theta_{j+1} - \theta_j), \tag{3.11}$$

for all the q change points to be detectable by MOSEG. Later in Section 3.3, we propose a multiscale extension of MOSEG which achieves consistency under a more relaxed condition than Assumption 3.5.

Theorem 3.1. Suppose that Assumptions 3.1, 3.2, 3.3, 3.4 and 3.5 hold. Let the tuning parameters satisfy $\lambda \geq 4C_{\mathrm{DEV}}\rho_{n,p}$, $r \in [1/G, 1/4)$, $\eta \in (4r, 1]$ and

$$\frac{48\sqrt{\mathfrak{s}}\lambda}{\omega} < D < \frac{\eta}{4\sqrt{2}} \min_{1\leq j\leq q} \delta_j \sqrt{G}. \tag{3.12}$$

Then on $\mathscr{D}^{(1)} \cap \mathscr{D}^{(2)} \cap \mathscr{R}^{(1)} \cap \mathscr{R}^{(2)}$, the following holds.

(i) Stage 1 of MOSEG returns $\widetilde{\Theta} = \{\widetilde{\theta}_j, 1 \leq j \leq \widehat{q} : \widetilde{\theta}_1 < \ldots < \widetilde{\theta}_{\widehat{q}}\}$ which satisfies

$$\widehat{q} = q \quad \text{and} \quad |\widetilde{\theta}_j - \theta_j| \leq \frac{48\sqrt{2\mathfrak{s}G}\lambda}{\omega\delta_j} + \lfloor rG \rfloor < \left\lfloor \frac{G}{2} \right\rfloor \text{ for each } j = 1, \ldots, q.$$

(ii) There exists a large enough constant $c_0 > 0$ such that Stage 2 of MOSEG returns $\widehat{\Theta} = \{\widehat{\theta}_j, 1 \leq j \leq \widehat{q} : \widehat{\theta}_1 < \ldots < \widehat{\theta}_{\widehat{q}}\}$ which satisfies

$$\max_{1\leq j\leq q} \delta_j^2 |\widehat{\theta}_j - \theta_j| \leq c_0 \max\left(\mathfrak{s}\rho_{n,p}^2, (\mathfrak{s}\log(p))^{\frac{1}{1-\tau}}\right).$$

Theorem 3.1 (i) establishes that Stage 1 of MOSEG correctly estimates the number of change points as well as identifying their locations by the pre-estimators with some accuracy. There is a trade-off between computational efficiency and theoretical consistency with respect to the choice of r. On one hand, increasing r leads to a coarser grid $\mathscr{T}$ with its cardinality $|\mathscr{T}| = O(n/(rG))$, and thus reduces the computational cost. On the other, the pre-estimators lie in the grid such that the best approximation to each change point θ_j can be as far from θ_j as $\lfloor rG \rfloor/2$, which is reflected on the localisation property of the pre-estimators. Theorem 3.1 (ii) derives the rate of estimation for the second-stage estimators $\widehat{\theta}_j$ which shows that the location estimation is more challenging when the size of change δ_j is small. Finally, we always have $\max_{1\leq j\leq q} \delta_j^{-2} \max(\mathfrak{s}\rho_{n,p}^2, (\mathfrak{s}\log(p))^{1/(1-\tau)}) \lesssim G \lesssim \min_{1\leq j\leq q+1}(\theta_j - \theta_{j-1})$ under Assumption 3.4.

3.2.3 Verification of Assumptions 3.2 and 3.3

Assumptions 3.2 and 3.3 generalise the deviation bound and the RSC condition which are often found in the high-dimensional M-estimation literature, to accommodate change points, serial dependence and heavy-tailedness. Condition 1 gives instances of $\{(\mathbf{x}_t, \varepsilon_t)\}_{t=1}^n$ that fulfil Assumptions 3.2 and 3.3 and specify the corresponding $\rho_{n,p}$ and τ.

Condition 1. Suppose that for i.i.d. random vectors $\boldsymbol{\xi}_t = (\xi_{1t}, \ldots, \xi_{p+1,t})^\top \in \mathbb{R}^{p+1}$, $t \in \mathbb{Z}$, with $\mathbb{E}(\boldsymbol{\xi}_t) = \mathbf{0}$ and $\mathrm{Cov}(\boldsymbol{\xi}_t) = \mathbf{I}$, we have

$$\begin{bmatrix} \mathbf{x}_t \\ \varepsilon_t \end{bmatrix} = \sum_{\ell=0}^{\infty} \mathbf{D}_\ell \boldsymbol{\xi}_{t-\ell} \quad \text{with} \quad \mathbf{D}_\ell = [D_{\ell,ik}, 1 \le i,k \le p+1] \in \mathbb{R}^{(p+1)\times(p+1)} \tag{3.13}$$

subject to $\mathbb{E}(\mathbf{x}_t \varepsilon_t) = \mathbf{0}$. Further, there exist constants $\Xi > 0$ and $\varsigma > 2$ such that

$$|D_{\ell,ik}| \le C_{ik}(1+\ell)^{-\varsigma} \quad \text{with} \quad \max\left\{ \max_{1\le k\le p+1} \sum_{i=1}^{p+1} C_{ik}, \max_{1\le i\le p+1} \sum_{k=1}^{p+1} C_{ik} \right\} \le \Xi \tag{3.14}$$

for all $\ell \ge 0$. Finally, we impose *either* of the two conditions on ξ_{it}.

(a) There exist some constants $C_\xi > 0$ and $\gamma \in (0,2]$ such that $(\mathbb{E}(|\xi_{it}|^\nu))^{1/\nu} = \|\xi_{it}\|_\nu \le C_\xi \nu^\gamma$ for all $\nu \ge 1$. In other words, $\|\xi_{it}\|_{\psi_\nu} := \sup_{\nu\ge 1} \nu^{-1/\gamma} \|\xi_{it}\|_\nu \le C_\xi$.

(b) $\xi_{it} \sim_{\text{iid}} \mathcal{N}(0,1)$.

Proposition 3.2. Suppose that Assumptions 3.1 and 3.4 and Condition 1 hold. Then, there exist some constants $c_1, c_2 > 0$ such that $\mathrm{P}(\mathscr{D}^{(1)} \cap \mathscr{D}^{(2)} \cap \mathscr{R}^{(1)} \cap \mathscr{R}^{(2)}) \ge 1 - c_1(p \vee n)^{-c_2}$, with $\omega = \Lambda_{\min}(\mathbf{\Sigma}_x)/2$, $\bar{\omega} = 3\Lambda_{\max}(\mathbf{\Sigma}_x)/2$, and τ and $\rho_{n,p}$ chosen as below.

(i) Under Condition 1 (a), we set $\tau = (4\gamma+2)/(4\gamma+3)$ and $\rho_{n,p} = \log^{2\gamma+3/2}(p \vee n)$.

(ii) Under Condition 1 (b), we set $\tau = 0$ and $\rho_{n,p} = \sqrt{\log(p \vee n)}$.

Under Condition 1, $\{(\mathbf{x}_t, \varepsilon_t)\}_{t=1}^n$ is a linear process with algebraically decaying serial dependence, according to the functional dependence measure of Zhang and Wu (2017). Also, Condition 1 (a) permits heavier tail behaviour than that allowed under sub-Gaussianity or sub-exponential distributions when $\gamma > 1/2$ and $\gamma > 1$, respectively. The proof of Proposition 3.2 follows using Lemma 12 of Loh and Wainwright (2012), which establishes a lower Restricted Strong Convexity bound.

Remark 3.1. The consistency of Lasso-type estimator (when $q = 0$) under serial dependence and non-Gaussianity, has been investigated under functional dependence or mixing conditions (Adamek et al., 2020; Han and Tsay, 2020; Wong et al., 2020; Wu and Wu, 2016). In the change point literature, Wang and Zhao (2022) propose a change point test and investigate its properties under β-mixing, and Xu et al. (2022) analyse the ℓ_0-penalised least squares estimation approach

when the functional dependence of $\{\mathbf{x}_t\}_{t=1}^n$ and $\{\varepsilon_t\}_{t=1}^n$ decays exponentially. Relaxing the Gaussianity, it is typically required that $\mathbf{x}_t$ is a sub-Weibull random vector, i.e. $\sup_{\mathbf{a}\in\mathbb{B}_2(1)} \|\mathbf{a}^\top \mathbf{x}_t\|_{\psi_\nu} < \infty$ for some $\gamma > 0$ (where $\mathbb{B}_d(r) = \{\mathbf{a} : |\mathbf{a}|_d \le r\}$) and similarly, $\|\varepsilon_t\|_{\psi_\gamma} < \infty$. Under these assumptions, the common approach is to verify the deviation bound and RSC conditions analogous to those made in Assumptions 3.2–3.3, with which the consistency of the Lasso estimator is derived (locally in the case of the change point detection problem).

Instead, we explicitly state the meta-assumptions and give Condition 1 as one scenario under which these assumptions are met. The proposed MOSEG achieves consistency in multiple change point detection as shown in Theorem 3.1 whenever Assumptions 3.2–3.3 are met, which can be verified using the arguments adopted in the aforementioned literature.

Corollary 3.3 follows immediately from Theorem 3.1 and Proposition 3.2.

Corollary 3.3. Suppose that Assumptions 3.1, 3.4 and 3.5 and Condition 1 hold, and λ, r, η and D are chosen as in Theorem 3.1. Then, there exist constants $c_i > 0$, $i = 0, 1, 2$, such that $\widehat{\Theta} = \{\widehat{\theta}_j, 1 \le j \le \widehat{q} : \widehat{\theta}_1 < \ldots < \widehat{\theta}_{\widehat{q}}\}$ returned by MOSEG satisfies the following.

(i) Under Condition 1 (a), we have

$$\mathsf{P}\left(\widehat{q} = q \ \text{ and } \ \max_{1\le j\le q} \delta_j^2 |\widehat{\theta}_j - \theta_j| \le c_0 (\mathfrak{s}\log(p\vee n))^{4\gamma+3}\right) \ge 1 - c_1 (p\vee n)^{-c_2}.$$

(ii) Under Condition 1 (b), we have

$$\mathsf{P}\left(\widehat{q} = q \ \text{ and } \ \max_{1\le j\le q} \delta_j^2 |\widehat{\theta}_j - \theta_j| \le c_0 \mathfrak{s}\log(p\vee n)\right) \ge 1 - c_1 (p\vee n)^{-c_2}.$$

Corollary 3.3 (ii) shows that under Gaussianity, the rate of localisation attained by MOSEG matches the minimax lower bound up to $\log(p \vee n)$, see Lemma 4 of Rinaldo et al. (2021). At the same time, Assumption 3.5 (b) translates to $\Delta^{(1)} \gtrsim \mathfrak{s}\log(p\vee n)$ in this setting, nearly matching the minimax lower bound on the separation rate derived in Lemma 3 of Rinaldo et al. (2021) up to the logarithmic term.

3.3 Multiscale methodology

The single-bandwidth methodology proposed in Section 3.2 enjoys theoretical consistency as well as computational efficiency, but faces the difficulty arising from identifying a bandwidth that satisfies Assumption 3.5 (a)–(b) simultaneously, that is, to identify one which is sufficiently large to identify the signal of the change but small enough that multiple regimes are not coverered by the active window. In this section, we propose MOSEG.MS, a multiscale extension of MOSEG, and show that it achieves consistency in a parameter space broader than that allowed by Assumption 3.5, and thus alleviates the difficulty associated with the choice of a single bandwidth.

In the context of a univariate series with changes in the mean where we observe $X_t = \mu_j + \varepsilon_t$, and $\delta_j = \mu_j - \mu_{j-1}$, Figure 2.7 demonstrates a divergence between $\Delta^{(1)}$ and $\Delta^{(2)}$. Large jumps δ_j occur at changes θ_j with small neighbouring regime lengths, such that $\min(\theta_j - \theta_{j-1}, \theta_{j+1} - \theta_j)$ is small. In terms of $\Delta^{(1)}$, the signal is very small, but in terms of $\Delta^{(2)}$, the signal is sufficiently large such that detection and localisation are possible with a multiscale algorithm.

3.3.1 MOSEG.MS: Multiscale extension of MOSEG

Similarly to MOSEG, MOSEG.MS consists of moving window-based data scanning and location refinement but it takes a set of bandwidths as an input. The key innovation lies in that for each change point, MOSEG.MS learns the bandwidth best-suited for its detection and localisation from the given set of bandwidths. While there exist multiscale extensions of moving sum procedures, they are mostly developed for univariate time series segmentation (Cho and Kirch, 2021b; Messer et al., 2014) and to the best of our knowledge, this is a first attempt at rigorously studying such an extension in a high-dimensional setting. Below we describe MOSEG.MS step-by-step.

Step 1: Pre-estimator generation. Given a set of bandwidths $\mathcal{G} = \{G_h, 1 \le h \le H : G_1 < \ldots < G_H\}$, we generate the coarse grid associated with each G_h and the parameter r by $\mathcal{T}_h = \mathcal{T}(r, G_h)$, see (3.6). As in Stage 1 of MOSEG, the sets of pre-estimators $\widetilde{\Theta}(G_h)$ are generated for $h = 1, \ldots, H$, and we denote by $\widetilde{\Theta}(\mathcal{G}) = \cup_{h=1}^{H} \widetilde{\Theta}(G_h)$ the pooled set of all such pre-estimators. By (3.7), at each $\widetilde{\theta} \in \widetilde{\Theta}(G_h)$, we have $T_{\widetilde{\theta}}(G_h) > D$ and $\widetilde{\theta} = \arg\max_{k \in \mathcal{I}_\eta(\widetilde{\theta}) \cap \mathcal{T}_h} T_k(G_h)$, where $\mathcal{I}_\eta(\widetilde{\theta}) = \{\widetilde{\theta} - \lfloor \eta G_h \rfloor + 1, \ldots, \widetilde{\theta} + \lfloor \eta G_h \rfloor\}$ denotes the detection interval associated with $\widetilde{\theta}$. For simplicity, we write $\mathcal{I}_1(\widetilde{\theta}) = \mathcal{I}(\widetilde{\theta})$. Below, we sometimes write $\widetilde{\theta}(G) \in \widetilde{\Theta}(G)$ to highlight that the pre-estimator is obtained with the bandwidth G, and denote by $G(\widetilde{\theta})$ the bandwidth involved in the detection of a pre-estimator $\widetilde{\theta}$. If some $\widetilde{\theta}$ is detected with more than one bandwidths, we distinguish between them.

Step 2: Anchor estimator identification. Next, we identify *anchor* change point estimators $\widetilde{\theta}^A(G) \in \widetilde{\Theta}(\mathcal{G})$ detected at some $G \in \mathcal{G}$ which satisfy

$$\bigcup_{h: G_h < G} \bigcup_{k \in \widetilde{\Theta}(G_h)} \left\{ \mathcal{I}(k) \cap \mathcal{I}(\widetilde{\theta}^A(G)) \right\} = \emptyset. \tag{3.15}$$

That is, each anchor change point estimator does not have its detection interval overlap with the detection interval of any pre-estimator that is detected with a finer bandwidth. Denote the set of all such anchor change point estimators by $\widetilde{\Theta}^A = \{\widetilde{\theta}^A_j, 1 \le j \le \widehat{q} : \widetilde{\theta}^A_1 < \ldots < \widetilde{\theta}^A_{\widehat{q}}\}$, with $\widehat{q} = |\widetilde{\Theta}^A|$ as an estimator of the number of change points q.

Step 3: Pre-estimator clustering. We find subsets of the pre-estimators in $\widetilde{\Theta}(\mathcal{G})$ denoted by $\mathcal{C}_j$, $j = 1, \ldots, \widehat{q}$, as described below. Initialised as $\mathcal{C}_j = \emptyset$, for each j, we add to $\mathcal{C}_j$ the jth anchor

estimator $\widetilde{\theta}_j^A$ as well as all $\widetilde{\theta} \in \widetilde{\Theta}(\mathscr{G})$ which simultaneously fulfil

$$\begin{aligned} &\mathscr{I}(\widetilde{\theta}) \cap \mathscr{I}(\widetilde{\theta}_j^A) \neq \emptyset, \quad \text{and} \\ &\{\widetilde{\theta} - G(\widetilde{\theta}) - \lfloor G(\widetilde{\theta})/2 \rfloor + 1, \ldots, \widetilde{\theta} + G(\widetilde{\theta}) + \lfloor G(\widetilde{\theta})/2 \rfloor\} \cap \mathscr{I}(\widetilde{\theta}_{j'}^A) = \emptyset \text{ for all } j' \neq j. \end{aligned} \tag{3.16}$$

Step 4: Location refinement. For each $\mathscr{C}_j$, $j = 1, \ldots, \widehat{q}$, we denote the smallest and the largest bandwidths associated with the detection of the pre-estimators in $\mathscr{C}_j$, by G_j^m and G_j^M, respectively, and the corresponding pre-estimators by $\widetilde{\theta}_j^m$ and $\widetilde{\theta}_j^M$ (when $|\mathscr{C}_j| = 1$, we have $\widetilde{\theta}_j^m = \widetilde{\theta}_j^M = \widehat{\theta}_j^A$ and $G_j^m = G_j^M$). Setting $G_j^* = \lfloor 3G_j^m/4 + G_j^M/4 \rfloor$, we identify the local minimiser of the objective function defined in (3.8), as

$$\begin{aligned} &\check{\theta}_j = \arg\min_{\widetilde{\theta}_j^m - G_j^* + 1 \leq k \leq \widetilde{\theta}_j^m + G_j^*} Q\left(k; \widetilde{\theta}_j^m - G_j^*, \widetilde{\theta}_j^m + G_j^*, \widehat{\boldsymbol{\beta}}_j^{\mathrm{L}}, \widehat{\boldsymbol{\beta}}_j^{\mathrm{R}}\right), \\ &\text{with} \quad \widehat{\boldsymbol{\beta}}_j^{\mathrm{L}} = \widehat{\boldsymbol{\beta}}_{(\widetilde{\theta}_j^m - G_j^m - G_j^*) \vee 0, \widetilde{\theta}_j^m - G_j^m} \quad \text{and} \quad \widehat{\boldsymbol{\beta}}_j^{\mathrm{R}} = \widehat{\boldsymbol{\beta}}_{\widetilde{\theta}_j^m + G_j^m, (\widetilde{\theta}_j^m + G_j^m + G_j^*) \wedge n}. \end{aligned} \tag{3.17}$$

Repeatedly performing (3.17) for $j = 1, \ldots, \widehat{q}$, we obtain $\check{\Theta} = \{\check{\theta}_j, 1 \leq j \leq \widehat{q}\}$.

An algorithmic description of MOSEG.MS is given in Algorithm 8. The identification of anchor

Algorithm 2: MOSEG.MS: Multiscale extension of MOSEG.

input : A set of bandwidths $\mathscr{G}$, grid resolution r, penalty λ, threshold D, $\eta \in (0, 1]$

initialise: $\widetilde{\Theta}^A = \emptyset$, $\check{\Theta} = \emptyset$, $\mathscr{C}_j = \emptyset$ for all j

// Pre-estimator generation

for $h = 1, \ldots, H$ **do**

  Initialise $\widetilde{\Theta}(G_h) = \emptyset$

  Compute $T_k(G_h)$ in (3.3) for all $k \in \mathscr{T}_h = \mathscr{T}(r, G_h)$

  Add all $\widetilde{\theta}$ satisfying $T_{\widetilde{\theta}}(G_h) > D$ and $\widetilde{\theta} = \arg\min_{k \in \mathscr{I}_\eta(\widetilde{\theta}) \cap \mathscr{T}_h} T_k(G_h)$, to $\widetilde{\Theta}(G_h)$

end

// Anchor change point estimator identification

Identify all $\widetilde{\theta}(G) \in \cup_{h=1}^H \widetilde{\Theta}(G_h)$ satisfying (3.15), and add all such estimators to $\widetilde{\Theta}^A$, which is denoted by $\widetilde{\Theta}^A = \{\widetilde{\theta}_j^A, 1 \leq j \leq \widehat{q} : \widetilde{\theta}_1^A < \ldots < \widetilde{\theta}_{\widehat{q}}^A\}$

for $j = 1, \ldots, \widehat{q}$ **do**

  // Pre-estimator clustering

  Identify all $\widetilde{\theta} \in \cup_{h=1}^H \widetilde{\Theta}(G_h)$ satisfying (3.16) and add it to $\mathscr{C}_j$

  // Location refinement

  Add $\check{\theta}_j$ obtained as in (3.17) to $\widehat{\Theta}$

end

return $\check{\Theta}$

change point estimators bears some resemblance with the bottom-up merging proposed in Messer

et al. (2014), but the anchor estimators do not come with a guaranteed rate of localisation. Instead, we cluster the pre-estimators and learn the bandwidth G_j^* well-suited for localising each θ_j in a data-driven way, with which we obtain a refined estimator. Figure 3.3 demonstrates the steps of the algorithm on simulated data.

Remark 3.2 (Bandwidth generation)**.** Cho and Kirch (2021b) propose to use $\mathscr{G}$ generated as a sequence of Fibonacci numbers, for a multiscale extension of the moving sum procedure proposed in Eichinger and Kirch (2018) in the context of univariate mean change point detection. For some finest bandwidth $G_0 = G_1$, we iteratively produce G_h, $h \geq 2$, as $G_h = G_{h-1} + G_{h-2}$. Equivalently, we set $G_h = F_h G_0$ where $F_h = F_{h-1} + F_{h-2}$ with $F_0 = F_1 = 1$. This is repeated until for some H, it holds that $G_H < \lfloor n/2 \rfloor$ while $G_{H+1} \geq \lfloor n/2 \rfloor$. By induction, $F_h = O(((1+\sqrt{2})/2)^h)$ such that the thus-generated bandwidth set $\mathscr{G}$ satisfies $|\mathscr{G}| = O(\log(n))$.

3.3.2 Consistency of MOSEG.MS

We make the following assumption on the size of change δ_j and the spacing between the neighbouring change points.

Assumption 3.5′. Let $\mathscr{G}$ denote the set of bandwidths generated as in Remark 3.2 with $G_1 \geq C_0 \max\{\rho_{n,p}^2, (\omega^{-1}\mathfrak{s}\log(p))^{1/(1-\tau)}\}$. Then, for each change point θ_j, $j = 1, \ldots, q$, there exists a bandwidth $G_{(j)} \in \mathscr{G}$ such that

(a) $4G_{(j)} \leq \min(\theta_{j+1} - \theta_j, \theta_j - \theta_{j-1})$, and

(b) $\delta_j^2 G_{(j)} \geq 4C_1 \max\left\{\omega^{-2}\mathfrak{s}\rho_{n,p}^2, \left(\omega^{-1}\mathfrak{s}\log(p)\right)^{1/(1-\tau)}\right\}$ with C_1 from Assumption 3.5.

If there are multiple such bandwidths, let $G_{(j)}$ denote the smallest one.

Assumption 3.5′ relaxes Assumption 3.5 by requiring that for each θ_j, there exists one bandwidth $G_{(j)} \in \mathscr{G}$ fulfilling the requirements imposed on a single bandwidth in the latter for all $j = 1, \ldots, q$. Assumption 3.5′ effectively places a condition on

$$\Delta^{(2)} = \min_{1 \leq j \leq q} \delta_j^2 \cdot \min(\theta_{j+1} - \theta_j, \theta_j - \theta_{j-1}) \tag{3.18}$$

for MOSEG.MS to detect all q changes. Compared to $\Delta^{(1)}$ defined in (3.11), we always have $\Delta^{(1)} \leq \Delta^{(2)}$ and, if frequent large changes and small changes over long stretches of stationarity are simultaneously present, the former can be considerably smaller than the latter, see also the discussion in Cho and Kirch (2021a). To the best of our knowledge, Theorem 3.4 below provides a first result obtained under the larger parameter space defined with $\Delta^{(2)}$, in establishing the consistency of a data segmentation methodology for the problem in (3.1). We refer to Section 3.1.1 for further discussions and comprehensive comparison between MOSEG, MOSEG.MS and competing methodologies.

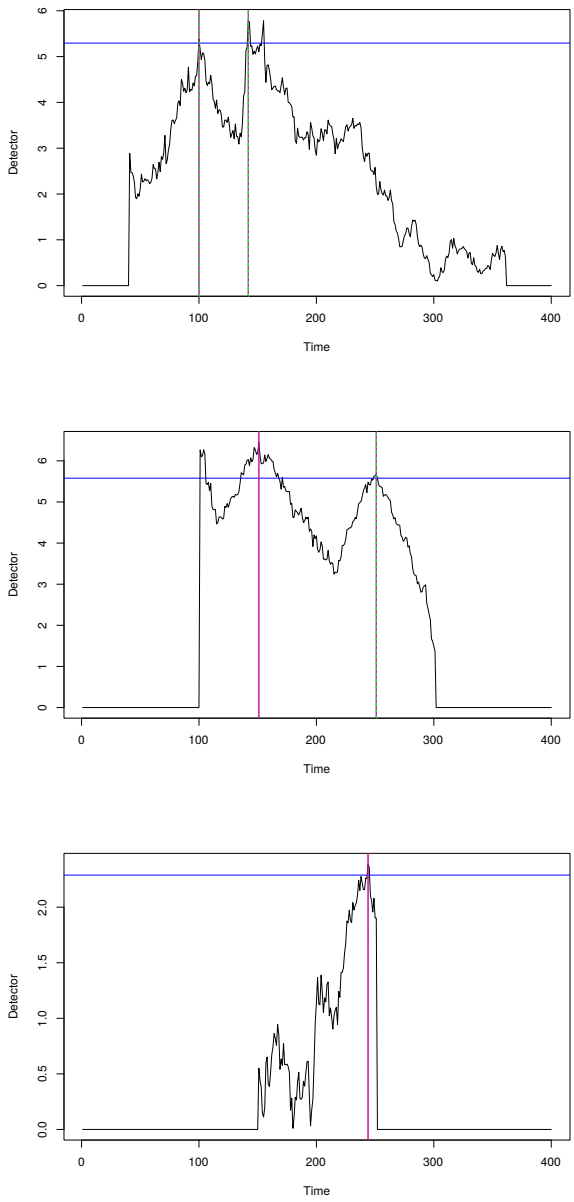


Figure 3.3: Results from the MOSEG.MS procedure on simulated data with changes located at $\theta_1 = 100, \theta_2 = 150$, and $\theta_3 = 250$. Each panel plots the outcome from a call with a given bandwidth. All candidate pre-estimators are marked with a vertical line, and the three selected as anchors are dashed. Those not selected as anchors are clustered as per Step 3, and used to determine the bandwidth in Step 4.

Theorem 3.4. Suppose that Assumptions 3.1, 3.2, 3.3, 3.4 and 3.5′ hold. Let the tuning parameters satisfy $\lambda \geq 4C_{\mathrm{DEV}}\rho_{n,p}$, $r \in [G_1^{-1}, 1/4)$, $\eta \in (4r, 1]$ and

$$\frac{48\sqrt{s}\lambda}{\omega} < D < \frac{\eta}{4\sqrt{2}} \min_{1\leq j\leq q} \delta_j \sqrt{G_{(j)}}. \tag{3.19}$$

Then, there exists a constant $c_0 > 0$ such that on $\mathscr{D}^{(1)} \cap \mathscr{D}^{(2)} \cap \mathscr{R}^{(1)} \cap \mathscr{R}^{(2)}$, MOSEG.MS returns

$\check{\Theta} = \{\check{\theta}_j, 1 \le j \le \widehat{q} : \check{\theta}_1 < \ldots < \check{\theta}_{\widehat{q}}\}$ which satisfies

$$\widehat{q} = q \quad \text{and} \quad \max_{1 \le j \le q} \delta_j^2 |\check{\theta}_j - \theta_j| \le c_0 \max\left(\mathfrak{s}\rho_{n,p}^2, (\mathfrak{s}\log(p))^{\frac{1}{1-\tau}}\right).$$

Corollary 3.5. Suppose that Assumptions 3.1, 3.4 and 3.5′ and Condition 1 hold, and λ, r and D are chosen as in Theorem 3.4. Then, there exist constants $c_i > 0, i = 0,1,2$, such that $\check{\Theta} = \{\check{\theta}_j, 1 \le j \le \widehat{q} : \check{\theta}_1 < \ldots < \check{\theta}_{\widehat{q}}\}$ returned by MOSEG.MS satisfies the following.

(i) Under Condition 1 (a), we have

$$\mathrm{P}\left(\widehat{q} = q \ \text{ and } \max_{1 \le j \le q} \delta_j^2 |\check{\theta}_j - \theta_j| \le c_0 (\mathfrak{s}\log(p \vee n))^{4+3\gamma}\right) \ge 1 - c_1 (p \vee n)^{-c_2}.$$

(ii) Under Condition 1 (b), we have

$$\mathrm{P}\left(\widehat{q} = q \ \text{ and } \max_{1 \le j \le q} \delta_j^2 |\check{\theta}_j - \theta_j| \le c_0 \mathfrak{s}\log(p \vee n)\right) \ge 1 - c_1 (p \vee n)^{-c_2}.$$

3.4 Numerical experiments

3.4.1 Choice of tuning parameters

We discuss the selection of tuning parameters involved in MOSEG and MOSEG.MS, namely the set of bandwidths $\mathscr{G}$, the grid $\mathscr{T}(r,G)$ in (3.6), $\eta \in (0,1]$ involved in the pre-estimation of the change points (see (3.7)), the penalty parameter λ and the threshold D.

Selection of $\mathscr{G}$. As described in Remark 3.2, the set of bandwidths $\mathscr{G}$ is determined once the finest bandwidth G_1 is chosen. To gain insights about the minimum bandwidth required for the reasonable performance of the local Lasso estimators, we conducted numerical experiments by simulating datasets under (6.4) with $q = 0$, $\mathbf{x}_t \sim_{\text{iid}} \mathcal{N}_p(\mathbf{0}, \mathbf{I}_p)$, $\varepsilon_t \sim_{\text{iid}} \mathcal{N}(0,1)$ and $\boldsymbol{\beta}_0 = (\beta_{0,1}, \ldots, \beta_{0,p})^\top$ where $\beta_{0,i}, 1 \le i \le \mathfrak{s}$, is sampled uniformly from $[-1,1]$ and $\beta_{0,i} = 0, \mathfrak{s}+1 \le i \le p$. Varying $(n,p,\mathfrak{s},G)$ and generating 100 realisations for each setting, we record the relative ℓ_2-error $\max_{0 \le k \le n-G} |\boldsymbol{\beta}_0|_2^{-1} |\widehat{\boldsymbol{\beta}}_{k,k+G} - \boldsymbol{\beta}_0|_2$ for each realisation. Then, we obtain a simple rule to determine the finest bandwidth as $G_1 = G_1(n,p) = \lfloor c_0^* \exp(c_1^* \log\log(n) + c_2^* \log\log(p)) \rfloor$ with pre-determined constants $c_i^*, i = 0,1,2$, which are chosen as transforms of the estimated regression coefficients from regressing the 90%-percentile of the logarithm of the estimation errors over 100 realisations, onto the corresponding $\log(G)$, $\log\log(p)$ and $\log\log(n)$ (with $R^2 = 0.8945$). Adopting the Fibonacci rule in Remark 3.2 sometimes gives a sequence of bandwidths that grows too quickly when the sample size n is small. Therefore, with the finest bandwidth G_1 chosen as above, we recommend generating bandwidths as $G_h = \lfloor (h+2)G_1/3 \rfloor$ for $h \ge 2$. Throughout the simulation studies and real data applications, we set $H = 3$.

Selection of D and λ. Theorems 3.1 and 3.4 provide ranges of values for λ and D for theoretical consistency, but they involve unknown parameters as is typically the case in the change point literature. For their simultaneous selection, we adopt a cross validation (CV) method motivated by Zou et al. (2020). Let Λ denote the grid of possible values for λ, which is chosen as an exponentially increasing sequence from $10^{-3}\lambda_{\max}$ up to $\lambda_{\max}$ with $\lambda_{\max} = \max_{0\le k\le n-G} |\sum_{t=k+1}^{k+G} \mathbf{x}_t Y_t|_\infty/\sqrt{G}$ the smallest value with which we obtain $\widehat{\boldsymbol{\beta}}_{k,G} = \mathbf{0}$ for all $0 \le k \le n-G$. For given $G \in \mathscr{G}$ and $\lambda \in \Lambda$, we generate $\widetilde{\Theta}(G,\lambda) = \{\widetilde{\theta}_j(G,\lambda), 1 \le j \le \widetilde{q}_0(G,\lambda)\}$, the set of pre-estimators with $D = 0$, i.e. we take all local maximisers of the MOSUM statistics according to (3.7); due to the detection rule, we always have $\widetilde{q}_0(G,\lambda) \le n/(2\eta G)$. Sorting the elements of $\widetilde{\Theta}(G,\lambda)$ in the decreasing order of the associated MOSUM detector values, we generate a sequence of nested change point models

$$\emptyset = \widetilde{\Theta}_{[0]}(G,\lambda) \subset \widetilde{\Theta}_{[1]}(G,\lambda) \subset \ldots \subset \widetilde{\Theta}_{[\widetilde{q}_0(G,\lambda)]}(G,\lambda) = \widetilde{\Theta}(G,\lambda).$$

Then, using the odd-indexed observations $(Y_t, \mathbf{x}_t)$, $t \in \mathscr{J}_1 = \{2t+1, t = 0, \ldots, \lfloor (n-1)/2 \rfloor\}$, we produce local estimators of the regression parameters and the even-indexed observations $(Y_t, \mathbf{x}_t)$, $t \in \mathscr{J}_0 = \{1, \ldots, n\} \setminus \mathscr{J}_1$, is used for validation. Specifically, we evaluate $\mathrm{CV}(G,\lambda,m) = \mathrm{RSS}_0(\widetilde{\Theta}_{[m]}(G,\lambda),\lambda)$, where for any $\mathscr{L} = \{\ell_j, 1 \le \ell_j \le L : 0 = \ell_0 < \ell_1 < \ldots < \ell_L < \ell_{L+1} = n\}$,

$$\mathrm{RSS}_0(\mathscr{L},\lambda) = \sum_{j=0}^{L} \sum_{\substack{t \in \mathscr{J}_0 \cap \\ \{\ell_j+1,\ldots,\ell_{j+1}\}}} \left(Y_t - \mathbf{x}_t^\top \widehat{\boldsymbol{\beta}}_j^{(1)}(\mathscr{L},\lambda)\right)^2.$$

Here, $\widehat{\boldsymbol{\beta}}_j^{(1)}(\mathscr{L},\lambda)$ denotes the Lasso estimator obtained using $(Y_t, \mathbf{x}_t)$, $t \in \mathscr{J}_1 \cap \{\ell_j, \ldots, \ell_{j+1}\}$ with the penalty parameter λ. Then for each $G_h \in \mathscr{G}$, we find

$$(\lambda^*, m^*) = \operatorname{arg\,min}_{\substack{(\lambda,m):\lambda\in\Lambda, \\ 0\le m\le \widetilde{q}_0(G_h,\lambda)}} \mathrm{CV}(G_h,\lambda,m)$$

and obtain the set of pre-estimators $\widetilde{\Theta}(G_h) = \widetilde{\Theta}_{[m^*]}(G_h,\lambda^*)$ using λ^* and m^*. This amounts to selecting the bandwidth-dependent threshold D at a value just below the m^*th largest MOSUM detector value. Such $\widetilde{\Theta}(G_h)$, $G_h \in \mathscr{G}$, serve as an input to Steps 2–4 of MOSEG.MS. In all numerical experiments reported in this chapter, we set $|\Lambda| = 5$.

Selection of other tuning parameters. For change point estimation, we recommend to use $\eta = 0.5$ in (3.7) based on extensive simulations, which show that the performance of MOSEG and MOSEG.MS is not too sensitive to its choice. As noted in Section 3.4.2, MOSEG.MS is highly competitive computationally against the existing methods even without adopting a coarse grid. Therefore, we report the results obtained with $r = G^{-1}$ (i.e. $\mathscr{T} = \{G, \ldots, n-G\}$ in (3.6)) in the main text and provide the results obtained with $r = 1/10$ in Section 3.5.1, where we observe that adopting a coarse grid does not undermine the performance of MOSEG.

3.4.2 Computational complexity and run time

Recall that Lasso(p) denotes the cost of solving a Lasso problem with p variables. For the coordinate descent algorithm (Friedman et al., 2010), each complete iteration of the coordinate descent has the cost $O(p^2)$. Then, the combined computational cost of Stages 1 and 2 of MOSEG is $O(n(rG)^{-1}\text{Lasso}(p))$, and the memory cost is $O(np)$. Similarly, with the set of bandwidths generated as described in Remark 3.2, the complexity of the multiscale extension MOSEG.MS is $O(n(rG_1)^{-1}\text{Lasso}(p))$ with G_1 denoting the finest scale, which follows from that $\sum_{h=1}^{H} n/(rG_h) \leq n/(rG_1)\sum_{h=1}^{\infty} F_h^{-1} = O(n(rG_1)^{-1})$ (see Remark 3.2 for the notations). The CV outlined in Section 3.4.1, we generate pre-estimators and evaluate the CV objective function on a sequence of nested models for each $\lambda \in \Lambda$, which brings the computational complexity of the complete MOSEG.MS methodology to $O(|\Lambda|n(rG_1)^{-1}\text{Lasso}(p))$.

We investigate the run time of change point detection methodologies for the problem in (3.1). [1] MOSEG (with $G = \lfloor n/6 \rfloor$) and MOSEG.MS are applied with the tuning parameters chosen as in Section 3.4.1 and the finest grid (i.e. $\mathscr{T} = \{G, \ldots, n-G\}$), and we include the CV procedure in run time. For comparison, we consider VPWBS (Wang et al., 2021a), DPDU (Xu et al., 2022) and ARBSEG (Kaul et al., 2019a) applied with the recommended tuning parameters. In particular, VPWBS and DPDU adopt a grid of size 3 for the Lasso tuning parameter while we use the set Λ with $|\Lambda| = 5$. We generate the data as described in the model (M3) in Section 3.4.3 below, with $\delta = 1.6$ and varying (n, p). Figure 3.1 reports the average execution time (in seconds) over 100 realisations for each setting, for the five methods in consideration. In the left panel, we fix $n = 450$ while varying $p \in \{80, 100, \ldots, 220\}$ and in the right, we fix $p = 100$ while varying $n \in \{240, 300, \ldots, 660\}$. Both MOSEG and MOSEG.MS take only a fraction of time taken by the competing methodologies in their computation even without the use of the coarse grid, and their run time does not vary much with increasing n or p in the ranges considered. As expected, MOSEG is faster than MOSEG.MS but the difference in execution time is much smaller than that between MOSEG.MS and other competitors.

3.4.3 Simulation settings

We apply MOSEG.MS to datasets simulated with varying (n, p, s) and change point configurations. In each setting, we generate $\mathbf{x}_t$ as i.i.d. Gaussian random vectors with mean $\mathbf{0}$ and the covariance matrix $\mathbf{\Sigma}_x$ which are specified below, and $\varepsilon_t \sim_{\text{iid}} \mathcal{N}(0, \sigma_\varepsilon^2)$; unless specified otherwise, we use $\sigma_\varepsilon = 1$. We report the results from non-Gaussian and serially dependent data in Section 3.5.2 where overall, the results are not sensitive to tail behaviour or temporal dependence. While we consider $p \geq 100$ in the main text, we report the results when $p = 1000$ in Section 3.5.3 which, together with Section 3.4.2, demonstrate the scalability of MOSEG.MS.

[1] All numerical work reported in this chapter was carried out using the computational facilities of the Advanced Computing Research Centre at the University of Bristol.

The models (M1)–(M3) below are taken from Wang et al. (2021a); in (M2), we adapt their model by randomly generating the set $\mathscr{S}$ on each realisation while in (M3), we consider a broader range of values for δ. In what follows, we assume that for given $\mathscr{S} \subset \{1,\ldots,p\}$ with $|\mathscr{S}| = \mathfrak{s}$, the parameter vector $\boldsymbol{\beta}_0 = (\beta_{0,1},\ldots,\beta_{0,p})^\top \in \mathbb{R}^p$ has $\beta_{0,i} \neq 0$ for $i \in \mathscr{S}$ and $\beta_{0,i} = 0$ otherwise, i.e. $\mathscr{S}$ is the support of $\boldsymbol{\beta}_0$. For each setting, we generate 100 realisations.

(M1) Setting $p = 100$, $q = 3$ and $\boldsymbol{\Sigma}_x = \mathbf{I}$, we vary $n \in \{480, 560, 640, 720, 800\}$ and the change points are located at $\theta_j = jn/4$, $j = 1,2,3$. Fixing $\mathscr{S} = \{1,\ldots,\mathfrak{s}\}$ with $\mathfrak{s} = 4$, we set $\beta_{0,i} = 0.4 \cdot (-1)^{i-1}$ for $i \in \mathscr{S}$ and $\boldsymbol{\beta}_j = (-1)^j \cdot \boldsymbol{\beta}_0$.

(M2) We set $n = 300$, $p = 100$ and $q = 2$, and $\boldsymbol{\Sigma}_x = [0.6^{|i-i'|}]_{i,i'=1}^p$. The change points are located at $\theta_j = jn/3$, $j = 1,2$, and we vary $\mathfrak{s} \in \{10, 20, 30\}$. For each realisation, we randomly draw $\mathscr{S} \subset \{1,\ldots,p\}$ of size $\mathfrak{s}$, and set $\beta_{0,i} = 1/\sqrt{4\mathfrak{s}}$ for $i \in \mathscr{S}$, $\boldsymbol{\beta}_j = (-1)^j \cdot \boldsymbol{\beta}_0$.

(M3) We have $n = 300$, $p = 100$, $q = 2$, $\mathfrak{s} = 10$ and $\boldsymbol{\Sigma}_x = [0.6^{|i-i'|}]_{i,i'=1}^p$. The change points are located at $\theta_j = jn/3$ and fixing $\mathscr{S} = \{1,\ldots,\mathfrak{s}\}$, we set $\beta_{0,i} = \delta \cdot (-1)^{i-1}$ for $i \in \mathscr{S}$ with varying $\delta \in \{0.2, 0.4, 0.8, 1.6\}/\sqrt{\mathfrak{s}}$, and $\boldsymbol{\beta}_j = (-1)^j \cdot \boldsymbol{\beta}_0$.

(M4) We set $n = 840$, $p = 50$, $q = 5$, $\mathfrak{s} = 10$ and $\boldsymbol{\Sigma}_x = \mathbf{I}$. The change points are located at $\theta_1 = 60$, $\theta_2 = 120$, $\theta_3 = 240$, $\theta_4 = 360$ and $\theta_5 = 600$ and fixing $\mathscr{S} = \{1,\ldots,\mathfrak{s}\}$, we set $\beta_{0,i} = \delta \cdot (-1)^{i-1}$ for $i \in \mathscr{S}$ with varying $\delta \in \{0.2, 0.4, 0.8, 1.6\}/\sqrt{\mathfrak{s}}$, and $\boldsymbol{\beta}_1 = -\boldsymbol{\beta}_2 = -2\boldsymbol{\beta}_0$, $\boldsymbol{\beta}_3 = -\boldsymbol{\beta}_4 = -\sqrt{2}\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}_5 = -\boldsymbol{\beta}_6 = -\boldsymbol{\beta}_0$.

(M5) The data is generated as in (M3) except for that $q = 0$, $\boldsymbol{\Sigma}_x = [10^2 \cdot 0.6^{|i-i'|}]_{i,i'=1}^p$ and $\sigma_\varepsilon = 10$, and we use $\delta \in \{1, 1.2, 1.4, 1.6\}$.

In setting (M4), the change points are multiscale in the sense that the size of change and spacing between the change points vary, but $\delta_j^2 \cdot \min(\theta_{j+1} - \theta_j, \theta_j - \theta_{j-1})$ is kept constant for $j = 1,3,5$ and $j = 2,4$, respectively. This results in $\Delta^{(1)}$ in (3.11) being much smaller than $\Delta^{(2)}$ in (3.18). (M5) is designed to test the performance of data segmentation methods when $q = 0$, where we scale the data to examine the sensitivity of the tuning parameter choices discussed in Section 3.4.1.

3.4.4 Simulation results

We apply MOSEG.MS with the tuning parameters selected as described in Section 6.3. For the purpose of illustration only, we also apply MOSEG with the bandwidth chosen with the knowledge of the minimum spacing between the change points; for (M1)–(M3) where change points are evenly spaced, we set $G = 3/4 \cdot \min_{0 \leq j \leq q}(\theta_{j+1} - \theta_j)$. For (M4) with multiscale change points, there does not exist a single bandwidth that works well in detecting all change points so we simply set $G = 125$. For (M5) with $q = 0$, we set $G = G_1$ selected as described in Section 6.3. For comparison, we apply the methods proposed by Wang et al. (2021a) (referred to as VPWBS) and Xu et al.

(2022) (DPDU). The VPWBS method learns the projections well-suited for the detection of the change points and applies the wild binary segmentation algorithm to the projected univariate series, and has been shown to outperform the methods proposed in Leonardi and Bühlmann (2016) and Lee et al. (2016). Based on dynamic programming, the DPDU algorithm minimises the ℓ_0-penalised cost function for multiple change point detection. Both methods have been applied with the default tuning parameters recommended by the authors. We also considered the method proposed by Kaul et al. (2019a) but omit the results due to its poor performance on the simulation models considered here.

In Tables 3.2–3.5, we report the distribution of the bias in change point number estimation ($\widehat{q} - q$) for each method over the 100 realisations generated under each setting. Additionally, we report the scaled Hausdorff distance between the sets of estimated ($\widehat{\Theta}$) and true (Θ) change points, i.e.

$$d_H(\widehat{\Theta}, \Theta) = \frac{1}{n} \max \left\{ \max_{\widehat{\theta} \in \widehat{\Theta}} \min_{\theta \in \Theta} |\widehat{\theta} - \theta|, \max_{\theta \in \Theta} \min_{\widehat{\theta} \in \widehat{\Theta}} |\widehat{\theta} - \theta| \right\}, \tag{3.20}$$

averaged over 100 realisations; by convention, we set $d_H(\emptyset, \Theta) = 1$. We remark that the Hausdorff distance tends to favour the cases when the change points are over-detected, than when they are under-detected. In Table 3.6 (considering the case $q = 0$), we report the proportion of realisations where any false positive is returned.

Generally, as expected, we observe better performance from all methods with increasing sample size in (M1) or increasing change size with δ in (M3)–(M4) while varying the sparsity level $\mathfrak{s}$ brings in less clear change in the performance. In the presence of homogeneous change points under (M1)–(M3), MOSEG performs as well as MOSEG.MS in terms of correctly estimating the number of change points, but it suffers from the lack of adaptivity in the presence of multiscale change points under (M4) where both large frequent shifts and small changes over long intervals are present. Here, we observe the benefit of the multiscale approach taken by MOSEG.MS particularly as δ grows, where it achieves better accuracy in detection and localisation against MOSEG. Comparing the performance of MOSEG.MS and VPWBS, we note that the former generally attains better detection power while the latter exhibits better localisation properties under (M2) and (M3) (when δ is large). DPDU tends to show good detection power in the more challenging scenarios, such as when n is small (under (M1)), $\mathfrak{s}$ is large (under (M2)) or the change size is small (see Table 3.8). At the same time, it is observed to over-estimate the number of change points across all scenarios.

Under (M5), where no changes are present, our methods are shown to control the number of false positives well. Here, we do not include VPWBS or DPDU in Table 3.6 as they tend to detect false positives in most cases.

Table 3.2: (M1) Performance of MOSEG, MOSEG.MS, VPWBS and DPDU over 100 realisations. The best performer in each setting is denoted in bold.

		$\widehat{q}-q$							
n	Method	-3	-2	-1	0	1	2	≥ 3	d_H
480	MOSEG	3	3	6	81	7	0	0	0.0852
	MOSEG.MS	1	6	7	**84**	2	0	0	0.0710
	VPWBS	1	3	14	58	16	5	3	0.0795
	DPDU	0	0	3	80	13	4	0	**0.0405**
560	MOSEG	2	3	5	72	17	1	0	0.0742
	MOSEG.MS	0	1	5	**93**	1	0	0	**0.0299**
	VPWBS	1	0	10	73	5	8	3	0.0579
	DPDU	3	0	1	79	15	2	1	0.0660
640	MOSEG	1	3	5	64	23	4	0	0.0652
	MOSEG.MS	0	1	2	**91**	6	0	0	**0.0203**
	VPWBS	0	1	3	89	3	2	2	0.0291
	DPDU	0	0	0	77	18	5	1	0.0344
720	MOSEG	1	3	1	76	18	1	0	0.0433
	MOSEG.MS	0	0	0	**97**	3	0	0	**0.0104**
	VPWBS	0	0	1	92	3	3	1	0.0190
	DPDU	1	0	0	75	22	2	0	0.0390
800	MOSEG	2	3	7	61	25	2	0	0.0753
	MOSEG.MS	0	0	0	**100**	0	0	0	**0.0073**
	VPWBS	0	0	2	92	3	2	1	0.0202
	DPDU	0	0	0	68	25	7	0	0.0385

Table 3.3: (M2) Performance of MOSEG, MOSEG.MS, VPWBS and DPDU over 100 realisations. The best performer in each setting is denoted in bold.

		$\widehat{q}-q$						
s	Method	-2	-1	0	1	2	≥ 3	d_H
10	MOSEG	27	28	35	10	0	0	0.4204
	MOSEG.MS	11	32	**44**	13	0	0	0.3117
	VPWBS	45	17	11	9	15	3	**0.2465**
	DPDU	55	5	33	7	0	0	0.5981
20	MOSEG	14	31	**50**	5	0	0	0.3016
	MOSEG.MS	8	32	48	12	0	0	0.2726
	VPWBS	44	13	18	13	9	3	**0.2302**
	DPDU	49	3	45	3	0	0	0.5300
30	MOSEG	14	25	50	11	0	0	0.2765
	MOSEG.MS	11	30	41	18	0	0	0.2848
	VPWBS	24	20	33	9	9	5	**0.1843**
	DPDU	26	9	**51**	14	0	0	0.3294

3.5 Additional simulations

3.5.1 Choice of the grid

We investigate the performance of MOSEG as the coarseness of the grid varies with $r \in \{1/3, 1/5, 1/10, 1/G\}$. Recall that when $r = 1/G$, we use the full grid $\mathcal{T} = \{G, \ldots, n-G\}$ in Stage 1

Table 3.4: (M3) Performance of MOSEG, MOSEG.MS, VPWBS and DPDU over 100 realisations. The best performer in each setting is denoted in bold.

		$\widehat{q}-q$						
$\sqrt{10}\delta$	Method	-2	-1	0	1	2	≥ 3	d_H
0.2	MOSEG	12	19	63	6	0	0	0.2367
	MOSEG.MS	4	16	**64**	15	1	0	**0.1609**
	VPWBS	77	9	5	6	1	2	0.3025
	DPDU	4	4	36	24	17	15	0.1779
0.4	MOSEG	7	9	**81**	3	0	0	**0.1401**
	MOSEG.MS	3	21	71	5	0	0	0.1488
	VPWBS	53	20	12	8	4	3	0.2681
	DPDU	0	0	21	23	34	22	0.1498
0.8	MOSEG	6	10	**82**	2	0	0	0.1242
	MOSEG.MS	2	14	77	6	1	0	0.1099
	VPWBS	13	7	58	14	6	2	**0.1061**
	DPDU	0	0	11	29	19	41	0.1644
1.6	MOSEG	3	5	**91**	1	0	0	0.0732
	MOSEG.MS	0	10	88	2	0	0	0.0737
	VPWBS	1	1	84	10	4	0	**0.0404**
	DPDU	0	0	14	12	25	49	0.1682

Table 3.5: (M4) Performance of MOSEG, MOSEG.MS, VPWBS and DPDU over 100 realisations. The best performer in each setting is denoted in bold.

		$\widehat{q}-q$							
$\sqrt{10}\delta$	Method	-3	-2	-1	0	1	2	≥ 3	d_H
0.2	MOSEG	45	6	4	5	21	9	10	0.4518
	MOSEG.MS	17	9	19	**12**	15	8	20	**0.2263**
	VPWBS	95	1	2	1	1	0	0	0.4073
	DPDU	99	0	0	0	1	0	0	0.9578
0.4	MOSEG	44	7	10	10	8	5	16	0.4347
	MOSEG.MS	15	12	16	**17**	7	16	17	**0.2015**
	VPWBS	73	2	7	6	7	5	0	0.3247
	DPDU	80	5	3	5	5	2	2	0.7317
0.8	MOSEG	4	23	31	27	9	5	1	0.1978
	MOSEG.MS	0	3	34	38	15	9	1	0.0834
	VPWBS	13	40	29	11	4	2	1	0.1165
	DPDU	0	0	0	**43**	36	21	8	**0.0629**
1.6	MOSEG	0	7	45	43	3	2	0	0.0970
	MOSEG.MS	0	1	32	**59**	7	1	0	**0.0387**
	VPWBS	3	35	38	19	1	3	1	0.0900
	DPDU	0	0	0	54	32	14	5	0.0444

of MOSEG, see (3.6). For this, we set $n = 300$, $p = 100$, $\mathfrak{s} = 2$ and $q = 1$, and generate the data under (3.1) with $\mathbf{x}_t \sim_{\text{iid}} \mathcal{N}_p(\mathbf{0}, \mathbf{I})$ and $\varepsilon_t \sim_{\text{iid}} \mathcal{N}(0,1)$. For each realisation, the change point θ_1 is randomly sampled from $\{51, \ldots, 250\}$. Varying $\delta \in \{0.1, 0.2, 0.4, 0.8\}$, we generate $\boldsymbol{\beta}_0 = (\beta_{0,1}, \ldots, \beta_{0,p})^\top$ with $\beta_{0,i} = \delta \cdot (-1)^{i-1}$ for $i \in \{1, \ldots, \mathfrak{s}\}$ and have $\boldsymbol{\beta}_1 = -\boldsymbol{\beta}_0$. Setting $G = 50$, we select the maximiser of the MOSUM statistic as the pre-estimator $\widetilde{\theta}_1$ in Stage 1 of MOSEG, which then is refined

Table 3.6: (M5) Proportions of detecting false positives when $q = 0$ for MOSEG and MOSEG.MS over 100 realisations.

	δ			
Method	1	1.2	1.4	1.6
MOSEG	0.05	0.01	0.01	0.02
MOSEG.MS	0.04	0.01	0.01	0.02

as in (3.10) in Stage 2. Table 3.7 reports the average and the standard error of $n^{-1}|\widetilde{\theta}_1 - \theta_1|$ and $n^{-1}|\widehat{\theta}_1 - \theta_1|$ over 100 realisations when different grids are used. See also Figure 3.4 which plots the Hausdorff distance d_H (see (3.20)) against r. When the size of change is very small, estimators from both Stages 1 and 2 perform equally poorly regardless of the choice of r. However, as δ increases, we quickly observe that the estimation error becomes close to zero for the estimators from both stages provided that r is not too large. Also, for $\delta \geq 0.2$, we observe that Stage 2 brings in small improvement in the localisation performance. From this, we conclude that the performance of MOSEG is robust to the choice of r provided that it is chosen reasonably small, say $r \leq 1/5$.

Table 3.7: Comparison of Hausdorff distance d_H for Stage 1 and Stage 2 estimators from MOSEG when different grids are used. The average and the standard error of estimation errors over 100 realisations are reported.

	$r = G^{-1}$				$r = 1/10$			
	Stage 1		Stage 2		Stage 1		Stage 2	
δ	Mean	SD	Mean	SD	Mean	SD	Mean	SD
0.1	0.2043	0.1561	0.2099	0.1670	0.2003	0.1525	0.2096	0.1683
0.2	0.1194	0.1367	0.1149	0.1442	0.1302	0.1402	0.1296	0.1549
0.4	0.0089	0.0104	0.0038	0.0053	0.0115	0.0179	0.0039	0.0050
0.8	0.0070	0.0089	0.0022	0.0052	0.0086	0.0096	0.0020	0.0049
	$r = 1/5$				$r = 1/3$			
	Stage 1		Stage 2		Stage 1		Stage 2	
δ	Mean	SD	Mean	SD	Mean	SD	Mean	SD
0.1	0.2142	0.1525	0.2232	0.1683	0.2179	0.1525	0.2255	0.1683
0.2	0.1383	0.1402	0.1349	0.1549	0.2165	0.1402	0.2061	0.1549
0.4	0.0142	0.0179	0.0039	0.0050	0.2696	0.0179	0.2359	0.0050
0.8	0.0126	0.0096	0.0016	0.0049	0.2303	0.0096	0.1929	0.0049

3.5.2 Heavy-tailedness and temporal dependence

We examine the performance of MOSEG.MS, VPWBS (Wang et al., 2021a), and DPDU (Xu et al., 2022) in the presence of heavy-tailed noise and temporal dependence. For this, we generate datasets with $n = 300$, $p = 100$, $\mathfrak{s} = 10$, $q = 1$ and the two change points are located at $\theta_j = jn/3$, $j = 1, 2$. We use $\boldsymbol{\beta}_0$ obtained as in Section 3.5.1 with $\delta \in \{0.2, 0.4, 0.8, 1.6\}$ and set $\boldsymbol{\beta}_j = (-1)^j \cdot \boldsymbol{\beta}_0$.

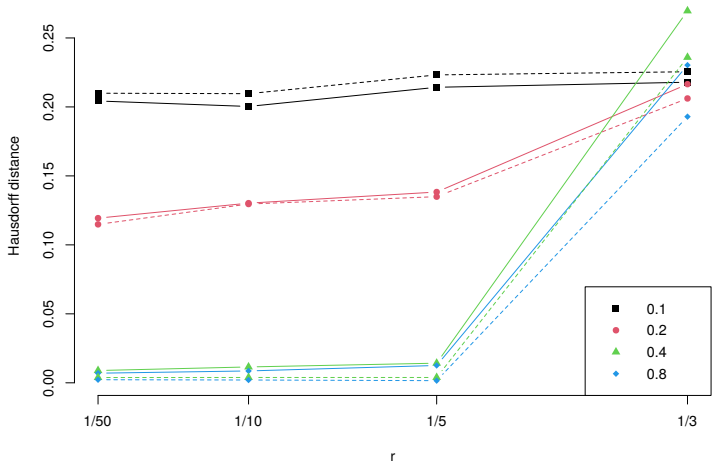


Figure 3.4: Hausdorff distance d_H against r for Stage 1 (solid line) and Stage 2 (dashed line) estimators from MOSEG, as the size of changes varies.

MOSEG.MS is applied with the recommended bandwidth set and the CV-based model selection discussed in Section 3.4.1. We consider the following three settings for the generation of $\mathbf{x}_t$ and ε_t.

(E1) $\mathbf{x}_t \sim_{\text{iid}} \mathcal{N}_p(\mathbf{0}, \mathbf{I})$ and $\varepsilon_t \sim_{\text{iid}} \mathcal{N}(0,1)$ for all t.

(E2) $X_{it} \sim_{\text{iid}} \sqrt{3/5} \cdot t_5$ for all i and t and $\varepsilon_t \sim_{\text{iid}} \sqrt{3/5} \cdot t_5$ for all t.

(E3) $\{(\mathbf{x}_t, \varepsilon_t)\}_{t=1}^n$ is generated as in (3.13) where $\mathbf{D}_1$ is a diagonal matrix with 0.3 on its diagonals, $\mathbf{D}_\ell = \mathbf{O}$ for $\ell \geq 2$ and $\boldsymbol{\zeta}_t \sim_{\text{iid}} \mathcal{N}_{p+1}(\mathbf{0}, \sqrt{1-0.3^2}\mathbf{I})$ for all t.

Under (E2)–(E3), the data is permitted to be heavy-tailed and serially correlated, respectively; (E1) serves as a benchmark. Table 3.8 reports the average and standard error of the Hausdorff distance in (3.20) and $\widehat{q} - q$ over 100 realisations. It shows that generally, neither method is sensitive to heavy-tailedness or temporal dependence. VPWBS shows good localisation performance, while MOSEG.MS tends to achieve better detection accuracy when the size of change is small. DPDU performs very well in the more challenging setting with small δ. However, as noted in Section 3.4, it is more prone to over-estimate the number of change points as δ increases with which the localisation performance also deteriorates.

3.5.3 When the dimensionality is large

We additionally examine the case where $p = 1000$, adopting the simulation setting (E1) from Section 3.5.2. We exclude VPWBS Wang et al. (2021a) which, as shown in Section 3.4.2, tends to take considerably longer time to run compared to MOSEG.MS. Table 3.9 shows that, in comparison to the the results under (E1) in Table 3.8 obtained when $p = 100$, the greater sample

Table 3.8: Performance of MOSEG.MS and VPWBS under (E1)–(E3) over 100 realisations. The best performer in each setting is denoted in bold.

			$\widehat{q}-q$						
δ	Setting	Method	-2	-1	0	1	2	≥ 3	d_H
0.2	(E1)	MOSEG.MS	10	35	46	9	0	0	0.2905
		VPWBS	56	10	9	14	9	2	0.2583
		DPDU	0	0	**86**	12	1	1	**0.0239**
	(E2)	MOSEG.MS	6	53	34	6	1	0	0.2779
		VPWBS	60	10	7	13	10	0	0.2663
		DPDU	0	0	**89**	11	0	0	**0.0168**
	(E3)	MOSEG.MS	8	30	44	17	1	0	0.2644
		VPWBS	10	15	18	13	28	16	0.1671
		DPDU	0	0	**85**	13	2	0	**0.0211**
0.4	(E1)	MOSEG.MS	0	9	86	4	1	0	0.0766
		VPWBS	1	3	**87**	7	2	0	0.0361
		DPDU	0	0	82	16	1	1	**0.0240**
	(E2)	MOSEG.MS	1	11	**83**	5	0	0	0.0830
		VPWBS	1	7	**83**	4	5	0	0.0567
		DPDU	0	0	82	16	2	0	**0.0248**
	(E3)	MOSEG.MS	1	9	**81**	9	0	0	0.0678
		VPWBS	0	1	80	14	4	1	0.0397
		DPDU	0	0	71	24	5	0	**0.0359**
0.8	(E1)	MOSEG.MS	0	0	**99**	1	0	0	0.0119
		VPWBS	0	0	97	3	0	0	**0.0095**
		DPDU	0	0	81	18	1	0	0.0227
	(E2)	MOSEG.MS	0	0	**98**	2	0	0	**0.0100**
		VPWBS	0	0	**98**	2	0	0	0.0104
		DPDU	0	0	80	17	1	2	0.0209
	(E3)	MOSEG.MS	0	1	96	3	0	0	0.0153
		VPWBS	0	0	**98**	2	0	0	**0.0103**
		DPDU	0	0	73	24	3	0	0.0285
1.6	(E1)	MOSEG.MS	0	0	97	3	0	0	0.0097
		VPWBS	0	0	**100**	0	0	0	**0.0037**
		DPDU	0	0	75	22	1	2	0.0309
	(E2)	MOSEG.MS	0	0	**100**	0	0	0	0.0036
		VPWBS	0	0	**100**	0	0	0	**0.0033**
		DPDU	0	0	78	18	1	3	0.0249
	(E3)	MOSEG.MS	0	1	96	3	0	0	0.0076
		VPWBS	0	0	**99**	1	0	0	**0.0045**
		DPDU	0	0	69	24	6	1	0.0307

size is required to detect smaller changes. Also, the localisation performance worsens as p increases. Nonetheless, MOSEG.MS demonstrates itself to be scalable as the dimensionality increases when the size of change is sufficiently large, which is in line with the theoretical requirements.

Table 3.9: Performance of MOSEG.MS under (E1) when $\mathbf{p} = \mathbf{1000}$ over 100 realisations.

	$\widehat{q} - q$						
δ	-2	-1	0	1	2	≥ 3	d_H
0.2	6	47	29	18	0	0	0.3051
0.4	10	34	44	11	1	0	0.2972
0.8	1	22	65	11	1	0	0.1391
1.6	4	3	92	1	0	0	0.0673

3.6 Real data application

There exists an extensive literature on the prediction of the equity premium, which is defined as the difference between the compounded return on the S&P 500 index and the three month Treasury bill rate. Using 14 macroeconomic and financial variables (see Table A.2.1 for full descriptions), Welch and Goyal (2008) demonstrate the difficulty of this prediction problem, in part due to the time-varying nature of the data. Koo et al. (2020) note that the majority of the variables are highly persistent with strong, positive autocorrelations, and develop an ℓ_1-penalised regression method that identifies co-integration relationships among the variables. Accordingly, we transform the data by taking the first difference of any variable labelled as being persistent by Koo et al. (2020), and scale each covariate series to have unit standard deviation. With the thus-transformed variables, we propose to model the monthly equity premium observed from 1927 to 2005 as Y_t, with the 14 variables at lags $1, 2, 3$ and 12 as regressors $\mathbf{x}_t$ via piecewise stationary linear regression; in total, we have $n = 936$ and $p = 57$ including the intercept.

We apply MOSEG.MS with $\mathscr{G} = \{72, 96, 120\}$ in line with the choice described in Section 3.4.1 but we select G_h to be multiples of 12 for interpretability as the observation frequency is monthly. MOSEG.MS returns $\widehat{q} = 7$ change point estimators reported in Table 3.10, and takes 45 seconds in total (including CV). When applied to the same dataset, DPDU takes 25 minutes and VPWBS takes 15 minutes, and neither detects any change point. In Figure 3.5, we plot the local parameter estimates obtained from each of the seven estimated segments. We can relate the change detected in 1954 to the findings reported in Rapach et al. (2010), where they attribute the instability in the pairwise relationships between the equity premium and each of the 14 variables to the Treasury-Federal Reserve Accord and the transition from the wartime economy. Dividend price ratio (`d/p`, at lag two) is active throughout the observation period which agrees with the observations made in Welch and Goyal (2008). They also remark that the recession from 1973 to 1975 due to the Oil Shock drives the good predictive performance of many models proposed for equity premium forecasting, and most perform poorly over the 30 year period (1975–2005) following the Oil Shock. The two last segments defined by the change point estimators reported in Table 3.10 are closely located with these important periods, which supports the validity of the segmentation returned by MOSEG.MS. We note that regardless of the choice of bandwidths, both of the two estimators in 1974 and 1975 defining the two periods are detected separately.

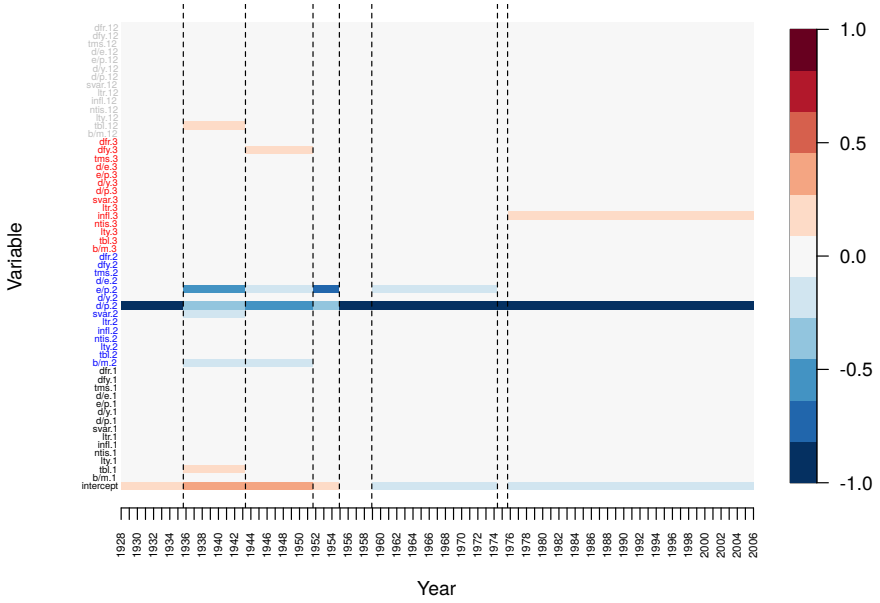


Figure 3.5: Equity premium data: Parameter estimates from each estimated segment obtained by MOSEG.MS. Variables at different lags are coloured differently in the y-axis.

Table 3.10: Equity premium data: Change point estimators detected by MOSEG.MS.

Estimator	$\widehat{\theta}_1$	$\widehat{\theta}_2$	$\widehat{\theta}_3$	$\widehat{\theta}_4$	$\widehat{\theta}_5$	$\widehat{\theta}_6$	$\widehat{\theta}_7$
Date	Oct 1935	Apr 1943	Aug 1951	Nov 1954	Nov 1958	May 1974	Aug 1975

3.7 Conclusions

In this chapter, we propose MOSEG, a high-dimensional data segmentation methodology for detecting multiple changes in the parameters under a linear regression model. It proceeds in two steps, first scanning the data for large changes in local parameter estimators over a moving window, followed by a computational efficient location refinement step. We further propose its multiscale extension, MOSEG.MS, which alleviates the necessity to select a single bandwidth. Both numerically and theoretically, we demonstrate the efficiency of the proposed methodologies. Computationally, they are highly competitive thanks to the careful design of the algorithms that limit the required number of Lasso estimators. theoretically, we show the consistency of MOSEG and MOSEG.MS in a general setting permitting serial dependence and heavy tails and establish their (near-)minimax optimality under Gaussianity. In particular, the consistency of MOSEG.MS is derived for a parameter space that simultaneously permits large changes over short intervals and small changes over long stretches of stationarity, which is much broader than that typically adopted in the literature. Comparative simulation studies and findings from the application of MOSEG.MS to equity premium data support its efficacy.

CHAPTER 4

MOVING SUM DATA SEGMENTATION FOR VECTOR AUTOREGRESSIVE TIME SERIES

4.1 Introduction

Time series data are often sampled in settings which are, by their nature, non-stationary. Settings studied in the literature include genomics (Olshen et al., 2004), macroeconomics (Stock and Watson, 1996), finance (Koo and Seo, 2015; Nasiadka et al., 2022), neuroscience (Bai et al., 2021), and remote sensing (Palm et al., 2018; Verbesselt et al., 2010). A simple way to account for changing model behaviour is to allow the model to vary in a piecewise constant manner between unknown change points; to estimate the number and locations of these is the problem of change point analysis, or data segmentation. See Niu et al. (2016) and Cho and Kirch (2021a) for reviews.

Real time series data tend to exhibit dependence in time and space, so it is of interest to identify changes in the second-order structure, such as the covariance (Aue et al., 2009) or dependence properties (Cho and Fryzlewicz, 2015). Vector autoregressive (VAR) models are popular parametric models for multiple time series, which capture linear dynamics. These are simple to interpret, estimate, and forecast from. Under stationarity assumptions they may used, for example, in biology for learning causal networks (Opgen-Rhein and Strimmer, 2007), in finance for forecasting (Ang and Piazzesi, 2003; Barigozzi et al., 2023), or in economics for impulse response analysis (Stock and Watson, 2001).

Testing for the presence of a single change point under a VAR model has been considered in many papers; Dvořák and Prášková (2013) and Dvořák (2017) use Wald-type statistics, and Kirch et al. (2015) use score statistics. In the multiple change setting, Bai (2000) uses a penalised likelihood approach for estimation, which is generalised by Qu and Perron (2007) to offer a range

of tests; the dynamic programming algorithm used for estimation incurs a cost which is quadratic in the length of the series (Bai and Perron, 2003). Other approaches include Bayesian methods (Ahelegbey et al., 2021) and regularised estimation (Li et al., 2020), where the low-dimensional problem is cast as solving a single high-dimensional regression problem. We briefly mention methods designed for the high-dimensional asymptotic setting where the number of series is allowed to grow with respect to the sample size, often under sparsity (Cho and Owens, 2022; Safikhani and Shojaie, 2022), low-rank (Enikeeva et al., 2023), or both (Bai et al., 2020b, 2023) assumptions on VAR parameter matrices. We also mention methods for univariate autoregressive series (Davis et al., 1995, 2006; Gombay, 2008; Gombay and Serban, 2009). See Section 2.1.2 for further discussion.

This chapter proposes a method which contrasts functions of the data on a G-sized moving window either side of a candidate change point $k \in \{G, G+1, \ldots, n-G\}$. Other moving window approaches have been used in, for example, Bauer and Hackl (1980); Cho and Owens (2022); Eichinger and Kirch (2018); Hušková (1990); McGonigle and Cho (2023); Preuss et al. (2015b); Yau and Zhao (2016). Kirch and Reckrühm (2022) propose a general moving sum (MOSUM) procedure based on estimating function. We adapt their moving window-based approach to change point detection under a time-varying VAR model. Their work develops a general framework for change point detection under generic estimating functions, providing theoretical guarantees for consistency of the estimating procedure.

We highlight the following contributions made in this work, in particular relatively to Kirch and Reckrühm (2022).

(i) **Multiple change points in a piecewise stationary VAR model.** We show that our method consistently estimates the number and locations (if any) of the change points, matching the best localisation rates available in the literature. We provide asymptotic guarantees for family-wise error control when no changes are present, and for consistency of detection when they are.

(ii) **Computational efficiency.** By scanning over a coarse grid $\mathscr{T} \subset \{1, \ldots, n\}$ we greatly reduce the number of times we must obtain parameter estimates, reducing the complexity to be sub-linear in the sample size. We then perform a cheap localisation step to obtain accurate estimators. For all the methods we propose, a C++ implementation is available that may be called in R.

(iii) **Algorithmic extensions.** We devise a range of extensions to the procedure, some of which apply to generic moving sum procedures. A multiple-bandwidth algorithm allows detection in the multiscale setting, where we have both large frequent and small infrequent changes in the same sample. Furthermore, we devise a recursive segmentation strategy to detect changes which are not detectable under a global inspection parameter. Specifically for the

VAR setting, we propose dimension reduction methods and a parametric bootstrap for the analysis of larger panels.

The Chapter is organised as follows: In Section 4.2 we introduce a piecewise stationary VAR model. In Section 4.3, we propose a moving window procedure with a score-type detector for data segmentation. We define estimators, prove these procedures have asymptotic size control and power 1, and give estimation guarantees for the number and locations of changes. In Section 4.4 we extend the procedure with data-adaptive algorithms, making changes detectable under a less stringent condition on the jump size, and allowing for multiscale changes. We make further proposals in Section 4.5 to address computational issues, scanning over a coarse grid to reduce cost and using projections to estimate models with many parameters. We validate our methods numerically in Section 4.6 with in-depth simulation studies and two applications to real datasets on air quality measurements and macroeconomic series.

Notation Let $\mathbb{R}$, $\mathbb{Z}$, and $\mathbb{N}$ denote the sets of real numbers, integers and natural numbers. We let $\boldsymbol{O}$ and $\mathbf{0}$ be a matrix and vector of zeros, respectively, and $\boldsymbol{I}_p$ be the $p \times p$ identity matrix. Let $\|\cdot\|$ denote the Euclidean norm of a vector or the Spectral norm of a matrix, and let $\|\cdot\|_F$ denote the Frobenius norm. We denote the Kronecker product of two matrices $\boldsymbol{A} = [a_{ij}] \in \mathbb{R}^{m\times n}$ and $\boldsymbol{B}$, by

$$\boldsymbol{A} \otimes \boldsymbol{B} = \begin{pmatrix} a_{11}\boldsymbol{B} & a_{12}\boldsymbol{B} & \dots & a_{1n}\boldsymbol{B} \\ a_{21}\boldsymbol{B} & a_{22}\boldsymbol{B} & & \vdots \\ \vdots & & \ddots & \\ a_{m1}\boldsymbol{B} & \dots & & a_{mn}\boldsymbol{B} \end{pmatrix}.$$

For a sequence of random variables $\{X_n\}_{n\in\mathbb{N}}$, let $X_n \xrightarrow{\mathscr{D}} X$ and $X_n \xrightarrow{P} X$ denote convergence in distribution and probability, respectively. We denote by Γ the Gamma function.

4.2 Piecewise stationary VAR model

We observe $\{\boldsymbol{X}_t\}_{t=1}^n$ consisting of time-ordered vectors $\boldsymbol{X}_t = (X_{1t}, X_{2t}, \dots, X_{pt})^\top \in \mathbb{R}^p$, which follow a piecewise stationary VAR model

$$\boldsymbol{X}_t = \begin{cases} \boldsymbol{X}_t^{(1)}, & k_0 + 1 = 1 \le t \le k_1, \\ \boldsymbol{X}_t^{(2)}, & k_1 + 1 \le t \le k_2, \\ \vdots & \\ \boldsymbol{X}_t^{(q+1)}, & k_q + 1 \le t \le k_{q+1} = n, \end{cases} \tag{4.1}$$

where each $\{\boldsymbol{X}_t^{(j)}\}_{t\in\mathbb{Z}}$ is a stationary VAR(d) process (in the sense of (2.4), or Equation (2.1.9) of Lütkepohl (2005)), i.e.

$$\boldsymbol{X}_t^{(j)} = \boldsymbol{a}_j \mathbb{X}_{t-1}^{(j)} + \boldsymbol{\varepsilon}_t, \text{ where } \boldsymbol{a}_j = \begin{bmatrix} \boldsymbol{a}_j(1)^\top \\ \vdots \\ \boldsymbol{a}_j(p)^\top \end{bmatrix} \in \mathbb{R}^{p\times(dp+1)} \text{ and } \mathbb{X}_{t-1}^{(j)} = \begin{bmatrix} 1 \\ \mathbb{X}_{1,t-1}^{(j)} \\ \vdots \\ \mathbb{X}_{p,t-1}^{(j)} \end{bmatrix} \in \mathbb{R}^{dp+1}$$

for $j = 1,\ldots,q+1$. Here, $\mathbb{X}_{i,t-1}^{(j)} = (X_{i,t-1}^{(j)},\ldots,X_{i,t-d}^{(j)})^\top$ collects the d lagged values of $X_{it}^{(j)}$, the i-th channel of $\boldsymbol{X}_t^{(j)}$, and $\boldsymbol{a}_j(i)$ collects the parameters involved in predicting the i-th channel. There are q change points at unknown locations k_j, $1 \le j \le q$, such that $\boldsymbol{a}_j \neq \boldsymbol{a}_{j+1}$ for all j. We assume that $k_j = \lfloor \lambda_j n \rfloor$ for $0 = \lambda_0 < \lambda_1 < \ldots, \lambda_q < \lambda_{q+1} = 1$, with q treated as being fixed. Our aim is to estimate the total number and the locations of the q change points. We require $\{\boldsymbol{\varepsilon}_t\}_{t=1}^n$ to be a zero-mean, independent process such that $\mathbb{E}(\boldsymbol{\varepsilon}_t) = \boldsymbol{0}$, $\mathrm{Cov}(\boldsymbol{\varepsilon}_t) = \boldsymbol{S}$ for some positive definite matrix $\boldsymbol{S} \in \mathbb{R}^{p\times p}$, and $\mathrm{Cov}(\boldsymbol{\varepsilon}_t, \boldsymbol{\varepsilon}_{t'}) = \boldsymbol{O}$ for any $t \neq t'$. We might allow the innovation covariance matrix $\boldsymbol{S}$ to vary from one segment to another, but our focus is on detecting changes due to shifts in the VAR parameters $\boldsymbol{a}_j$.

4.3 Data segmentation methodology

4.3.1 Moving sum procedure

We introduce a score-type statistic computed on a moving window for detecting changes under the model (4.1). Referred to as a MOSUM score statistic, the detector statistic is adopted in Kirch and Reckrühm (2022) for multiple change point detection in general data segmentation problems, which we apply here to the piecewise stationary VAR process. This scans for large discrepancies in the moving sums of the estimating function for the method of least squares. We begin by specifying the estimating function

$$\boldsymbol{H}_t(\widetilde{\boldsymbol{a}}) = \boldsymbol{H}(\boldsymbol{X}_t, \mathbb{X}_{t-1}, \widetilde{\boldsymbol{a}}) = -(\boldsymbol{X}_t - \widetilde{\boldsymbol{a}}\mathbb{X}_{t-1}) \otimes \mathbb{X}_{t-1} \in \mathbb{R}^{p(dp+1)} \tag{4.2}$$

which corresponds to the least squares objective. We let $\widetilde{\boldsymbol{a}}$ be an inspection parameter which may be data-dependent, i.e. one which we will "plug in" to the function in order to reveal changes. We propose to scan the data with the statistic

$$\widehat{T}_k(G, \widetilde{\boldsymbol{a}}) = \frac{1}{\sqrt{2G}} \left\| (\widehat{\boldsymbol{\Sigma}}_k(\widetilde{\boldsymbol{a}}))^{-1/2} \boldsymbol{m}_k(G, \widetilde{\boldsymbol{a}}) \right\| \text{ for } k = G+d,\ldots,n-G, \text{ where} \tag{4.3}$$

$$\boldsymbol{m}_k(G, \widetilde{\boldsymbol{a}}) = \sum_{t=k+1}^{k+G} \boldsymbol{H}_t(\widetilde{\boldsymbol{a}}) - \sum_{t=k-G+1}^{k} \boldsymbol{H}_t(\widetilde{\boldsymbol{a}}),$$

and $\widehat{\boldsymbol{\Sigma}}_k(\widetilde{\boldsymbol{a}})$ estimates $\boldsymbol{\Sigma}_k(\widetilde{\boldsymbol{a}}) = \boldsymbol{\Sigma}_{(j)}(\widetilde{\boldsymbol{a}}) = \mathrm{Cov}(\boldsymbol{H}(\boldsymbol{X}_1^{(j)}, \mathbb{X}_0^{(j)}, \widetilde{\boldsymbol{a}}))$ for $k \in \{k_j+1,\ldots,k_{j+1}\}$. The proposed detector statistic $\widehat{T}_k(G, \widetilde{\boldsymbol{a}})$ compares the scaled sums of the estimating function evaluated at the inspection parameter $\widetilde{\boldsymbol{a}}$. We discuss the estimation of $\boldsymbol{\Sigma}_k(\widetilde{\boldsymbol{a}})$ in Section 4.3.2.

With the maximum detector statistic $\widehat{T}(G,\widetilde{\boldsymbol{a}}) = \max_{G\le k\le n-G}\widehat{T}_k(G,\widetilde{\boldsymbol{a}})$, we test the null hypothesis of no change, $H_0 : q = 0$, against the alternative $H_1 : q \ge 1$, as:

$$\text{Reject } H_0 \text{ if } \widehat{T}(G,\widetilde{\boldsymbol{a}}) > D(G,\alpha). \tag{4.4}$$

Here, the critical value $D(G,\alpha)$ denotes the critical value derived from the asymptotic null distribution of $\widehat{T}(G,\widetilde{\boldsymbol{a}})$ at a given significance level $\alpha \in (0,1)$, see (4.11) below.

When H_0 is rejected and there is evidence for one or more change points, our aim is to estimate their number and locations. With an appropriately chosen $\widetilde{\boldsymbol{a}}$, we expect $\widehat{T}_k(G,\widetilde{\boldsymbol{a}})$ to take large values around the change points as illustrated in Figure 4.1. As such, we estimate k_j with the locations of significant local maximisers. To automatically identify these, we adopt a criterion proposed in Eichinger and Kirch (2018). We consider all pairs of indices (v_j, w_j) that simultaneously satisfy

$$\begin{aligned} &\widehat{T}_k(G,\widetilde{\boldsymbol{a}}) > D(G,\alpha) \text{ for } v_j \le k \le w_j, \text{ and} \\ &\widehat{T}_k(G,\widetilde{\boldsymbol{a}}) \le D(G,\alpha) \text{ for } k = v_j - 1, w_j + 1, \end{aligned} \tag{4.5}$$

with $w_j - v_j \ge \epsilon G$ for some $\epsilon \in (0, 1/2)$. We take the total number of these pairs as an estimator of q:

$$\widehat{q} = \widehat{q}(\widetilde{\boldsymbol{a}}) = \text{ number of pairs } (v_j, w_j),$$

and for each $j = 1,\dots,\widehat{q}$, we estimate the location of a change point by

$$\widehat{k}_j = \widehat{k}_j(\widetilde{\boldsymbol{a}}) = \arg\max_{v_j\le k\le w_j}\widehat{T}_k(G,\widetilde{\boldsymbol{a}}).$$

The criterion in (4.5) selects local maximisers of $\widehat{T}_k(G,\widetilde{\boldsymbol{a}})$, over intervals of length greater than ϵG, as change point estimators. An alternative approach is to take any local maximiser of $\widehat{T}_k(G,\widetilde{\boldsymbol{a}})$ over an interval of length proportional to G, at which the MOSUM statistic exceeds the critical value. This criterion, referred to as the η-criterion in Meier et al. (2021), is shown to be less conservative in comparison with ϵ-criterion; for further discussion, see Appendix B.4.

It remains to select an inspection parameter $\widetilde{\boldsymbol{a}}$. Let $\widehat{\boldsymbol{a}}_{s,e}$ be the unique solution satisfying $\sum_{t=s}^{e}\boldsymbol{H}_t(\widehat{\boldsymbol{a}}_{s,e}) = \boldsymbol{0}$ for any $1 \le s < e \le n$ with $e - s + 1 \ge p(dp+1)$. This contains solutions to the least squares problem for each channel, which is

$$\widehat{\boldsymbol{a}}_{s,e}(i) = \left(\sum_{t=s}^{e} X_{it}\mathbb{X}_{t-1}^{\top}\right)\left(\sum_{t=s}^{e}\mathbb{X}_{t-1}\mathbb{X}_{t-1}^{\top}\right)^{-1}, i = 1,\dots,p. \tag{4.6}$$

Kirch and Reckrühm (2022) suggest to adopt $\widehat{\boldsymbol{a}}_{1,n}$, the global parameter estimator, as the inspection parameter. This choice guarantees the family-wise error control under $H_0 : q = 0$, but it may lack power to detect some changes as discussed in Remark 4.1 below. In Section 4.4.1, we describe an adaptation of the MOSUM procedure which adaptively selects the local parameter estimates as the inspection parameter. An alternative approach is to directly compare local VAR parameter estimates to scan for multiple change points. Referred to as the MOSUM Wald procedure, we describe this method in Appendix B.1.

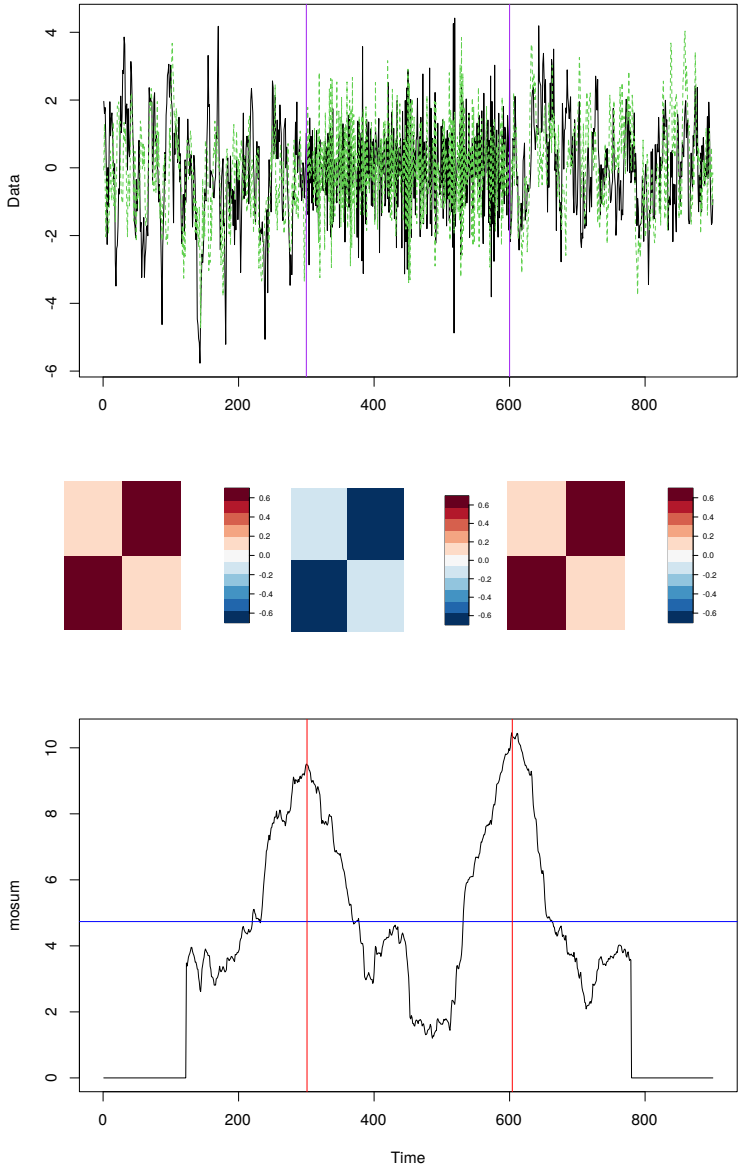


Figure 4.1: Top: A realisation from a piecewise stationary bivariate VAR(1) model with changes at $k_1 = 300$ and $k_2 = 600$ (denoted by vertical lines), where each series is differently coloured. Middle: The time-varying VAR parameters in each regime. Bottom: MOSUM score statistic $\widehat{T}_k(G, \widetilde{a}), G \leq k \leq n - G$ with $G = 120$ and the inspection parameter $\widetilde{a}$ obtained as the global least squares estimator. The threshold $D(G, \alpha)$ with $\alpha = 0.05$ is denoted by the horizontal line and the change point estimators $\widehat{k}_1$ and $\widehat{k}_2$ by the vertical lines.

4.3.2 Estimation of $\boldsymbol{\Sigma}_k(\widetilde{\boldsymbol{a}})$

We propose two estimators of $\boldsymbol{\Sigma}_k(\widetilde{\boldsymbol{a}})$. The first estimator is formed by combining an estimator for $\boldsymbol{C}_{(j)} = \mathrm{Cov}(\mathbb{X}_t^{(j)})$ and that for $\boldsymbol{S}(\widetilde{\boldsymbol{a}}) = \mathrm{Cov}(\widehat{\boldsymbol{\varepsilon}}_t(\widetilde{\boldsymbol{a}}))$ where $\widehat{\boldsymbol{\varepsilon}}_t(\widetilde{\boldsymbol{a}}) = \boldsymbol{X}_t - \widetilde{\boldsymbol{a}}\mathbb{X}_{t-1}$, such that

$$\widehat{\boldsymbol{\Sigma}}_k^{(1)}(\widetilde{\boldsymbol{a}}) = \widehat{\boldsymbol{S}}_k(\widetilde{\boldsymbol{a}}) \otimes \widehat{\boldsymbol{C}}_{k-G+1,k+G}. \tag{4.7}$$

This choice is motivated by noting that, when $\widetilde{\boldsymbol{a}} = \boldsymbol{a}_1$ and $q = 0$, we have that $\mathbb{X}_t$ and $\widehat{\boldsymbol{\varepsilon}}_t = \boldsymbol{\varepsilon}_t$ are independent. Let the local estimator of $\mathrm{Cov}(\mathbb{X}_t)$ be

$$\widehat{\boldsymbol{C}}_{s,e} = \frac{1}{e-s+1}\sum_{t=s}^{e} \mathbb{X}_{t-1}\mathbb{X}_{t-1}^\top.$$

For an estimator of $\boldsymbol{S}(\widetilde{\boldsymbol{a}})$, we consider

$$\begin{aligned}\widehat{\boldsymbol{S}}_k(\widetilde{\boldsymbol{a}}) = \frac{1}{2G}\Bigg(&\sum_{t=k-G+1}^{k} \big(\widehat{\boldsymbol{\varepsilon}}_t(\widetilde{\boldsymbol{a}}) - \bar{\boldsymbol{\varepsilon}}_{k-G+1,k}(\widetilde{\boldsymbol{a}})\big)\big(\widehat{\boldsymbol{\varepsilon}}_t(\widetilde{\boldsymbol{a}}) - \bar{\boldsymbol{\varepsilon}}_{k-G+1,k}(\widetilde{\boldsymbol{a}})\big)^\top \\ &+ \sum_{t=k+1}^{k+G} \big(\widehat{\boldsymbol{\varepsilon}}_t(\widetilde{\boldsymbol{a}}) - \bar{\boldsymbol{\varepsilon}}_{k+1,k+G}(\widetilde{\boldsymbol{a}})\big)\big(\widehat{\boldsymbol{\varepsilon}}_t(\widetilde{\boldsymbol{a}}) - \bar{\boldsymbol{\varepsilon}}_{k+1,k+G}(\widetilde{\boldsymbol{a}})\big)^\top\Bigg), \\ &\text{where } \bar{\boldsymbol{\varepsilon}}_{s,e}(\widetilde{\boldsymbol{a}}) = \frac{1}{e-s+1}\sum_{t=s}^{e}\widehat{\boldsymbol{\varepsilon}}_t(\widetilde{\boldsymbol{a}}).\end{aligned} \tag{4.8}$$

Alternatively, we estimate $\boldsymbol{\Sigma}_k(\widetilde{\boldsymbol{a}})$ directly by the sample covariance of $\boldsymbol{H}_t(\widetilde{\boldsymbol{a}})$ locally with

$$\begin{aligned}\widehat{\boldsymbol{\Sigma}}_k^{(2)}(\widetilde{\boldsymbol{a}}) = \frac{1}{2G}\Bigg[\Bigg(&\sum_{t=k-G+1}^{k} \big(\boldsymbol{H}_t(\widetilde{\boldsymbol{a}}) - \bar{\boldsymbol{H}}_{k-G+1,k}(\widetilde{\boldsymbol{a}})\big)\big(\boldsymbol{H}_t(\widetilde{\boldsymbol{a}}) - \bar{\boldsymbol{H}}_{k-G+1,k}(\widetilde{\boldsymbol{a}})\big)^\top \\ &+ \sum_{t=k+1}^{k+G} \big(\boldsymbol{H}_t(\widetilde{\boldsymbol{a}}) - \bar{\boldsymbol{H}}_{k+1,k+G}\big)\big(\boldsymbol{H}_t(\widetilde{\boldsymbol{a}}) - \bar{\boldsymbol{H}}_{k+1,k+G}(\widetilde{\boldsymbol{a}})\big)^\top\Bigg)\Bigg], \\ &\text{where } \bar{\boldsymbol{H}}_{s,e}(\widetilde{\boldsymbol{a}}) = \frac{1}{e-s+1}\sum_{t=s}^{e}\boldsymbol{H}_t(\widetilde{\boldsymbol{a}}).\end{aligned} \tag{4.9}$$

We demonstrate that these estimators are consistent in the sense of Assumption 4.7 in Appendix B.3.5.

4.3.3 Theoretical properties

In this section we show that under general conditions, the score procedure consistently estimates the number and locations of changes. These results are consequences of the more general framework developed in Kirch and Reckrühm (2022).

We begin by stating our modelling assumptions.

Assumption 4.1. The data are generated according to (4.1). Moreover, there exists $\widetilde{\nu} > 0$ such that for each $j = 1, \ldots, q+1$, it holds that $0 < \mathbb{E}\|\mathbb{X}_1^{(j)}\|^{4+\widetilde{\nu}} < \infty$, and $\mathrm{Cov}(\mathbb{X}_1^{(j)}) = \boldsymbol{C}_{(j)}$ is a positive definite matrix.

Assumption 4.2. $\{\boldsymbol{\varepsilon}_t\}_{t=1}^n$ is a sequence of independent random vectors in $\mathbb{R}^p$ such that $\mathbb{E}(\boldsymbol{\varepsilon}_t) = \boldsymbol{0}$, and $\boldsymbol{S} = \mathbb{E}(\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}_t^\top)$ is a $p \times p$ nonsingular, symmetric, positive semi-definite matrix.

Assumption 4.3. Let $\mathscr{F}_s^e$ be the filtration of the sequence $\{\boldsymbol{X}_t\}_{t=s}^e$. Defining

$$\alpha(n) = \sup_{j\in\mathbb{N}} \sup_{A\in\mathscr{F}_1^j, B\in\mathscr{F}_{j+n}^\infty} |\mathsf{P}(A\cap B) - \mathsf{P}(A)\mathsf{P}(B)|,$$

we let $\alpha(n) = O(n^{-b})$ for some $b > 1 + 2/\widetilde{\nu}$.

Assumption 4.1 requires that $\mathbb{X}_t^{(j)}$ has at least four finite moments. By Lemma E.1 of Barigozzi et al. (2023), this condition will hold if the moment condition is made on the innovation process, as well a bound on $\|\boldsymbol{S}\|$, and that the VAR parameter permits a moving average representation with absolutely summable coefficients. Assumption 4.2 states that the errors form an independent process, which permits asymmetry and heavy tails. As a direct consequence of these conditions, each series $\{\boldsymbol{H}(\boldsymbol{X}_t^{(j)}, \mathbb{X}_{t-1}^{(j)}, \widetilde{\boldsymbol{a}})\}_{t=1}^n$ has a positive definite covariance matrix $\boldsymbol{\Sigma}_{(j)}(\widetilde{\boldsymbol{a}}) = \boldsymbol{S}(\widetilde{\boldsymbol{a}}) \otimes \boldsymbol{C}_{(j)}$. Assumption 4.3 states that the observed series is strongly mixing at a rate which depends on the number of moments of $\mathbf{X}_t$.

Assumption 4.4. Let the bandwidth G depend on n, i.e. $G = G(n)$.

(a) For $\widetilde{\nu} > 0$ assume that

$$\frac{n}{G} \to \infty \text{ and } \frac{n^{\frac{2}{2+\widetilde{\nu}}}\log(n)}{G} \to 0 \text{ for } n \to \infty.$$

(b) The minimum distance between change points is larger than $2G$ as $n \to \infty$, so that

$$\liminf_{n\to\infty} \min_{j=1,\ldots,q+1} \frac{k_j - k_{j-1}}{G} > 2.$$

Assumption 4.5. Let $\widetilde{\boldsymbol{a}}$ be fixed. For $j = 1,\ldots,q$, define the jump size $\delta_j(\widetilde{\boldsymbol{a}}) = \|\boldsymbol{d}_j(\widetilde{\boldsymbol{a}})\|$, where

$$\boldsymbol{d}_j(\widetilde{\boldsymbol{a}}) = \mathbb{E}\left(\boldsymbol{H}(\boldsymbol{X}_t^{(j+1)}, \mathbb{X}_{t-1}^{(j+1)}, \widetilde{\boldsymbol{a}})\right) - \mathbb{E}\left(\boldsymbol{H}(\boldsymbol{X}_t^{(j)}, \mathbb{X}_{t-1}^{(j)}, \widetilde{\boldsymbol{a}})\right). \tag{4.10}$$

We let $\min_{1\le j\le q} \delta_j(\widetilde{\boldsymbol{a}}) \ge c_{\delta,n} > 0$, where $c_{\delta,n} \cdot \sqrt{\frac{G}{\log(n/G)}} \to \infty$.

Assumption 4.4 requires that the bandwidth for the MOSUM procedure grows with respect to the sample size and inversely to the number of moments, and guarantees that, asymptotically, a detector at any $k = G,\ldots,n-G$ draws observations from at most two segments. Assumption 4.5 allows the size of each jump to tend to zero.

Remark 4.1 (Detectability)**.** Changes can only be detected by the score procedure (using the inspection parameter $\widetilde{\boldsymbol{a}}$) when a change in the expectation of the estimating function occurs, i.e. $\delta_j(\widetilde{\boldsymbol{a}}) \ge c_{\delta,n}$ as per Assumption 4.5. We can contrive examples where this is not the case, e.g. $\boldsymbol{a}_j \neq \boldsymbol{a}_{j+1}$ but

$$\boldsymbol{d}_j(\widetilde{\boldsymbol{a}}) = (\widetilde{\boldsymbol{a}} - \boldsymbol{a}_{j+1})\boldsymbol{C}_{(j+1)} - (\widetilde{\boldsymbol{a}} - \boldsymbol{a}_j)\boldsymbol{C}_{(j)} = \boldsymbol{0},$$

so $\delta_j(\widetilde{\boldsymbol{a}}) = 0$. For example, in the univariate case this is solved with the inspection parameter

$$\widetilde{a} = \left(\frac{1}{1-a_{j+1}^2} - \frac{1}{1-a_j^2}\right)^{-1}\left(\frac{a_{j+1}}{1-a_{j+1}^2} - \frac{a_j}{1-a_j^2}\right).$$

If we use $\widetilde{\boldsymbol{a}} = \boldsymbol{a}_{1,n}$, the weighted average of $\boldsymbol{a}_j, j = 1,\dots,q+1$ defined in (4.12), by Lemma B.11 it holds that $\delta_j(\widetilde{\boldsymbol{a}}) > 0$ for at least one $j \in \{1,\dots,q\}$. In practice, we may use $\widetilde{\boldsymbol{a}} = \widehat{\boldsymbol{a}}_{1,n}$ in (4.6), which by Lemma B.9 is $\sqrt{n}$-consistent for $\boldsymbol{a}_{1,n}$. This motivates the data-adaptive extension proposed in Section 4.4.1, which allows for the detection of a broader class of changes with a relaxed condition on the jump size.

Assumption 4.6. Define $c_\alpha = -\log\log(1-\alpha)^{-1/2}$, which is the $(1-\alpha)$ quantile of the Gumbel type 2 distribution. Let the sequence $\{\alpha_n\}_{n\in\mathbb{N}}$ fulfil

$$\alpha_n \to 0 \quad \text{and} \quad \frac{c_{\alpha_n}}{a(n/G)\sqrt{G}} = o(1).$$

Assumption 4.6 is a technical condition in which the significance level is not fixed but converging to 0.

Assumption 4.7. (a) For all $j = 1,\dots,q+1$, it holds that

$$\max_{k_{j-1}+G\le k\le k_j-G} \left\|(\widehat{\boldsymbol{\Sigma}}_k(\widetilde{\boldsymbol{a}}))^{-1/2} - (\boldsymbol{\Sigma}_k(\widetilde{\boldsymbol{a}}))^{-1/2}\right\|_F = o_P\left(\log(n/G)^{-1}\right).$$

(b) For each $j = 1,\dots,q$, it holds that

$$\max_{k:|k-k_j|<G} \left\|(\widehat{\boldsymbol{\Sigma}}_k(\widetilde{\boldsymbol{a}}))^{1/2}\right\|_F < \infty \text{ and } \max_{k:|k-k_j|<G} \left\|(\widehat{\boldsymbol{\Sigma}}_k(\widetilde{\boldsymbol{a}}))^{-1/2}\right\|_F < \infty,$$

At k sufficiently far from any change point (i.e. $|k-k_j| \ge G$), the estimators $\widehat{\boldsymbol{\Sigma}}_k(\widetilde{\boldsymbol{a}})$ are assumed to be consistent, while for k satisfying $|k-k_j| < G$, we only require its finiteness. The estimators proposed in Section 4.3.2 are shown to meet Assumption 4.7 in Appendix B.3.5.

We define the detector

$$T(G,\widetilde{\boldsymbol{a}}) = \max_{G\le k\le n-G} T_k(G,\widetilde{\boldsymbol{a}}), \quad T_k(G,\widetilde{\boldsymbol{a}}) = \frac{1}{\sqrt{2G}}\left\|(\boldsymbol{\Sigma}_k(\widetilde{\boldsymbol{a}}))^{-1/2}\boldsymbol{m}_k(G,\widetilde{\boldsymbol{a}})\right\|,$$

which differs from (4.3) by using the infeasible $\boldsymbol{\Sigma}_k(\widetilde{\boldsymbol{a}})$. Proposition 4.1 derives the asymptotic null distribution of the proposed test statistic. Proofs can be found in Section B.3.1; we note that the proof is structured as a verification of the conditions for the corresponding Theorem 2.1 of Kirch and Reckrühm (2022), which establishes the result more generally.

Proposition 4.1. Let Assumptions 4.1–4.4 hold. Let $\widetilde{\boldsymbol{a}}$ be some fixed parameter.

(a) Under H_0, we have

$$a(n/G)T(G,\widetilde{\boldsymbol{a}}) - b(n/G) \xrightarrow{\mathscr{D}} G_2,$$

where G_2 is a Gumbel-distributed random variable such that $\mathsf{P}(G_2 \leq x) = \exp(-2\exp(-x))$, $a(x) = \sqrt{2\log(x)}$, and

$$b(x) = 2\log(x) + \frac{p(dp+1)}{2}\log(\log(x)) - \log\left(\frac{2}{3}\Gamma\left(\frac{p(dp+1)}{2}\right)\right).$$

(b) Using the least squares estimator $\widehat{\boldsymbol{a}}_{1,n}$ in (4.6),

$$a(n/G)T(G, \widehat{\boldsymbol{a}}_{1,n}) - b(n/G) \xrightarrow{\mathscr{D}} G_2.$$

(c) We can replace $\boldsymbol{\Sigma}_{(1)}(\widetilde{\boldsymbol{a}})$ with an estimator $\widehat{\boldsymbol{\Sigma}}_k(\widetilde{\boldsymbol{a}})$ as in (4.7) or (4.9) and the results of parts (a) and (b) hold.

With the critical value c_α, we identify the asymptotic distribution of the transformed statistics under H_0. As a result of Proposition 4.1, (4.4) defines a testing procedure with asymptotic level α, where we define

$$D(G, \alpha) = \frac{b(n/G) + c_\alpha}{a(n/G)}. \tag{4.11}$$

Theorem 4.2 establishes that the score procedure consistently estimates both the number and locations of detectable changes.

Theorem 4.2 (MOSUM score procedure consistency)**.** Let Assumptions 4.1-4.6 hold.

(a) Using the score procedure, we have that

$$\mathsf{P}\left(\widehat{q} = q; \max_{1\leq j\leq q} |\widehat{k}_j - k_j| < G\right) \to 1$$

as $n \to \infty$.

(b) Letting $w_n \to \infty$ and $0 < w_n < G(n) \cdot \min_{j=1,\ldots,q} \delta_j^2(\widetilde{\boldsymbol{a}})$, there exists some $\gamma > 2$ such that

$$\mathsf{P}\left(\max_{1\leq j\leq q} \left|\widehat{k}_j - k_j\right| \delta_j^2(\widetilde{\boldsymbol{a}}) > w_n\right) = O(w_n^{-\gamma/2}) + o(1).$$

(c) The results hold using the least squares estimator $\widehat{\boldsymbol{a}}_{1,n}$ in (4.6), or an estimator $\widehat{\boldsymbol{\Sigma}}_k(\widetilde{\boldsymbol{a}})$ as in (4.7) or (4.9) in place of the true $\boldsymbol{\Sigma}_k(\widetilde{\boldsymbol{a}})$.

In part (b) of Theorem 4.2 we establish the rate of estimation which involves $(\delta_j(\widetilde{\boldsymbol{a}}))^{-2}$ defined in Assumption 4.5. This matches the results in Qu and Perron (2007), where the jump size is assumed either constant or to tend to zero such that $c_{\delta,n} \cdot \sqrt{\frac{n}{\log(n)}} \to \infty$, and matches the $O_P(1)$ rate established in Theorem 3.6 of Kirch and Reckrühm (2022) when $c_{\delta,n} = c_\delta$ is a constant. This rate depends on the selection of a parameter $\widetilde{\boldsymbol{a}}$ which is 'good enough' in the sense of Assumption 4.5. We discuss the relaxation of this condition in Section 4.4.1.

4.4 Extensions for improved detection power

4.4.1 MOSUM recursive segmentation

As we discuss in Remark 4.1, when using the MOSUM score procedure with a global inspection parameter, there may exist changes which are not detectable in the sense of Assumption 4.5. Adapting the Binary Segmentation (BS) algorithm (Vostrikova, 1981), we propose to recursively apply the MOSUM score procedure on estimated stationary segments with data-dependent inspection parameters, giving a procedure which asymptotically detects all change points as we show in Theorem 4.3 below.

Algorithm 9 describes the MOSUMBS algorithm. For a given segment $\{s,\dots,e\}$ for some $1 \le s \le e \le n$, a local parameter estimator $\widehat{\boldsymbol{a}}_{s,e}$ is obtained as in (4.6) with which the MOSUM score procedure is performed and change point estimators are added to $\widehat{\mathcal{K}}$. This is repeatedly performed on the segments defined by the consecutive elements of $\widehat{\mathcal{K}}$ until no further change point is detected. Initialised with $(s,e) = (1,n)$ and $\widehat{\mathcal{K}} = \emptyset$, the call MOSUMBS$(\widehat{k}_j+1, \widehat{k}_{j+1}, D, G, \widehat{\mathcal{K}})$ returns the final set of estimators.

Algorithm 3: MOSUMBS$(s,e,D,G,\widehat{\mathcal{K}})$ Recursive Segmentation Algorithm

input : Start and end indices s and e, Threshold D, Bandwidth G, Change point set $\widehat{\mathcal{K}}$

if $e - s > 2G$ **then**
- Compute parameter $\widehat{\boldsymbol{a}}_{s,e}$
- Compute statistic $T \leftarrow \max_{s \le k \le e} T_k(G, \widehat{\boldsymbol{a}}_{s,e})$ as in (4.3)
- **if** $T > D$ **then**
  - Locate $\widehat{\mathcal{K}}_{s,e} \leftarrow \{\widehat{k}_{j,s,e} : s < \widehat{k}_{1,s,e} < \cdots < \widehat{k}_{\widehat{q}_{s,e},s,e} < e\}$ with (4.5)
  - Update global change point set $\widehat{\mathcal{K}} \leftarrow \widehat{\mathcal{K}} \cup \widehat{\mathcal{K}}_{s,e}$
  - **for** $j = 1,\dots,\widehat{q}_{s,e}+1$ **do**
    - MOSUMBS$(\widehat{k}_{j-1,s,e}+1, \widehat{k}_{j,s,e}, D, G, \widehat{\mathcal{K}})$
  - **end**

end

return $\widehat{\mathcal{K}}$

Define $\boldsymbol{a}_{s,e}$ as the unique solution of

$$\sum_{j=1}^{q+1} \left(\min\left(k_j, e\right) - \max\left(k_{j-1}, s\right) + 1\right)_+ \mathbb{E}\boldsymbol{H}\left(\boldsymbol{X}_1^{(j)}, \mathbb{X}_0^{(j)}, \boldsymbol{a}_{s,e}\right) = \boldsymbol{0}, \tag{4.12}$$

where $x_+ = \max(x,0)$. For a given pair (s,e) such that $s < e$, define

$$\mathbb{K}_{s,e} = \left\{k_j, 1 \le j \le q : \min\{k_j - s + 1, e - k_j\} \ge c_{\delta,n}^{-2} \cdot w_n\right\} \tag{4.13}$$

for some $w_n \to \infty$.

Assumption 4.8. For all pairs (s,e) such that $s < e$, we have:

(a) Using $\widetilde{\boldsymbol{a}} = \boldsymbol{a}_{s,e}$ in (4.12), if $\mathbb{K}_{s,e}$ in (4.13) is non-empty, we have that for at least one $k_j \in \mathbb{K}_{s,e}$ and $\delta_j(\widetilde{\boldsymbol{a}}) \geq c_{\delta,n} > 0$, where $c_{\delta,n} \cdot \sqrt{\frac{G}{\log(n/G)}} \to \infty$.

(b) For any $s \leq s' < k_j < e' \leq e$ such that $\min\{k_j - s', e' - k_j\} \leq \delta_j^{-2}(\boldsymbol{a}_{s,e}) \cdot w_n$, we have that $\delta_j^{-2}(\boldsymbol{a}_{s,e}) \cdot \delta_j^2(\boldsymbol{a}_{s',e'}) = O(1)$.

Assumption 4.5 requires the existence of $\widetilde{\boldsymbol{a}}$ that ensures the detectability of all change points. In place of this, we make Assumption 4.8 defined with local inspection parameters. This is motivated by the observation made in Lemma B.11 that, over any interval $\{s, \ldots, e\}$ containing change points which are well within the boundaries of the interval, at least one of them is detectable with the local parameter $\boldsymbol{a}_{s,e}$ in the sense of Assumption 4.5. Part (b) requires that the ratio of jump sizes at a given change point close to the active boundary may vary at a constant rate given different inspection parameters, where the intervals are nested and contain the change point of interest. Under Assumption 4.8, Theorem 4.3 demonstrates that the recursive segmentation procedure detects a broader class of changes than the MOSUM score procedure. Proofs can be found in Section B.3.3.

Theorem 4.3 (Recursive segmentation consistency)**.** Let Assumptions 4.1–4.4, 4.6, and 4.8 hold.

(a) Using the recursive segmentation procedure, we have that

$$\mathrm{P}\left(\widehat{q} = q; \max_{1 \leq j \leq q} |\widehat{k}_j - k_j| < G\right) \to 1$$

as $n \to \infty$.

(b) Let $\widetilde{\boldsymbol{a}}_j$ be the inspection parameter with which $\widehat{k}_j$ is detected. Letting $w_n \to \infty$ and $0 < w_n < G \cdot \min_{j=1,\ldots,q} \delta_j^2(\widetilde{\boldsymbol{a}}_j)$, there exists some $\gamma > 2$ such that

$$\mathrm{P}\left(\max_{1 \leq j \leq q} \left|\widehat{k}_j - k_j\right| \delta_j^2(\widetilde{\boldsymbol{a}}_j) > w_n\right) = O(w_n^{-\gamma/2}) + o(1).$$

(c) The result holds using an estimator $\widehat{\boldsymbol{\Sigma}}_k(\widetilde{\boldsymbol{a}})$ as in (4.7) or (4.9) in place of the true $\boldsymbol{\Sigma}_k(\widetilde{\boldsymbol{a}})$.

4.4.2 Multiscale MOSUM procedure

The theory justifying our method relies on Assumption 4.4, where the chosen bandwidth G becomes small enough such that each change point may be isolated within a window of length $2G$. This suggests that G cannot be too large, while a larger G leads to greater detection power due to Assumption 4.5. The true minimum spacing between changes is always unobservable, however, and often hard to make reasonable prior choices for in practice. Moreover, the signal may be multiscale, such that large frequent and small infrequent changes occur in the same series. To account for these issues, we propose a bottom-up multiscale algorithm which runs the single-scale procedure with bandwidths of increasing size and merges the resulting sets of estimated change

points. Similar ideas are proposed in Messer et al. (2014) and Meier et al. (2021) for the mean change point detection problem.

Define the bandwidth set $\mathscr{G} = \{G_h : 1 \le h \le H, G_1 < \ldots < G_H\}$ with strictly increasing elements. The null hypothesis is rejected if any of test statistics evaluated with bandwidth $G_h \in \mathscr{G}$ exceed their respective thresholds $D(G_h, \alpha)$. For the smallest bandwidth at which we reject, we take the estimated change points as the initial set of changes. For each subsequent bandwidth $G_{h'}$, a detected point is added to the set only if the point is at least $\pi G_{h'}$ away from any point already in the change point set, for some $\pi \in (0,1)$. We formalise this in Algorithm 8.

The procedure is consistent under single-scale changes by repeated application of Theorem 4.2. It is possible to combine the multiscale and recursive segmentation discussed in Section 4.4.1 such that, at each scale, change points are located with Algorithm 9 and combined in the bottom-up fashion.

Algorithm 4: Multiscale MOSUM Algorithm

input : Bandwidth set $\mathscr{G}$, Threshold set $\{D(G_h, \alpha), h = 1, \ldots, H\}$, Localisation parameter π

initialise : $\widehat{\mathcal{K}} = \emptyset$

for $T_h \in \mathscr{G}, h = 1, \ldots, H$, **do**
  Calculate $\widehat{T}_k(G_h, \widetilde{\boldsymbol{a}})$ for $k = G, \ldots, n - G$
  if $\widehat{T}(G_h, \widetilde{\boldsymbol{a}}) > D(G_h, \alpha)$ **then**
    Locate $\widehat{\mathcal{K}}(G_h) \leftarrow \{\widehat{k}_j(G_h)\}_{j=1}^{\widehat{q}}$ as in (4.5)
    for $\widehat{k}_j \in \widehat{\mathcal{K}}(G_h)$ **do**
      if $\min_{k \in \widehat{\mathcal{K}}} |\widehat{k}_j(G_h) - k| \ge \pi G_h$ **then**
        add $\widehat{k}_j(G_h)$ to $\widehat{\mathcal{K}}$
    end
end
return $\widehat{\mathcal{K}}$

4.5 Extensions based on computational considerations

4.5.1 Grid-based procedure

Due mostly to the cost of inverting each of the chosen $\widehat{\boldsymbol{\Sigma}}_k$, evaluating the detector at a given k has a cost which grows polynomially in d and p, and doing this for all $k = G, \ldots, n - G$ incurs this cost $O(n)$ times (we discuss computational complexity in Section 4.6.1); to avoid this fast growth in complexity, we propose to adopt a coarse grid over which the detector $\widehat{T}_k(G, \widehat{\boldsymbol{a}})$ is computed. We define the grid $\mathscr{T}$ with resolution constant $r \in [G^{-1}, 1]$ as

$$\mathscr{T}(r, G) = \left\{ t : t = G + d + m\lfloor rG \rfloor, 0 \le m \le \left\lfloor \frac{n - 2G}{rG} \right\rfloor \right\}. \tag{4.14}$$

This set collects every $\lfloor rG \rfloor$-th point in the index set starting at $t = G + d$, the earliest at which the MOSUM statistic can be evaluated. When $r = G^{-1}$, we have the finest grid $\mathscr{T} = \{G + d, \ldots, n - G\}$, and a larger value of r gives a coarser grid. After scanning over the grid, intervals with detectors exceeding the threshold are filled in with score statistics, where the inspection parameter is calculated over each contiguous interval of exceedance. To locate changes, we can use either of the ϵ- or η-criteria as in (4.5) or (B.4.1); our theoretical results are derived with the former but will hold with the latter by the same arguments as Lemma B.28. This reduces the overall cost by up to a factor of G (see Table 4.1) while asymptotically achieving the same estimation guarantees (see Theorem 4.4).

We describe the procedure as follows.

MOSUM score grid-based procedure

1. Identify all $k \in \mathscr{T}$ such that $\widehat{T}_k(G, \widehat{\boldsymbol{a}}_{1,n}) \geq D(G, \alpha_n)$.

2. Collect these k into contiguous intervals such that $k = s_i, s_i + 1, \ldots, e_i$, $i = 1, \ldots, Q$ and extend the intervals outwards to define the sets

$$\mathscr{T}_i = \{s_i - G, s_i - G + 1, \ldots, e_i + G\}.$$

3. Compute the estimator $\widehat{\boldsymbol{a}}_{\mathscr{T}_i}$ as in (4.6) with $s = s_i - G$ and $e = e_i + G$.

4. For each i, calculate the statistic $\widehat{T}_k(G, \widehat{\boldsymbol{a}}_{\mathscr{T}_i})$ for each k in the set $\mathscr{T}_i$. For each k, assign T_k to be the pointwise maximum of $\widehat{T}_k(G, \widehat{\boldsymbol{a}}_{\mathscr{T}_i})$ over i.

5. Locate changes with the ϵ-criterion as in (4.5).

In Theorem 4.4 we collect a set of results stating that the previous consistency results hold when using the grid-based procedure. Proofs can be found in Section 4.5.1.

Theorem 4.4 (Grid-based procedure consistency)**.** The results of Theorem 4.2 holds for the output $\{\widehat{k}_j : 1 \leq j \leq \widehat{q}\}$ from the score grid-based procedure.

4.5.2 Dimension reduction

For a VAR model, the number of parameters grows linearly in the order d and quadratically in the number of series p, making estimation difficult for moderately large VAR order and dimensionality. We propose to use zero-restrictions on parameters, and a projection onto univariate autoregressive models, both of which allow the analysis of larger data panels.

4.5.2.1 Parameter restrictions

To reduce the number of parameters, we consider the proposal of Kirch et al. (2015) in which any parameter believed a priori to have no influence on the response is set equal to 0 in the estimation step. We consider the vector of parameters which are involved in modelling channel $i = 1, \ldots, p$:

$$\boldsymbol{a}(i) = \left(\omega_i, a_{1i}, \ldots, a_{pi}, \ldots, a_{dpi}\right)^\top \in \mathbb{R}^{dp+1}.$$

Denote the index set of the elements of $\boldsymbol{a}(i)$ whose values are equal to zero (i.e. no influence on the ith channel) by $\mathscr{I}(i) = \{r : a(i,r) = 0\}$, and the corresponding projection operator by $\mathscr{P}_{\mathscr{I}} : \mathbb{R}^{dp+1} \to \mathbb{R}^{dp+1-|\mathscr{I}|}$ such that for any $\boldsymbol{Y} \in \mathbb{R}^{dp+1}$, we have $\mathscr{P}_{\mathscr{I}}(\boldsymbol{Y}) = (Y_r : r \notin \mathscr{I})^\top$. We further write $\boldsymbol{a}_{I(i)}(i) = \mathscr{P}_{\mathscr{I}(i)}(\boldsymbol{a}(i))$ and regressors $\mathbb{X}_{I(i),t-1} = \mathscr{P}_{I(i)}(\mathbb{X}_{t-1})$ for each channel. This gives the reduced models

$$X_{it} = \begin{cases} X_{it}^{(1)}, & k_0 = 1 \le t \le k_1, \\ X_{it}^{(2)}, & k_1 + 1 \le t \le k_2, \\ \vdots & \\ X_{it}^{(q+1)}, & k_q + 1 \le t \le k_{q+1} = n, \end{cases}$$

where $X_{it}^{(j)} = \boldsymbol{a}_{I(i),j}(i)^\top \mathbb{X}_{I(i),t-1}^{(j)} + \varepsilon_{it}$. Note that the change points k_j are common across channels.

4.5.2.2 Projection

We propose a projection method to further reduce the number of parameters. Rewriting the model (4.1) as

$$\boldsymbol{X}_t = \boldsymbol{\omega} + \sum_{l=1}^{d} \boldsymbol{A}_l \boldsymbol{X}_{t-l} + \boldsymbol{\varepsilon}_t,$$

we first remove cross-dependence in $\{\boldsymbol{X}_t\}_{t=1}^n$ by constructing the series

$$\boldsymbol{Z}_t = \boldsymbol{X}_t - \sum_{l=1}^{d} \mathrm{OffDiag}(\boldsymbol{A}_l)\boldsymbol{X}_{t-l},$$

where $\mathrm{OffDiag} : \mathbb{R}^{p\times p} \to \mathbb{R}^{p\times p}$ is a function such that $\mathrm{OffDiag}(\boldsymbol{A})$ sets the diagonal elements of $\boldsymbol{A}$ to 0. In practice, we use the global least squares estimators in (4.6) for $\boldsymbol{A}_l$. We then regress each Z_{it} on lagged observations $\mathbb{Z}_{it-1} = (1, Z_{i,t-1}, \ldots, Z_{i,t-d})^\top$, so a given pair (s,e) such that $s < e$ has the parameter estimator

$$\widehat{\boldsymbol{a}}_{s,e}^{[z]} = \begin{pmatrix} \widehat{\boldsymbol{a}}_{s,e}^{[z]}(1)^\top \\ \widehat{\boldsymbol{a}}_{s,e}^{[z]}(2)^\top \\ \cdots \\ \widehat{\boldsymbol{a}}_{s,e}^{[z]}(p)^\top \end{pmatrix}, \quad \widehat{\boldsymbol{a}}_{s,e}^{[z]}(i) = \left(\sum_{t=s}^{e} Z_{it}\mathbb{Z}_{t-1}^\top\right)\left(\sum_{t=s}^{e} \mathbb{Z}_{t-1}\mathbb{Z}_{t-1}^\top\right)^{-1}, i = 1, \ldots, p.$$

To ensure that the limit of the maximum of the detector in (4.3) under the projection is pivotal, the residuals $\widehat{\boldsymbol{\varepsilon}}_t(\widetilde{\boldsymbol{a}})$ from $\boldsymbol{X}_t$ are plugged in to the estimating functions (4.2) in the MOSUM score

procedure, so that

$$\boldsymbol{H}_t(\widetilde{\boldsymbol{a}}) = -\widehat{\boldsymbol{\varepsilon}}_t(\widetilde{\boldsymbol{a}}) \otimes \mathbb{Z}_{t-1},$$

where $\widetilde{\boldsymbol{a}} = \widehat{\boldsymbol{a}}_{s,e}^{[z]}$. We also use estimators of $\boldsymbol{\Sigma}_k^{(1)}(\widetilde{\boldsymbol{a}})$ designed for the univariate setting, adapting $\widehat{\boldsymbol{\Sigma}}_k^{(1)}(\widetilde{\boldsymbol{a}})$ in (4.7) by plugging in $\mathbb{Z}_{t-1}$. This choice of function is motivated by the empirical observation that series often depend on themselves more so than on others; in principle, we can remove the contributions of any component of $\boldsymbol{A}$ using an appropriate function in place of OffDiag; for example, we could set one row to zero, isolating the effect of that variable on other channels.

4.5.3 Threshold bootstrap

We show in Proposition 4.1 that the size of the MOSUM score procedure is controlled asymptotically with the threshold in (4.11). Due to the slow convergence of the Gumbel distribution, however, we cannot expect good performance when the sample size is small relative to the model dimensions. Therefore we propose a parametric bootstrap method for obtaining a threshold, which is of particular interest in combination with the projection method in Section 4.5.2.

We assume a correctly specified model under the null hypothesis. Fit a VAR model to the entire series, computing the parameter estimators in (4.6) and the error covariance estimators in (B.1.6). For each bootstrap iteration $m = 1, \ldots, M$:

1. Using Gaussian innovations, simulate a series $\{\boldsymbol{X}_t^m\}_{t=1}^n$ with the estimated parameters.

2. Store the maximum test statistic $\widehat{T}^m(G, \widehat{\boldsymbol{a}}_{1,n}^{[m]}) = \max_{G \le k \le n-G} \widehat{T}_k^m(G, \widehat{\boldsymbol{a}}_{1,n}^{[m]})$, computed with $\{\boldsymbol{X}_t^m\}_{t=1}^n$ as in (4.3), where $\widehat{\boldsymbol{a}}_{1,n}^{[m]}$ is estimated over the same data.

We then take the $(1-\alpha)$-th quantile of the M recorded statistics as the threshold. Under the null hypothesis, this will produce quantiles of the test statistic from realisations of a correctly-specified model. When $q \ge 1$, the realisation are drawn from a model with parameters approximating a weighted mixture of the parameters from different regimes. This will not overly inflate the threshold, but may make the procedure less sensitive to small changes. The bootstrap procedure is computationally intensive, so to balance execution time with the quality of the estimate of the quantile, we recommend setting $M = 1000$.

4.6 Numerical results

In this section, we discuss computational complexity and tuning parameter selection. We describe the performance of our methods in Monte Carlo experiments, we compare our methods to others available in the literature, and apply our methods to real data sets.

4.6.1 Complexity

Table 4.1 lists the time complexity of procedure/estimator combinations. Denote the size of $\mathscr{G}$ as $|\mathscr{G}|$. The computation of the MOSUM score detector statistics can be carried out sequentially, see Appendix B.5. For the localisation step of the grid-based procedure of Section 4.5.1, since q is fixed and asymptotically estimated correctly, the cost is dominated by the grid search.

We can see that the cost grows faster with respect dimension for direct estimators than the combined forms, and that the MOSUM score procedure is cheaper than the Wald due to the sequential updating procedure. Grid-based methods are sub-linear in n, while the recursive segmentation method incurs a $\log(n)$ factor, and the multiscale MOSUM procedure grows with $|\mathscr{G}|$.

Table 4.1: Computational complexity of proposed procedures

Procedure	Estimator $\widehat{\mathbf{\Sigma}}_k$	Time Complexity
MOSUM score	(4.7)	$O\left(d^2p^2\max(p^2,d)+ndp\right)$
	(4.9)	$O\left(d^3p^6+ndp^2\right)$
MOSUM Wald	(B.1.5)	$O\left(np^3d^2\right)$
(Section B.1)	(B.1.7)	$O\left(np^6d^3\right)$
Recursive segmentation	(4.7)	$O\left(\lvert\mathscr{G}\rvert\cdot(d^2p^2\max(p^2,d)+ndp)\right)$
(Section 4.4.1)	(4.9)	$O\left(\lvert\mathscr{G}\rvert\cdot(d^3p^6+ndp^2)\right)$
Multiscale MOSUM	(4.7)	$O\left(\log(n)\cdot(d^2p^2\max(p^2,d)+ndp)\right)$
(Section 4.4.2)	(4.9)	$O\left(\log(n)\cdot(d^3p^6+ndp^2)\right)$
Grid-based (score)	(4.7)	$O\left(d^2p^2\max(p^2,d)+\frac{n}{rG}dp\right)$
(Section 4.5.1)	(4.9)	$O\left(d^3p^6+\frac{n}{rG}dp^2\right)$
Grid-based (Wald)	(B.1.5)	$O\left(\frac{n}{rG}p^3d^2\right)$
(Section B.1.2.1)	(B.1.7)	$O\left(\frac{n}{rG}p^6d^3\right)$

4.6.2 Tuning parameters

Threshold The transformed critical value $D(G,\alpha)$ from (4.11) involves an evaluation of the Gamma function, meaning $D(G,\alpha)$ grows very slowly with respect to d and p, and can even take negative values. The slow convergence is discussed with simulation evidence in Section 5 of Dvořák (2017). We recommend fixing $\alpha = 0.05$. We hence propose using the adjusted critical value

$$\widetilde{D}_n(G,\alpha) = \max\left\{D(G,\alpha), \sqrt{2\log(n)} + \frac{c_\alpha}{\sqrt{2\log(n)}}\right\} \tag{4.15}$$

which is guaranteed to be positive.

VAR order In the simulations in Section 4.6.3, we treat the lag order d as known. For data-driven selection, we suggest minimising the Schwartz Information Criterion (SIC) (Lütkepohl,

2005)

$$\mathrm{SIC}(d) = \frac{G}{2}\log\left(\frac{1}{G}\|\sum_{t=d+1}^{G}\left(\boldsymbol{X}_t - \widehat{\boldsymbol{a}}_{1,G}[d]\mathbb{X}_{t-1}[d]\right)\|^2\right) + d\log(G),$$

for models fit with all available samples from $t = 1,\ldots,G$, where $\widehat{\boldsymbol{a}}_{1,G}[d]$ is the estimator corresponding to $\mathbb{X}_{t-1}[d]$ which consists of $\mathbb{X}_{i,t-1} = \left(X_{i,t-1},\ldots,X_{i,t-d}\right)^\top, i = 1,\ldots,p$.

Bandwidth To select G, we want to control the parameter estimation error, while also meeting the requirement that $G \leq \min_j(k_j - k_{j-1})/2$. To relate G to p and n, we simulated datasets under (4.16) with $q = 0$, $d = 1$, and varying (n,p,G) (the entries $a'_{ii'}$ are chosen uniformly on $[-1,1]$ and ρ uniformly on $[0.2,0.7]$), recording the relative ℓ_2-error $\max_{0\leq k\leq n-G}\|\boldsymbol{a}\|_2^{-1}\|\widehat{\boldsymbol{a}}_{k+1,k+G} - \boldsymbol{a}\|_2$ for each realisation. Then, regressing the 90%-percentile of the estimation errors over 100 realisations onto $\log(G)$, $\log\log(\sqrt{dp^2})$ and $\log\log(n)$ (giving $R^2 = 0.8447$), we obtain the rule to determine the finest bandwidth as $G = G(n,p,d) = \exp(c_0 - c_1\log\log(n) + c_2\log\log(\sqrt{dp^2}))$ with $c_i > 0$, $i = 0,1,2$, obtained from the regression coefficients.

Both the MOSUM score and Wald procedures require $G \geq p(dp+1) + p(p+1)/2$ due to the degrees of freedom expended by using the estimators $\widehat{\boldsymbol{a}}_{1,n}$ and $\widehat{\boldsymbol{\Sigma}}_k$. To ensure reasonable estimation we hence recommend that $G \geq \max\{p(dp+1)\log(p(dp+1)),(4/3)n^{2/3}\}$; if the above recommended bandwidth $G(n,p,d)$ does not satisfy the requirement, we recommend the dimension reduction methods discussed in Section 4.5.2. For the multiscale algorithm, we recommend using $G_1 = G(n,p,d)$, $G_2 = \lfloor(4/3)G_1\rfloor$, and $G_3 = \lfloor(5/3)G_1\rfloor$.

Localisation Based on simulation, we recommend localising with $\epsilon = 0.5$ (per (4.5)) with the MOSUM score procedure, $\eta = 0.25$ with the MOSUM Wald procedure (per (B.4.2)), $\epsilon = 0.7$ with the recursive segmentation procedure, and $\pi = 0.5$ in the multiscale MOSUM procedure. These prevent spurious detection while retaining good localisation properties.

4.6.3 Simulation studies

We examine how the proposed methods perform under varying model parameters, namely the dimensionality, lag, distance between changes, detectability of changes, and sample size. Each extension is tested in a specifically designed setting. We use the level of significance $\alpha = 0.05$. Each setting uses $N = 100$ simulations[1].

Performance evaluation metrics When $q \geq 1$, we report the distribution of $\widehat{q} - q$. We quantify the quality of the estimated segmentation using the Covering Metric (CM, Arbelaez et al. (2010); van den Burg and Williams (2020)) as follows. The true change points $\{k_j\}_{j=1}^q$ define a partition

[1]Simulations are performed in the R language (R Core Team, 2020) with the 'Rcpp' package (Eddelbuettel et al., 2011) and 'armadillo' library (Sanderson and Curtin, 2016) on an Intel Core i7-8650U CPU @ 1.90GHz (8 Cores) processor. This work was carried out using the computational facilities of the Advanced Computing Research Centre, University of Bristol.

$\mathscr{P}$ of $\{1,\ldots,n\}$ into disjoint sets $\mathscr{S}_j = \{k_{j-1}+1,\ldots,k_j\}$. We denote the estimated equivalents for $\{\widehat{k}_j\}_{j=1}^{\widehat{q}}$ as $\widehat{\mathscr{P}}$ and $\widehat{\mathscr{S}}_j$. The CM is then

$$\mathscr{C}(\widehat{\mathscr{P}},\mathscr{P}) = \frac{1}{n}\sum_{\mathscr{S}\in\mathscr{P}}|\mathscr{S}|\max_{\widehat{\mathscr{S}}\in\widehat{\mathscr{P}}}\left\{\frac{\mathscr{S}\cap\widehat{\mathscr{S}}}{\mathscr{S}\cup\widehat{\mathscr{S}}}\right\}.$$

We have $\mathscr{C}(\widehat{\mathscr{P}},\mathscr{P}) \in [0,1]$ with 1 denoting a perfect segmentation. As opposed to the Hausdorff metric (3.20) used in Chapter 3, this metric does not systematically favour over-estimation of the number, and is perhaps an easier summary to interpret. When $q = 0$, we report the empirical size, i.e. the proportion of the realisations for which $\widehat{q} \geq 1$.

4.6.3.1 Settings

We simulate from $\boldsymbol{A}_l^{(j)}$ under each regime so that for $t = k_j+1,\ldots,k_{j+1}$,

$$\boldsymbol{X}_t = \sum_{l=1}^{d}\boldsymbol{A}_l^{(j)}\boldsymbol{X}_{t-l} + \boldsymbol{\varepsilon}_t. \tag{4.16}$$

The parameters are defined as follows: letting $\boldsymbol{A}'$ have diagonal entries $a'_{ii} = 0.7$ and off-diagonal entries $a'_{ii'} = -0.1, i \neq i'$, we set $\boldsymbol{A}_l^{(1)} = \rho_j \boldsymbol{A}'/\|\boldsymbol{A}'\|_F^2$, for $l = 1,\ldots,d$, where $\rho_j \in \mathbb{R}$ is a scalar controlling the signal strength. We generate errors so that $\boldsymbol{\varepsilon}_t \sim N_p(\boldsymbol{0},\boldsymbol{I}_p)$. For each setting, we simulate under $q = 0$, and for some $q \geq 1$ as specified. In all settings, we report results from the MOSUM Wald and score procedures.

(M1) We test a design with varying dimensionality $p \in \{3,4,5\}$. We have $n = 900$, $d = 1$, and $q = 2$ change points at $k_1 = 300$ and $k_2 = 600$. When $q \geq 1$, we set $\rho_1 = -\rho_2 = \rho_3 = 0.7$. We also report results with the covariance estimators (4.9) and (B.1.7)

(M2) We test a design with varying lags $d \in \{1,2,3\}$. We have $n = 900$, $p = 2$, and $q = 2$ change points at $k_1 = 300$ and $k_2 = 600$. When $q \geq 1$, we set $\rho_1 = -\rho_2 = \rho_3 = 0.7$.

(M3) We test a design with varying $n \in \{3000,6000,9000\}$. We fix $p = 3, d = 1$, and have $q = 2$ change points at $k_1 = n/3$ and $k_2 = 2n/3$. When $q \geq 1$, we set $\rho_1 = -\rho_2 = \rho_3 = 0.5$. We compare the grid-based approaches with the resolution $r = 1/10$

(M4) We test a design with multiscale changes. We vary $n \in \{1400,2100,2800\}$. We fix $p = 3, d = 1$, and have $q = 5$ change points at $k_1 = n/14, k_2 = n/7, k_3 = 2n/7, k_4 = 3n/7$, and $k_5 = 5n/7$, and we set $\rho_1 = -\rho_2 = 1, \rho_3 = -\rho_4 = 1/\sqrt{2}$, and $\rho_5 = -\rho_6 = 1/2$. The distance between changes increases with j but the size of the change decreases. Here we compare multiscale MOSUM procedures.

(M5) We test a design with varying dimensionality $p \in \{10,15,20\}$. We have $n = 900$, $d = 1$, and $q = 2$ change points at $k_1 = 300$ and $k_2 = 600$. When $q \geq 1$, we set $\rho_1 = -\rho_2 = \rho_3 = 1$. We compare the dimension reduction procedure per Section 4.5.2; the threshold is determined by the bootstrap procedure in Section 4.5.3.

(M6) We test a design with varying signal strength $\rho_1 = -\rho_2 = \rho_3 \in \{0.4, 0.6, 0.8\}$. We have $n = 900$, $p = 3$, $d = 1$, and $q = 2$ change points at $k_1 = 300$ and $k_2 = 600$. We also report the recursive segmentation procedure.

(M7) We investigate how the procedure handles model misspecification, using a moving average design so that $\boldsymbol{X}_t = \boldsymbol{\varepsilon}_t + \sum_{j=1}^{q+1} \boldsymbol{A}^{(j)} \mathbb{I}\{k_j + 1 \leq t \leq k_{j+1}\} \cdot \boldsymbol{\varepsilon}_{t-1}$, setting $n = 900$, $q = 2$ change points at $k_1 = 300$ and $k_2 = 600$, and $\rho_1 = -\rho_2 = \rho_3 = 1.5$. We vary $p \in \{3, 4, 5\}$.

(M8) We test a design with heavier tails, simulating the errors from a scaled multivariate t-distribution so that $\varepsilon_{it} \sim_{iid} \sqrt{\nu/(\nu-2)} \cdot t_\nu$, varying the degrees of freedom $\nu \in \{3, 5\}$. We have $n = 900$, $d = 1$, and $q = 2$ change points at $k_1 = 300$ and $k_2 = 600$. When $q \geq 1$, we set $\rho_1 = -\rho_2 = \rho_3 = 0.7$.

4.6.3.2 Results

We report the results in Tables 4.2–4.3. Setting (M1) shows that performance is generally worse with increasing p, and that the MOSUM score procedure far outperforms the MOSUM Wald procedure. The direct covariance estimators can give better localisation with the MOSUM Wald procedure, but fail to control size with the MOSUM score procedure. Setting (M2) shows that increasing the lag has a very serious impact on detection and localisation performance. In idealised designs, the MOSUM Wald procedure can localise better than the MOSUM score procedure. Setting (M3) shows the grid-based MOSUM score procedure is more reliable than the grid-based MOSUM Wald procedure at smaller sample sizes, but the performance of the two converges with n. Setting (M4) shows that, in a multiscale setting, the score and recursive methods tend to perform better than the MOSUM Wald procedure. Indeed, the short segments lead to worse estimation of the regression parameter, explaining the poorer performance of the MOSUM Wald procedure for smaller n, though this difference is less pronounced for $n = 2800$; the systematic over-estimation of the number here is due to the smallest bandwidth generating spurious estimators when the jumps are large. Setting (M5) shows that the dimension reduction and bootstrap have good size control, and has good localisation properties for p between 10 and 15, though the detection properties diminish as p grows. Setting (M6) shows that, for smaller signals, the recursive segmentation procedure is more sensitive than the others while still maintaining size control. Setting (M7) shows that the MOSUM Wald procedure performs particularly well under model misspecification. Setting (M8) shows that both methods are fairly robust to heavy tails, comparing to the results in Setting (M1).

We here mention that, since each method is implemented in C++, the orders of the runtimes for each is sufficiently small (on the order of 10^{-3} seconds for the settings we consider in this section) that a comparison is uninformative. We should expect that the MOSUM score procedure is faster than the MOSUM Wald procedure, and that this comes at the expense of computation

time, though this is not the case in our results. This phenomenon may occur, however, when translating our ideas to more expensive models.

The overall similar performance of the MOSUM score and Wald prcoedures, given we should expect the Wald to perform better, can also be explained by the short data series we use examine in our study. The small values for n imply a small value for G through our choice in Section 4.6.2. This will incur error in the estimation of the regression parameter, which is carried over to the estimation of change numbers and locations. Larger values for n are uncommon in macroeconomic and financial applications, and as such we concentrate on settings with smaller n.

4.6.4 Comparative simulations

We compare to the simulation results of McGonigle and Cho (2023) for the detection of changes in temporal dependence. They propose a non-parametric moving window method for the segmentation of dependent time series, which measures for differences in joint characteristic functions. We report their results when the method considers multiple lags (NP-MOJO-$\mathscr{L}$). Also reported are results from the wavelet-based wild binary segmentation (WBSTS) method proposed in Korkas and Fryzlewicz (2017) when $p = 1$, and the sparsified binary segmentation (SBS) method proposed in Cho and Fryzlewicz (2015) when $p = 2$. We looked for implementations of methods which are designed for the same settings as our methods, such as Bai (2000), but none of these were readily available in R.

For the autoregressive models (C1) and (C3), we use $d = 1$, and for the moving average models (C2) and (C4), we select the order adaptively. We have $q = 2$, $(k_1, k_2) = (333, 667)$ and $\varepsilon_t \sim_{\text{i.i.d.}} N(0,1)$.

(C1) $X_t = X_t^{(j)} = a^{(j)} X_{t-1}^{(j)} + \varepsilon_t$ for $k_j + 1 \le t \le k_{j+1}$, where $(a^{(1)}, a^{(2)}, a^{(3)}) = (-0.8, 0.8, -0.8)$.

(C2) $X_t = \varepsilon_t + \sum_{j=1}^{q+1} a^{(j)} \mathbb{I}\{k_j + 1 \le t \le k_{j+1}\} \cdot \varepsilon_{t-2}$, where $(a^{(1)}, a^{(2)}, a^{(3)}) = (-0.7, 0.7, -0.7)$.

(C3) $\boldsymbol{X}_t = \boldsymbol{X}_t^{(j)} = \boldsymbol{A}_j \boldsymbol{X}_{t-1}^{(j)} + \boldsymbol{\varepsilon}_t$ for $k_j + 1 \le t \le k_{j+1}$, where $\boldsymbol{A}^{(1)} = \boldsymbol{A}^{(3)} = \left(\begin{smallmatrix} 0.5 & 0.1 \\ 0.1 & 0.5 \end{smallmatrix}\right)$ and $\boldsymbol{A}^{(2)} = \left(\begin{smallmatrix} -0.5 & 0.1 \\ 0.1 & -0.5 \end{smallmatrix}\right)$.

(C4) $\boldsymbol{X}_t = \boldsymbol{\varepsilon}_t + \sum_{j=1}^{q+1} \boldsymbol{A}^{(j)} \mathbb{I}\{k_j + 1 \le t \le k_{j+1}\} \cdot \boldsymbol{\varepsilon}_{t-1}$, where $\boldsymbol{A}^{(1)} = \boldsymbol{A}^{(3)} = \left(\begin{smallmatrix} 1 & 0.1 \\ 0.1 & 1 \end{smallmatrix}\right)$ and $\boldsymbol{A}^{(2)} = \left(\begin{smallmatrix} -1 & 0.1 \\ 0.1 & -1 \end{smallmatrix}\right)$.

Model (C1) was studied in Korkas and Fryzlewicz (2017), while models similar to (C3) and (C4) were considered in Preuss et al. (2015b).

We report results in Table 4.4. We can see that our methods are competitive. In Setting (C1), the MOSUM Wald procedure performs almost identically to NP-MOJO-$\mathscr{L}$; we would expect to be the best here due to the parametric specification, however this is not possible since the performance is almost perfect. In Setting (C2), under a misspecified model, the MOSUM Wald procedure performs well though the score-based methods lack power. As we would expect, our methods excel in Setting (C3), and though the model is again misspecified, ours are almost perfect

Table 4.2: (M1)–(M5): we report the distribution of the estimated number of change points and the average CM over 1000 realisations. The best performer for each metric is given in bold.

				$\widehat{q}-q$					
Setting	Method	Variable	CM	≤-2	-1	0	1	2	Size
(M1)	MOSUM score	$p=3$	**0.9694**	0	2	**98**	0	0	**0**
		4	**0.9481**	0	4	**91**	5	0	0.02
		5	**0.9146**	0	2	**76**	20	2	0.17
	MOSUM Wald	3	0.9195	1	15	84	0	0	**0**
		4	0.8044	6	39	55	0	0	**0**
		5	0.7771	9	37	52	2	0	**0.03**
	MOSUM score	3	0.9599	0	0	89	10	1	0.11
	(with (4.9))	4	0.9066	0	1	69	24	6	0.89
		5	0.7884	0	43	46	10	1	1
	MOSUM Wald	3	0.9438	1	8	91	0	0	**0**
	(with (B.1.7))	4	0.8385	6	28	65	1	0	**0**
		5	0.8074	6	32	58	4	0	**0.03**
(M2)	MOSUM score	$d=1$	0.9654	0	0	87	13	0	**0**
		2	**0.8220**	2	37	**52**	9	0	**0**
		3	0.4948	49	42	7	2	0	**0**
	MOSUM Wald	1	**0.9832**	0	0	**99**	0	1	0.02
		2	0.6923	5	66	26	3	0	0.24
		3	**0.5359**	33	53	**13**	1	0	0.37
(M3)	MOSUM score	$n=3000$	**0.9914**	0	0	**100**	0	0	**0**
	(grid-based)	6000	**0.9955**	0	0	**100**	0	0	**0**
		9000	**0.9974**	0	0	**100**	0	0	**0**
	MOSUM Wald	3000	0.9806	0	3	94	3	0	**0**
	(grid-based)	6000	0.9850	0	3	95	2	0	**0**
		9000	0.9944	0	1	99	0	0	**0**
(M4)	MOSUM score	$n=1400$	**0.8438**	11	80	**9**	0	0	**0**
	(multiscale)	2100	**0.8941**	0	72	26	2	0	**0**
		2800	**0.9620**	0	0	39	60	1	**0**
	MOSUM Wald	1400	0.4635	98	2	0	0	0	**0**
	(multiscale)	2100	0.6916	61	35	4	0	0	**0**
		2800	0.8078	1	24	**47**	28	0	**0**
	Recursive	1400	0.8296	13	82	5	0	0	**0**
	(multiscale)	2100	0.8847	1	68	**31**	0	0	**0**
		2800	0.9500	0	2	40	58	0	**0**
(M5)	MOSUM score	$p=10$	0.7769	9	40	51	0	0	**0**
	(dimension reduction)	15	0.4147	75	23	2	0	0	**0**
		20	0.3393	98	2	0	0	0	**0**
	MOSUM Wald	10	**0.9363**	0	3	**94**	3	0	0.06
	(dimension reduction)	15	**0.7098**	13	46	**39**	2	0	0.05
		20	**0.5000**	48	45	**7**	0	0	0.06

Table 4.3: (M6)–(M8): we report the distribution of the estimated number of change points and the average CM over 1000 realisations. The best performer for each metric is given in bold.

				$\widehat{q}-q$					
Setting	Method	Variable	CM	≤-2	-1	0	1	2	Size
(M6)	MOSUM score	$\rho = 0.4$	0.3967	81	18	1	0	0	**0**
		0.6	0.8557	3	30	66	1	0	**0**
		0.8	**0.9806**	0	0	**100**	0	0	**0**
	MOSUM Wald	0.4	0.3426	97	3	0	0	0	**0**
		0.6	0.6846	21	49	30	0	0	**0**
		0.8	0.9745	0	1	99	0	0	**0**
	Recursive	0.4	**0.6509**	25	47	**26**	2	0	**0**
		0.6	**0.9596**	0	2	**93**	5	0	**0**
		0.8	0.9691	0	0	91	9	0	**0**
(M7)	MOSUM score	$p = 3$	0.4012	68	32	0	0	0	**0**
		4	0.4533	48	44	7	1	0	**0**
		5	0.5354	24	44	21	10	1	0.1
	MOSUM Wald	3	**0.6610**	0	15	**85**	0	0	**0**
		4	**0.6371**	0	40	**60**	0	0	**0**
		5	**0.6271**	0	55	**44**	1	0	**0**
(M8)	MOSUM score	$\nu = 3$	**0.9398**	0	5	**90**	5	0	**0**
		5	**0.9632**	0	1	**95**	4	0	**0**
	MOSUM Wald	3	0.8559	0	27	67	6	0	**0**
		5	0.8684	2	27	70	1	0	**0**

under Setting (C4). This particular moving average model is likely well approximated by the VAR, and we note that all the methods perform well here.

4.6.5 Applications

4.6.5.1 Bristol air quality data

According to the World Health Organisation, air pollution kills an estimated seven million people worldwide every year. Many different particulates and chemicals are present in the air, particularly in urban areas, and understanding how the concentrations of these change over time can aid us in designing and evaluating public policy. We propose to use our methods to identify changes in the second-order structure for measurements of NO_x (Nitric oxide and nitrogen dioxide) levels in Bristol, UK.

The methods we propose are suitable here, since serial and spatial dependence is present in these levels, but abrupt changes can occur, perhaps due to policy choices or exogenous shocks. Moreover, the piecewise structure of the resulting model is simple and easy to interpret. Changes

Table 4.4: (C1)–(C4): we report the distribution of the estimated number of change points and the average CM over 1000 realisations. The best performer for each metric is given in bold.

			$\widehat{q}-q$				
Model	Method	CM	-2	-1	0	1	≥ 2
(C1)	Score	0.945	0	1	741	225	33
	Wald	0.977	0	1	977	22	0
	Recursive	0.965	0	0	919	75	6
	NP-MOJO-$\mathscr{L}$	**0.980**	0	0	**986**	12	0
	WBSTS	0.904	0	0	414	299	287
(C2)	Score	0.418	760	219	21	0	0
	Wald	**0.955**	2	79	917	2	0
	Recursive	0.362	914	84	2	0	0
	NP-MOJO-$\mathscr{L}$	0.950	1	51	**942**	6	0
	WBSTS	0.896	7	21	899	62	11
(C3)	Score	**0.980**	0	2	991	7	0
	Wald	0.977	0	13	984	3	0
	Recursive	0.978	0	3	**993**	4	0
	NP-MOJO-$\mathscr{L}$	0.907	4	165	818	13	0
	SBS	0.903	70	0	911	19	0
(C4)	Score	**0.985**	0	0	**999**	1	0
	Wald	**0.985**	0	7	993	0	0
	Recursive	0.981	0	2	998	0	0
	NP-MOJO-$\mathscr{L}$	0.976	0	12	979	9	0
	SBS	0.967	6	0	961	33	0

in cross-channel correlations might indicate changing spatial patterns, while changes in serial dependence could indicate differences in persistence, and intercept changes will indicate changes in the conditional average levels.

Similar analyses to ours are made in Fassò (2013), where a policy intervention is evaluated with a known change point, and in Voynikova et al. (2015) where a forecasting model is built with exogenous regressors. Rinaldo et al. (2021) look for parameter changes under a high-dimensional regression model, while Cho and Fryzlewicz (2020) apply a method for univariate mean changes under serial dependence. Yin et al. (2021) use a stationary network autoregressive model for an aggregated air quality index.

Over the period from January 2019 to January 2022 we have hourly readings of NO_x levels at five locations around Bristol[2]; see Figure 4.2. To improve the signal quality of the data and handle missing observations, we take daily averages, giving $n = 1099$ and $p = 5$. We control for

[2]Data are available from Open Data Bristol.

meteorological and seasonal effects by regressing $\sqrt{NO_{xt}}$ onto 6 covariates: Temperature, Wind speed, Wind direction, Atmospheric pressure, and factors for the Day of the Week and the Season. NO_x readings are bounded below and demonstrate large excess kurtosis, so we use a square root transform.

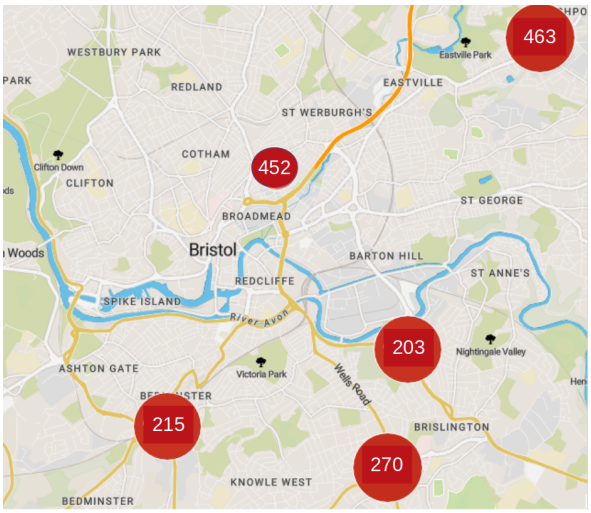


Figure 4.2: Map of Bristol, UK with air quality detectors (labelled with site IDs) located at AURN St Pauls (452); Brislington Depot (203); Parson Street School (215); Wells Road (270); Fishponds Road (463).

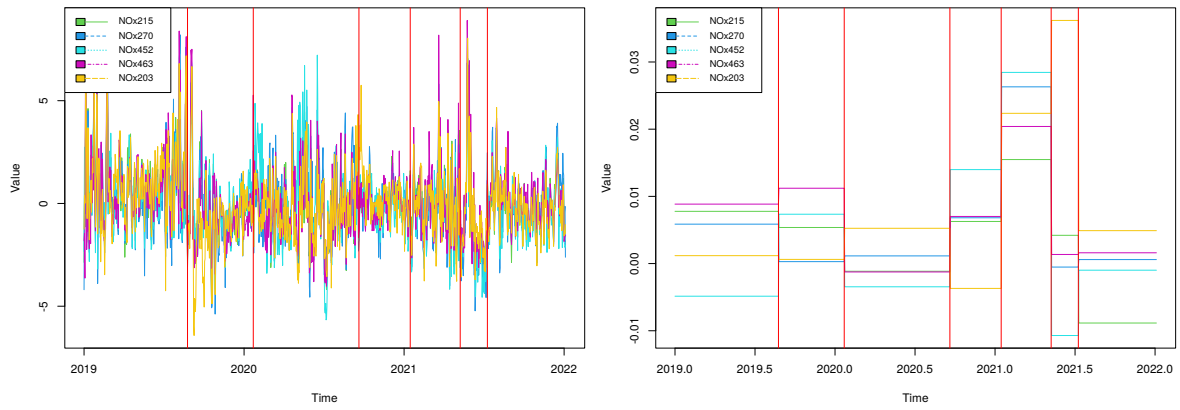
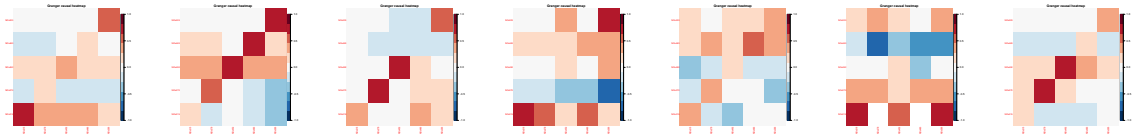


Figure 4.3: Left: $\sqrt{NO_x}$ levels (controlling for meteorological and seasonal effects, in $\mu g/m^3$), January 2019 - September 2022, in Bristol, UK. Different colours and line types indicate different detector locations. Right: Estimated intercept values for each estimated segment. Estimated change points marked with vertical lines.

We report results from the MOSUM score procedure described in Section 4.3. The recursive and multiscale procedures gave similar results. The MOSUM Wald procedure returned three change point estimators which are a subset of the estimators returned by the MOSUM score procedure. Tuning parameters chosen as recommended in Section 4.6.2: we use a bandwidth $G = 140$ covering twenty weeks; using the ϵ-criterion here would be unsuitable as the detector



01/2019–08/2019 08/2019–01/2020 01/2020–09/2020 09/2020–01/2021 01/2021–05/2021 05/2021–07/2021 07/2021–01/2022

Figure 4.4: Parameter heatmaps for each estimated segment for the air quality data studied in Section 4.6.5.1. Red hues denote large positive values and blue hues denote large negative values, within the interval $[-1, 1]$.

uniformly exceeds the threshold, so changes are located with $\eta = 0.25$.

We detect $\widehat{q} = 6$ change points located at August 24th 2019 ($\widehat{k}_1$), January 22 2020 ($\widehat{k}_2$), September 20 2020 ($\widehat{k}_3$), January 15 2021 ($\widehat{k}_4$), May 09 2021 ($\widehat{k}_5$), and July 10 2021 ($\widehat{k}_6$). Changes $\widehat{k}_2$ and $\widehat{k}_3$ mark the start and end of the initial Covid-19 lockdown, and the effect on emissions through reduced activity can be seen in the sharp drop in the intercept terms plotted in Figure 4.3. This is consistent with the findings of the street-level analyses of Jenkins et al. (2020).

Figure 4.4 plots heatmaps of the parameters for the estimated segments. In each segment, each series tends to exhibit strong positive autocorrelations, but the pattern of dependence across channels is subject to change. The cross-dependence is particular weak in the third estimated segment, corresponding to the Covid-19 lockdown, which also indicates that airborne matter was less mobile during this period.

Persistent dependence over long time scales is a commonly noted feature of air quality data. Chelani (2016) ascribe this to long memory and give physical justifications. However as noted by e.g. Norwood and Killick (2018), slow ACF decay can also be caused by structural breaks. To confirm whether the slow ACF decay is in fact due to structural breaks or long memory, we inspect the ACF on stationary segments. We look at site 452, the closest to the city centre, although our findings are similar across the sites. In Figure 4.5, we plot the empirical ACF estimated on the entire series, and on the first estimated segment $\{1, \ldots, \widehat{k}_1\}$. The ACF decays much faster in the latter, which provides an alternative explanation for the persistence observed in the air quality data.

4.6.5.2 Macroeconomic data

VAR models are often used in macroeconomics for forecasting and identifying shock transmissions. Aastveit et al. (2017) evaluate the evidence for parameter stability since the 2008 global financial crisis, allowing for smoothly-varying parameters or regime-switching behaviour. We analyse an extension of the same data using our methods. Our panel consists of $p = 4$ series: GDP, unemployment, core PCE inflation, and the federal funds rate, with quarterly data from 1959Q2 to 2022Q2, giving $n = 253$. We take the growth rate of GDP (i.e. GDP_t/GDP_{t-1}), and the first

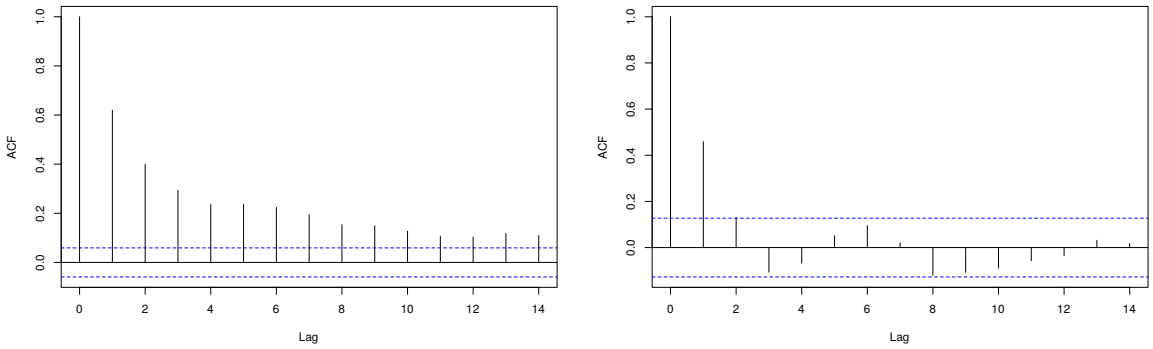


Figure 4.5: Empirical ACF of $\sqrt{NO_x}$ levels (controlling for meteorological and seasonal effects) at site 452. Left: Estimated on $t = 1,\dots,n$. Right: Estimated on $t = 1,\dots,\widehat{k}_1$.

differences of the other series, to account for unit-root behaviour[3].

We use the MOSUM score procedure, in combination with the dimension reduction procedure and bootstrap threshold discussed in Section 4.5. We set the VAR order as $d = 2$ following Aastveit et al. (2017) and adopt the bandwidth $G = 32$ corresponding to 8 years which is considered large enough to ensure reasonable estimation of $\boldsymbol{a}_{1,n}$. Changes are localised with the $\eta-$criterion, as the detector mostly exceeds the threshold, possibly due to model misspecification. The $\epsilon-$criterion would only return a single estimator here, so is not appropriate. Figure 4.6 plots the series and the $\widehat{q} = 3$ estimated change points, which are located at 1970Q3 ($\widehat{k}_1$), 1983Q2 ($\widehat{k}_2$), 1989Q1 ($\widehat{k}_3$), and 2013Q4 ($\widehat{k}_4$). Under varying specifications, the authors of Aastveit et al. (2017) find evidence for parameter breaks in 1980Q4, 1992Q4, and 2008Q4; the first two of these approximate our $\widehat{k}_2$ and $\widehat{k}_3$. They also find evidence for changes in the error covariance, at 1985Q1 and Q2, 1987Q2, and 2013Q1. For their method, this depends on the regression parameter, so we may also correspond the last of these to our $\widehat{k}_4$. We would expect to see an estimated change associated with the global financial crisis of 2008-2009, however the nearby estimated break $\widehat{k}_4$ takes precedence due to our choice of G and the use of the η-criterion. In Figure 4.6 the series visibly change in behaviour around 2020, likely due to the Covid-19 pandemic, though we do not have enough data post-event to detect any changes associated with this.

Considering the series of residuals from the model fitted on each (estimated) stationary segment, the multivariate (adjusted) portmanteau test (Lütkepohl, 2005, Section 4.4.4) rejects the null hypothesis of independence only for the final segment, using the multiplicity-corrected significance $\alpha = 0.05/5 = 0.01$. Figure 4.7 plots heatmaps of the parameters for each estimated segment. We can see that the parameters vary across segments, for example the cross-dependence structure changes from segment two to segment three. This observation corresponds to varying levels of volatility that can be seen in Figure 4.6. The penultimate segment lies mostly within the

[3]These are available from the FRED website `fred.stlouisfed.org/series`, and for simplicity we consider the most recent vintage.

so-called 'Great Moderation', a period of low volatility which ended in 2007, and this is reflected by small estimated parameter values. The large parameter values in the final segment reflect the period of volatility around the outbreak of the novel coronavirus in 2020, and explain the rejection by the portmanteau test.

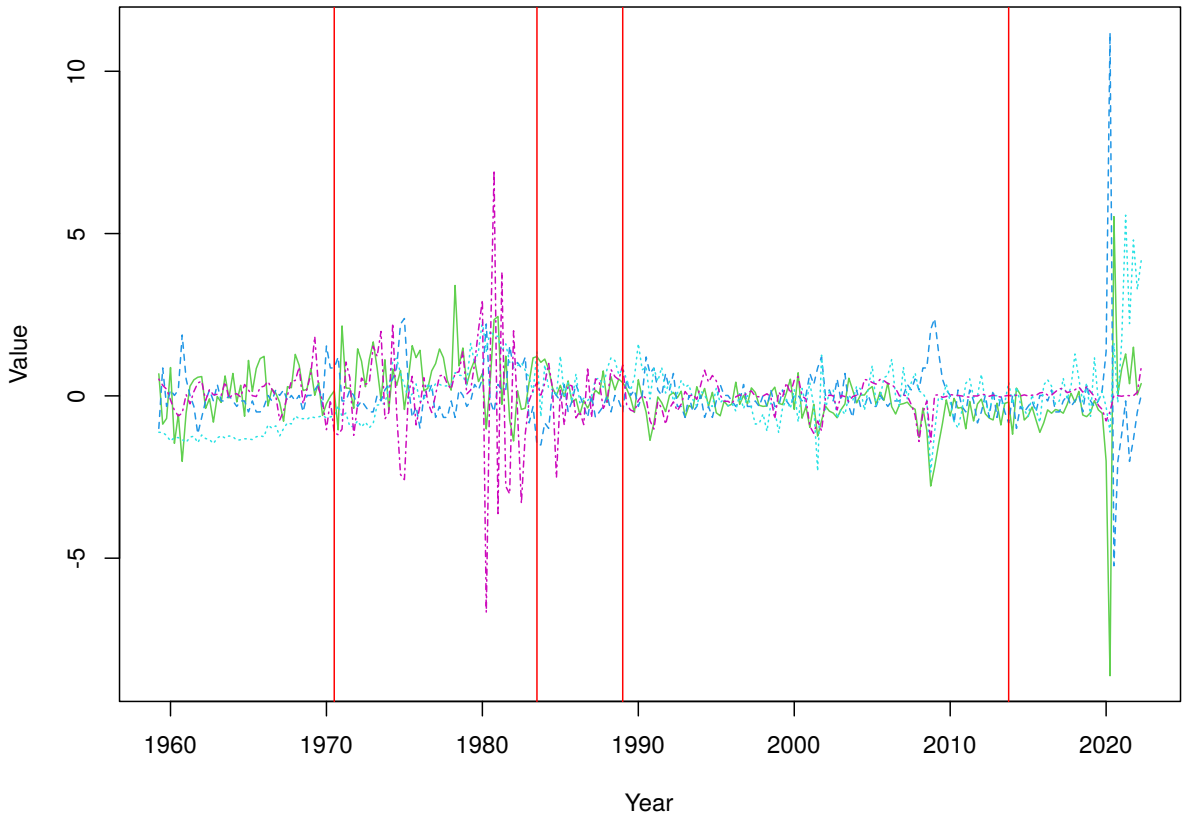


Figure 4.6: Macroeconomic panel time series studied in Section 4.6.5.2. Different colours and line types indicate different series. Estimated change points marked with vertical lines.

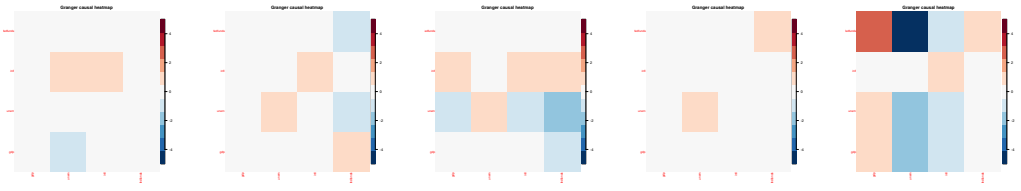


Figure 4.7: Parameter heatmaps for each estimated segment for the macroeconomic data studied in Section 4.6.5.2. Red hues denote large positive values and blue hues denote large negative values, inside the interval $[-5, 5]$.

4.7 Conclusion

We propose methods to detect and locate multiple change points in multivariate time series under a vector autoregressive model. We derive a score-type moving window procedure specific to this problem, and establish its consistency in multiple change point detection. Further, we propose extensions of the procedure which are designed to improve challenges associated with improving detection power, efficient computation and handling moderately large dimensions. To ease computation for large samples, we propose grid-based procedures, which reduce the dependence of the time complexity on the sample size to sub-linear order, but have the same asymptotic properties as the full-resolution procedure. A recursive segmentation method is proposed for change points which are otherwise undetectable, and a bottom-up merging procedure is used to detect multiscale changes. We also use a projection for dimension reduction, which can be combined with a parametric bootstrap, to allow the procedure to be used with panels of larger dimension. The methods are empirically validated with extensive simulation studies, and two applications to air quality and macroeconomic datasets. The extensions we have made to the MOSUM procedure are generic, so can be used for other models by specifying a different estimating function.

We have chosen the method discussed in this chapter, in comparison to a method designed for high-dimensional data such as that discussed in Chapter 3, as the current presentation allows precise asymptotic statements on consistency in the fixed-dimensional regime, which allows the choice of a threshold in the procedure which is motivated by theory rather than the data. Moreover, we have no need for structural assumptions such as sparsity, which can be restrictive or unrealistic, and have fewer tuning parameters to choose as a result.

CHAPTER 5

SEGMENTING AND FORECASTING NONSTATIONARY FACTOR-AUGMENTED REGRESSION MODELS

5.1 Introduction

Macroeconomic and financial time series data often exhibit two properties. First, information is shared across series via strong correlations (Stock and Watson, 2002a). Second, instabilities occur in the generating process (Rossi, 2021; Stock and Watson, 1996, 2002b). In light of this, how should we produce forecasts? The first point is often addressed by using factor models, which assume that the shared variation in large panels can be explained by a low-dimensional model. For the second, a simple, interpretable explanation is the piecewise-stationary assumption, where model parameters are constant between unknown change points. The task for us is then to identify the number and locations of changes, and incorporate this knowledge into factor model forecasts.

In this work, we study the diffusion index forecasting model (Bai and Ng, 2006; Stock and Watson, 2002a,c), a regression model with observable and latent (factor) regressors. We observe tuples $\{(y_t, \boldsymbol{X}_t, \boldsymbol{u}_t)\}_{t=1}^n$, where y_t is the diffusion index of interest, $\boldsymbol{X}_t = (X_{1t}, X_{2t}, \ldots, X_{pt})^\top \in \mathbb{R}^p$ is a vector of dependent time series, and $\boldsymbol{u}_t \in \mathbb{R}^{p_y - r}$ is a vector of covariates.

One goal is to predict $\boldsymbol{X}_{n+1}$ given the history $\{\boldsymbol{X}_t\}_{t=1}^n$. For this, we assume $\boldsymbol{X}_t$ follows a static factor model

$$\boldsymbol{X}_t = \boldsymbol{\chi}_t + \boldsymbol{\varepsilon}_t. \tag{5.1}$$

The common component $\{\boldsymbol{\chi}_t\}_{t\in\mathbb{Z}}$, where $\boldsymbol{\chi}_t = \boldsymbol{\Lambda}\boldsymbol{F}_t$, consists of latent factors $\boldsymbol{F}_t = (F_{1t}, F_{2t}, \ldots, F_{rt})^\top \in \mathbb{R}^r$, multiplied by $\boldsymbol{\Lambda} \in \mathbb{R}^{p\times r}$, a fixed matrix of loadings. The factor dynamics follow a piecewise stationary vector autoregressive (VAR) process for which the parameters are piecewise constant between change points; we can use the VAR structure to forecast $\boldsymbol{F}_{n+1}$, and hence forecast $\boldsymbol{X}_{n+1}$.

The idiosyncratic component $\{\boldsymbol{\varepsilon}_t\}_{t=1}^n$ is a zero-mean, white noise process. We describe the model in detail in Section 5.2.

Another goal is to predict y_n using $\boldsymbol{X}_n$ and $\boldsymbol{u}_n$. For this we consider a factor-augmented regression model for the diffusion index y_t, such that

$$y_t = \boldsymbol{\beta}^\top \boldsymbol{z}_t + \varepsilon_t^y \tag{5.2}$$

where $\boldsymbol{z}_t = \left(\boldsymbol{F}_t^\top, \boldsymbol{u}_t^\top\right)^\top \in \mathbb{R}^{p_y}$. In this work we also treat the regression parameter $\boldsymbol{\beta} \in \mathbb{R}^{p_y}$ as piecewise constant between change points. $\{\varepsilon_t^y\}_{t=1}^n$ is a univariate zero-mean, white noise process with $\mathrm{Var}(\varepsilon_t^y) = \sigma^2$. For the final goal, to predict y_{n+1} we require a forecast for $\boldsymbol{F}_{n+1}$, and so the two other tasks are complementary.

To account for piecewise-stationarity, we develop data segmentation methods to identify structural breaks in (i) the low-dimensional VAR parameters, and (ii) the diffusion index regression parameters, while not assuming common breaks across the two. The estimated breaks are incorporated into short-term forecasts by re-weighting the data during model estimation.

In simulations we show that our method has competitive performance for change point detection, and highlight the benefit of weighted forecasts under discrete breaks. In an application to real macroeconomic data with economic indicators and a risk premium, we identify changes corresponding to real events and show our forecasts compare favourably to the widely-used rolling window approach.

Data segmentation for factor models There is a vast literature on instability in factor models. Broadly, these can be divided according to the part (or parts) of the model in which instability is present. These are the (i) loadings, (ii) factor number, and (iii) second-order structure. For a single break in (i) and (ii), tests and estimators are proposed by e.g. Breitung and Eickmeier (2011), Chen et al. (2014), Corradi and Swanson (2014), Han and Inoue (2015), Bai et al. (2020a), Duan et al. (2023), and Koo et al. (2023), and multiple changes are considered in Su and Wang (2017), Ma and Su (2018), and Liu and Zhang (2021b). Less focus, however, has been given to changes in (iii). Barigozzi et al. (2018) consider a static factor model with changes in both the factor and idiosyncratic components. Cho et al. (2022) do the same, assuming a generalised dynamic factor model and a VAR structure in the idiosyncratic component, while Barigozzi and Trapani (2020) propose a sequential method. Kim et al. (2021) consider testing for a single change in the low-dimensional VAR parameters of a static factor model. This is a similar but distinct problem to segmentation under a low-rank VAR model (Bai et al., 2020b, 2023; Enikeeva et al., 2023).

Under the factor-augmented regression model, Corradi and Swanson (2014) and Massacci (2019) perform inference for a single break, while Wang et al. (2015) perform estimation. For forecasting, Banerjee et al. (2008) allow for instability in the loadings of the factor model, while Stock and Watson (2009) allow for instability in the regression relationship.

Estimation and forecasting under structural breaks When forecasting with parameter instability, the problem is to identify which window(s) of observations to use, and how to weight the resulting data. This has been considered in a range of sources such as Pesaran et al. (2013), which considers optimal forecasts under sudden or smooth changes. Intuitively, if changes occur sufficiently far in the past, the model can be estimated well on post-break observations. When breaks are recent, however, as with the online setting, Pesaran and Timmermann (2007) make the case that the bias-variance trade-off can be exploited by using pre-break data, provided breaks are not too large. Hännikäinen (2017) compares window-based methods available in the literature, namely rolling windows, exponentially-weighted moving averages, and the average window method of Pesaran and Pick (2011). Assenmacher-Wesche and Pesaran (2008) and Pesaran et al. (2009) focus on VARs in particular. Stock and Watson (2009) consider forecasting from a Dynamic Factor Model (DFM) in the presence of instabilities. Remarkably, they find that using the entire sample for estimation but only a subsample for forecasting improves the forecasting performance when compared to using subsamples for both. Bates et al. (2013) discusses estimation of the DFM under instability, showing that potentially 'large' breaks can be safely ignored for the consistent estimation of the loadings, but not in estimating the factor number. Giraitis et al. (2015) allows varying memory processes, generalising the adaptive window methods and discussing data-driven choices for the window size. Massacci and Kapetanios (2023) study the effect of a structural break on forecasts from a factor-augmented regression.

Often in practice, either rolling windows of fixed length or the entire sample are used for estimation and prediction. As demonstrated in Rossi (2021), this may lead to a deterioration in forecasting performance.

The broader concept of learning in non-stationary environments, known as *concept drift* (Krempl et al., 2021; Lu et al., 2018), has recently received much attention. Our proposal offers a method for learning under a specific model when discrete shifts occur, and this could be easily adapted to other models.

Diffusion indices Diffusion index (or factor-augmented regression) forecasting is well-studied, and often applied to GDP and inflation data (Bai and Ng, 2009; Stock and Watson, 1999). This is related to the *nowcasting* problem, where the goal is to predict a response observed at a low frequency given access to higher-frequency regressors. Bańbura et al. (2013) reviews the problem, and mentions a stationary version of the model in (5.3). Bell et al. (2014) mention parameter instability as a challenge to be addressed.

We highlight the following contributions made in this work.

(i) **Multiple changes in factor dynamics and factor-augmented regressions.** We are the first to propose a method to detect multiple changes in the latent VAR parameters of a factor model, or in the parameters of a factor-augmented regression model. We also give recursive segmentation schemes and multi-scale algorithms for data-adaptive detection.

(ii) **Computational efficiency.** By scanning over a coarse grid $\mathscr{T} \subset \{1, \dots, n\}$ we greatly reduce the number of times we must obtain parameter estimates, reducing the complexity to be sub-linear in the sample size. We then perform a cheap localisation step to obtain accurate estimators. This is accompanied by a fast implementation combining R and C++.

(iii) **Prediction in high dimensions under piecewise-stationarity.** We extend the literature on weighted forecasting in the presence of discrete breaks to two high dimensional regression settings, which combine to give a competitive method for forecasting real financial and macroeconomic panels and diffusion indices.

Notation Let $\mathbb{R}$, $\mathbb{Z}$, and $\mathbb{N}$ denote the sets of real numbers, integers and natural numbers. We let $\boldsymbol{O}$ and $\mathbf{0}$ be a matrix and vector of zeros, respectively, and $\boldsymbol{I}_p$ be the $p \times p$ identity matrix. Let $\|\cdot\|$ denote the Euclidean norm of a vector or the Spectral norm of a matrix, and let $\|\cdot\|_F$ denote the Frobenius norm. We denote the Kronecker product of two matrices $\boldsymbol{A} = [a_{ij}] \in \mathbb{R}^{m \times n}$ and $\boldsymbol{B}$, by

$$\boldsymbol{A} \otimes \boldsymbol{B} = \begin{pmatrix} a_{11}\boldsymbol{B} & a_{12}\boldsymbol{B} & \dots & a_{1n}\boldsymbol{B} \\ a_{21}\boldsymbol{B} & a_{22}\boldsymbol{B} & & \vdots \\ \vdots & & \ddots & \\ a_{m1}\boldsymbol{B} & \dots & & a_{mn}\boldsymbol{B} \end{pmatrix}.$$

For a sequence of random variables $\{X_n\}_{n \in \mathbb{N}}$, let $X_n \xrightarrow{\mathscr{D}} X$ and $X_n \xrightarrow{P} X$ denote convergence in distribution and probability, respectively. Let N_p denote the p-dimensional normal distribution. We denote by Γ the Gamma function. Finally, we write $a \vee b = \max(a, b)$ and $a \wedge b = \min(a, b)$.

5.2 Piecewise stationary models

5.2.1 Piecewise stationary factor VAR model

In this section we discuss the piecewise-stationary factor model with VAR dynamics, contained in (5.1). The model assumes a specific *static* representation such that the observations depend only contemporaneously on the factors, as opposed to the DFM proposed in Forni et al. (2000) which allows $\boldsymbol{X}_t$ to depend on $\boldsymbol{F}_t$ at previous lags. The factors follow a piecewise stationary VAR model

$$\boldsymbol{F}_t = \begin{cases} \boldsymbol{F}_t^{(1)}, & k_0 + 1 = 1 \leq t \leq k_1, \\ \boldsymbol{F}_t^{(2)}, & k_1 + 1 \leq t \leq k_2, \\ \vdots & \\ \boldsymbol{F}_t^{(q+1)}, & k_q + 1 \leq t \leq k_{q+1} = n, \end{cases} \tag{5.3}$$

where each $\{\boldsymbol{F}_t^{(j)}\}_{t\in\mathbb{Z}}$ follows a stationary VAR(d) process (in the sense of (2.4), or Equation (2.1.9) of Lütkepohl (2005)), i.e.

$$\boldsymbol{F}_t^{(j)} = \boldsymbol{a}_j \mathbb{F}_{t-1}^{(j)} + \boldsymbol{\eta}_t, \quad \text{where} \quad \boldsymbol{a}_j = \begin{bmatrix} \boldsymbol{a}_j(1)^\top \\ \vdots \\ \boldsymbol{a}_j(r)^\top \end{bmatrix} \in \mathbb{R}^{r\times dp} \text{ and } \mathbb{F}_{t-1}^{(j)} = \begin{bmatrix} \mathbb{F}_{1,t-1}^{(j)} \\ \vdots \\ \mathbb{F}_{r,t-1}^{(j)} \end{bmatrix} \in \mathbb{R}^{dr}$$

for $j = 1,\ldots,q+1$. Here, $\mathbb{F}_{i,t-1}^{(j)} = (F_{i,t-1}^{(j)},\ldots,F_{i,t-d}^{(j)})^\top$ collects the d lagged values of $F_{it}^{(j)}$, the i-th channel of $\boldsymbol{F}_t^{(j)}$, and $\boldsymbol{a}_j(i)$ collects the parameters involved in predicting the i-th channel. There are a fixed number q of change points at unknown locations k_j, $1 \le j \le q$, which obey a linear spacing such that $k_j = \lfloor \lambda_j n \rfloor$ for $0 = \lambda_0 < \lambda_1 < \ldots, \lambda_q < \lambda_{q+1} = 1$, and such that $\boldsymbol{a}_j \neq \boldsymbol{a}_{j+1}$ for all j. Our aim is to estimate the total number and the locations of the q change points. We assume $\{\boldsymbol{\eta}_t\}_{t=1}^n$ is a zero-mean, independent process such that $\mathbb{E}(\boldsymbol{\eta}_t) = \boldsymbol{0}$, $\mathrm{Cov}(\boldsymbol{\eta}_t) = \boldsymbol{S}$ for some positive definite matrix $\boldsymbol{S} \in \mathbb{R}^{p\times p}$, and $\mathrm{Cov}(\boldsymbol{\eta}_t, \boldsymbol{\eta}_{t'}) = \boldsymbol{O}$ for any $t \neq t'$.

The measurement equation (5.1) and the state equation (5.3) form a state space model. Ours generalises the globally stationary representation of the *approximate DFM* considered for example in Giannone et al. (2008) and Forni et al. (2009). We note that our model is contained within the factor-augmented VAR (Bai et al., 2016; Bernanke et al., 2005), though ours is distinguished by the fact that we do not require identification restrictions for the purposes of forecasting.

Remark 5.1 (Characterising changes)**.** In the most flexible factor models, structural changes may occur in (a) the loadings, (b) the (auto)correlation structure, or (c) the factor number. As discussed in e.g. Barigozzi et al. (2018), Appendix A, and Duan et al. (2023), these changes are not identifiable from the data; we discuss their representation by (5.1) and (5.3) as follows.

(a) Rotational changes in the loadings, where the rank of the factor series is equal before and after the change, can be represented by a rotation in the factor space. Suppose that for $t = 1,\ldots,k_1$ we have $\boldsymbol{\chi}_t^{(1)} = \boldsymbol{\Lambda}^{(1)}\boldsymbol{F}_t^{(1)}$ and for $t = k_1+1,\ldots,n$ we have $\boldsymbol{\chi}_t^{(2)} = \boldsymbol{\Lambda}^{(2)}\boldsymbol{F}_t^{(z,2)}$, where $\boldsymbol{\Lambda}^{(2)} = \boldsymbol{\Lambda}^{(1)}\boldsymbol{Z}$ for some nonsingular $\boldsymbol{Z} \in \mathbb{R}^{r\times r}$, and $\boldsymbol{F}_t^{(z,2)}$ is the latent series. We can define the factors of interest as $\boldsymbol{F}_t^{(2)} = \boldsymbol{Z}\boldsymbol{F}_t^{(z,2)}$, and hence we have a representation with a constant loading matrix. Since $\boldsymbol{F}_t^{(z,2)} = \sum_{l=1}^d \boldsymbol{A}_l^{(z,2)}\boldsymbol{F}_{t-l}^{(z,2)} + \boldsymbol{\eta}_t$, we have the representation $\boldsymbol{F}_t^{(2)} = \sum_{l=1}^d \boldsymbol{A}_l^{(2)}\boldsymbol{F}_{t-l}^{(2)} + \boldsymbol{Z}\boldsymbol{\eta}_t$ where $\boldsymbol{A}_l^{(2)} = \boldsymbol{Z}\boldsymbol{A}_l^{(z,2)}$.

(b) Changes in the autocorrelation structure will be accounted for in the autoregression parameters in (5.3). We may allow $\boldsymbol{S}$, the covariance for the innovation process, to change, but these will not be detectable. We could extend the methodology to include $\boldsymbol{S}$ in the estimating function (see Section 5.3.1) to account for this.

(c) If the factor number changes, we would need to represent the newly-arriving (or leaving) factor as a degenerate variable outside of any active regime. This violates Assumption 5.1 (ii), and so cannot be represented. Indeed, any change which enlarges the factor space over the whole sample cannot be represented by our model.

5.2.2 Piecewise stationary factor-augmented regression model

Recalling (5.2), $y_t, t = 1, \ldots, n$ are generated as

$$y_t = y_t^{(j)} = \boldsymbol{z}_t^\top \boldsymbol{\beta}_j + \varepsilon_t^y \tag{5.4}$$

for $k_{j-1}^y + 1 \leq t \leq k_j^y$, where $j = 1, \ldots, q^y + 1$. As in (5.3), the number q^y and locations $k_j^y = \lfloor \lambda_j^y n \rfloor, 1 \leq j \leq q^y$ of changes are unknown and to be estimated from the data. As noted, the number and locations need not be the same as q and $k_j, 1 \leq j \leq q$, which allows for potentially strongly divergent structures between y_t and the components of $\boldsymbol{X}_t$, as well as a very flexible forecasting model.

Remark 5.2 (Nowcasting)**.** The factor-augmented regression model can be generalised to the case where y_t is sampled at a lower frequency than $\boldsymbol{X}_t$, for example quarterly samples relative to monthly samples. This is relevant for the nowcasting problem (Bańbura et al., 2013), and the model can be used to produce intra-period forecasts and nowcasts. Under infrequent sampling we may only have enough data to assume stationarity of the regression relationship (i.e. $q^y = 0$).

5.3 Methodology

5.3.1 Data segmentation methodology

We now propose a method to detect change points for the model described in (5.1)–(5.3). The idea is to first recover the factor series (up to rotation) by using Principal Components Analysis (PCA) on the sample covariance matrix, selecting the first $r \leq p$ components. We then supply the resulting series to the `mosumvar` algorithm proposed in Chapter 4. Moving sum (MOSUM) procedures aim to detect and locate changes the expectation of a given series, comparing sums over G-length intervals before and after a candidate change point $k \in \{G, \ldots, n - G\}$. Similar moving window-based approaches have been used in, for example, Bauer and Hackl (1980); Cho and Owens (2022); Eichinger and Kirch (2018); Hušková (1990); Preuss et al. (2015b); Yau and Zhao (2016). We formally describe the procedure in Algorithm 5.

Algorithm 5: `mosumfvar`: Moving window data segmentation under a factor VAR model

input : Data $\{\boldsymbol{X}_t\}_{t=1}^n$, Bandwidth G, Threshold D, VAR order d, Factor number r
initialise: $\widehat{\mathcal{K}} = \emptyset$

Step 1: Recover $\widehat{\boldsymbol{\Lambda}}$, $\widehat{\boldsymbol{F}}_t = (F_{jt}, 1 \leq j \leq r)^\top$ for $t = 1, \ldots, n$ using PCA
Step 2: $\widehat{\mathcal{K}} \leftarrow$ `mosumvar`$(\{\widehat{\boldsymbol{F}}_t\}_{t=1}^n, G, D, d)$ (Algorithm 6)
return $\widehat{\mathcal{K}}$

Step 1: Factor analysis For the sample covariance matrix $\widehat{\boldsymbol{\Gamma}}_x = n^{-1} \sum_{t=1}^n \mathbf{X}_t \mathbf{X}_t^\top$, let $\widehat{\mathbf{w}}_{x,j}$ denote the normalised eigenvector corresponding to its j-th largest eigenvalue $\widehat{\mu}_{x,j}$, with its entries

$\widehat{w}_{x,ij}, i = 1,\dots,p$. Then, for a given number of factors $r \geq 1$, the factors are estimated by $\widehat{\mathbf{F}}_t = \left(\widehat{F}_{1t},\dots,\widehat{F}_{rt}\right)^\top$, up to an orthogonal rotation matrix $\boldsymbol{R}$ with $\widehat{F}_{jt} = \widehat{\mathbf{w}}_{x,j}^\top \mathbf{X}_t/\sqrt{p}$. Then, the common components are estimated by $\widehat{\chi}_{it} = \sum_{j=1}^r \widehat{\lambda}_{ij}\widehat{F}_{jt}$ with the estimated loadings $\widehat{\lambda}_{ij} = \sqrt{p}\,\widehat{w}_{x,ij}$, and the idiosyncratic components by $\widehat{\varepsilon}_{it} = X_{it} - \widehat{\chi}_{it}$. If r is unknown, we may instead use a data-driven choice $\widehat{r}$ (see Section 5.5.3).

Step 2: VAR model segmentation We look for changes in the second-order structure of $\boldsymbol{F}_t$ by detecting changes in the parameters of the VAR structure of $\widehat{\boldsymbol{F}}_t$. By the mosumvar methodology of Chapter 4, this amounts to finding changes in the expectation of an estimating function. We specify the estimating function

$$\boldsymbol{H}_t(\widetilde{\boldsymbol{a}}) = \boldsymbol{H}\left(\widehat{\boldsymbol{F}}_t, \widehat{\mathbb{F}}_{t-1}, \widetilde{\boldsymbol{a}}\right) = -(\widehat{\boldsymbol{F}}_t - \widetilde{\boldsymbol{a}}\widehat{\mathbb{F}}_{t-1}) \otimes \widehat{\mathbb{F}}_{t-1},$$

which corresponds to the least squares objective, for estimated factors, where $\widetilde{\boldsymbol{a}}$ is a data-dependent inspection parameter. We scan the data with the score detector

$$\begin{aligned} \widehat{T}_k(G,\widetilde{\boldsymbol{a}}) &= \frac{1}{\sqrt{2G}} \left\| (\widehat{\boldsymbol{\Sigma}}_k(\widetilde{\boldsymbol{a}}))^{-1/2} \widehat{\boldsymbol{m}}_k(G,\widetilde{\boldsymbol{a}}) \right\|, \text{ for } k = G,\dots,n-G, \text{ where} \\ \widehat{\boldsymbol{m}}_k(G,\widetilde{\boldsymbol{a}}) &= \sum_{t=k+1}^{k+G} \boldsymbol{H}_t(\widetilde{\boldsymbol{a}}) - \sum_{t=k-G+1}^{k} \boldsymbol{H}_t(\widetilde{\boldsymbol{a}}), \end{aligned} \tag{5.5}$$

and $\widehat{\boldsymbol{\Sigma}}_k(\widetilde{\boldsymbol{a}})$ is an estimator for $\boldsymbol{\Sigma}_{(j)}(\widetilde{\boldsymbol{a}}) = \mathrm{Cov}(\boldsymbol{H}(\boldsymbol{X}_1^{(j)}, \mathbb{X}_0^{(j)}, \widetilde{\boldsymbol{a}}))$ for $k_j + 1 \leq k \leq k_{j+1}$. We discuss the selection of $\widehat{\boldsymbol{\Sigma}}_k(\widetilde{\boldsymbol{a}})$ in Appendix C.1.1.1. Consider the maximum detector statistic $\widehat{T}(G,\widetilde{\boldsymbol{a}}) = \max_{G+d\leq k\leq n-G} \widehat{T}_k(G,\widetilde{\boldsymbol{a}})$. With this, we test the null hypothesis of no change, $H_0 : q = 0$, against the alternative $H_1 : q \geq 1$:

$$\text{Reject } H_0 \text{ if } \widehat{T}(G,\widetilde{\boldsymbol{a}}) > D(G,\alpha). \tag{5.6}$$

Here, the critical value $D(G,\alpha)$ is derived from the asymptotic null distribution of $\widehat{T}(G,\widetilde{\boldsymbol{a}})$ at a given significance level $\alpha \in (0,1)$, see Assumption 5.10 below.

When H_0 is rejected, there is evidence for one or more change points, and we wish to estimate their number and locations. With an appropriately chosen $\widetilde{\boldsymbol{a}}$, we expect $\widehat{T}_k(G,\widetilde{\boldsymbol{a}})$ to take large values around the change points. As such, we estimate k_j with the locations of significant local maximisers. To automatically identify these, we adopt a criterion proposed in Eichinger and Kirch (2018). We consider all pairs of indices $\left(v_j, w_j\right)$ such that for some $\epsilon \in (0, 1/2)$,

$$\begin{aligned} &\widehat{T}_k(G,\widetilde{\boldsymbol{a}}) \geq D\left(\alpha, G\right) \text{ for } v_j \leq k \leq w_j, \text{ and} \\ &\widehat{T}_k(G,\widetilde{\boldsymbol{a}}) < D\left(\alpha, G\right) \text{ for } k = v_j - 1, w_j + 1, \end{aligned} \tag{5.7}$$

with $w_j - v_j \geq \epsilon G$. We take the number of these pairs as an estimator for the number of changes:

$$\widehat{q} = \widehat{q}(\widetilde{\boldsymbol{a}}) = \text{ number of pairs } \left(v_j, w_j\right),$$

and for each $j = 1, \ldots, \widehat{q}$, we estimate the location of a change point by

$$\widehat{k}_j = \widehat{k}_j(\widetilde{\boldsymbol{a}}) = \arg\max_{v_j \le k \le w_j} \widehat{T}_k(G, \widetilde{\boldsymbol{a}}).$$

The procedure is presented in Algorithm 6. Define the estimator

$$\widehat{\boldsymbol{a}}_w = (\sum_{t=1}^{n} w_t \widehat{\mathbb{F}}_{t-1} \widehat{\mathbb{F}}_{t-1}^\top)^{-1} (\sum_{t=1}^{n} w_t \widehat{\mathbb{F}}_{t-1} \widehat{\boldsymbol{F}}_t^\top) \tag{5.8}$$

where w_t are estimator weights. When the factors are observed, Chapter 4 suggests to use $\widetilde{\boldsymbol{a}} = \widehat{\boldsymbol{a}}_{1,n}$, the estimator uniquely solving $\sum_{t=1}^{n} \boldsymbol{H}_t(\widehat{\boldsymbol{a}}_{1,n}) = \boldsymbol{0}$, which is equivalent to (5.8) with $w_t = 1$ for all $t = 1, \ldots, n$.

Algorithm 6: `mosumvar`: Moving sum data segmentation under a VAR model

input : Data $\{\widehat{\boldsymbol{F}}_t\}_{t=1}^n$, Bandwidth G, Threshold D, VAR order d
initialise: $\widehat{\mathscr{K}} = \emptyset$

Compute inspection parameter $\widetilde{\boldsymbol{a}} = \widehat{\boldsymbol{a}}_{1,n}$
Calculate $\widehat{T}_k(G, \widetilde{\boldsymbol{a}}), k = G + d, \ldots, n - G$ as in (5.5)
if $\widehat{T}(G, \widetilde{\boldsymbol{a}}) > D$ **then**
  Locate $\widehat{\mathscr{K}} \leftarrow \{\widehat{k}_j, 1 \le j \le \widehat{q}\}$ according to (5.7)
return $\widehat{\mathscr{K}}$

Step 3: Regression model segmentation Let $\widehat{\boldsymbol{z}}_t = \left(\widehat{\boldsymbol{F}}_t^\top, \boldsymbol{u}_t^\top\right)^\top$. Here, the observed estimating function is

$$\boldsymbol{H}^y\left(y_t, \widehat{\boldsymbol{z}}_t, \widetilde{\boldsymbol{\beta}}\right) = -(y_t - \widehat{\boldsymbol{z}}_t^\top \widetilde{\boldsymbol{\beta}}) \widehat{\boldsymbol{z}}_t.$$

The detector here is

$$\widehat{T}^y(G, \widetilde{\boldsymbol{\beta}}) = \max_{G \le k \le n-G} \widehat{T}_k^y(G, \widetilde{\boldsymbol{\beta}}), \quad \widehat{T}_k^y(G, \widetilde{\boldsymbol{\beta}}) = \frac{1}{\sqrt{2G}} \left\| (\widehat{\boldsymbol{\Sigma}}_k^y(\widetilde{\boldsymbol{\beta}}))^{-1/2} \widehat{\boldsymbol{m}}_k(G, \widetilde{\boldsymbol{\beta}}) \right\|,$$

where $\widehat{\boldsymbol{\Sigma}}_k^y(\widetilde{\boldsymbol{\beta}})$ is an estimator for $\boldsymbol{\Sigma}_{(j)}^y(\widetilde{\boldsymbol{\beta}}) = \mathrm{Cov}(\boldsymbol{H}(y_1, \boldsymbol{z}_1^{(j)}, \widetilde{\boldsymbol{\beta}}))$ for $k_j^y + 1 \le k \le k_{j+1}^y$. We discuss the selection of $\widehat{\boldsymbol{\Sigma}}_k^y(\widetilde{\boldsymbol{\beta}})$ in Appendix C.1.2. The difference vector at time k, evaluated with inspection parameter $\widetilde{\boldsymbol{\beta}}$, is

$$\widehat{\boldsymbol{m}}_k^y(G, \widetilde{\boldsymbol{\beta}}) = \sum_{t=k+1}^{k+G} \boldsymbol{H}^y\left(y_t, \widehat{\boldsymbol{z}}_t, \widetilde{\boldsymbol{\beta}}\right) - \sum_{t=k-G+1}^{k} \boldsymbol{H}^y\left(y_t, \widehat{\boldsymbol{z}}_t, \widetilde{\boldsymbol{\beta}}\right).$$

The segmentation procedure is then similar to Step 2; see Algorithm 7. Define the estimator

$$\widehat{\boldsymbol{\beta}}_w = (\sum_{t=1}^{n} w_t \widehat{\boldsymbol{z}}_t \widehat{\boldsymbol{z}}_t^\top)^{-1} (\sum_{t=1}^{n} w_t \widehat{\boldsymbol{z}}_t y_t). \tag{5.9}$$

We may choose $\widetilde{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}_{1,n}$, the estimator uniquely solving $\sum_{t=1}^{n} \boldsymbol{H}^y(y_t, \widehat{\boldsymbol{z}}_t, \widehat{\boldsymbol{\beta}}_{1,n}) = \boldsymbol{0}$, which equivalent to (5.9) with $w_t = 1$ for all $t = 1, \ldots, n$.

Algorithm 7: `mosumlm`: Moving sum data segmentation under a linear regression model

input : Data $\{\widehat{\boldsymbol{F}}_t\}_{t=1}^n$, Bandwidth G, Threshold D
initialise: $\widehat{\mathscr{K}}^y = \emptyset$

Compute inspection parameter $\widetilde{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}_{1,n}$
Calculate $\widehat{T}_k^y(G, \widetilde{\boldsymbol{\beta}}), k = G, \ldots, n-G$
if $\widehat{T}^y(G, \widetilde{\boldsymbol{\beta}}) > D$ **then**
  Locate $\widehat{\mathscr{K}}^y \leftarrow \{\widehat{k}_j^y, 1 \le j \le \widehat{q}^y\}$ similarly to (5.7)
return $\widehat{\mathscr{K}}^y$

5.3.1.1 Extensions

In Chapter 4, we propose a range of methodological extensions which can also be used in combination with Step 2. While these are proposed for use with the VAR segmentation methodology, they can be applied just as easily to regression segmentation.

Grid-based procedure We can reduce the procedure's computational complexity (see Section 5.5.1) by evaluating the detector not on the whole sample, but on a coarse grid

$$\mathscr{T}(r,G) = \left\{ t : t = G + d + m\lfloor rG \rfloor, 0 \le m \le \left\lfloor \frac{n-2G}{rG} \right\rfloor \right\}.$$

for some constant $r \in [G^{-1}, 1]$. After scanning over the grid, intervals with detectors exceeding the threshold are filled in with score statistics with a inspection parameter calculated over each contiguous significant interval.

Multiscale algorithm The theory justifying our method relies on Assumption 5.7, where the chosen bandwidth G becomes small enough asymptotically such that each change point will be isolated within a window of length $2G$. This suggests we want to select G to be as large as possible so that this holds, as to maximise detection power and give the best possible location estimators. The true minimum spacing between changes is always unobservable, however, and often hard to make reasonable prior choices for in practice. Moreover, multiscale changes may be present in the data, where large frequent and small infrequent changes occur in the same series. To account for these issues, we propose a bottom-up multiscale algorithm which runs the single-scale procedure with bandwidths of increasing size and merges the resulting sets of change point estimates. Similar ideas are proposed in Messer et al. (2014) and Meier et al. (2021) for the mean-change problem.

Define the bandwidth set $\mathscr{G} = \{G_h, 1 \le h \le H : G_1 < \ldots < G_H\}$ with strictly increasing elements. The null hypothesis is rejected if any of test statistics evaluated with bandwidth $G_h \in \mathscr{G}$ exceed their respective thresholds $D(G_h, \alpha)$. For the smallest bandwidth at which we reject, we take the estimated change points as the initial set of changes. For each subsequent bandwidth $G_{h'}$, a

detected point is added to the set only if the point is at least $\pi G_{h'}$ away from any point already in the change point set, for some $\pi \in (0,1)$. We formalise this in Algorithm 8.

Algorithm 8: Multiscale MOSUM Algorithm

input : Data $\{\widehat{\boldsymbol{F}}_t\}_{t=1}^n$, Bandwidth set $\mathcal{G}$, Threshold set $\{D(G_h,\alpha), h = 1,\ldots,H\}$, Localisation parameter π

initialise : $\widehat{\mathcal{K}} = \emptyset$

Compute inspection parameter $\widetilde{\boldsymbol{a}} = \widehat{\boldsymbol{a}}_{1,n}$ **for** $h = 1,\ldots,H$, **do**
  Calculate $\widehat{T}_k(G_h, \widetilde{\boldsymbol{a}})$ as in (5.5)
  if $\widehat{T}(G_h, \widetilde{\boldsymbol{a}}) > D(G_h,\alpha)$ **then**
    Locate $\widehat{\mathcal{K}}(G_h) \leftarrow \{\widehat{k}_j(G_h), 1 \le j \le \widehat{q}_h\}$ with (5.7)
    for $\widehat{k}_j \in \widehat{\mathcal{K}}(G_h)$ **do**
      if $\min_{k\in\widehat{\mathcal{K}}} |\widehat{k}_j(G_h) - k| \ge \epsilon G_h$ **then**
        add $\widehat{k}_j(G_h)$ to $\widehat{\mathcal{K}}$
    end
end
return $\widehat{\mathcal{K}}$

Recursive segmentation As discussed in Remark 4.1, when using the MOSUM score procedure with a global inspection parameter, there may exist changes which are not detectable in the sense of Assumption 5.8. By combining the MOSUM score procedure with the Binary Segmentation (BS) algorithm (Vostrikova, 1981), we recursively apply the MOSUM score procedure with data-dependent inspection parameters, giving a procedure which asymptotically detects all change points.

Algorithm 9 describes the MOSUMBS algorithm. For a given segment $\{s,\ldots,e\}$ for some $1 \le s \le e \le n$, a local parameter estimator $\widehat{\boldsymbol{a}}_{s,e}$ is obtained as in (5.8) (computed only over the local sample) with which the MOSUM score procedure is performed and change point estimators are added to $\widehat{\mathcal{K}}$. This is repeatedly performed on the segments defined by the consecutive elements of $\widehat{\mathcal{K}}$ until no further change point is detected. Initialised with $(s,e) = (1,n)$ and $\widehat{\mathcal{K}} = \emptyset$, the call MOSUMBS$(\widehat{k}_j + 1, \widehat{k}_{j+1}, D, G, \widehat{\mathcal{K}})$ returns the final set of estimators.

5.3.2 Forecasting

Having obtained estimates for break points in the model, it remains to produce forecasts or nowcasts. With access to population quantities, the optimal h-step ahead forecast at time $k_{j-1}+1 \le t \le k_j$ is

$$\boldsymbol{X}_{t+h|t} = \boldsymbol{\Lambda}\boldsymbol{F}^{(j)}_{t+h|t}, \tag{5.10}$$

where $\boldsymbol{F}^{(j)}_{t+1|t} = \boldsymbol{a}_j \mathbb{F}^{(j)}_t$. For $h \ge 2$, this can be defined recursively. For y_t, we have

$$y_{t+h|t} = (\boldsymbol{z}^{(j)}_{t+h|t})^\top \boldsymbol{\beta}_j, \tag{5.11}$$

Algorithm 9: MOSUMBS$(s,e,D,G,\widehat{\mathcal{K}})$ Recursive Segmentation Algorithm

input : Start and end indices s and e, Threshold D, Bandwidth G, Change point set $\widehat{\mathcal{K}}$

if $e-s>2G$ **then**
  Compute parameter $\widehat{\boldsymbol{a}}_{s,e}$
  Compute statistic $\widehat{T} \leftarrow \max_{s\le k\le e} \widehat{T}_k(G,\widehat{\boldsymbol{a}}_{s,e})$ as in (5.5)
  if $\widehat{T}>D$ **then**
    Locate $\widehat{\mathcal{K}}_{s,e} \leftarrow \{\widehat{k}_j : s=\widehat{k}_{0,s,e}+1<\widehat{k}_{1,s,e}<\cdots<\widehat{k}_{\widehat{q}_{s,e},s,e}<e=\widehat{k}_{\widehat{q}_{s,e}+1,s,e}\}$ with (5.7)
    Update global change point set $\widehat{\mathcal{K}} \leftarrow \widehat{\mathcal{K}} \cup \widehat{\mathcal{K}}_{s,e}$
    for $j=1,\ldots,\widehat{q}_{s,e}+1$ **do**
      MOSUMBS$(\widehat{k}_{j-1,s,e}+1,\widehat{k}_{j,s,e},D,G,\widehat{\mathcal{K}})$
    end
end
return $\widehat{\mathcal{K}}$

where $\boldsymbol{z}^{(j)}_{t+h|t}=\left((\boldsymbol{F}^{(j)}_{t+h|t})^\top,\boldsymbol{u}^\top_{t+h}\right)^\top$. With $h=0$, this defines a nowcast, as per Remark 5.2. In general for $h\ge 0$, we do not have access to $\boldsymbol{u}_{t+h}$ nor a corresponding forecasted version, so we study the case where $\boldsymbol{z}_{t+h|t}=\boldsymbol{F}_{t+h|t}$. We define the sample versions of (5.10) and (5.11) as

$$\widehat{\boldsymbol{X}}_{t+h|t}=\widehat{\boldsymbol{\Lambda}}\widehat{\boldsymbol{F}}_{t+h|t}, \text{ and } \widehat{y}_{t+h|t}=\widehat{\boldsymbol{F}}^\top_{t+h|t}\widehat{\boldsymbol{\beta}}, \tag{5.12}$$

where $\widehat{\boldsymbol{F}}_{t+1|t}=\widehat{\boldsymbol{a}}\widehat{\mathbb{F}}_t$. We can obtain $\widehat{\boldsymbol{\Lambda}}$ and $\widehat{\boldsymbol{F}}_t$ with PCA as per Step 1 of Section 5.3.1. The problem then is to obtain weighted estimators $\widehat{\boldsymbol{a}}_w$ in (5.8) and $\widehat{\boldsymbol{\beta}}_w$ in (5.9) with which to produce forecasts.

Under piecewise stationarity we use a weighting scheme, as per Pesaran et al. (2013), which accounts for the presence of change points. Using only data from the most recent estimated segment will produce an unbiased parameter estimate, and hence an unbiased forecast. However, when the most recent segment is short, the estimator and the forecast will have high variance. The weighting schemes introduce a small amount of bias into the forecast in exchange for a reduction in the variance. Moreover, uncertainty in the location of the true change point can amplify the bias and inefficiency problems, and so we can design our weights to account for this. The forecast error-optimal weights are not analytically available, so we consider the weight choices in Table 5.1, where the Linear and Robust schemes are as proposed in Pesaran et al. (2013). We ignore observations before $\widehat{k}_{\widehat{q}-1}$ for simplicity. These are normalised so that $w_t=\widetilde{w}_t/\sum_{t=1}^n \widetilde{w}_t$.

5.4 Theoretical results

In this section, we make modelling assumptions and derive consistency properties of the data segmentation procedure.

Table 5.1: Forecast weight choices for weighted estimators $\widehat{\boldsymbol{a}}_w$ (5.8) and $\widehat{\boldsymbol{\beta}}_w$ (5.9).

Title	Weight $\widetilde{w}_t$	Description
Expanding	1	The entire dataset.
Current	$\mathbb{I}_{\widehat{k}+1\le t\le n}$	Data after the most recent change point.
Linear	$\mathbb{I}_{\widehat{k}+1\le t\le n} + \frac{t-\widehat{k}_{\widehat{q}-1}}{\widehat{k}_{\widehat{q}}-\widehat{k}_{\widehat{q}-1}}\mathbb{I}_{\widehat{k}_{\widehat{q}-1}+1\le t\le \widehat{k}_{\widehat{q}}}$	Linearly decaying weights between the last two estimated change points.
Robust	$\mathbb{I}_{\widehat{k}+1\le t\le n} + [\frac{\log(1-t/\widehat{k}_{\widehat{q}})}{\log(1-(\widehat{k}_{\widehat{q}}-1)/\widehat{k}_{\widehat{q}})} \vee 1]\cdot \mathbb{I}_{\widehat{k}_{\widehat{q}-1}+1\le t\le \widehat{k}_{\widehat{q}}}$	Robust weights between the last two estimated change points.
Rolling	$\mathbb{I}_{n-N+1\le t\le n}$	A window of fixed length.

5.4.1 Assumptions

5.4.1.1 Factor model assumptions

Assumption 5.1 (Latent series error distribution)**.** (i) $\mathbb{E}(\boldsymbol{\eta}_t) = \mathbf{0}$.

(ii) $\mathrm{Cov}(\boldsymbol{\eta}_t) = \boldsymbol{S}$ for some positive definite matrix $\boldsymbol{S} \in \mathbb{R}^{r\times r}$.

(iii) $\boldsymbol{\eta}_t$ and $\boldsymbol{\eta}_{t'}$ are independent for $t \neq t'$, so $\mathrm{Cov}(\boldsymbol{\eta}_t, \boldsymbol{\eta}_{t'}) = \boldsymbol{O}$.

(iv) There exist constants $c_\eta, b_\eta \in (0,\infty)$ such that for any $z > 0$ and $j = 1,\dots,r$, we have $\mathsf{P}(|\eta_{it}| > z) \le \exp(1-(z/c_\eta)^{b_\eta})$.

(v) For each $j = 1,\dots,q+1$, there exists $\widetilde{\nu} > 0$ such that $0 < \mathbb{E}\|\mathbb{F}_1^{(j)}\|^{4+\widetilde{\nu}} < \infty$.

Remark 5.3 (Factor distribution)**.** By Wold's Decomposition Theorem, each $\boldsymbol{F}_t^{(j)}, j = 1,\dots,q+1, t = 1,\dots,n$ admits a moving average representation

$$\boldsymbol{F}_t^{(j)} = \sum_{l=0}^{\infty} \boldsymbol{B}_l^{(j)} \boldsymbol{\eta}_{t-l},$$

where $\boldsymbol{B}_l^{(j)}$ is square-summable over $l \in \mathbb{N}$, and $\boldsymbol{B}_0^{(j)} = \boldsymbol{I}_r$. Hence,

(i) For each $j = 1,\dots,q+1$, and each $\ell = 0,1,2,\dots$, $\mathrm{Cov}(\boldsymbol{F}_t^{(j)}, \boldsymbol{F}_{t+\ell}^{(j)}) = \sum_{l=0}^{\infty} \boldsymbol{B}_l^{(j)} \boldsymbol{S} (\boldsymbol{B}_{l+h}^{(j)})^\top = \boldsymbol{\Gamma}_F^{(j)}(\ell)$ for some matrix $\boldsymbol{\Gamma}_F^{(j)}(\ell) \in \mathbb{R}^{r\times r}$. We have $\boldsymbol{\Gamma}_F^{(j)}(-\ell) = (\boldsymbol{\Gamma}_F^{(j)}(\ell))^\top$, and we denote the covariance $\boldsymbol{\Gamma}_F^{(j)}(0) = \boldsymbol{\Gamma}_F^{(j)}$.

(ii) There exist constants $c_F, b_F \in (0,\infty)$ such that for any $z > 0$ and $i = 1,\dots,r$, we have $\mathsf{P}(|F_{it}| > z) \le \exp(1-(z/c_F)^{b_F})$.

Assumption 5.2 (Loadings)**.** (i) There exists a positive definite $r \times r$ matrix $\boldsymbol{L}$ with distinct eigenvalues such that $n^{-1}\boldsymbol{\Lambda}^\top\boldsymbol{\Lambda} \to \boldsymbol{L}$ as $n \to \infty$.

(ii) There exists $\bar{\lambda} \in (0,\infty)$ such that $|\lambda_{ij}| \le \bar{\lambda}$ for all i, j.

Assumption 5.3 (Idiosyncratic component)**.** (i) There exists a constant C_ε such that

$$\max_{1\le s\le e\le n} \frac{1}{e-s+1} \|\sum_{t=s}^{e} \mathbb{E}(\boldsymbol{\varepsilon}_t \boldsymbol{\varepsilon}_t^\top)\| < C_\varepsilon.$$

(ii) There exist constants $c_\varepsilon, b_\varepsilon \in (0,\infty)$ such that for any $z > 0$ and $i = 1,\dots,p$, we have $\mathsf{P}(|\varepsilon_{it}| > z) \le \exp(1-(z/c_\varepsilon)^{b_\varepsilon})$.

Assumption 5.1 controls the innovation process for the latent VAR, requiring independent white noise behaviour and exponential-type tails, and Remark 5.3 summarises the autocovariances and tail behaviour of the factor series. Assumption 5.1 (v) would hold as a consequence of (iv) and a further assumption on the autoregression parameter permitting a moving average representation with absolutely summable coefficients, by Lemma E.1 of Barigozzi et al. (2023). Assumption 5.2 controls the factor loadings, and is standard in the literature; see for example Barigozzi (2022). Assumption 5.3 controls the innovation process for the factor model, allowing for second-order non-stationarity (i.e. heteroscedasticity) and exponential-type tails.

Remark 5.4 (Eigenvalues)**.** Under Assumptions 5.1–5.3, we can describe the behaviour of the population eigenvalues. Define

$$\boldsymbol{\Gamma}_\chi = \boldsymbol{\Lambda}\boldsymbol{\Gamma}_F\boldsymbol{\Lambda}^\top, \quad \boldsymbol{\Gamma}_F = \frac{1}{n}\sum_{j=1}^{q+1}(k_j - k_{j-1})\boldsymbol{\Gamma}_F^{(j)}, \quad \boldsymbol{\Gamma}_\varepsilon = \frac{1}{n}\sum_{t=1}^{n}\mathbb{E}(\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}_t^\top),$$

and $\boldsymbol{\Gamma}_x = \boldsymbol{\Gamma}_\chi + \boldsymbol{\Gamma}_\varepsilon$, and denote the eigenvalues (in non-decreasing order) of each respectively as $\mu_{x,i}, \mu_{\chi,i}$, and $\mu_{\varepsilon,i}$. Then

(i) There exist constants $\underline{\gamma}, \bar{\gamma}$ such that for each $i = 1,\dots,r$,

$$0 < \underline{\gamma} < \liminf_{n\to\infty} \frac{\mu_{\chi,i}}{p} \le \limsup_{n\to\infty} \frac{\mu_{\chi,i}}{p} < \bar{\gamma} < \infty.$$

(ii) $\mu_{\varepsilon,1} < C_\varepsilon$.

(iii) $\mu_{x,i}, i = 1,\dots,r$ diverge linearly as $p \to \infty$.

(iv) $\mu_{x,r+1}$ is bounded.

Assumption 5.4 (Dimension)**.** $p \to \infty$ as $n \to \infty$ such that $n = O(p^2)$ and $p = O(n^\kappa)$ for some $1/2 \le \kappa < \infty$.

Remark 5.4 indicates we will obtain good estimation results when p grows large relative to n, suggesting the polynomial growth in Assumption 5.4.

Assumption 5.5 (Joint distribution)**.** (i) $\{\boldsymbol{\eta}_t\}_{t=1}^n$ and $\{\boldsymbol{\varepsilon}_t\}_{t=1}^n$ are independent.

(ii) Let $\mathscr{F}_s^e$ be the filtration of the sequence $\{(\boldsymbol{\eta}_t, \boldsymbol{\varepsilon}_t)\}_{t=s}^e$. Defining

$$\alpha(n) = \sup_{j\in\mathbb{N}} \sup_{A\in\mathscr{F}_1^j, B\in\mathscr{F}_{j+n}^\infty} |\mathsf{P}(A\cap B) - \mathsf{P}(A)\mathsf{P}(B)|,$$

there exist constants $c_a, b \in (0,\infty)$ such that $3\min\{b_F^{-1}, b_\varepsilon^{-1}\} + b^{-1} > 1$, and $\alpha(n) \le \exp(-c_a n^b)$.

Assumption 5.5 controls the requires the innovation processes to be independent and α-mixing.

Assumption 5.6 (Regression)**.** Let $\boldsymbol{z}_t^{(j)} = \left((\boldsymbol{F}_t^{(j)})^\top, \boldsymbol{u}_t^\top\right)^\top$ for $j = 1,\dots,q+1$.

(FAR1) $\{\varepsilon_t^y\}_{t=1}^n$ is i.i.d. with $\mathbb{E}(\varepsilon_t^y) = 0, \mathbb{E}((\varepsilon_t^y)^2) = \sigma^2 < \infty$.

(FAR2) $\{\boldsymbol{z}_t^{(j)}\}_{t=1}^n$ and $\{\varepsilon_t^y\}_{t=1}^n$ are independent for $j = 1,\dots,q+1$.

Assumption 5.6 places independent white noise restrictions on the innovation series of the diffusion index.

5.4.1.2 Segmentation assumptions

Assumption 5.7 (Bandwidth)**.** Let the bandwidth G depend on n, i.e. $G = G(n)$.

(a) For $\widetilde{\nu} > 0$ assume that

$$\frac{n}{G} \to \infty \text{ and } \frac{n^{\frac{2}{2+\nu}}\log(n)}{G} \to 0 \text{ for } n \to \infty.$$

(b) The minimum distance between change points is larger, as $n\to\infty$, than $2G$, so that

$$(i) \liminf_{n\to\infty} \min_{j=1,\dots,q+1} \frac{k_j - k_{j-1}}{G} > 2$$

and

$$(ii) \liminf_{n\to\infty} \min_{j=1,\dots,q^y+1} \frac{k_j^y - k_{j-1}^y}{G} > 2.$$

Assumption 5.8 (Jump size)**.** (a) For $j = 1,\dots,q$, define the jump size as $\delta_j(\widetilde{\boldsymbol{a}}) = \|\boldsymbol{d}_j(\widetilde{\boldsymbol{a}})\|$, where

$$\boldsymbol{d}_j(\widetilde{\boldsymbol{a}}) = \mathbb{E}\left(\boldsymbol{H}(\boldsymbol{F}_t^{(j+1)}, \mathbb{F}_{t-1}^{(j+1)}, \widetilde{\boldsymbol{a}})\right) - \mathbb{E}\left(\boldsymbol{H}(\boldsymbol{F}_t^{(j)}, \mathbb{F}_{t-1}^{(j)}, \widetilde{\boldsymbol{a}})\right). \tag{5.13}$$

As $n\to\infty$, we let $\min_{1\le j\le q}\delta_j(\widetilde{\boldsymbol{a}}) > c_\delta(n) > 0$, where $c_\delta(n)\cdot\sqrt{\frac{G}{\log(n/G)}} \to \infty$.

(b) For $j = 1,\dots,q^y$, define the jump size as $\delta_j^y(\widetilde{\boldsymbol{\beta}}) = \|\boldsymbol{d}_j^y(\widetilde{\boldsymbol{\beta}})\|$, where

$$\boldsymbol{d}_j^y(\widetilde{\boldsymbol{\beta}}) = \mathbb{E}\left(\boldsymbol{H}^y(y_t^{(j+1)}, \boldsymbol{z}_t, \widetilde{\boldsymbol{\beta}})\right) - \mathbb{E}\left(\boldsymbol{H}^y(y_t^{(j)}, \boldsymbol{z}_t, \widetilde{\boldsymbol{\beta}})\right). \tag{5.14}$$

As $n\to\infty$, we let $\min_{1\le j\le q^y}\delta_j^y(\widetilde{\boldsymbol{\beta}}) > c_\delta(n) > 0$, where $c_\delta(n)\cdot\sqrt{\frac{G}{\log(n/G)}} \to \infty$.

Assumption 5.7 requires that the bandwidth for the MOSUM procedure grows with respect to the sample size and the number of moments, and guarantees that, asymptotically, a detector at any $k = G, \ldots, n-G$ draws observations from at most two regimes. Assumption 5.8 requires that the size of each jump is non-zero and possibly shrinking. Together, these assumptions define the *separation rate* $2Gc_\delta^2$, which lower bounds the *signal-to-noise ratio* $\min_{1\leq j\leq q} \delta_j^2 \cdot \min_{0\leq j\leq q}(k_{j+1} - k_j)$, or for the regression setting, $\min_{1\leq j\leq q^y}(\delta_j^y)^2 \cdot \min_{0\leq j\leq q^y}(k_{j+1}^y - k_j^y)$.

Assumption 5.9. Define $c_\alpha = -\log\log(1-\alpha)^{-1/2}$ to be the $(1-\alpha)$ quantile of the Gumbel type 2 distribution. Let the sequence $\{\alpha_n\}_{n\in\mathbb{N}}$ fulfil

$$\alpha_n \to 0 \quad \text{and} \quad \frac{c_{\alpha_n}}{a(n/G)\sqrt{G}} = o(1).$$

Assumption 5.10. Let the threshold D be such that

$$D(G,\alpha) = \frac{b(n/G) + (c_\alpha + c_D \log^{4v}(n))}{a(n/G)}$$

for $c_D > 0$, where

$$a(x) = \sqrt{2\log(x)}, \text{ and } b(x) = 2\log(x) + \frac{\mathfrak{d}}{2}\log(\log(x)) - \log\left(\frac{2}{3}\Gamma\left(\frac{\mathfrak{d}}{2}\right)\right).$$

In the VAR case, we have $\mathfrak{d} = r^2 d$, and in the regression case, $\mathfrak{d} = r$.

Assumption 5.9 is a technical condition in which the significance level is not fixed but converging to 0. Assumption 5.10 controls the threshold for the MOSUM procedure.

5.4.1.3 Estimator assumptions

Assumption 5.11. The estimator $\widehat{\boldsymbol{\Sigma}}_k(\widetilde{\boldsymbol{a}})$ of the covariance matrix $\boldsymbol{\Sigma}_k(\widetilde{\boldsymbol{a}})$ satisfies

(a)
$$\max_{G\leq k\leq n-G} \left\| (\widehat{\boldsymbol{\Sigma}}_k(\widetilde{\boldsymbol{a}}))^{-1/2} \right\|_F = O_P(\log^{2v}(n)).$$

(b) For any $j = 1, \ldots, q$ it holds that
$$\max_{k:|k-k_j|<G} \left\| (\widehat{\boldsymbol{\Sigma}}_k(\widetilde{\boldsymbol{a}}))^{1/2} \right\|_F = O_P(\log^{2v}(n)).$$

Assumption 5.12. The estimator $\widehat{\boldsymbol{\Sigma}}_k^y(\widetilde{\boldsymbol{\beta}})$ of the covariance matrix $\boldsymbol{\Sigma}_k^y(\widetilde{\boldsymbol{\beta}})$ satisfies

(a)
$$\max_{G\leq k\leq n-G} \left\| (\widehat{\boldsymbol{\Sigma}}_k^y(\widetilde{\boldsymbol{\beta}}))^{-1/2} \right\|_F = O_P(\log^{2v}(n)).$$

(b) For any $j = 1, \ldots, q^y$ it holds that
$$\max_{k:|k-k_j^y|<G} \left\| (\widehat{\boldsymbol{\Sigma}}_k^y(\widetilde{\boldsymbol{\beta}}))^{1/2} \right\|_F = O_P(\log^{2v}(n)).$$

Assumption 5.13. (a) $\|\widetilde{\boldsymbol{a}}\| = O(1)$. (b) $\|\widetilde{\boldsymbol{\beta}}\| = O(1)$.

Assumptions 5.11 and 5.12 require that $\widehat{\boldsymbol{\Sigma}}_k(\widetilde{\boldsymbol{a}})$ (respectively $\widehat{\boldsymbol{\Sigma}}^y_k(\widetilde{\boldsymbol{\beta}})$) is a sufficiently good estimator of $\boldsymbol{\Sigma}_k(\widetilde{\boldsymbol{a}})$ (respectively $\boldsymbol{\Sigma}^y_k(\widetilde{\boldsymbol{\beta}})$), in the sense that the size does not diverge too quickly. By Remark 3.4 of Kirch and Reckrühm (2022), we do not need that the estimator is uniformly consistent away from change points when Assumption 5.10 holds. Assumption 5.13 requires that the inspection parameter is bounded in size.

5.4.2 Factor consistency

Our goal is to demonstrate that the factor analysis procedure (Step 1 of Section 5.3.1) consistently recovers the factors, and that the difference between the sample-estimated and latent population statistic is bounded. In Proposition 5.1 we show that asymptotically, we select the correct factor number r. Using this, Lemma 5.1 establishes a bound on the difference between the factor-recovered and true detector series for the piecewise stationary factor VAR, which implies we can use the estimated series for change point detection. Lemma 5.2 does the same for the piecewise factor-augmented regression model.

Proposition 5.1. Let $g(n,p) \to 0$ as $n \to \infty$ while $(p \wedge n) \cdot g(n,p) \to \infty$. Under Assumptions 5.1–5.5, $\widehat{r}$ returned by (5.16) satisfies $\mathsf{P}(\widehat{r} = r) \to 1$.

Lemma 5.1. Let the conditions of Proposition 5.1 hold. Let Assumption 5.11 hold for $\widehat{\boldsymbol{\Sigma}}_k$, and Assumption 5.13 (a) hold for some $\widetilde{\boldsymbol{a}}$. Then,

$$\max_{G \le k \le n-G} |\widehat{T}_k(G, \widetilde{\boldsymbol{a}}) - T_k(G, \widetilde{\boldsymbol{a}})| = O_P(\log^{4\upsilon}(n)).$$

Lemma 5.2. Let the conditions of Proposition 5.1 hold, as well as Assumption 5.6. Let Assumption 5.12 hold for $\widehat{\boldsymbol{\Sigma}}^y_k$, and Assumption 5.13 (b) hold for some $\widetilde{\boldsymbol{\beta}}$. Then,

$$\max_{G \le k \le n-G} |\widehat{T}^y_k(G, \widetilde{\boldsymbol{\beta}}) - T^y_k(G, \widetilde{\boldsymbol{\beta}})| = O_P(\log^{4\upsilon}(n)).$$

As per Lemma C.7, the results of Lemmas 5.1 and 5.2 depend on the existence of an $r \times r$-orthogonal matrix $\boldsymbol{R}$, and the consistency holds up to a rotation by $\boldsymbol{R}$ in the factor space.

5.4.3 Segmentation consistency

Theorem 5.3 demonstrates that the data segmentation procedure for the piecewise stationary factor VAR model is consistent when the factor series is latent. These results are inherited from the results for observed series given in Chapter 4. In Theorem 5.4 we give a similar result for the regression model.

Theorem 5.3 (VAR segmentation consistency)**.** Let the conditions of Lemma 5.1 hold. Let Assumptions 5.7–5.10 hold. Then we have that

(a)
$$\mathsf{P}\left(\widehat{q} = q; \max_{1\le j\le q} |\widehat{k}_j - k_j| < G\right) \to 1$$
as $n \to \infty$.

(b) The result holds using the least squares estimator $\widehat{\boldsymbol{a}}_{1,n}$ in (5.8), or an estimator $\widehat{\boldsymbol{\Sigma}}_k(\widetilde{\boldsymbol{a}})$ as in (C.1.1) in place of the true $\boldsymbol{\Sigma}_k(\widetilde{\boldsymbol{a}})$.

Theorem 5.4 (Regression segmentation consistency)**.** Let the conditions of Lemma 5.2 hold. Let Assumptions 5.7–5.10 hold. Then we have that

(a)
$$\mathsf{P}\left(\widehat{q}^{y} = q^{y}; \max_{1\le j\le q^{y}} |\widehat{k}^{y}_j - k^{y}_j| < G\right) \to 1$$
as $n \to \infty$.

(b) The result holds using the least squares estimator $\widehat{\boldsymbol{\beta}}_{1,n}$ in (5.9), or an estimator $\widehat{\boldsymbol{\Sigma}}^{y}_k(\widetilde{\boldsymbol{\beta}})$ as in (C.1.8) in place of the true $\boldsymbol{\Sigma}^{y}_k(\widetilde{\boldsymbol{\beta}})$.

Remark 5.5 (Localisation rate)**.** Due to the latency of the factor series, we cannot localise changes at the rate presented in Kirch and Reckrühm (2022). It may be possible to derive a localisation rate which depends on the error bounds in Lemmata 5.1–5.2, giving tighter localisation than parts (a) of Theorems 5.3–5.4.

5.5 Computation

In this section we discuss computational aspects of the method.

5.5.1 Complexity

Table 5.2 lists the time complexity of procedure/estimator combinations. We let Assumption 5.7 (a) hold, and denote the size of $\mathscr{G}$ as $|\mathscr{G}|$. We denote the complexity of principal component analysis with dimensions n and p as $\mathrm{PCA}(n,p)$, which is typically $O(p^2(n \vee p))$.

Table 5.2: Computational complexity of proposed factor VAR segmentation procedures (Section 5.3.1).

Procedure	Time Complexity
`mosumfvar`	$O\left(\mathrm{PCA}(n,p) + d^2p^2\max(p^2,d) + ndp\right)$
`mosumfvar` grid-based	$O\left(\mathrm{PCA}(n,p) + d^2p^2\max(p^2,d) + \frac{n}{\rho G}dp\right)$
`mosumfvar` recursive segmentation	$O\left(\mathrm{PCA}(n,p) + \log(n)(d^2p^2\max(p^2,d) + ndp)\right)$
`mosumfvar` multiscale	$O\left(\mathrm{PCA}(n,p) + \lvert\mathscr{G}\rvert(d^2p^2\max(p^2,d) + ndp)\right)$

5.5.2 Online segmentation

When using the proposed segmentation methods in practice, we may be in the online setting, in which new data points arrive in real time. For problems with small dimensions (n, p, r) it is feasible to simply rerun the segmentation algorithm for each new point that arrives. When the dimensions are large, however, we are faced with a large computational burden in recalculating all of the detector statistics. Instead, we may use online updates to the factor decomposition (Cardot and Degras, 2018).

5.5.3 Tuning parameter selection

Factor number For r, we consider two estimators. We may use the ratio-based estimator

$$\mathrm{ER}(b) = \widehat{\mu}_b / \widehat{\mu}_{b+1} \tag{5.15}$$

of Ahn and Horenstein (2013), selecting $\widehat{r} = \arg\max_{1 \le b \le \bar{r}} \mathrm{ER}(b)$ where $\bar{r}$ is an upper bound on the factor number. We may also use the Information Criterion-based estimator proposed by Alessi et al. (2010), an extension of that proposed in Bai and Ng (2002). We consider

$$\mathrm{IC}(b, c) = \log\left(\frac{1}{p} \sum_{j=r+1}^{p} \widehat{\mu}_j\right) + b \cdot c \cdot g(n, p) \tag{5.16}$$

where $c > 0$ is a constant, and $g(n, p) = (n \wedge p)^{-1} \log(n \wedge p)$ is the penalty function; other choices can be found in Bai and Ng (2002). IC in (5.16) consistently selects r for arbitrary values of c, so a data-driven method is used for to select $\widehat{r}$ and c simultaneously, wherein $\widehat{r}$ is evaluated for varying c over a given range, and a value of c is selected in the second region such that $\widehat{r}$ is stable. See Alessi et al. (2010); Owens et al. (2023) for further details. We prove that this consistently estimates r in Proposition 5.1.

Threshold In practice we use the threshold

$$\widetilde{D}(G, \alpha) = \max\left\{D(G, \alpha), \sqrt{2\log(n)} + \frac{c_\alpha}{\sqrt{2\log(n)}}\right\}, \tag{5.17}$$

where $D(G, \alpha)$ meets Assumption 5.10. With the large values of p for which our method is designed, we find $c_D = 0$ is a suitable choice.

Order To select the lag order d, we suggest minimising the Schwartz Information Criterion (SIC)

$$\mathrm{SIC}(d) = \frac{n_0}{2} \log\left(\frac{1}{n_0} \|\widehat{\boldsymbol{F}}_t - \widehat{\boldsymbol{a}}_{1,n_0}[d]\widehat{\mathbb{F}}_{t-1}[d]\|_F^2\right) + d\log(n_0), \tag{5.18}$$

for models fit on $t = 1, \ldots, n_0$, where $\widehat{\boldsymbol{a}}_{1,G}[d]$ is the estimator corresponding to $\widehat{\mathbb{F}}_{t-1}[d]$ consisting of regressors up to lag d, and $n_0 < n$ determines the size of the initial sample. We recommend setting $n_0 = G$ if this is available, and $n_0 = \lfloor n/20 \rfloor$ otherwise.

Bandwidth We set $G = G(n,r,d) = \exp(c_0 - c_1 \log\log(n) + c_2 \log\log(\mathfrak{d}))$ with pre-specified $c_i > 0$, $i = 0,1,2$ as described in Chapter 4.

Localisation Based on simulation, we recommend localising with $\epsilon = 0.3$.

5.6 Simulations

In this section, in simulations we validate the finite sample properties of the change point detection and forecasting methods.

5.6.1 Factor VAR

We evaluate Step 2 of the `mosumfvar` method (Section 5.3.1) for segmenting and forecasting factor model dynamics.

5.6.1.1 Data segmentation

First we evaluate `mosumfvar` in terms of detecting change points. We refer to Chapter 4 for an extensive investigation of the `mosumvar` algorithm's computational properties, particularly those of the methodological extensions in Section 5.3.1.1. We examine the method under more general factor models in Appendix C.3.

Metrics When $q \geq 1$, we report the distribution of $\widehat{q} - q$. We quantify the quality of the estimated segmentation using the Covering Metric (CM, Arbelaez et al. (2010); van den Burg and Williams (2020)) as follows. The true change points $\{k_j\}_{j=1}^{q}$ define a partition $\mathscr{P}$ of $\{1,\ldots,n\}$ into disjoint sets $\mathscr{S}_j = \{k_{j-1}+1,\ldots,k_j\}$. We denote the estimated equivalents for $\{\widehat{k}_j\}_{j=1}^{\widehat{q}}$ as $\widehat{\mathscr{P}}$ and $\widehat{\mathscr{S}}_j$. The CM is then

$$\mathscr{C}(\widehat{\mathscr{P}}, \mathscr{P}) = \frac{1}{n} \sum_{\mathscr{S} \in \mathscr{P}} |\mathscr{S}| \max_{\widehat{\mathscr{S}} \in \widehat{\mathscr{P}}} \left\{ \frac{\mathscr{S} \cap \widehat{\mathscr{S}}}{\mathscr{S} \cup \widehat{\mathscr{S}}} \right\}.$$

We have that $\mathscr{C}(\widehat{\mathscr{P}}, \mathscr{P}) \in [0,1]$ with 1 denoting a perfect segmentation. When $q = 0$, we report the empirical size, i.e. the proportion for which $\widehat{q} \geq 1$.

Settings Loadings $\lambda_{ii'}, 1 \leq i \leq r, 1 \leq i' \leq p$, are generated uniformly on $[0.2, 0.8]$. For the factor series, under the alternative we simulate from $\boldsymbol{A}_l^{(j)}$ under each regime so that for $t = k_j + 1,\ldots,k_{j+1}$,

$$\boldsymbol{F}_t = \sum_{l=1}^{d} \boldsymbol{A}_l^{(j)} \boldsymbol{F}_{t-l}^{(j)} + \boldsymbol{\varepsilon}_t, \tag{5.19}$$

while under the null we simulate with parameters $\boldsymbol{A}_l^{(1)}, l = 1,\ldots,d$ for all observations. The parameters are defined as follows: letting $\boldsymbol{A}'$ have diagonal entries $a'_{ii} = 0.7$ and off-diagonal entries

$a'_{ii'} = -0.1$, we set $\boldsymbol{A}_l^{(1)} = \rho_j \boldsymbol{A}'/\|\boldsymbol{A}'\|_F^2$, for $l = 1,\ldots,d$, where $\rho_j \in [-1,1]$ is a scalar controlling the signal strength. We generate errors so that $\boldsymbol{\eta}_t \sim N_r(\boldsymbol{0},\boldsymbol{I}_r)$ and $\boldsymbol{\varepsilon}_t \sim N_p(\boldsymbol{0},\boldsymbol{I}_p)$. We have $n = 2000$, $d = 1$, and $q = 3$ change points at $k_1 = 500, k_2 = 1000$, and $k_3 = 1500$. We set $\rho_1 = -\rho_2 = \rho_3 = 0.5$.

(V1) We test a design with varying dimensionality $p \in \{25, 50, 100, 150\}$. We fix $r = 3$.

(V2) We test a design with a varying factor number $r \in \{2,3,4,5\}$. We fix $p = 100$.

(V3) We test a design with heavier-tailed error distributions. We generate errors so that each η_{it} and ε_{it} is an i.i.d. draw from the normalised t-distribution $\sqrt{\nu/(\nu-2)} \cdot t_\nu$ with $\nu \in \{3,5,7\}$ degrees of freedom. We fix $p = 100, r = 3$.

Competing methods We use our own method both with automatic parameter selection (selecting r with (5.16) and fixing $n_0 = 100$ in (5.18) to select d), and given access to the true parameters, and we compare to the factor segmentation method proposed by Cho et al. (2022), using the default tuning parameters.[1]

Results We report the results in Table 5.3. In Setting (V1), we can see that the method performs better as p increases. `mosumfvar` is generally conservative, sometimes lacking power but giving perfect size control. There is a small loss in performance with automatic tuning parameter selection, relative to when the parameters are given. `fvarseg` gives better segmentations, but does not control size at the nominal level. In Setting (V2), we see that the performance of `mosumfvar` improves as r grows. `fvarseg` is very strong for small r but deteriorates as r grows, again struggling to control size. Setting (V3) shows `mosumfvar` is seemingly robust to mild tails, where large deviations actually improve detection power. `fvarseg` does not control size but offers good segmentation here.

5.6.1.2 Forecasting

Next we evaluate our methods for prediction.

Methods For the panel we consider seven forecasting methods. The first two use linear and robust weights given the true change point and factor number, similarly to Pesaran and Timmermann (2007). The next four methods use linear, robust, and segment weights, based on change points estimated from the data via the `mosumfvar` methodology with automatic tuning parameter selection. We also report forecasts with Expanding weighting, and Rolling weights with $N = 100, 200$.[2]

[1] We attempted to compare to Kim et al. (2021), but the implementation does not offer model selection.

[2] We also attempted the method of Stock and Watson (2002c), forecasting by regressing y_{t+1} directly onto $\widehat{\boldsymbol{F}}_t$, although the results from this were poor and we do not report them.

Metrics We report for $\boldsymbol{X}_{t+1}$ the normalised square errors $\mathrm{FE}_x^{avg} = \|\widehat{\boldsymbol{X}}_{t+1|t} - \boldsymbol{X}_{t+1}\|_F^2 / \|\boldsymbol{X}_{t+1}\|_F^2$, normalised absolute errors $\mathrm{FE}_x^{abs} = \|\widehat{\boldsymbol{X}}_{t+1|t} - \boldsymbol{X}_{t+1}\|_1 / \|\boldsymbol{X}_{t+1}\|_1$, and the normalised maximum error $\mathrm{FE}_x^{max} = \|\widehat{\boldsymbol{X}}_{t+1|t} - \boldsymbol{X}_{t+1}\|_\infty / \|\boldsymbol{X}_{t+1}\|_\infty$.

Settings We have $n = 450, p = 100$, and a change point at $k_1 = 300$. We use a pseudo-real time design, so that the sample $t = 1, \ldots, T$ for $T = 200, \ldots, 449$ is made available to the segmentation and forecasting algorithms. We repeat 30 times and average the results. We use the following models to simulate $\boldsymbol{X}_t$:

(F1) From (5.19) with $r = 2$ and $\rho = 0.7$.

(F2) A GDFM with Moving Average loadings (GDFM1).

(F3) A GDFM with Autoregressive loadings (GDFM2).

Results We report the results in Table 5.4. Under Setting (F1), the Current and Rolling ($N = 100$) weightings are the best performers. Interestingly, forecasts produced with estimated, rather than given, change points are often better performers. Under (F2), the Rolling ($N = 100$) and Robust oracle methods are best, and under (F3) the Robust oracle method is always best.

5.6.2 Factor-augmented regression

We evaluate Step 3 of the `mosumfvar` method for factor-augmented regression models.

5.6.2.1 Data segmentation

Settings The factor component simulated from (5.19) under stationarity. The regression data is generated according to (5.4) with $\boldsymbol{\beta}_j = (-1)^{j-1} \cdot \sqrt{r} \cdot (1, \ldots, 1)^\top \in \mathbb{R}^r$. Each ε_t^y is an i.i.d. draw from a standard Normal distribution. We have $n = 1000$, and under the alternative we have $q^y = 3$ change points at $k_1^y = 250, k_2^y = 500$, and $k_3^y = 750$. We compare to the `moseg` algorithm of Cho and Owens (2022), designed for detecting changes in sparse regression parameters. The default tuning parameters are used.

(R1) We test a design with varying dimensionality $p \in \{100, 200, 300, 400\}$. We fix $r = 4$.

(R2) We test a design with a varying factor number $r \in \{2, 4, 6, 8, 10\}$. We fix $p = 100$.

(R3) We test a design with heavier-tailed error distributions. We generate errors so that each η_{it}, ε_{it}, and ε_t^y is an i.i.d. draw from the normalised t-distribution $\sqrt{\nu/(\nu-2)} \cdot t_\nu$ with $\nu \in \{3, 5, 7\}$ degrees of freedom. We fix $p = 100, r = 2$.

Results We report results in Table 5.5. Under Setting (R1), the method performs consistently well for $p = 200$ or greater. There is a loss in performance for $p = 100$ with automatic selection due to the performance of the IC factor number selection method. `moseg` struggles as expected as it is not designed for strong correlations. Under (R2), there is a lack of detection power for $r = 2$, and IC struggles for $r = 10$. There is a very strong performance for `mosumfvar` when the number of factors is large. Under (R3) there is a lack of power for the heaviest tails but performance is reasonable for milder tails.

5.6.2.2 Forecasting

Settings We have $n = 450, p = 100$, and change points at $k_1^y = 250$ and $k_2^y = 350$. We use a pseudo-real time design, so that the sample $t = 1, \ldots, T$ for $T = 200, \ldots, 449$ is made available to the segmentation and forecasting algorithms. We repeat 30 times and summarise the results. To simulate $\boldsymbol{X}_t$, we use Settings (F1), (F2), and (F3), with no change points in the factor structure. The regression data is generated according to (5.4) with $\boldsymbol{\beta}_j = (-1)^{j-1} \cdot \sqrt{r} \cdot (1, \ldots, 1)^\top \in \mathbb{R}^r$, fixing $r = 2$.

Metrics For y_{t+1}, we report the square error $\mathrm{FE}_y^{avg} = (\widehat{y}_{t+1|t} - y_{t+1})^2$, the absolute error $\mathrm{FE}_y^{abs} = |\widehat{y}_{t+1|t} - y_{t+1}|$, and the sign error $\mathrm{FE}_y^{sign} = \mathbb{I}[\mathrm{sign}(\widehat{y}_{t+1|t}) \neq \mathrm{sign}(y_{t+1})]$.

Results We report results in Table 5.6. Clearly the Rolling ($N = 200$) method is in general the strongest performer, though our methods are competitive in many metrics. We note that the metrics here are not on the same scale as each other.

5.7 Application to real data

In this section, we use our methods for data segmentation and forecasting on real macroeconomic data. We use the FRED-MD database proposed in McCracken and Ng (2016). The constituents are indicators covering e.g. the labour market, prices, and financial markets in the United States. The series are transformed as per the paper, e.g. with differencing or by taking logarithms, to attain stationarity. All series are observed monthly, and we have data from 1961:06-2019:12, meaning we have $n = 703$ and $p = 122$. The logarithm of the excess return of holding an N-year bond ($b^{(N)}$) from month $t - 24$ to t, when its remaining maturity is $N - 1$ as 24 months have passed, can then be expressed as

$$xr_t^{(N)} = -(N-1)(b_t^{(N-1)} - b_{t-24}^{(N)}) + (b_{t-24}^{(N)} - b_{t-24}^{(1)}),$$

as used in Liu and Wu (2021)[3]. We use the panel to model $xr_t^{(2)}$, the US government bond risk premium at $N = 2$ years, using a factor-augmented regression. Massacci (2019) analyse a similar

[3]Data are available at `https://sites.google.com/view/jingcynthiawu`.

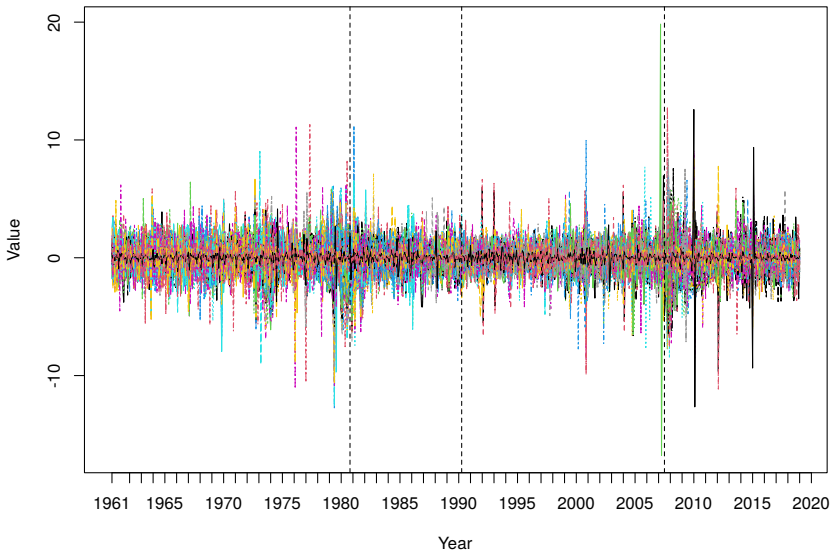


Figure 5.1: FRED-MD panel with stationarity transforms studied in Section 6.6.2. Estimated changes in the VAR structure are denoted by dashed lines.

dataset for a single break, covering 1965:01–2007:12, while Giovannelli et al. (2021) consider a similar panel for forecasting the equity premium, accounting for breaks in a factor model.

5.7.1 Data segmentation

First, we use our segmentation methods on the entire dataset. Using IC (5.16) we find $\widehat{r} = 5$ factors. For comparison, Stock and Watson (2012) select $r = 5$ a priori on a shorter dataset, while Massacci (2019) use $r = 2$. In the VAR structure we find $\widehat{q} = 3$ three changes, dated 1981:09, 1991:03, and 2008:06, plotted in Figure 5.1. These can be corresponded to significant events, namely the oil shock, the 1990 western recession, and the 2008 global financial crisis. The `fvarseg` method of Cho et al. (2022) returns $\widehat{q} = 5$ changes at 1969:10, 1978:02, 1984:08, 2005:8, and 2013:5. In other studies, Cheng et al. (2016) locate a single change in 2007, while on quarterly data, Barigozzi et al. (2018) find changes in 1983, 2007, and 2009.

In the predictive regression for the bond risk premium, we find $\widehat{q}^{y} = 2$ changes, dated 1981:01 and 2008:10, plotted in Figure 5.2. These are similarly located to changes $\widehat{k}_1$ and $\widehat{k}_3$ in the VAR structure. These results complement the single break found in Massacci (2019) when $N = 5$.

5.7.2 Forecasting

Setting We aim to predict the excess bond return $xr^{(2)}_{t+1|t}$, and the panel $\boldsymbol{X}_{t+1|t}$. We perform the same pseudo-real time exercise as in Section 5.6.1.2, so that the sample $t = 1, \dots, T$ for $T = 200, \dots, 761$ is made available to the segmentation and forecasting algorithms. We estimate r using

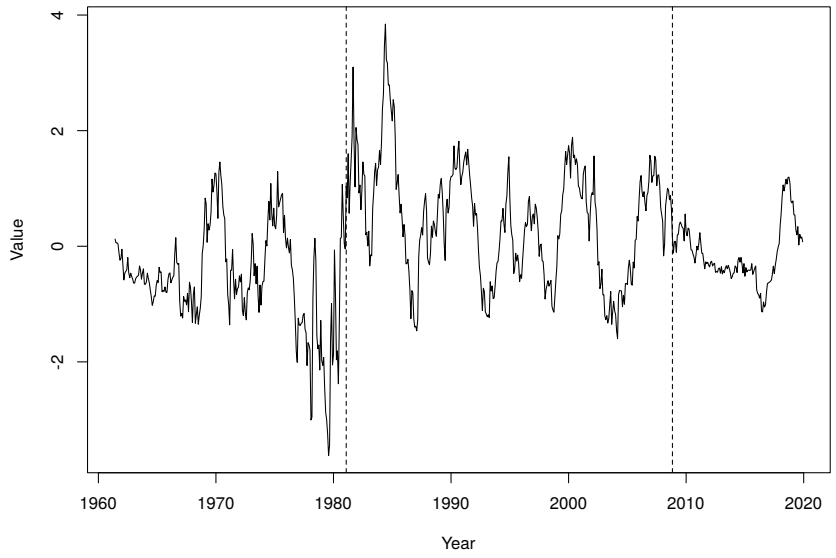


Figure 5.2: Excess bond return $xr_t^{(2)}$ studied in Section 6.6.2. Estimated changes in the factor-augmented regression structure are denoted by dashed lines.

IC at each time step, but fix the order $d = 1$ to ensure stability and prevent the dimensionality growing too large.

For $\boldsymbol{X}_{t+1|t}$, similarly to the exercise in Giovannelli et al. (2021) for stock returns, we compare to forecasts from a GDFM. Using the implementation of Owens et al. (2023), we report forecasts using the following methods: (i) Restricted forecasts from a static model as per Stock and Watson (2002a) (*SW*); (ii) Restricted forecasts from a dynamic model as per Forni et al. (2000) (*FHLR*); (iii) Unrestricted forecasts from a dynamic model as per Forni et al. (2015, 2017) (*FHLZ*). Tuning parameters are automatically selected by the package defaults.[4]

Results We report the results for $\boldsymbol{X}_{t+1|t}$ in Table 5.7, and plot them in Figures 5.3–5.4 subtracting the expectation from each metric to highlight the relative performance over time. We plot the weighted and GDFM methods separately as the former set almost uniformly outperform the latter set. The GDFM methods often have lower forecast variance, but only FHLR is competitive in point estimation. The forecasts for $\boldsymbol{X}_{t+1|t}$ are clearly correlated across the weighted methods. Using the current segment tends to perform poorly, as we would expect. Rolling window methods perform well, though the best performers are always our adaptively-weighted methods. The superior predictive power becomes clearer as time continues.

Our stated goal is for our algorithms to outperform the rolling windows method, as this is widely used, simple to implement and interpret, and has a low computational cost. We can clearly say this is true for the panel forecasting method.

[4]We attempted to compare to the method of Massacci and Kapetanios (2023) for $xr_{t+1|t}^{(2)}$ though this requires the knowledge of the location of a single structural break for fair comparison.

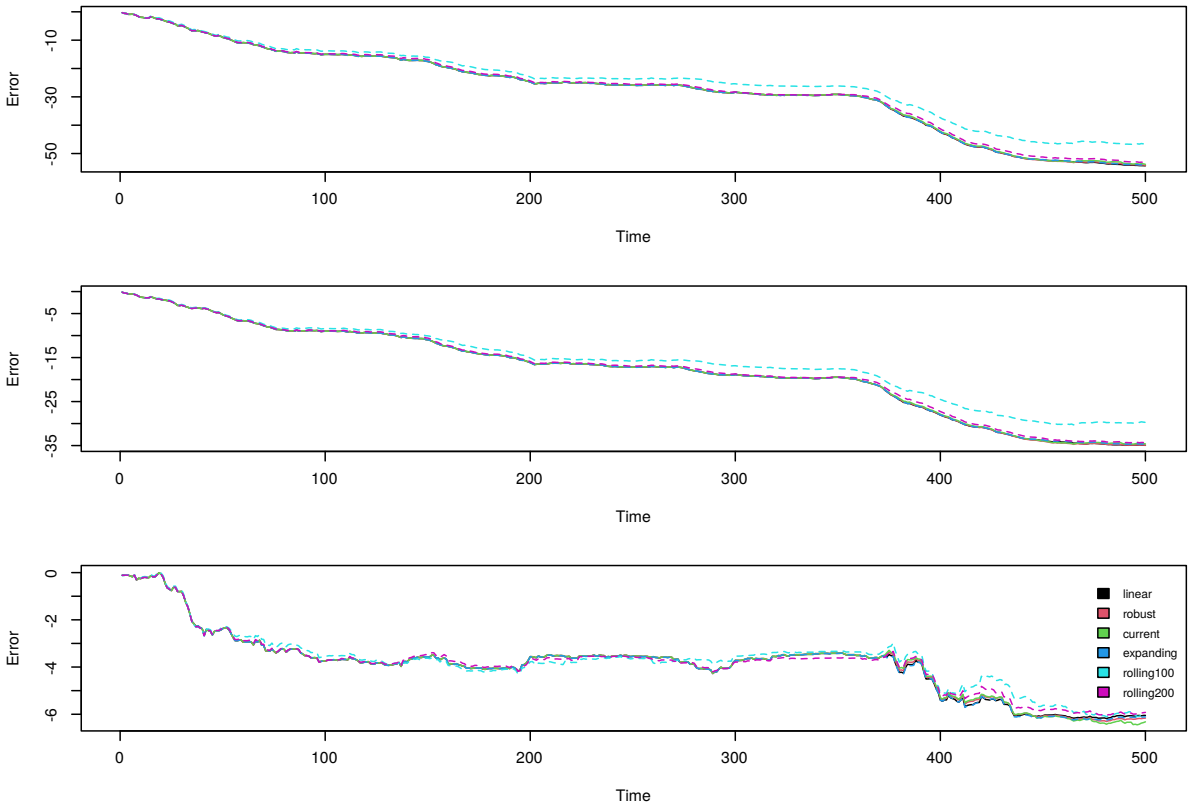


Figure 5.3: Cumulative relative forecast errors ($\mathrm{FE}_x^{avg}-t, \mathrm{FE}_x^{abs}-t$ and $\mathrm{FE}_x^{max}-t$ respectively) for $\boldsymbol{X}_{t+1|t}$, using weighting methods from Table 5.1, for the FRED-MD data described in Section 6.6.2. Each colour corresponds to a different forecast weighting method. Solid lines denote methods accounting for change points, while dashed lines denote those which do not.

We report forecasting results for $xr^{(2)}_{t+1|t}$ in Table 5.8, and Figure 5.5 (again adjusting for t). The $N = 200$ rolling window methods performs well across the metrics, though the $N = 100$ method consistently performs poorly. Without a method to select the rolling window length, this poses a problem for the analyst in practice. The weighted methods tend to outperform the $N = 100$ method, and are broadly competitive with the $N = 200$ window method.

5.8 Conclusion

We propose a comprehensive method for forecasting diffusion indices and time series panels in the presence of non-stationarity. This relies on two moving sum data segmentation methods to detect changes in two factor model components: in the latent VAR parameters, and in factor-augmented regressions. We show that these consistently detect and locate multiple changes. We describe data-adaptive extensions to the methodology, to address undetectable or multiscale changes, and a coarse grid procedure to reduce computational cost. After the data has been segmented, we propose to make forecasts with data-adaptive weights, which take the estimated change points

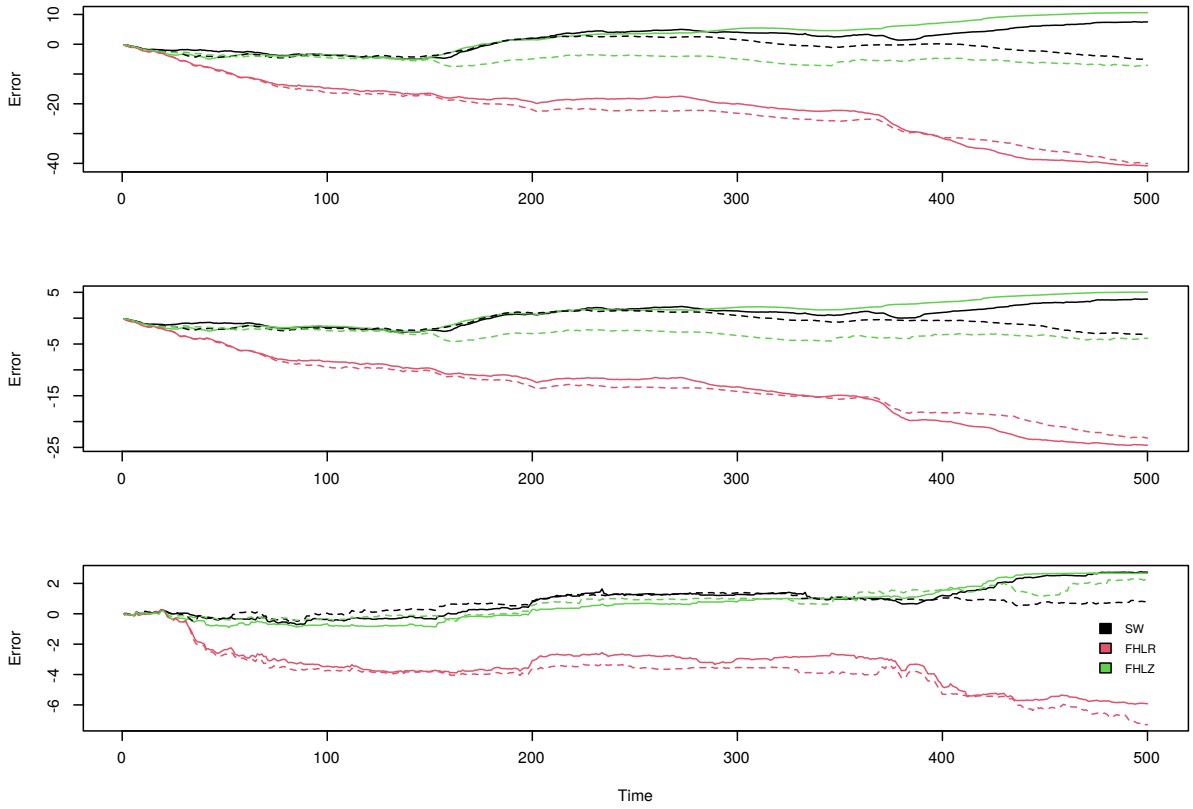


Figure 5.4: Cumulative relative forecast errors ($\mathrm{FE}_x^{avg}-t, \mathrm{FE}_x^{abs}-t$ and $\mathrm{FE}_x^{max}-t$ respectively) for $\boldsymbol{X}_{t+1|t}$, using GDFM forecasts, for the FRED-MD data described in Section 6.6.2. Each colour corresponds to a different forecast method. Solid lines denote expanding estimation windows, while dashed lines denote a rolling $N = 200$ window.

into account.

We report the performance of our methods on simulated data, and find the VAR segmentation method has favourable performance against competitors when the model is well-specified. With an application to a real macroeconomic dataset we show that our methodology can give superior forecast performance to the popularly-used rolling window method, while adding little computational burden.

It would be straightforward to extend the proposed method to the factor-augmented vector autoregression model, or to account for changes in the low-dimensional innovation covariance, allowing for the detection of all the changes categorised in Remark 5.1. The proposed weighted schemes could be used with the GDFM forecast methods, and these could be compared empirically. Finally, the estimated change points could be incorporated into Kalman filter estimation (Kim, 1994).

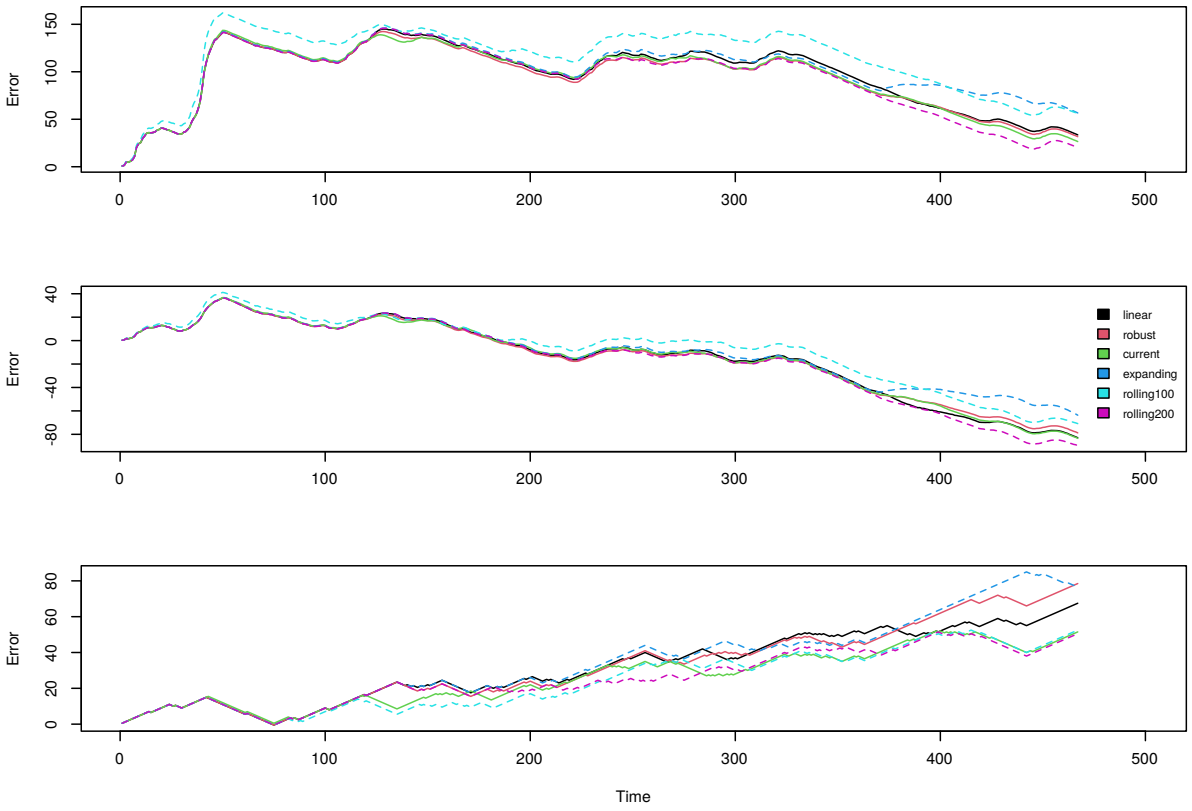


Figure 5.5: Cumulative forecast errors ($\mathrm{FE}_y^{avg} - t, \mathrm{FE}_y^{abs} - t$ and $\mathrm{FE}_y^{sign} - t/2$ respectively), using weighting methods from Table 5.1, for the excess bond return $xr_t^{(2)}$ described in Section 6.6.2. Each colour corresponds to a different forecast weighting method. Solid lines denote methods accounting for change points, while dashed lines denote those which do not.

Table 5.3: (V1)–(V3): Distributions of $\widehat{q} - q$ and the covering metric $\mathscr{C}(\widehat{\mathscr{P}}, \mathscr{P})$ of the estimated segmentations when $q = 3$, and the empirical size when $q = 0$ returned by `mosumfvar`, with or without automatic parameter selection, and `fvarseg`. The best performer for each metric is given in bold.

			$\widehat{q}-q$								
Model	Method	Variable	-3	-2	-1	0	1	2	3	CM	Size
(V1)	`mosumfvar`	$p = 25$	90	10	0	0	0	0	0	0.2744	**0**
	(automatic)	50	80	19	1	0	0	0	0	0.2983	**0**
		100	64	29	6	1	0	0	0	0.3561	**0**
		150	28	40	27	4	1	0	0	0.5114	0.01
		200	15	41	30	14	0	0	0	**0.5919**	0
	`mosumfvar`	25	93	7	0	0	0	0	0	0.2671	**0**
	(fixed)	50	81	18	1	0	0	0	0	0.2958	**0**
		100	27	57	14	2	0	0	0	0.4682	**0**
		150	25	41	30	3	1	0	0	0.5213	**0**
		200	16	40	32	12	0	0	0	0.5852	**0**
	`fvarseg`	25	31	26	23	**13**	5	2	0	**0.5196**	0.07
		50	31	30	23	**15**	1	0	0	**0.5166**	0.11
		100	26	29	30	**13**	2	0	0	**0.5341**	0.12
		150	27	22	34	**15**	2	0	0	**0.5458**	0.08
		200	38	26	20	**16**	0	0	0	0.4955	0.04
(V2)	`mosumfvar`	$r = 2$	87	5	8	0	0	0	0	0.3011	**0**
	(automatic)	3	64	29	6	1	0	0	0	0.3561	**0**
		4	56	30	13	1	0	0	0	0.3875	**0**
		5	38	17	20	10	10	3	2	0.4994	**0.14**
	`mosumfvar`	2	14	39	36	10	1	0	0	0.6012	**0**
	(fixed)	3	27	57	14	2	0	0	0	0.4682	**0**
		4	49	34	15	2	0	0	0	**0.4089**	**0**
		5	5	25	32	**20**	15	3	0	**0.6359**	0.2
	`fvarseg`	2	0	5	14	**67**	14	0	0	**0.8621**	0.01
		3	26	29	30	**13**	2	0	0	**0.5341**	0.12
		4	61	22	11	**6**	0	0	0	0.3604	0.1
		5	70	19	10	0	1	0	0	0.3283	0.26
(V3)	`mosumfvar`	$v = 3$	10	16	19	**33**	10	7	5	**0.7334**	0.13
	(automatic)	5	35	16	29	20	0	0	0	0.5728	**0**
		7	72	10	12	6	0	0	0	0.3761	**0**
	`mosumfvar`	3	26	35	29	8	2	0	0	0.5380	**0**
	(fixed)	5	3	23	44	30	0	0	0	0.7342	**0**
		7	4	33	40	22	1	0	0	0.6900	**0**
	`fvarseg`	3	17	6	11	20	28	9	9	0.6039	0.75
		5	0	2	8	**51**	27	9	3	**0.8361**	0.7
		7	0	1	14	**52**	28	5	0	**0.8622**	0.27

Table 5.4: Forecast errors for $\boldsymbol{X}_{T+1}$ in terms of FE_x^{avg},FE_x^{abs}, and FE_x^{max} under (F1)–(F3) for $T = 200,\ldots,449$ over 30 realisations. Forecast weights are as described in Table 5.1, and the change points are either given ('Oracle') or estimated by `mosumfvar`. The best performer for each metric is given in bold.

			Oracle		Estimated					
Model	Metric	Summary	Linear	Robust	Linear	Robust	Current	Expanding	Rolling (100)	Rolling (200)
(F1)	avg	mean	0.7774	0.7767	0.6990	0.6990	**0.6902**	0.6990	0.6904	0.6935
		median	0.9716	0.9713	0.8579	0.8579	0.8403	0.8579	**0.8386**	0.8478
		se	0.3955	0.3952	0.3767	0.3767	**0.3735**	0.3767	0.3737	0.3746
	abs	mean	0.7872	0.7869	0.7416	0.7416	**0.7372**	0.7416	0.7373	0.7387
		median	0.9852	0.9847	0.9206	0.9206	0.9137	0.9206	**0.9116**	0.9155
		se	0.3958	0.3956	0.3800	0.3800	**0.3781**	0.3800	0.3782	0.3788
	max	mean	0.7921	0.7917	0.7610	0.7610	0.7552	0.7610	**0.7550**	0.7577
		median	0.9819	0.9819	0.9194	0.9194	**0.9128**	0.9194	0.9133	0.9150
		se	0.4001	0.3999	0.3980	0.3980	0.3948	0.3980	**0.3946**	0.3963
(F2)	avg	mean	0.8352	**0.8017**	0.8627	0.8627	0.8380	0.8627	0.8046	0.8263
		median	0.9953	0.9710	1.0035	1.0035	0.9728	1.0035	**0.9522**	0.9832
		se	0.4501	**0.4284**	0.4749	0.4749	0.4774	0.4749	0.4499	0.4486
	abs	mean	0.8163	0.7983	0.8301	0.8301	0.8151	0.8301	**0.7978**	0.8112
		median	0.9977	0.9852	1.0012	1.0012	0.9863	1.0012	**0.9741**	0.9914
		se	0.4180	**0.4077**	0.4283	0.4283	0.4256	0.4283	0.4137	0.4163
	max	mean	0.8103	0.7988	0.8186	0.8186	0.8092	0.8186	**0.7983**	0.8068
		median	0.9964	0.9841	1.0001	1.0001	0.9827	1.0001	**0.9741**	0.9895
		se	0.4107	**0.4045**	0.4165	0.4165	0.4148	0.4165	0.4080	0.4095
(F3)	avg	mean	0.7485	**0.7396**	0.7583	0.7583	0.7493	0.7583	0.7505	0.7497
		median	0.9058	**0.8906**	0.9213	0.9213	0.9007	0.9213	0.8968	0.9061
		se	0.3939	**0.3908**	0.3983	0.3983	0.3983	0.3983	0.4023	0.3958
	abs	mean	0.7710	**0.7663**	0.7758	0.7758	0.7710	0.7758	0.7714	0.7714
		median	0.9500	**0.9417**	0.9580	0.9580	0.9479	0.9580	0.9443	0.9495
		se	0.3915	**0.3897**	0.3937	0.3937	0.3928	0.3937	0.3940	0.3922
	max	mean	0.7782	**0.7733**	0.7844	0.7844	0.7784	0.7844	0.7785	0.7791
		median	0.9452	**0.9387**	0.9524	0.9524	0.9432	0.9524	0.9414	0.9464
		se	0.4012	**0.3989**	0.4050	0.4050	0.4028	0.4050	0.4034	0.4025

Table 5.5: (R1)–(R3): Distributions of $\widehat{q}^y - q^y$ and the covering metric $\mathscr{C}(\widehat{\mathscr{P}}, \mathscr{P})$ of the estimated segmentations when $q^y = 3$, and the empirical size when $q^y = 0$ returned by `mosumfvar`, with or without automatic parameter selection, and `moseg`. The best performer for each metric is given in bold.

			$\widehat{q}^y - q^y$								
Model	Method	Variable	-3	-2	-1	0	1	2	3	CM	null
(R1)	`mosumfvar`	$p = 100$	8	2	15	74	1	0	0	0.8654	**0**
	(automatic)	200	0	0	9	90	1	0	0	**0.9426**	**0**
		300	0	1	6	93	0	0	0	**0.9483**	**0**
		400	0	2	7	91	0	0	0	**0.9369**	**0**
	`mosumfvar`	100	0	1	15	**83**	1	0	0	**0.9270**	**0**
	(fixed)	200	0	0	9	**90**	1	0	0	**0.9426**	**0**
		300	0	1	6	**93**	0	0	0	**0.9483**	**0**
		400	0	2	7	**91**	0	0	0	**0.9369**	**0**
	`moseg`	100	28	20	5	16	10	8	13	0.5064	0.61
		200	31	20	12	9	8	11	9	0.5257	0.45
		300	31	20	10	7	18	8	6	0.5458	0.52
		400	24	14	16	9	25	11	1	0.6249	0.47
(R2)	`mosumfvar`	$r = 2$	96	3	1	0	0	0	0	0.2623	**0**
	(automatic)	4	8	2	15	74	1	0	0	0.8654	**0**
		6	3	1	0	95	1	0	0	0.9384	0.01
		8	0	2	1	86	3	4	4	0.8854	0.24
		10	0	0	0	29	1	1	69	0.6736	0.82
	`mosumfvar`	2	95	4	1	0	0	0	0	0.2647	**0**
	(fixed)	4	0	1	15	**83**	1	0	0	**0.9270**	**0**
		6	0	0	0	**100**	0	0	0	**0.9675**	**0**
		8	0	0	0	**99**	1	0	0	**0.9585**	**0.12**
		10	0	0	0	**100**	0	0	0	**0.9618**	**0.35**
	`moseg`	2	38	15	8	**10**	6	9	14	0.4773	0.53
		4	28	20	5	16	10	8	13	0.5064	0.61
		6	21	16	14	10	9	15	15	0.5290	0.54
		8	30	19	9	13	7	8	14	0.4889	0.56
		10	31	12	8	6	11	4	28	0.5131	0.63
(R3)	`mosumfvar`	$v = 3$	98	2	0	0	0	0	0	0.2546	**0**
	(automatic)	5	17	33	36	**14**	0	0	0	0.5910	**0**
		7	45	40	13	2	0	0	0	0.4208	**0**
	`mosumfvar`	3	98	2	0	0	0	0	0	0.2549	**0**
	(fixed)	5	14	36	36	**14**	0	0	0	**0.6048**	**0**
		7	41	42	15	2	0	0	0	0.4351	**0**
	`moseg`	3	47	14	6	**11**	5	5	12	**0.3960**	0.47
		5	39	25	5	8	14	5	4	0.4582	0.53
		7	34	20	6	**11**	10	7	12	**0.4739**	0.47

Table 5.6: Forecast errors for y_{T+1} in terms of FE_y^{avg},FE_y^{abs}, and FE_y^{sign} under (R1)–(R3) for $T = 200,\ldots,449$ over 30 realisations. Forecast weights are as described in Table 5.1, and the change points are either given ('Oracle') or estimated by `mosumfvar`. The best performer for each metric is given in bold.

		Oracle		Estimated					
Metric	Summary	Linear	Robust	Linear	Robust	Current	Expanding	Rolling (100)	Rolling (200)
avg	mean	64.0287	5.8744	16.1094	20.0395	39.3405	20.4682	9.4892	**3.3583**
	median	0.9797	0.9269	0.9292	0.9425	0.9421	**0.9051**	0.9550	0.9578
	se	3023.9654	177.8788	613.5313	693.3084	1685.7670	682.4135	335.4904	**68.3457**
abs	mean	1.2593	0.9814	1.0994	1.1270	1.2324	1.1488	0.9909	**0.9055**
	median	0.9898	**0.9627**	0.9639	0.9708	0.9706	0.9514	0.9772	0.9787
	se	7.9026	2.2163	3.8604	4.3326	6.1503	4.3762	2.9169	**1.5933**
sign	mean	0.4603	0.3819	0.3968	0.4136	0.4236	0.3945	0.3987	**0.3735**
	se	0.4985	0.4859	0.4893	0.4925	0.4942	0.4888	0.4897	**0.4838**

Table 5.7: Forecast errors for $\boldsymbol{X}_{t+1}$ measured by FE_x^{avg}, FE_x^{abs}, and FE_x^{max}, using weighting methods from Table 5.1 and factor model forecasts, for the FRED-MD data described in Section 6.6.2. The best performer for each metric is given in bold.

						Rolling		Expanding			Rolling		
Metric		Linear	Robust	Current	Expanding	$N = 100$	$N = 200$	SW	FHLR	FHLZ	SW	FHLR	FHLZ
avg	mean	**0.8914**	0.8918	0.8924	0.8919	0.9066	0.8937	1.0150	0.9184	1.0212	0.9900	0.9199	0.9859
	median	0.9211	**0.9201**	0.9191	0.9201	0.9322	0.9246	1.0112	0.9321	1.0215	0.9840	0.9379	0.9781
	se	0.1478	0.1485	0.1505	0.1483	0.1526	0.1481	0.1049	0.1347	**0.0837**	0.0888	0.1264	0.0992
abs	mean	**0.9301**	0.9302	0.9308	0.9307	0.9405	0.9313	1.0074	0.9509	1.0100	0.9939	0.9536	0.9923
	median	0.9477	**0.9443**	0.9443	0.9454	0.9519	0.9479	1.0063	0.9560	1.0092	0.9897	0.9647	0.9871
	se	0.0918	0.0921	0.0931	0.0921	0.0942	0.0919	0.0626	0.0844	**0.0484**	0.0531	0.0800	0.0632
max	mean	0.9879	0.9877	**0.9874**	0.9878	0.9880	0.9882	1.0055	0.9881	1.0053	1.0016	0.9854	1.0045
	median	0.9990	0.9989	0.9990	0.9990	0.9998	0.9993	1.0023	**0.9985**	1.0011	1.0021	0.9995	1.0016
	se	0.0756	0.0757	0.0761	0.0767	0.0809	0.0765	0.0532	0.0759	**0.0443**	0.0482	0.0792	0.0623

Table 5.8: Forecast errors for y_{t+1} measured by FE_y^{avg}, FE_y^{abs}, and FE_y^{sign}, using weighting methods from Table 5.1, for the excess bond return $xr_t^{(2)}$ described in Section 6.6.2. The best performer for each metric is given in bold.

Metric	Summary	Linear	Robust	Current	Expanding	Rolling (100)	Rolling (200)
avg	mean	1.0718	1.0678	1.0568	1.1213	1.1201	**1.0440**
	median	0.4936	0.5124	0.4835	0.6125	0.5182	**0.4540**
	se	1.7584	1.7453	1.7593	1.7422	1.9251	**1.7581**
abs	mean	0.8224	0.8313	0.8213	0.8637	0.8483	**0.8082**
	median	0.7026	0.7158	0.6953	0.7826	0.7199	**0.6738**
	se	0.6296	0.6144	0.6189	**0.6133**	0.6335	0.6258
sign	mean	0.6445	0.6681	0.6103	0.6638	0.6124	**0.6081**
	se	0.4792	0.4714	0.4882	**0.4729**	0.4877	0.4887

CHAPTER 6

FACTOR-ADJUSTED NETWORK ESTIMATION AND FORECASTING FOR HIGH-DIMENSIONAL TIME SERIES

6.1 Introduction

Vector autoregressive (VAR) models are popularly adopted for modelling time series datasets collected in many disciplines including economics (Koop, 2013), finance (Barigozzi and Brownlees, 2019), neuroscience (Kirch et al., 2015) and systems biology (Shojaie and Michailidis, 2010), to name a few. By fitting a VAR model to the data, we can infer dynamic interdependence between the variables and forecast future values. In particular, estimating the non-zero elements of the VAR parameter matrices recovers directed edges between the components of vector time series in a Granger causality network. Besides, by estimating the precision matrix (inverse of the covariance matrix) of the VAR innovations, we can define a network representing their contemporaneous dependencies by means of partial correlations. Finally, the inverse of the long-run covariance matrix of the data simultaneously captures lead-lag and contemporaneous co-movements of the variables. For further discussions on the network interpretation of VAR modelling, we refer to Dahlhaus (2000), Eichler (2007), Billio et al. (2012) and Barigozzi and Brownlees (2019).

Fitting VAR models to the data quickly becomes a high-dimensional problem as the number of parameters grows quadratically with the dimensionality of the data. There exists a mature literature on ℓ_1-regularisation methods for estimating VAR models in high dimensions under suitable sparsity assumptions on the VAR parameters (Basu and Michailidis, 2015; Han et al., 2015; Kock and Callot, 2015; Liu and Zhang, 2021a; Medeiros and Mendes, 2016; Nicholson et al., 2020). Consistency of such methods is derived under the assumption that the spectral density matrix of the data has bounded eigenvalues. However, in many applications, the datasets exhibit

strong serial and cross-sectional correlations which leads to the violation of this assumption. As a motivating example, we introduce a dataset of node-specific prices in the PJM (Pennsylvania, New Jersey and Maryland) power pool area in the United States, see Section 6.6.1 for further details. Figure 6.1 demonstrates that the leading eigenvalue of the long-run covariance matrix (i.e. spectral density matrix at frequency 0) increases linearly as the dimension of the data increases, which implies the presence of latent common factors in the panel data (Forni et al., 2000). Additionally, the left panel of Figure 6.2 shows the inadequacy of fitting a VAR model to such data under the sparsity assumption via ℓ_1-regularisation methods, unless the presence of strong correlations is accounted for by a *factor-adjustment* step as in the right panel. Similar behaviour is demonstrated in Barigozzi et al. (2023) for the panel of equity volatility measures, also analysed in this chapter.

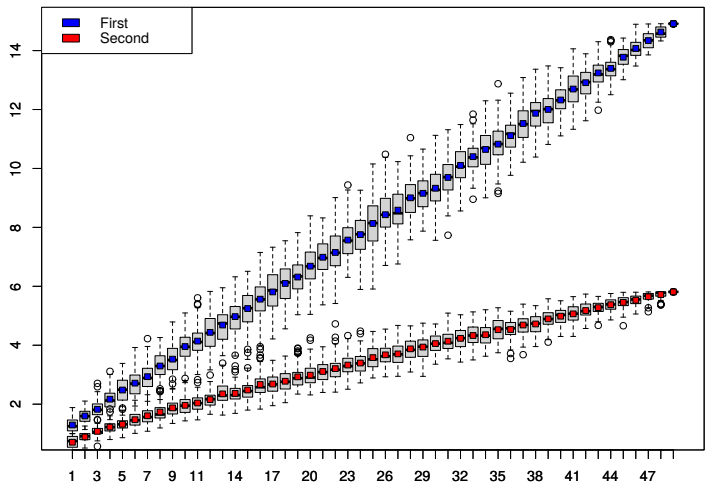


Figure 6.1: Box plots of the two largest eigenvalues (y-axis) of the long-run covariance matrix estimated from the energy price data collected between 01/01/2021 and 19/07/2021 ($n = 200$), see Section 6.6.2 for further details. Cross-sections of the data are randomly sampled 100 times for each given dimension $p \in \{2, \ldots, 50\}$ (x-axis) to produce the box plots.

Barigozzi et al. (2023) propose the FNETS methodology for factor-adjusted VAR modelling of high-dimensional, second-order stationary time series. Under their proposed model, the data is decomposed into two latent components such that the *factor-driven* component accounts for pervasive leading, lagging or contemporaneous co-movements of the variables, while the remaining *idiosyncratic* dynamic dependence between the variables is modelled by a sparse VAR process. Then, FNETS provides tools for inferring the networks underlying the latent VAR process and forecasting.

In this chapter, we present an R package named **fnets** which implements the FNETS methodology. It provides a range of user-friendly tools for estimating and visualising the networks representing the interconnectedness of time series variables, and for producing forecasts. In addition, **fnets** thoroughly addresses the problem of selecting tuning parameters ranging from

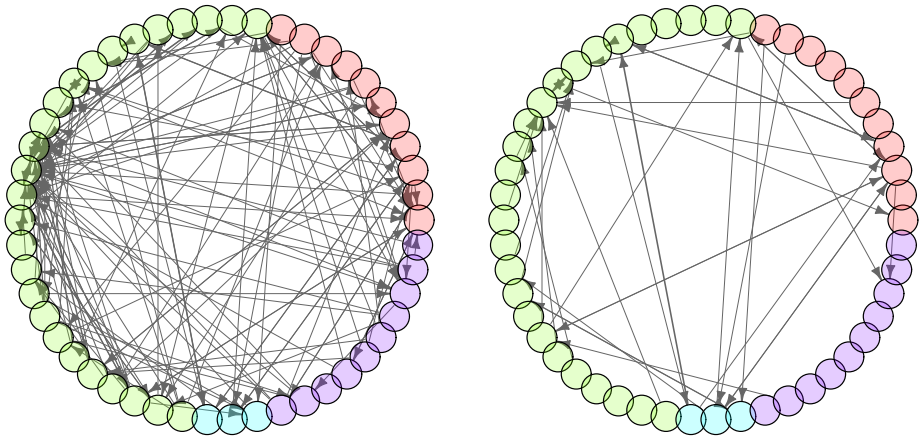


Figure 6.2: Granger causal networks defined in (6.5) obtained from fitting a VAR(1) model to the energy price data analysed in Figure 6.1, without (left) and with (right) the factor adjustment step outlined in Section 6.2.3. Edge weights (proportional to the size of coefficient estimates) are visualised by the width of each edge, and the nodes are coloured according to their groupings, see Section 6.6.2

the number of factors and the VAR order, to regularisation and thresholding parameters adopted for producing sparse and interpretable networks. As such, a simple call of the main routine of **fnets** requires the input data only, and it outputs an object of S3 class `fnets` which is supported by a `plot` method for network visualisation and a `predict` method for time series forecasting.

There exist several packages for fitting VAR models and their extensions to high-dimensional time series, see **lsvar** (Bai, 2021), **sparsevar** (Vazzoler, 2021), **nets** (Brownlees, 2020), **mgm** (Haslbeck and Waldorp, 2020), **graphicalVAR** (Epskamp et al., 2018), **bigVAR** (Nicholson et al., 2017), and **bigtime** (Wilms et al., 2021). There also exist R packages for time series factor modelling such as **dfms** (Krantz and Bagdziunas, 2023) and **sparseDFM** (Mosley et al., 2023), and **FAVAR** (Bernanke et al., 2005) for Bayesian inference of factor-augmented VAR models. The package **fnets** is clearly distinguished from, and complements, the above list by handling strong cross-sectional and serial correlations in the data via factor-adjustment step performed in frequency domain. In addition, the FNETS methodology operates under the most general approach to high-dimensional time series factor modelling termed the Generalised Dynamic Factor Model (GDFM), first proposed in Forni et al. (2000) and further investigated in Forni et al. (2015). Accordingly, **fnets** is the first R package to provide tools for high-dimensional panel data analysis under the GDFM, such as fast computation of spectral density and autocovariance matrices via the Fast Fourier Transform, but it is flexible enough to allow for more restrictive static factor models. While there exist some packages for network-based time series modelling (e.g. **GNAR**, Knight et al., 2020), we highlight that the goal of **fnets** is to learn the networks underlying a time series and does not require a network as an input.

6.2 FNETS methodology

In this section, we introduce the factor-adjusted VAR model and describe the FNETS methodology proposed in Barigozzi et al. (2023) for network estimation and forecasting of high-dimensional time series. We limit ourselves to describing the key steps of FNETS and refer to the above paper for its comprehensive treatment, both methodologically and theoretically.

6.2.1 Factor-adjusted VAR model

A zero-mean, p-variate process $\boldsymbol{\xi}_t$ follows a VAR(d) model if it satisfies

$$\boldsymbol{\xi}_t = \sum_{\ell=1}^{d} \mathbf{A}_\ell \boldsymbol{\xi}_{t-\ell} + \boldsymbol{\Gamma}^{1/2} \boldsymbol{\varepsilon}_t, \tag{6.1}$$

where $\mathbf{A}_\ell \in \mathbb{R}^{p\times p}$, $1 \le \ell \le d$, determine how future values of the series depend on their past. For the p-variate random vector $\boldsymbol{\varepsilon}_t = (\varepsilon_{1t},\ldots,\varepsilon_{pt})^\top$, we assume that ε_{it} are independently and identically distributed (i.i.d.) for all i and t with $\mathbb{E}(\varepsilon_{it}) = 0$ and $\mathsf{Var}(\varepsilon_{it}) = 1$. Then, the positive definite matrix $\boldsymbol{\Gamma} \in \mathbb{R}^{p\times p}$ is the covariance matrix of the innovations $\boldsymbol{\Gamma}^{1/2}\boldsymbol{\varepsilon}_t$.

In the literature on factor modelling of high-dimensional time series, the factor-driven component exhibits strong cross-sectional and/or serial correlations by 'loading' finite-dimensional vectors of factors linearly. Among many time series factor models, the GDFM (Forni et al., 2000) provides the most general approach where the p-variate factor-driven component $\boldsymbol{\chi}_t$ admits the following representation

$$\boldsymbol{\chi}_t = \mathscr{B}(L)\mathbf{u}_t = \sum_{\ell=0}^{\infty} \mathbf{B}_\ell \mathbf{u}_{t-\ell} \;\text{ with }\; \mathbf{u}_t = (u_{1t},\ldots,u_{qt})^\top \text{ and } \mathbf{B}_\ell \in \mathbb{R}^{p\times q}, \tag{6.2}$$

for some fixed q, where L stands for the lag operator. The q-variate random vector $\mathbf{u}_t$ contains the common factors which are loaded across the variables and time by the filter $\mathscr{B}(L) = \sum_{\ell=0}^{\infty} \mathbf{B}_\ell L^\ell$, and it is assumed that u_{jt} are i.i.d. with $\mathbb{E}(u_{jt}) = 0$ and $\mathsf{Var}(u_{jt}) = 1$. The model (6.2) reduces to a static factor model (Bai, 2003; Fan et al., 2013; Stock and Watson, 2002a), when $\mathscr{B}(L) = \sum_{\ell=0}^{s} \mathbf{B}_\ell L^\ell$ for some finite integer $s \ge 0$. Then, we can write

$$\boldsymbol{\chi}_t = \boldsymbol{\Lambda}\mathbf{F}_t \;\text{ where }\; \mathbf{F}_t = (\mathbf{u}_t^\top,\ldots,\mathbf{u}_{t-s}^\top)^\top \;\text{ and }\; \boldsymbol{\Lambda} = [\mathbf{B}_0,\ldots,\mathbf{B}_s] \tag{6.3}$$

with $r = q(s+1)$ as the dimension of static factors $\mathbf{F}_t$. Throughout, we refer to the models (6.2) and (6.3) as *unrestricted* and *restricted* to highlight that the latter imposes more restrictions on the model.

Barigozzi et al. (2023) propose a factor-adjusted VAR model under which we observe a zero-mean, second-order stationary process $\mathbf{X}_t = (X_{1t},\ldots,X_{pt})^\top$ for $t = 1,\ldots,n$, that permits a decomposition into the sum of the unobserved components $\boldsymbol{\xi}_t$ and $\boldsymbol{\chi}_t$, i.e.

$$\mathbf{X}_t = \boldsymbol{\xi}_t + \boldsymbol{\chi}_t. \tag{6.4}$$

We assume that $\mathbb{E}(\varepsilon_{it}u_{jt'}) = 0$ for all i, j, t and t' as is commonly assumed in the literature, such that $\mathbb{E}(\xi_{it}\chi_{i't'}) = 0$ for all $1 \le i,i' \le p$ and $t,t' \in \mathbb{Z}$.

6.2.2 Networks

Under (6.4), it is of interest to infer three types of networks representing the interconnectedness of $\mathbf{X}_t$ after factor adjustment. Let $\mathcal{V} = \{1, \ldots, p\}$ denote the set of vertices representing the p cross-sections. Then, the VAR parameter matrices, $\mathbf{A}_\ell = [A_{\ell,ii'}, 1 \leq i, i' \leq p]$, encode the directed network $\mathcal{N}^{\mathrm{G}} = (\mathcal{V}, \mathcal{E}^{\mathrm{G}})$ representing Granger causal linkages, where the set of edges are given by

$$\mathcal{E}^{\mathrm{G}} = \left\{(i, i') \in \mathcal{V} \times \mathcal{V} : A_{\ell,ii'} \neq 0 \text{ for some } 1 \leq \ell \leq d\right\}. \tag{6.5}$$

Here, the presence of an edge $(i, i') \in \mathcal{E}^{\mathrm{G}}$ indicates that $\xi_{i',t-\ell}$ Granger causes ξ_{it} at some lag $1 \leq \ell \leq d$ (Dahlhaus, 2000).

The second network contains undirected edges representing contemporaneous cross-sectional dependence in VAR innovations $\mathbf{\Gamma}^{1/2}\boldsymbol{\varepsilon}_t$, denoted by $\mathcal{N}^{\mathrm{C}} = (\mathcal{V}, \mathcal{E}^{\mathrm{C}})$. We have $(i, i') \in \mathcal{E}^{\mathrm{C}}$ if and only if the partial correlation between the i-th and i'-th elements of $\mathbf{\Gamma}^{1/2}\boldsymbol{\varepsilon}_t$ is non-zero, which in turn is given by $-\delta_{ii'}/\sqrt{\delta_{ii} \cdot \delta_{i'i'}}$ where $\mathbf{\Gamma}^{-1} = \mathbf{\Delta} = [\delta_{ii'}, 1 \leq i, i' \leq p]$ (Peng et al., 2009). Hence, the set of edges for $\mathcal{N}^{\mathrm{C}}$ is given by

$$\mathcal{E}^{\mathrm{C}} = \left\{(i, i') \in \mathcal{V} \times \mathcal{V} : i \neq i' \text{ and } -\frac{\delta_{ii'}}{\sqrt{\delta_{ii} \cdot \delta_{i'i'}}} \neq 0\right\}, \tag{6.6}$$

Finally, we can summarise the aforementioned lead-lag and contemporaneous relations between the variables in a single, undirected network $\mathcal{N}^{\mathrm{L}} = (\mathcal{V}, \mathcal{E}^{\mathrm{L}})$ by means of the long-run partial correlations of $\boldsymbol{\xi}_t$. Let $\mathbf{\Omega} = [\omega_{ii'}, 1 \leq i, i' \leq p]$ denote the inverse of the zero-frequency spectral density (a.k.a. long-run covariance) of $\boldsymbol{\xi}_t$, which is given by $\mathbf{\Omega} = 2\pi\mathscr{A}^\top(1)\mathbf{\Delta}\mathscr{A}(1)$ with $\mathscr{A}(z) = \mathbf{I} - \sum_{\ell=1}^{d} \mathbf{A}_\ell z^\ell$. Then, the long-run partial correlation between the i-th and i'-th elements of $\boldsymbol{\xi}_t$, is obtained as $-\omega_{ii'}/\sqrt{\omega_{ii} \cdot \omega_{i'i'}}$ (Dahlhaus, 2000), so the edge set of $\mathcal{N}^{\mathrm{L}}$ is given by

$$\mathcal{E}^{\mathrm{L}} = \left\{(i, i') \in \mathcal{V} \times \mathcal{V} : i \neq i' \text{ and } -\frac{\omega_{ii'}}{\sqrt{\omega_{ii} \cdot \omega_{i'i'}}} \neq 0\right\}. \tag{6.7}$$

6.2.3 FNETS: Network estimation

We describe the three-step methodology for estimating the networks $\mathcal{N}^{\mathrm{G}}$, $\mathcal{N}^{\mathrm{C}}$ and $\mathcal{N}^{\mathrm{L}}$. Throughout, we assume that the number of factors, either q under the more general model in (6.2) or r under the restricted model in (6.3), and the VAR order d are known, and discuss its selection in Section 6.3.

6.2.3.1 Step 1: Factor adjustment

The autocovariance (ACV) matrices of $\boldsymbol{\xi}_t$, denoted by $\mathbf{\Gamma}_\xi(\ell) = \mathbb{E}(\boldsymbol{\xi}_{t-\ell}\boldsymbol{\xi}_t^\top)$ for $\ell \geq 0$ and $\mathbf{\Gamma}_\xi(\ell) = (\mathbf{\Gamma}_\xi(-\ell))^\top$ for $\ell < 0$, play a key role in network estimation. Since $\boldsymbol{\xi}_t$ is not directly observed, we propose to adjust for the presence of the factor-driven $\boldsymbol{\chi}_t$ and estimate $\mathbf{\Gamma}_\xi(\ell)$. For this, we adopt a frequency domain-based approach and perform dynamic principal component analysis (PCA). Spectral density matrix $\mathbf{\Sigma}_x(\omega)$ of a time series $\{\mathbf{X}_t\}_{t\in\mathbb{Z}}$ aggregates information of its

ACV $\mathbf{\Gamma}_x(\ell)$, $\ell \in \mathbb{Z}$, at a specific frequency $\omega \in [-\pi, \pi]$, and is obtained by the Fourier transform $\mathbf{\Sigma}_x(\omega) = (2\pi)^{-1} \sum_{\ell=-\infty}^{\infty} \mathbf{\Gamma}_x(\ell) \exp(-\iota \ell \omega)$ where $\iota = \sqrt{-1}$. Denoting the sample ACV matrix of $\mathbf{X}_t$ at lag ℓ by

$$\widehat{\mathbf{\Gamma}}_x(\ell) = \frac{1}{n} \sum_{t=\ell+1}^{n} \mathbf{X}_{t-\ell} \mathbf{X}_t^\top \text{ when } \ell \geq 0 \quad \text{and} \quad \widehat{\mathbf{\Gamma}}_x(\ell) = (\widehat{\mathbf{\Gamma}}_x(-\ell))^\top \text{ when } \ell < 0,$$

we estimate the spectral density of $\mathbf{X}_t$ by

$$\widehat{\mathbf{\Sigma}}_x(\omega_k) = \frac{1}{2\pi} \sum_{\ell=-m}^{m} K\left(\frac{\ell}{m}\right) \widehat{\mathbf{\Gamma}}_x(\ell) \exp(-\iota \ell \omega_k), \tag{6.8}$$

where $K(\cdot)$ denotes a kernel, m the kernel bandwidth (for its choice, see Section 6.3) and $\omega_k = 2\pi k/(2m+1)$ the Fourier frequencies. We adopt the Bartlett kernel as $K(\cdot)$ which ensures positive semi-definiteness of $\widehat{\mathbf{\Sigma}}_x(\omega)$ and also $\widehat{\mathbf{\Gamma}}_\xi(\ell)$ estimating $\mathbf{\Gamma}_\xi(\ell)$ obtained as described below.

Performing PCA on $\widehat{\mathbf{\Sigma}}_x(\omega_k)$ at each ω_k, we obtain the estimator of the spectral density matrix of $\boldsymbol{\chi}_t$ as $\widehat{\mathbf{\Sigma}}_\chi(\omega_k) = \sum_{j=1}^{q} \widehat{\mu}_{x,j}(\omega_k) \widehat{\mathbf{e}}_{x,j}(\omega_k) (\widehat{\mathbf{e}}_{x,j}(\omega_k))^*$, where $\widehat{\mu}_{x,j}(\omega_k)$ denotes the j-th largest eigenvalue of $\widehat{\mathbf{\Sigma}}_x(\omega_k)$, $\widehat{\mathbf{e}}_{x,j}(\omega_k)$ its associated eigenvector, and for any vector $\mathbf{a} \in \mathbb{C}^n$, we denote its transposed complex conjugate by $\mathbf{a}^*$. Then taking the inverse Fourier transform of $\widehat{\mathbf{\Sigma}}_\chi(\omega_k)$, $-m \leq k \leq m$, leads to an estimator of $\mathbf{\Gamma}_\chi(\ell)$, the ACV matrix of $\boldsymbol{\chi}_t$, as

$$\widehat{\mathbf{\Gamma}}_\chi(\ell) = \frac{2\pi}{2m+1} \sum_{k=-m}^{m} \widehat{\mathbf{\Sigma}}_\chi(\omega_k) \exp(\iota \ell \omega_k) \quad \text{for} \quad -m \leq \ell \leq m.$$

Finally, we estimate the ACV of $\boldsymbol{\xi}_t$ by

$$\widehat{\mathbf{\Gamma}}_\xi(\ell) = \widehat{\mathbf{\Gamma}}_x(\ell) - \widehat{\mathbf{\Gamma}}_\chi(\ell). \tag{6.9}$$

When we assume the restricted factor model in (6.3), the factor-adjustment step is simplified as it suffices to perform PCA in the time domain, i.e. eigenanalysis of the sample covariance matrix $\widehat{\mathbf{\Gamma}}_x(0)$. Denoting the eigenvector of $\widehat{\mathbf{\Gamma}}_x(0)$ associated with its j-th largest eigenvalue by $\widehat{\mathbf{e}}_{x,j}$, we obtain $\widehat{\mathbf{\Gamma}}_\xi(\ell) = \widehat{\mathbf{\Gamma}}_x(\ell) - \widehat{\mathbf{E}}_x \widehat{\mathbf{E}}_x^\top \widehat{\mathbf{\Gamma}}_x(\ell) \widehat{\mathbf{E}}_x \widehat{\mathbf{E}}_x^\top$ where $\widehat{\mathbf{E}}_x = [\widehat{\mathbf{e}}_{x,j}, 1 \leq j \leq r]$.

6.2.3.2 Step 2: Estimation of $\mathcal{N}^{\mathbf{G}}$

Recall from (6.5) that $\mathcal{N}^{\mathrm{G}}$ representing Granger causal linkages, has its edge set determined by the VAR transition matrices $\mathbf{A}_\ell$, $1 \leq \ell \leq d$. By the Yule-Walker equation, we have $\boldsymbol{\beta} = [\mathbf{A}_1, \ldots, \mathbf{A}_d]^\top = \mathbf{G}(d)^{-1} \mathbf{g}(d)$, where

$$\mathbf{G}(d) = \begin{bmatrix} \mathbf{\Gamma}_\xi(0) & \mathbf{\Gamma}_\xi(-1) & \ldots & \mathbf{\Gamma}_\xi(-d+1) \\ \mathbf{\Gamma}_\xi(1) & \mathbf{\Gamma}_\xi(0) & \ldots & \mathbf{\Gamma}_\xi(-d+2) \\ & & \ddots & \\ \mathbf{\Gamma}_\xi(d-1) & \mathbf{\Gamma}_\xi(d-2) & \ldots & \mathbf{\Gamma}_\xi(0) \end{bmatrix} \quad \text{and} \quad \mathbf{g}(d) = \begin{bmatrix} \mathbf{\Gamma}_\xi(1) \\ \mathbf{\Gamma}_\xi(2) \\ \vdots \\ \mathbf{\Gamma}_\xi(d) \end{bmatrix}. \tag{6.10}$$

We propose to estimate $\boldsymbol{\beta}$ as a regularised Yule-Walker estimator based on $\widehat{\mathbf{G}}(d)$ and $\widehat{\mathbf{g}}(d)$, each of which is obtained by replacing $\mathbf{\Gamma}_\xi(\ell)$ with $\widehat{\mathbf{\Gamma}}_\xi(\ell)$ (see (6.9)) in the definition of $\mathbf{G}(d)$ and $\mathbf{g}(d)$.

For any matrix $\mathbf{M} = [m_{ij}] \in \mathbb{R}^{n_1 \times n_2}$, let $|\mathbf{M}|_1 = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} |m_{ij}|$, $|\mathbf{M}|_\infty = \max_{1 \le i \le n_1} \max_{1 \le j \le n_2} |m_{ij}|$ and $\mathrm{tr}(\mathbf{M}) = \sum_{i=1}^{n_1} m_{ii}$ when $n_1 = n_2$. We consider two estimators of $\boldsymbol{\beta}$. Firstly, we adopt a Lasso-type estimator which solves an ℓ_1-regularised M-estimation problem

$$\widehat{\boldsymbol{\beta}}^{\mathrm{las}} = \underset{\mathbf{M} \in \mathbb{R}^{pd \times p}}{\arg\min}\, \mathrm{tr}\left(\mathbf{M}^\top \widehat{\mathbf{G}}(d)\mathbf{M} - 2\mathbf{M}^\top \widehat{\mathbf{g}}(d)\right) + \lambda |\mathbf{M}|_1 \tag{6.11}$$

with a tuning parameter $\lambda > 0$. In the implementation, we solve (6.11) via the fast iterative shrinkage-thresholding algorithm (FISTA, Beck and Teboulle, 2009). Alternatively, we adopt a constrained ℓ_1-minimisation approach closely related to the Dantzig selector (DS, Candes and Tao, 2007):

$$\widehat{\boldsymbol{\beta}}^{\mathrm{DS}} = \underset{\mathbf{M} \in \mathbb{R}^{pd \times p}}{\arg\min}\, |\mathbf{M}|_1 \quad \text{subject to} \quad \left|\widehat{\mathbf{G}}(d)\mathbf{M} - \widehat{\mathbf{g}}(d)\right|_\infty \le \lambda \tag{6.12}$$

for some tuning parameter $\lambda > 0$. We divide (6.12) into p sub-problems and obtain each column of $\widehat{\boldsymbol{\beta}}^{\mathrm{DS}}$ via the simplex algorithm (using the function `lp` in **lpSolve**).

Barigozzi et al. (2023) establish the consistency of both $\widehat{\boldsymbol{\beta}}^{\mathrm{las}}$ and $\widehat{\boldsymbol{\beta}}^{\mathrm{DS}}$ but, as is typically the case for ℓ_1-regularisation methods, they do not achieve exact recovery of the support of $\boldsymbol{\beta}$. Hence we propose to estimate the edge set of $\mathcal{N}^{\mathrm{G}}$ by thresholding the elements of $\widehat{\boldsymbol{\beta}}$ with some threshold $\mathrm{t} > 0$, where either $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}^{\mathrm{las}}$ or $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}^{\mathrm{DS}}$, i.e.

$$\widetilde{\boldsymbol{\beta}}(\mathrm{t}) = \left[\widehat{\beta}_{ij} \cdot \mathbb{I}_{\{|\widehat{\beta}_{ij}| > \mathrm{t}\}}, 1 \le i \le pd, 1 \le j \le p\right]. \tag{6.13}$$

We discuss cross validation and information criterion methods for selecting λ, and a data-driven choice of t, in Section 6.3.

6.2.3.3 Step 3: Estimation of $\mathcal{N}^{\mathbf{C}}$ and $\mathcal{N}^{\mathbf{L}}$

From the definitions of $\mathcal{N}^{\mathrm{C}}$ and $\mathcal{N}^{\mathrm{L}}$ given in (6.6) and (6.7), their edge sets are obtained by estimating $\boldsymbol{\Delta} = \boldsymbol{\Gamma}^{-1}$ and $\boldsymbol{\Omega} = 2\pi \mathscr{A}^\top(1)\boldsymbol{\Delta}\mathscr{A}(1)$. Given $\widehat{\boldsymbol{\beta}} = [\widehat{\mathbf{A}}_1, \ldots, \widehat{\mathbf{A}}_d]^\top$, some estimator of the VAR parameter matrices obtained as in either (6.11) or (6.12), a natural estimator of $\boldsymbol{\Gamma}$ arises from the Yule-Walker equation $\boldsymbol{\Gamma} = \boldsymbol{\Gamma}_\xi(0) - \sum_{\ell=1}^{d} \mathbf{A}_\ell \boldsymbol{\Gamma}_\xi(\ell) = \boldsymbol{\Gamma}_\xi(0) - \boldsymbol{\beta}^\top \mathbf{g}$, as $\widehat{\boldsymbol{\Gamma}} = \widehat{\boldsymbol{\Gamma}}_\xi(0) - \widehat{\boldsymbol{\beta}}^\top \widehat{\mathbf{g}}$. In high dimensions, it is not feasible or recommended to directly invert $\widehat{\boldsymbol{\Gamma}}$ to estimate $\boldsymbol{\Delta}$. Therefore, we adopt a constrained ℓ_1-minimisation method motivated by the CLIME methodology of Cai et al. (2011). Specifically, the CLIME estimator of $\boldsymbol{\Delta}$ is obtained by first solving

$$\check{\boldsymbol{\Delta}} = \arg\min_{\mathbf{M} \in \mathbb{R}^{p \times p}} |\mathbf{M}|_1 \quad \text{subject to} \quad \left|\widehat{\boldsymbol{\Gamma}}\mathbf{M} - \mathbf{I}\right|_\infty \le \eta, \tag{6.14}$$

and applying a symmetrisation step to $\check{\boldsymbol{\Delta}} = [\check{\delta}_{ii'}, 1 \le i, j \le p]$ as

$$\widehat{\boldsymbol{\Delta}} = [\widehat{\delta}_{ii'}, 1 \le i, i' \le p] \text{ with } \widehat{\delta}_{ii'} = \check{\delta}_{ii'} \cdot \mathbb{I}_{\{|\check{\delta}_{ii'}| \le |\check{\delta}_{i'i}|\}} + \check{\delta}_{i'i} \cdot \mathbb{I}_{\{|\check{\delta}_{i'i}| < |\check{\delta}_{ii'}|\}}. \tag{6.15}$$

for some tuning parameter $\eta > 0$.

Cai et al. (2016) propose ACLIME, which improves the CLIME estimator by selecting the parameter η in (6.15) adaptively; we describe this here. Let $\widehat{\boldsymbol{\Gamma}}^* = \widehat{\boldsymbol{\Gamma}} + n^{-1}\mathbf{I}$ and $\eta_1 = 2\sqrt{\log(p)/n}$.

Step 1: Let $\check{\mathbf{\Delta}}^{(1)} = [\check{\delta}^{(1)}_{ii'}]$ be the solution to

$$\check{\mathbf{\Delta}}^{(1)}_{\cdot i'} = \arg\min_{\mathbf{m}\in\mathbb{R}^p} |\mathbf{m}|_1 \quad \text{subject to} \tag{6.16}$$
$$\left|(\widehat{\mathbf{\Gamma}}^* \mathbf{m} - \mathbf{e}_{i'})_i\right| \le \eta_1(\widehat{\gamma}_{ii} \vee \widehat{\gamma}_{i'i'}) m_{i'} \ \forall\ 1 \le i \le p \ \text{ and } \ m_{i'} > 0,$$

for $i' = 1, \ldots, p$. Then we obtain truncated estimates

$$\widehat{\delta}^{(1)}_{ii} = \check{\delta}^{(1)}_{ii} \cdot \mathbb{I}_{\{|\widehat{\gamma}_{ii}| \le \sqrt{n/\log(p)}\}} + \sqrt{\frac{\log(p)}{n}} \cdot \mathbb{I}_{\{|\widehat{\gamma}_{ii}| > \sqrt{n/\log(p)}\}}.$$

Step 2: We obtain

$$\check{\mathbf{\Delta}}^{(2)}_{\cdot i'} = \arg\min_{\mathbf{m}\in\mathbb{R}^p} |\mathbf{m}|_1 \quad \text{subject to} \quad \left|(\widehat{\mathbf{\Gamma}}^* \mathbf{m} - \mathbf{e}_{i'})_i\right| \le \eta_2 \sqrt{\widehat{\gamma}_{ii} \widehat{\delta}^{(1)}_{i'i'}} \quad \forall\ 1 \le i \le p,$$

where $\eta_2 > 0$ is a tuning parameter. Since $\check{\mathbf{\Delta}}^{(2)}$ is not guaranteed to be symmetric, the final estimator is obtained after a symmetrisation step:

$$\widehat{\mathbf{\Delta}}_{ada} = [\widehat{\delta}_{ii'}, 1 \le i, i' \le p] \text{ with } \widehat{\delta}^{(2)}_{ii'} = \check{\delta}^{(2)}_{ii'} \cdot \mathbb{I}_{\{|\check{\delta}^{(2)}_{ii'}| \le |\check{\delta}^{(2)}_{i'i}|\}} + \check{\delta}^{(2)}_{i'i} \cdot \mathbb{I}_{\{|\check{\delta}^{(2)}_{i'i}| < |\check{\delta}^{(2)}_{ii'}|\}}. \tag{6.17}$$

The constraints in (6.16) incorporate the parameter in the right-hand side. To use linear programming software to solve this, we formulate the constraints for each $1 \le i' \le p$ as

$$\begin{aligned} \forall 1 \le i \le p, \quad & ((\widehat{\mathbf{\Gamma}}^* - \boldsymbol{Q}^{i'})\mathbf{m} - \mathbf{e}_{i'})_i \le 0, \\ \forall 1 \le i \le p, \quad & -((\widehat{\mathbf{\Gamma}}^* + \boldsymbol{Q}^{i'})\mathbf{m} - \mathbf{e}_{i'})_i \le 0, \\ & m_{i'} > 0. \end{aligned}$$

where $\boldsymbol{Q}^{i'}$ has entries $q_{ii'} = \eta_1(\widehat{\gamma}_{ii} \vee \widehat{\gamma}_{i'i'})$ in column i' and 0 elsewhere.

Given the estimators $\widehat{\mathscr{A}}(1) = \mathbf{I} - \sum_{\ell=1}^{d} \widehat{\mathbf{A}}_\ell$ and $\widehat{\mathbf{\Delta}}$, we estimate $\mathbf{\Omega}$ by $\widehat{\mathbf{\Omega}} = 2\pi \widehat{\mathscr{A}}^\top(1) \widehat{\mathbf{\Delta}} \widehat{\mathscr{A}}(1)$. In Barigozzi et al. (2023), $\widehat{\mathbf{\Delta}}$ and $\widehat{\mathbf{\Omega}}$ are shown to be consistent in ℓ_∞- and ℓ_1-norms under suitable sparsity assumptions. However, an additional thresholding step as in (6.13) is required to guarantee consistency in estimating the support of $\mathbf{\Delta}$ and $\mathbf{\Omega}$ and consequently the edge sets of $\mathscr{N}^{\mathrm{C}}$ and $\mathscr{N}^{\mathrm{L}}$. We discuss data-driven selection of these thresholds and η in Section 6.3.

6.2.4 FNETS: Forecasting

Following the estimation procedure, FNETS performs forecasting by estimating the best linear predictor of $\mathbf{X}_{n+a}$ given $\mathbf{X}_t$, $t \le n$, for a fixed integer $a \ge 1$. This is achieved by separately producing the best linear predictors of $\boldsymbol{\chi}_{n+a}$ and $\boldsymbol{\xi}_{n+a}$ as described below, and then combining them.

6.2.4.1 Forecasting the factor-driven component

For given $a \ge 0$, the best linear predictor of $\boldsymbol{\chi}_{n+a}$ given $\mathbf{X}_t$, $t \le n$, under (6.2) is

$$\boldsymbol{\chi}_{n+a|n} = \sum_{\ell=0}^{\infty} \mathbf{B}_{\ell+a} \mathbf{u}_{n-\ell}.$$

Forni et al. (2015) show that the model (6.2) admits a low-rank VAR representation with $\mathbf{u}_t$ as the innovations under mild conditions, and Forni et al. (2017) propose the estimators of $\mathbf{B}_\ell$ and $\mathbf{u}_t$ based on this representation which make use of the estimators of the ACV of $\boldsymbol{\chi}_t$ obtained as described in Section 6.2.3.1. Then, a natural estimator of $\boldsymbol{\chi}_{n+a|n}$ is

$$\widehat{\boldsymbol{\chi}}^{\text{unr}}_{n+a|n} = \sum_{\ell=0}^{K} \widehat{\mathbf{B}}_{\ell+a}\widehat{\mathbf{u}}_{n-\ell} \tag{6.18}$$

for some truncation lag K. We refer to $\widehat{\boldsymbol{\chi}}^{\text{unr}}_{n+a|n}$ as the *unrestricted* estimator of $\boldsymbol{\chi}_{n+a|n}$ as it is obtained without imposing any restrictions on the factor model (6.2).

When $\boldsymbol{\chi}_t$ admits the static representation in (6.3), we can show that $\boldsymbol{\chi}_{n+a|n} = \boldsymbol{\Gamma}_\chi(-a)\mathbf{E}_\chi \mathscr{M}_\chi^{-1}\mathbf{E}_\chi^\top \boldsymbol{\chi}_n$, where $\mathscr{M}_\chi \in \mathbb{R}^{r\times r}$ is a diagonal matrix with the r eigenvalues of $\boldsymbol{\Gamma}_\chi(0)$ on its diagonal and $\mathbf{E}_\chi \in \mathbb{R}^{p\times r}$ the matrix of the corresponding eigenvectors; see Section 4.1 of Barigozzi et al. (2023) and also Forni et al. (2005). This suggests an estimator

$$\widehat{\boldsymbol{\chi}}^{\text{res}}_{n+a|n} = \widehat{\boldsymbol{\Gamma}}_\chi(-a)\widehat{\mathbf{E}}_\chi \widehat{\mathscr{M}}_\chi^{-1}\widehat{\mathbf{E}}_\chi^\top \mathbf{X}_n, \tag{6.19}$$

where $\widehat{\mathscr{M}}_\chi$ and $\widehat{\mathbf{E}}_\chi$ are obtained from the eigendecomposition of $\widehat{\boldsymbol{\Gamma}}_\chi(0)$. We refer to $\widehat{\boldsymbol{\chi}}^{\text{res}}_{n+a|n}$ as the *restricted* estimator of $\boldsymbol{\chi}_{n+a|n}$. As a by-product, we obtain the in-sample estimators of $\boldsymbol{\chi}_t$, $t \le n$, as $\widehat{\boldsymbol{\chi}}_{t|n} = \widehat{\boldsymbol{\chi}}_t$, with either of the two estimators in (6.18) and (6.19).

6.2.4.2 Forecasting the latent VAR process

Once the VAR parameters are estimated either as in (6.11) or (6.12), we produce an estimator of $\boldsymbol{\xi}_{n+a|n} = \sum_{\ell=1}^{d} \mathbf{A}_\ell \boldsymbol{\xi}_{n+a-\ell}$, the best linear predictor of $\boldsymbol{\xi}_{n+a}$ given $\mathbf{X}_t$, $t \le n$, as

$$\widehat{\boldsymbol{\xi}}_{n+a|n} = \sum_{\ell=1}^{\max(1,a)-1} \widehat{\mathbf{A}}_\ell \widehat{\boldsymbol{\xi}}_{n+a-\ell|n} + \sum_{\ell=\max(1,a)}^{d} \widehat{\mathbf{A}}_\ell \widehat{\boldsymbol{\xi}}_{n+a-\ell}. \tag{6.20}$$

Here, $\widehat{\boldsymbol{\xi}}_{n+1-\ell} = \mathbf{X}_{n+1-\ell} - \widehat{\boldsymbol{\chi}}_{n+1-\ell}$ denotes the in-sample estimator of $\boldsymbol{\xi}_{n+1-\ell}$, which may be obtained with either of the two (in-sample) estimators of the factor-driven component in (6.18) and (6.19).

6.3 Tuning parameter selection

6.3.1 Factor numbers q and r

The estimation and forecasting tools of the FNETS methodology require the selection of the number of factors, i.e. q under the unrestricted factor model in (6.2), and r under the restricted, static factor model in (6.3). Under (6.2), there exists a large gap between the q leading eigenvalues of the spectral density matrix of $\mathbf{X}_t$ and the remainder which diverges with p (see also Figure 6.1 We provide two methods for selecting the factor number q, which make use of the postulated eigengap using $\widehat{\mu}_{x,j}(\omega_k)$, $1 \le j \le p$, the eigenvalues of the spectral density estimator of $\mathbf{X}_t$ in (6.8) at a given Fourier frequency ω_k, $-m \le k \le m$.

Hallin and Liška (2007) propose an information criterion for selecting the number of factors under the model (6.2) and further, a methodology for tuning the multiplicative constant in the penalty. Define

$$\mathrm{IC}(b,c) = \log\left(\frac{1}{p}\sum_{j=b+1}^{p}\frac{1}{2m+1}\sum_{k=-m}^{m}\widehat{\mu}_{x,j}(\omega_k)\right) + b\cdot c\cdot \mathrm{pen}(n,p), \tag{6.21}$$

where $\mathrm{pen}(n,p) = \min(p, m^2, \sqrt{n/m})^{-1/2}$ by default (for other choices of the information criterion, see Section D.1) and $c > 0$ a constant. Provided that $\mathrm{pen}(n,p) \to 0$ sufficiently slowly, for an arbitrary value of c, the factor number q is consistently estimated by the minimiser of $\mathrm{IC}(b,c)$ over $b \in \{0,\ldots,\bar{q}\}$, with some fixed $\bar{q}$ as the maximum allowable number of factors. However, this is not the case in finite sample, and Hallin and Liška (2007) propose to simultaneously select q and c. First, we identify $\widehat{q}(n_l,p_l,c) = \arg\min_{0\le b\le\bar{q}}\mathrm{IC}(n_l,p_l,b,c)$ where $\mathrm{IC}(n_l,p_l,b,c)$ is constructed analogously to $\mathrm{IC}(b,c)$, except that it only involves the sub-sample $\{X_{it}, 1\le i\le p_l, 1\le t\le n_l\}$, for sequences $0 < n_1 < \ldots < n_L = n$ and $0 < p_1 < \ldots < p_L = p$. Then, denoting the sample variance of $\widehat{q}(n_l,p_l,c)$, $1\le l\le L$, by $S(c)$, we select $\widehat{q} = \widehat{q}(n,p,\widehat{c})$ with $\widehat{c}$ corresponding to the second interval of stability with $S(c) = 0$ for the mapping $c \mapsto S(c)$ as c increases from 0 to some $c_{\max}$ (the first stable interval is where $\bar{q}$ is selected with a very small value of c). Figure 6.3 plots $\widehat{q}(n,p,c)$ and $S(c)$ for varying values of c obtained from a dataset simulated in Section 6.4.1. In the implementation of this methodology, we set $n_l = n-(L-l)\lfloor n/20\rfloor$ and $p_l = \lfloor 3p/4 + lp/40\rfloor$ with $L = 10$, and $\bar{q} = \min(50, \lfloor\sqrt{\min(n-1,p)}\rfloor)$.

Alternatively, we can adopt the ratio-based estimator $\widehat{q} = \arg\min_{1\le b\le\bar{q}}\mathrm{ER}(b)$ proposed in Avarucci et al. (2022), where

$$\mathrm{ER}(b) = \left(\sum_{k=-m}^{m}\widehat{\mu}_{x,b+1}(\omega_k)\right)^{-1}\left(\sum_{k=-m}^{m}\widehat{\mu}_{x,b}(\omega_k)\right). \tag{6.22}$$

These methods are readily modified to select the number of factors r under the restricted factor model in (6.3), by replacing $(2m+1)^{-1}\sum_{k=-m}^{m}\widehat{\mu}_{x,j}(\omega_k)$ with $\widehat{\mu}_{x,j}$, the j-th largest eigenvalues of the sample covariance matrix $\widehat{\mathbf{\Gamma}}_x(0)$. We refer to Bai and Ng (2002) and Alessi et al. (2010) for the discussion of the information criterion-based method in this setting, and Ahn and Horenstein (2013) for that of the eigenvalue ratio-based method.

6.3.2 Threshold $\mathfrak{t}$

Motivated by Liu et al. (2021), we propose a method for data-driven selection of the threshold $\mathfrak{t}$, which is applied to the estimators of $\mathbf{A}_\ell$, $1\le\ell\le d$, $\mathbf{\Delta}$ or $\mathbf{\Omega}$ for estimating the edge sets of $\mathcal{N}^{\mathrm{G}}$, $\mathcal{N}^{\mathrm{C}}$ or $\mathcal{N}^{\mathrm{L}}$, respectively; see also (6.13).

Let $\mathbf{B} = [b_{ij}] \in \mathbb{R}^{m\times n}$ denote a matrix for which a threshold is to be selected, i.e. $\mathbf{B}$ may be either $\widehat{\boldsymbol{\beta}} = [\widehat{\mathbf{A}}_1,\ldots,\widehat{\mathbf{A}}_d]^\top$, $\widehat{\mathbf{\Delta}}_0$ ($\widehat{\mathbf{\Delta}}$ with diagonals set to zero) or $\widehat{\mathbf{\Omega}}_0$ ($\widehat{\mathbf{\Omega}}$ with diagonals set to zero) obtained from Steps 2 and 3 of FNETS. We work with $\widehat{\mathbf{\Delta}}_0$ and $\widehat{\mathbf{\Omega}}_0$ since we do not threshold the

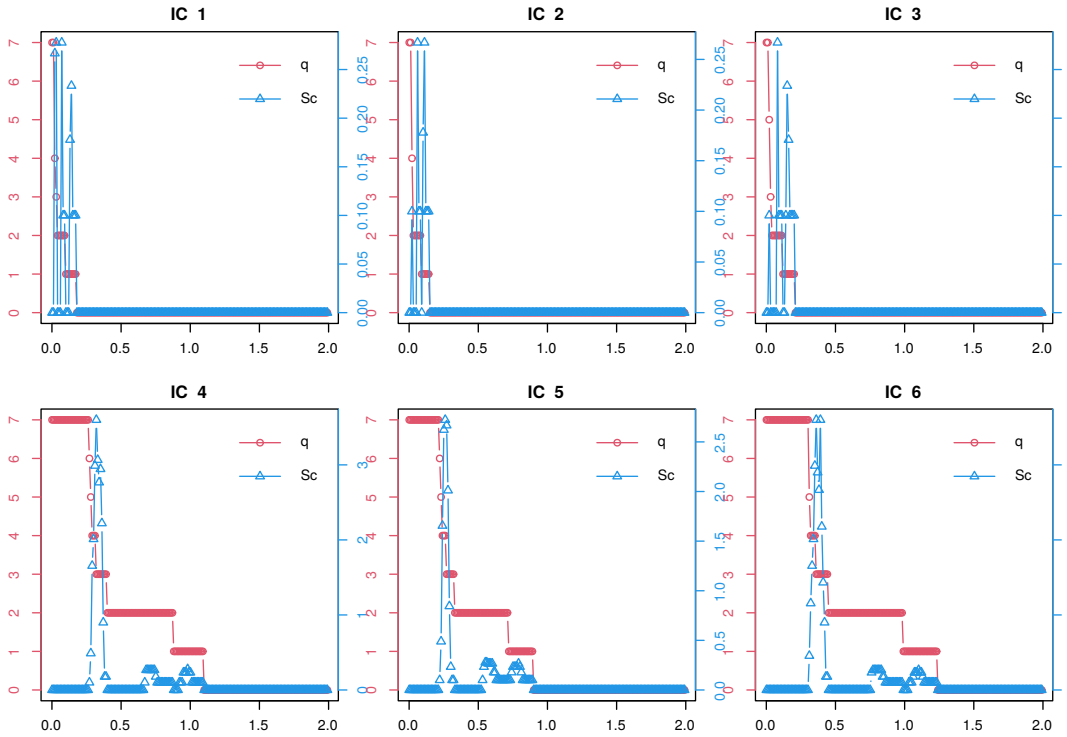


Figure 6.3: Plots of c against $\widehat{q}(n,p,c)$ (in circle, y-axis on the left) and $S(c)$ (in triangle, y-axis on the right) with the six IC (see Section D.1) implemented in the function `factor.number` of **fnets**, on a dataset simulated as in Section 6.4.1 (with $n = 500$, $p = 50$ and $q = 2$). With the default choice of IC in (6.21) (IC_5), we obtain $\widehat{q} = \widehat{q}(n,p,\widehat{c}) = 2$ correctly estimating $q = 2$.

diagonal entries of $\widehat{\mathbf{\Delta}}$ and $\widehat{\mathbf{\Omega}}$. As such estimators have been shown to achieve consistency in ℓ_∞-norm, we expect there exists a large gap between the entries of $\mathbf{B}$ corresponding to true positives and false positives. Further, it is expected that the number of edges reduces at a faster rate when increasing the threshold from 0 towards this (unknown) gap, compared to when increasing the threshold from the gap to $|\mathbf{B}|_\infty$. Therefore, we propose to identify this gap by casting the problem as that of locating a single change point in the trend of the ratio of edges to non-edges,

$$\text{Ratio}_k = \frac{|\mathbf{B}(\mathrm{t}_k)|_0}{\max(N - |\mathbf{B}(\mathrm{t}_k)|_0, 1)}, \quad k = 1,\ldots,M.$$

Here, $\mathbf{B}(\mathrm{t}) = [b_{ij} \cdot \mathbb{I}_{\{|b_{ij}|>\mathrm{t}\}}]$, $|\mathbf{B}(\mathrm{t})|_0 = \sum_{i=1}^{m_1}\sum_{j=1}^{m_2} \mathbb{I}_{\{|b_{ij}|>\mathrm{t}\}}$ and $\{\mathrm{t}_k, 1 \le k \le M : 0 = \mathrm{t}_1 < \mathrm{t}_2 < \cdots < \mathrm{t}_M = |\mathbf{B}|_\infty\}$ denotes a sequence of candidate threshold values. We recommend using an exponentially growing sequence for $\{\mathrm{t}_k\}_{k=1}^M$ since the size of the false positive entries tends to be very small. The quantity N in the denominator of Ratio_k is set as $N = p^2 d$ when $\mathbf{B} = \widehat{\boldsymbol{\beta}}$, and $N = p(p-1)$ when $\mathbf{B} = \widehat{\mathbf{\Delta}}_0$ or $\mathbf{B} = \widehat{\mathbf{\Omega}}_0$. Then, from the difference quotient

$$\text{Diff}_k = \frac{\text{Ratio}_k - \text{Ratio}_{k-1}}{\mathrm{t}_k - \mathrm{t}_{k-1}}, \quad k = 2,\ldots,M,$$

we compute the cumulative sum (CUSUM) statistic

$$\text{CUSUM}_k = \sqrt{\frac{k(M-k)}{M}} \left| \frac{1}{k} \sum_{l=2}^{k} \text{Diff}_l - \frac{1}{M-k} \sum_{l=k+1}^{M} \text{Diff}_l \right|, \quad k = 2, \ldots, M-1,$$

and select $\mathrm{t}_{\text{ada}} = \mathrm{t}_{k^*}$ with $k^* = \arg\max_{2 \le k \le M-1} \text{CUSUM}_k$. For illustration, Figure 6.4 plots Ratio_k and CUSUM_k against candidate thresholds for the dataset simulated in Section 6.4.1.

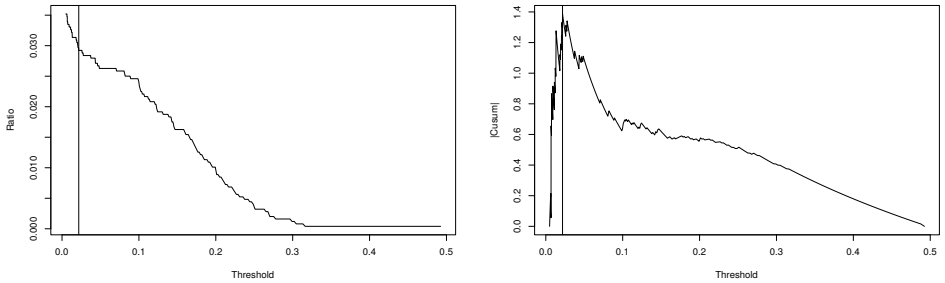


Figure 6.4: Ratio_k (left) and CUSUM_k (right) plotted against t_k when $\mathbf{B} = \widehat{\boldsymbol{\beta}}^{\text{las}}$ obtained from the data simulated in Section 6.4.1 with $n = 500$ and $p = 50$, as a Lasso estimator of the VAR parameter matrix, with the selected t_{ada} denoted by the vertical lines.

6.3.3 VAR order d, λ and η

Steps 2 and 3 of the network estimation methodology of FNETS involve the selection of the tuning parameters λ and η (see (6.11), (6.12) and (6.14)) and the VAR order d. While there exist a variety of methods available for VAR order selection in fixed dimensions (Lütkepohl, 2005, Chapter 4), the data-driven selection of d in high dimensions remains largely unaddressed with a few exceptions (Krampe and Margaritella, 2021; Nicholson et al., 2020; Zheng, 2022). We suggest two methods for jointly selecting λ and d for Step 2. The first method is also applicable for selecting η in Step 3.

6.3.3.1 Cross validation

Cross validation (CV) methods have popularly been adopted for tuning parameter and model selection. While some works exist which justify the usage of conventional CV procedure in the time series setting in the absence of model mis-specification (Bergmeir et al., 2018), such arguments do not apply to our problem due to the latency of component time series. Instead, we propose to adopt a modified CV procedure that bears resemblance to out-of-sample evaluation or rolling forecasting validation (Wang and Tsay, 2021), for simultaneously selecting d and λ in Step 2. For this, the data is partitioned into L folds, $\mathscr{I}_l = \{n_l^\circ + 1, \ldots, n_{l+1}^\circ\}$ with $n_l^\circ = \min(l\lceil n/L \rceil, n)$, $1 \le l \le L$, and each fold is split into a training set $\mathscr{I}_l^{\text{train}} = \{n_l^\circ + 1, \ldots, \lceil (n_l^\circ + n_{l+1}^\circ)/2 \rceil\}$ and a test set $\mathscr{I}_l^{\text{test}} = \mathscr{I}_l \setminus \mathscr{I}_l^{\text{train}}$. On each fold, $\boldsymbol{\beta}$ is estimated from $\{\mathbf{X}_t, t \in \mathscr{I}_l^{\text{train}}\}$ as either the Lasso (6.11) or the Dantzig

selector (6.12) estimators with λ as the tuning parameter and some b as the VAR order, say $\widehat{\boldsymbol{\beta}}_l^{\text{train}}(\lambda, b)$, using which we compute the CV measure

$$\text{CV}(\lambda, b) = \sum_{l=1}^{L} \text{tr}\Big(\widehat{\boldsymbol{\Gamma}}_{\xi,l}^{\text{test}}(0) - (\widehat{\boldsymbol{\beta}}_l^{\text{train}}(\lambda, b))^\top \widehat{\mathbf{g}}_l^{\text{test}}(b) - (\widehat{\mathbf{g}}_l^{\text{test}}(b))^\top \widehat{\boldsymbol{\beta}}_l^{\text{train}}(\lambda, b) + (\widehat{\boldsymbol{\beta}}_l^{\text{train}}(\lambda, b))^\top \widehat{\mathbf{G}}_l^{\text{test}}(b) \widehat{\boldsymbol{\beta}}_l^{\text{train}}(\lambda, b)\Big),$$

where $\widehat{\boldsymbol{\Gamma}}_{\xi,l}^{\text{test}}(\ell), \widehat{\mathbf{G}}_l^{\text{test}}(b)$ and $\widehat{\mathbf{g}}_l^{\text{test}}(b)$ are generated analogously as $\widehat{\boldsymbol{\Gamma}}_\xi(\ell)$, $\widehat{\mathbf{G}}(b)$ and $\widehat{\mathbf{g}}(b)$, respectively, from the test set $\{\mathbf{X}_t, t \in \mathscr{I}_l^{\text{test}}\}$. Although we do not directly observe $\boldsymbol{\xi}_t$, the measure $\text{CV}(\lambda, b)$ gives an approximation of the prediction error. Then, we select $(\widehat{\lambda}, \widehat{d}) = \arg\min_{\lambda \in \Lambda, 1 \le b \le \bar{d}} \text{CV}(\lambda, b)$, where Λ is a grid of values for λ, and $\bar{d} \ge 1$ is a pre-determined upper bound on the VAR order. A similar approach is taken for the selection of η with a Burg matrix divergence-based CV measure:

$$\text{CV}(\eta) = \sum_{l=1}^{L} \text{tr}\left(\widehat{\boldsymbol{\Delta}}_l^{\text{train}}(\eta) \widehat{\boldsymbol{\Gamma}}_l^{\text{test}}\right) - \log\left|\widehat{\boldsymbol{\Delta}}_l^{\text{train}}(\eta) \widehat{\boldsymbol{\Gamma}}_l^{\text{test}}\right| - p.$$

Here, $\widehat{\boldsymbol{\Delta}}_l^{\text{train}}(\eta)$ denotes the estimator of $\boldsymbol{\Delta}$ with η as the tuning parameter from $\{\mathbf{X}_t, t \in \mathscr{I}_l^{\text{train}}\}$, and $\widehat{\boldsymbol{\Gamma}}_l^{\text{test}}$ the estimator of $\boldsymbol{\Gamma}$ from $\{\mathbf{X}_t, t \in \mathscr{I}_l^{\text{test}}\}$, see Section 6.2.3.3 for the descriptions of the estimators. In the numerical results reported in Section 6.5, the sample size is relatively small (ranging between $n = 200$ and $n = 500$ while $p \in \{50, 100, 200\}$ and the number of parameters increasing with p^2), and we set $L = 1$ which returns reasonably good performance. When a larger number of observations are available relative to the dimensionality, we may use the number of folds greater than one.

6.3.3.2 Extended Bayesian information criterion

Alternatively, to select the pair (λ, d) in Step 2, we propose to use the extended Bayesian information criterion (eBIC) of Chen and Chen (2008), originally proposed for variable selection in high-dimensional linear regression. Let $\widetilde{\boldsymbol{\beta}}(\lambda, b, \mathrm{t}_{\text{ada}})$ denote the thresholded version of $\widehat{\boldsymbol{\beta}}(\lambda, b)$ as in (6.13) with the threshold t_{ada} chosen as described in Section 6.3.2. Then, letting $s(\lambda, b) = |\widetilde{\boldsymbol{\beta}}(\lambda, b, \mathrm{t}_{\text{ada}})|_0$, we define

$$\text{eBIC}_\alpha(\lambda, b) = \frac{n}{2} \log(\mathscr{L}(\lambda, b)) + s(\lambda, b) \log(n) + 2\alpha \log \binom{bp^2}{s(\lambda, b)}, \quad \text{where} \tag{6.23}$$

$$\mathscr{L}(\lambda, b) = \text{tr}\big(\widehat{\mathbf{G}}(b) - (\widetilde{\boldsymbol{\beta}}(\lambda, b))^\top \widehat{\mathbf{g}}(b) - (\widehat{\mathbf{g}}(b))^\top \widetilde{\boldsymbol{\beta}}(\lambda, b) + (\widetilde{\boldsymbol{\beta}}(\lambda, b))^\top \widehat{\mathbf{G}}(b) \widetilde{\boldsymbol{\beta}}(\lambda, b)\big).$$

Then, we select $(\widehat{\lambda}, \widehat{d}) = \arg\min_{\lambda \in \Lambda, 1 \le b \le \bar{d}} \text{eBIC}_\alpha(\lambda, b)$. The constant $\alpha \in (0, 1)$ determines the degree of penalisation which may be chosen from the relationship between n and p. Preliminary simulations suggest that $\alpha = 0$ is a suitable choice for the dimensions (n, p) considered in our numerical studies.

6.3.4 Other tuning parameters

Motivated by theoretical results reported in Barigozzi et al. (2023), we select the kernel bandwidth for Step 1 of FNETS as $m = \lfloor 4(n/\log(n))^{1/3} \rfloor$. In forecasting the factor-driven component as in (6.18), we set the truncation lag at $K = 20$, as it is expected that the elements of $\mathbf{B}_\ell$ decay rapidly as ℓ increases for short-memory processes.

6.4 Package overview

fnets is available from the Comprehensive R Archive Network (CRAN). The main function, `fnets`, implements the FNETS methodology for the input data and returns an object of S3 class `fnets`. `fnets.var` implements Step 2 of the FNETS methodology estimating the VAR parameters only, and is applicable directly for VAR modelling of high-dimensional time series; its outputs are of class `fnets`. `fnets.factor.model` performs factor modelling under either of the two models (6.2) and (6.3), and returns an object of class `fm`. We provide `predict` methods for the objects of classes `fnets` and `fm`, and a `plot` method for the objects of the `fnets` class. We recommend that the input time series for the above functions are to be transformed to stationarity (if necessary) after a unit root test. In this section, we demonstrate how to use the functions included with the package.

6.4.1 Data generation

For illustration, we generate an example dataset of $n = 500$ and $p = 50$ following the model (6.4). **fnets** provides functions for this purpose. For given n and p, the function `sim.var` generates the VAR(1) process following (6.1) with $d = 1$, $\mathbf{\Gamma}$ as supplied to the function ($\mathbf{\Gamma} = \mathbf{I}$ by default), and $\mathbf{A}_1$ generated as described in Section 6.5. The function `sim.unrestricted` generates the factor-driven component under the unrestricted factor model in (6.2) with q dynamic factors ($q = 2$ by default) and the filter $\mathscr{B}(L)$ generated as in model (C1) of Section 6.5.

```
set.seed(111)
n <- 500
p <- 50
x <- sim.var(n, p)$data + sim.unrestricted(n, p)$data
```

Throughout this section, we use the thus-generated dataset in demonstrating **fnets** unless specified otherwise. There also exists `sim.restricted` which generates the factor-driven component under the restricted factor model in (6.3). For all data generation functions, the default is to use the standard normal distribution for generating $\mathbf{u}_t$ and $\boldsymbol{\varepsilon}_t$, while supplying the argument `heavy = TRUE`, the innovations are generated from $\sqrt{3/5} \cdot t_5$, the t-distribution with 5 degrees of freedom scaled to have unit variance. The package also comes attached with pre-generated datasets `data.restricted` and `data.unrestricted`.

6.4.2 Calling `fnets` with default parameters

The function `fnets` can be called with the $n \times p$ data matrix `x` as the only input, which sets all other arguments to their default choices. Then, it performs the factor-adjustment under the unrestricted model in (6.2) with q estimated by minimising the IC in (6.21). The VAR parameter matrix is estimated via the Lasso estimator in (6.11) with $d = 1$ as the VAR order and the tuning parameters λ and η chosen via CV, and no thresholding is performed. This returns an object of class `fnets` whose entries are described in Table 6.1, and is supported by a `print` method as below.

```
fnets(x)


Factor-adjusted vector autoregressive model with
n: 500, p: 50
Factor-driven common component ---------
Factor model: unrestricted
Factor number: 2
Factor number selection method: ic
Information criterion: IC5
Idiosyncratic VAR component ---------
VAR order: 1
VAR estimation method: lasso
Tuning method: cv
Threshold: FALSE
Non-zero entries: 95/2500
Long-run partial correlations ---------
LRPC: TRUE
```

6.4.3 Calling `fnets` with optional parameters

We can also specify the arguments of `fnets` to control how Steps 1–3 of FNETS are to be performed. The full model call is as follows:

```
out <- fnets(x, center = TRUE, fm.restricted = FALSE,
  q = c("ic", "er"), ic.op = NULL, kern.bw = NULL,
  common.args = list(factor.var.order = NULL, max.var.order = NULL, trunc.lags = 20,
  n.perm = 10), var.order = 1, var.method = c("lasso", "ds"),
  var.args = list(n.iter = NULL, n.cores = min(parallel::detectCores() - 1, 3)),
  do.threshold = FALSE, do.lrpc = TRUE, lrpc.adaptive = FALSE,
  tuning.args = list(tuning = c("cv", "bic"), n.folds = 1, penalty = NULL,
```

Table 6.1: Entries of S3 objects of class `fnets`

Name	Description	Type
`q`	Factor number	integer
`spec`	Spectral density matrices for $\mathbf{X}_t$, $\boldsymbol{\chi}_t$ and $\boldsymbol{\xi}_t$ (when `fm.restricted = FALSE`)	list
`acv`	Autocovariance matrices for $\mathbf{X}_t$, $\boldsymbol{\chi}_t$ and $\boldsymbol{\xi}_t$	list
`loadings`	Estimates of $\mathbf{B}_\ell$, $0 \le \ell \le K$ (when `fm.restricted = FALSE`) or $\boldsymbol{\Lambda}$ (when `fm.restricted = TRUE`)	array
`factors`	Estimates of $\{\mathbf{u}_t\}$ (when `fm.restricted = FALSE`) or $\{\mathbf{F}_t\}$ (when `fm.restricted = TRUE`)	array
`idio.var`	Estimates of $\mathbf{A}_\ell$, $1 \le \ell \le d$ and $\boldsymbol{\Gamma}$, and d and λ used	list
`lrpc`	Estimates of $\boldsymbol{\Delta}$, $\boldsymbol{\Omega}$, (long-run) partial correlations and η used	list
`mean.x`	Sample mean vector	vector
`var.method`	Estimation method for $\mathbf{A}_\ell$ (input parameter)	string
`do.lrpc`	Whether to estimate the long-run partial correlations (input parameter)	Boolean
`kern.bw`	Kernel bandwidth (when `fm.restricted = FALSE`, input parameter)	double

```
  path.length = 10)
)
```

Here, we discuss a selection of input arguments. The `center` argument will de-mean the input. `fm.restricted` determines whether to perform the factor-adjustment under the restricted factor model in (6.3) or not. If the number of factors is known, we can specify `q` with a non-negative integer. Otherwise, it can be set as `"ic"` or `"er"` which selects the factor number estimator to be used between (6.21) and (6.22). When `q = "ic"`, setting the argument `ic.op` as an integer between 1 and 6 specifies the choice of the IC (see Section D.1) where the default is `ic.op = 5`. `kern.bw` takes a positive integer which specifies the bandwidth to be used in Step 1 of FNETS. The list `common.args` specifies arguments for estimating $\mathbf{B}_\ell$ and $\mathbf{u}_t$ under (6.2), and relates to the low-rank VAR representation of $\boldsymbol{\chi}_t$ under the unrestricted factor model. `var.order` specifies a vector of positive integers to be considered in VAR order selection. `var.method` determines the method for VAR parameter estimation, which can be either `"lasso"` (for the estimator in (6.11)) or `"ds"` (for that in (6.12)). The list `var.args` takes additional parameters for Step 2 of FNETS, such as the number of gradient descent steps (`n.iter`, when `var.method = "lasso"`) or the number of cores to use for parallel computing (`n.cores`, when `var.method = "ds"`). `do.threshold` selects whether to threshold the estimators of $\mathbf{A}_\ell$, $1 \le \ell \le d$, $\boldsymbol{\Delta}$ and $\boldsymbol{\Omega}$. It is possible to perform Steps 1–2 of FNETS only without estimating $\boldsymbol{\Delta}$ and $\boldsymbol{\Omega}$ by setting `do.lrpc = FALSE`. If `do.lrpc = TRUE`, `lrpc.adaptive` specifies whether to use the non-adaptive estimator in (6.14) or the ACLIME estimator. The list `tuning.args` supplies arguments to the CV or eBIC procedures, including the number of folds L (`n.folds`), the eBIC parameter α (`penalty`, see (6.23)) and the length of the grid of values for λ and/or η (`path.length`). Finally, it is possible to set only a subset of the arguments of `common.args`, `var.args` and `tuning.args` whereby the unspecified arguments are set to their default values.

The factor adjustment (Step 1) and VAR parameter estimation (Step 2) functionalities can be

accessed individually by calling `fnets.factor.model` and `fnets.var`, respectively. The latter is equivalent to calling `fnets` with `q = 0` and `do.lrpc = FALSE`. The former returns an object of class `fm` which contains the entries of the `fnets` object in Table 6.1 that relate to the factor-driven component only.

6.4.4 Network visualisation

Using the `plot` method available for the objects of class `fnets`, we can visualise the Granger network $\mathcal{N}^{\mathrm{G}}$ induced by the estimated VAR parameter matrices, see the left panel of Figure 6.5.

```
plot(out, type = "granger", display = "network")
```

With `display = "network"`, it plots an `igraph` object from **igraph** (Csardi et al., 2006). Setting the argument `type` to `"pc"` or `"lrpc"`, we can visualise $\mathcal{N}^{\mathrm{C}}$ given by the partial correlations of VAR innovations or $\mathcal{N}^{\mathrm{L}}$ given by the long-run partial correlations of $\boldsymbol{\xi}_t$. We can instead visualise the networks as a heat map, with the edge weights colour-coded by setting `display = "heatmap"`. We plot $\mathcal{N}^{\mathrm{L}}$ as a heat map in the right panel of Figure 6.5 using the following command.

```
plot(out, type = "lrpc", display = "heatmap")
```

It is possible to directly produce an `igraph` object from the objects of class `fnets` via `network` method as:

```
g <- network(out, type = "granger")$network
plot(g, layout = igraph::layout_in_circle(g),
     vertex.color = grDevices::rainbow(1, alpha = 0.2), vertex.label = NA,
     main = "Granger causal network")
```

This produces a plot identical to the left panel of Figure 6.5 using the `igraph` object `g`.

6.4.5 Forecasting

The `fnets` objects are supported by the `predict` method with which we can forecast the input data `n.ahead` steps. For example, we can produce a one-step ahead forecast of $\mathbf{X}_{n+1}$ as

```
pr <- predict(out, n.ahead = 1, fc.restricted = TRUE)
pr$forecast
```

The argument `fc.restricted` specifies whether to use the estimator $\widehat{\boldsymbol{\chi}}^{\mathrm{res}}_{n+h|n}$ in (6.19) generated under a restricted factor model (6.3), or $\widehat{\boldsymbol{\chi}}^{\mathrm{unr}}_{n+h|n}$ in (6.18) generated without such a restriction. Table 6.2 lists the entries from the output from `predict.fnets`. We can similarly produce forecasts from `fnets` objects output from `fnets.var`, or `fm` objects from `fnets.factor.model`.

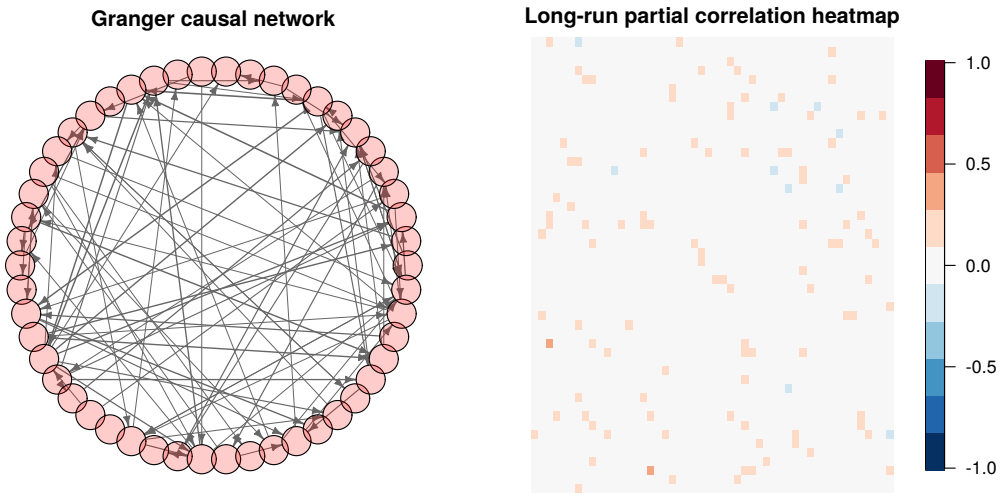


Figure 6.5: Estimated networks for data simulated as in Section 6.4.1. Left: Granger causal network $\mathcal{N}^{\mathrm{G}}$. A directed arrow from node i to node i' indicates that variable i Granger causes node i', and the edge weights proportional to the size of estimated coefficients are visualised by the edge width. Right: Long-run partial correlation network $\mathcal{N}^{\mathrm{L}}$ where the edge weights (i.e. partial correlations) are visualised by the colour.

Table 6.2: Entries of the output from `predict.fnets`

Name	Description	Type
`forecast`	$h \times p$ matrix containing the h-step ahead forecasts of $\mathbf{X}_t$	matrix
`common.predict`	A list containing	list
`$is`	$n \times p$ matrix containing the in-sample estimator of $\boldsymbol{\chi}_t$	
`$fc`	$h \times p$ matrix containing the h-step ahead forecasts of $\boldsymbol{\chi}_t$	
`$h`	Input parameter	
`$r`	Factor number (only produced when `fc.restricted = TRUE`)	
`idio.predict`	A list containing `is`, `fc` and `h`, see `common.predict`	list
`mean.x`	Sample mean vector	vector

6.4.6 Factor number estimation

It is of independent interest to estimate the number of factors (if any) in the input dataset. The function `factor.number` provides access to the two methods for selecting q described in Section 6.3.1. The following code calls the information criterion-based factor number estimation method in (6.21), and prints the output:

```
fn <- factor.number(x, fm.restricted = FALSE)
print(fn)

Factor number selection
Factor model: unrestricted
Method: Information criterion
```

```
Number of factors:
IC1: 2
IC2: 2
IC3: 3
IC4: 2
IC5: 2
IC6: 2
```

Calling `plot(fn)` returns Figure 6.3 which visualises the factor number estimators from six information criteria implemented. Alternatively, we call the eigenvalue ratio-based method in (6.22) as

```
fn <- factor.number(x, method = "er", fm.restricted = FALSE)
```

In this case, `plot(fn)` produces a plot of $\mathrm{ER}(b)$ against the candidate factor number $b \in \{1, \ldots, \bar{q}\}$.

6.4.7 Visualisation of tuning parameter selection procedures

The method for threshold selection discussed in Section 6.3.2 is implemented by the `threshold` function, which returns objects of `threshold` class supported by `print` and `plot` methods.

```
th <- threshold(out$idio.var$beta)
th


Thresholded matrix
Threshold: 0.0297308643
Non-zero entries: 62/2500
```

The call `plot(th)` generates Figure 6.4. Additionally, we provide tools for visualising the tuning parameter selection results adopted in Steps 2 and 3 of FNETS (see Section 6.3.3). These tools are accessible from both `fnets` and `fnets.var` by calling the `plot` method with the argument `display = "tuning"`, e.g.

```
set.seed(111)
n <- 500
p <- 10
x <- sim.var(n, p)$data
out1 <- fnets(x, q = 0, var.order = 1:3, tuning.args = list(tuning = "cv"))
plot(out1, display = "tuning")
```

This generates the two plots reported in Figure 6.6 which visualise the CV errors computed as described in Section 6.3.3.1 and, in particular, the left plot shows that the VAR order is correctly

selected by this approach. When `tuning.args` contains `tuning = "bic"`, the results from the eBIC method described in Section 6.3.3.2 adopted in Step 2, is similarly visualised in place of the left panel of Figure 6.6.

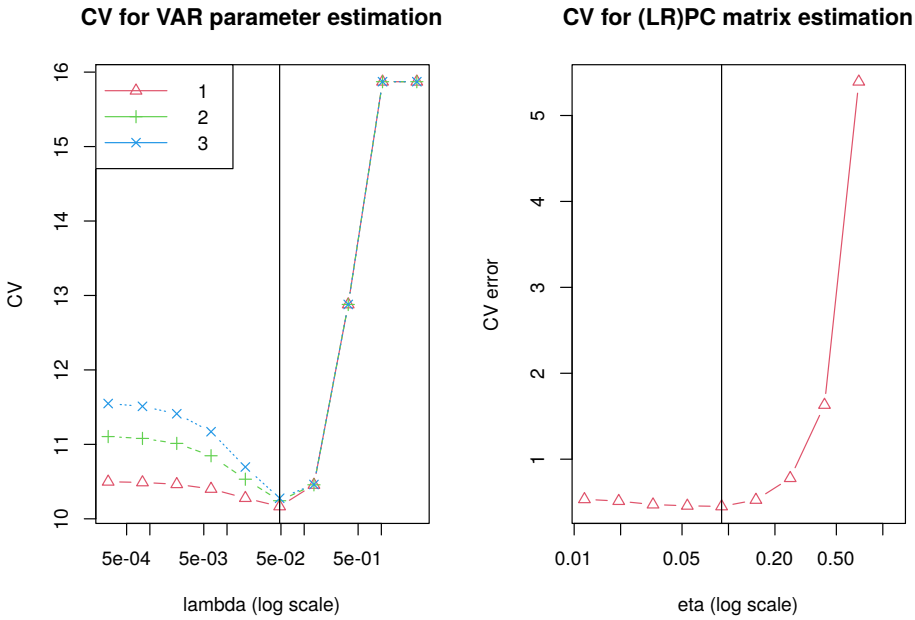


Figure 6.6: Plots of CV(λ, b) against λ with $b \in \{1,2,3\}$ (left) and CV(η) against η (right). Vertical lines denote where the minimum CV measure is attained with respect to λ and η, respectively.

6.5 Simulations

Here we apply the network estimation and forecasting methodologies from FNETS to datasets simulated under a variety of settings, from Gaussian innovations $\mathbf{u}_t$ and $\boldsymbol{\varepsilon}_t$ with (E1) $\boldsymbol{\Delta} = \mathbf{I}$ and (E2) $\boldsymbol{\Delta} \neq \mathbf{I}$, to (E4) heavy-tailed ($t_5$) innovations with $\boldsymbol{\Delta} = \mathbf{I}$, and when $\boldsymbol{\chi}_t$ is generated from (C1) fully dynamic or (C2) static factor models. In addition, we consider the 'oracle' setting (C0) $\boldsymbol{\chi}_t = \mathbf{0}$ where, in the absence of the factor-driven component, the results obtained can serve as a benchmark. We also include the factor-adjusted regression method of Fan et al. (2021) which is referred to as FARM, and present the performance of their estimator $\widehat{\boldsymbol{\beta}}^{\text{FARM}}$ of VAR parameters and forecasts (see Appendix 6.5.1 for full descriptions). For each setting, 100 realisations are generated. The complete simulation results are provided in Appendix D.3.

We also assess the performance of the methods for selecting tuning parameters such as the threshold and VAR order discussed in Section 6.3. Additionally, we compare the adaptive and the non-adaptive estimators in estimating $\boldsymbol{\Delta}$ and also investigate how their performance is carried over to estimating $\boldsymbol{\Omega}$.

6.5.1 Settings

For generating a VAR(d) process $\boldsymbol{\xi}_t$, we first generate a directed Erdős-Rényi random graph $\mathscr{N} = (\mathscr{V}, \mathscr{E})$ on $\mathscr{V} = \{1, \dots, p\}$ with the link probability $1/p$, and set entries of $\mathbf{A}_d$ such that $A_{d,ii'} = 0.275$ when $(i, i') \in \mathscr{E}$ and $A_{d,ii'} = 0$ otherwise. Also, we set $\mathbf{A}_\ell = \mathbf{O}$ for $\ell < d$. The VAR innovations are generated as below.

(E1) Gaussian with $\boldsymbol{\Gamma} = \mathbf{I}$.

(E2) Gaussian with $\boldsymbol{\Gamma} = \boldsymbol{\Delta}^{-1}$, where $\delta_{ii} = 1.5$ for $1 \le i \le p$, $\delta_{ii'} = -1/\sqrt{d_i d_{i'}}$ if $(i, i') \in \mathscr{E}^{\mathrm{C}}$ and $\delta_{ii'} = 0$ otherwise. Here, $\mathscr{N}^{\mathrm{C}}$ is an undirected Erdős-Rényi random graph on $\mathscr{V}$ with the link probability $1/p$, and d_i denotes the degree of the i-th node in $\mathscr{E}^{\mathrm{C}}$. This model is taken from Barigozzi and Brownlees (2019).

(E3) Gaussian with the covariance matrix $\boldsymbol{\Gamma} = \boldsymbol{\Delta}^{-1}$ such that $\delta_{ii} = 1$, $\delta_{i,i+1} = \delta_{i+1,i} = 0.6$, $\delta_{i,i+2} = \delta_{i+2,i} = 0.3$, and $\delta_{ii'} = 0$ for $|i - i'| \ge 3$.

(E4) Heavy-tailed with $\sqrt{5/3} \cdot \varepsilon_{it} \sim_{\mathrm{iid}} t_5$ (such that $\mathsf{Var}(\varepsilon_{it}) = 1$) and $\boldsymbol{\Gamma} = \mathbf{I}$.

We consider two models for the generation of factor-driven common component:

(C1) Taken from Forni et al. (2017), χ_{it} is generated as sum of q AR processes $\chi_{it} = \sum_{j=1}^{q} a_{ij}(1 - \alpha_{ij}L)^{-1} u_{jt}$, where $a_{ij} \sim_{\mathrm{iid}} \mathscr{U}[-1, 1]$ and $\alpha_{ij} \sim_{\mathrm{iid}} \mathscr{U}[-0.8, 0.8]$ with $\mathscr{U}[a, b]$ denoting a uniform distribution. This model does not admit a static factor model representation, and we consider $q = 2$.

(C2) χ_{it} admits a static factor model representation as $\chi_{it} = a_i \sum_{\ell=1}^{2} \boldsymbol{\lambda}_{i\ell}^{\top} \mathbf{f}_{t-\ell+1}$ with $\mathbf{f}_t = \mathbf{D}\mathbf{f}_{t-1} + \mathbf{u}_t$; here, $\mathbf{F}_t = (\mathbf{f}_t^{\top}, \mathbf{f}_{t-1}^{\top})^{\top} \in \mathbb{R}^r$ denotes the static factor with $r = 2q$, $\mathbf{f}_t \in \mathbb{R}^q$ the dynamic factor and and $\mathbf{u}_t = (u_{1t}, \dots, u_{qt})^{\top}$ the common shocks. The entries of the loadings $\boldsymbol{\lambda}_{i\ell} \in \mathbb{R}^q$ are generated i.i.d. from $\mathscr{N}(0, 1)$, and $\mathbf{D} = 0.7 \cdot \mathbf{D}_0 / \Lambda_{\max}(\mathbf{D}_0)$ where the off-diagonal entries of $\mathbf{D}_0 \in \mathbb{R}^{q \times q}$ are generated i.i.d. from $\mathscr{U}[0, 0.3]$ and its diagonal entries from $\mathscr{U}[0.5, 0.8]$. The multiplicative factor a_i is chosen for each realisation to keep sample estimate of $\mathsf{Var}(\chi_{it})/\mathsf{Var}(\xi_{it})$ at one. We fix $q = 2$ (such that $r = 4$).

Additionally, we consider the following 'oracle' setting:

(C0) $\boldsymbol{\chi}_t = 0$, i.e. the idiosyncratic VAR process is directly observed as $\mathbf{X} = \boldsymbol{\xi}_t$.

We vary $(n, p) \in \{(100, 50), (100, 100), (200, 50), (200, 100), (500, 100), (500, 200)\}$. According to the distribution of $\boldsymbol{\varepsilon}_t$, we also vary the distribution of $\mathbf{u}_t$; under (E1) (E2) or (E3), $u_{jt} \sim_{\mathrm{iid}} \mathscr{N}(0, 1)$ while under (E4), $\sqrt{5/3} \cdot u_{jt} \sim_{\mathrm{iid}} t_5$. For each setting, we generate 100 realisations.

For comparison, we consider the FARM methodology of Fan et al. (2021), for factor-adjusted regression modelling under a static factor model. We implement the factor-adjustment step with the information criterion-based factor number estimator of Alessi et al. (2010) and, to the

residuals from removing factors, we apply the Lasso to estimate the VAR parameters using the R package **glmnet** (Friedman et al., 2010); the resulting estimator is referred to as $\widehat{\boldsymbol{\beta}}^{\mathrm{FARM}}$. Also, Fan et al. (2021) propose a forecasting methodology based on VAR modelling of the estimated factors. We report the performance of FARM in estimating the VAR parameters and forecasting $\boldsymbol{\chi}_{n+1}$, $\boldsymbol{\xi}_{n+1}$ and $\mathbf{X}_{n+1}$.

6.5.2 Estimation of $\boldsymbol{\beta}^0$ and $\boldsymbol{\Omega}$

In Tables D.3.1–D.3.4, we report the errors of $\widehat{\boldsymbol{\beta}}^{\mathrm{las}}$, $\widehat{\boldsymbol{\beta}}^{\mathrm{DS}}$ and $\widehat{\boldsymbol{\beta}}^{\mathrm{FARM}}$ in estimating $\boldsymbol{\beta}^0$, and $\widehat{\boldsymbol{\Omega}}^{\mathrm{las}}$ and $\widehat{\boldsymbol{\Omega}}^{\mathrm{DS}}$ ($\widehat{\boldsymbol{\Omega}}$ obtained with $\widehat{\boldsymbol{\beta}}^{\mathrm{las}}$ and $\widehat{\boldsymbol{\beta}}^{\mathrm{DS}}$, respectively) in estimating $\boldsymbol{\Omega}$. To assess the support recovery performance of these estimators, we also report the true positive rate (TPR) when the false positive rate (FPR) is set to be 0.05, and produce the corresponding receiver operating characteristic (ROC) curves averaged over 100 realisations, see Figure 6.7 and also Figures D.3.1–D.3.2 in Appendix D.3.

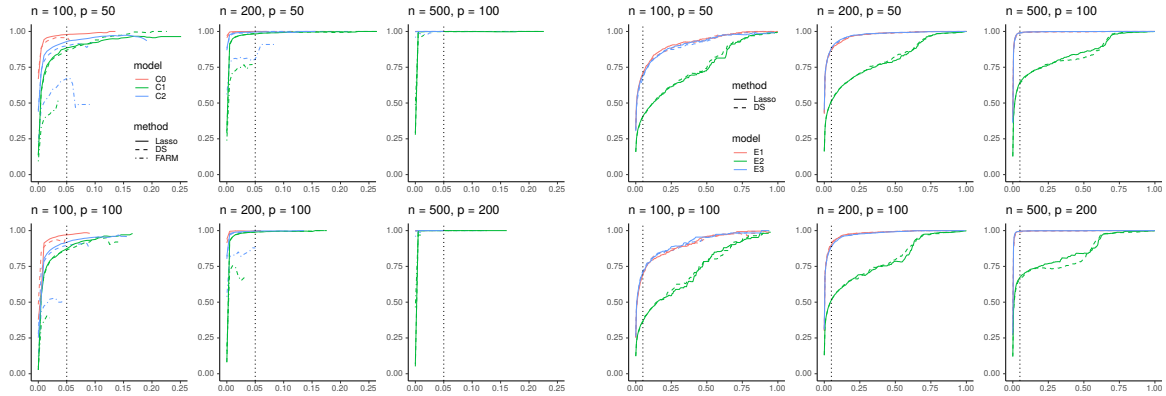


Figure 6.7: Left: ROC curves of TPR against FPR for $\widehat{\boldsymbol{\beta}}^{\mathrm{las}}$, $\widehat{\boldsymbol{\beta}}^{\mathrm{DS}}$ and $\widehat{\boldsymbol{\beta}}^{\mathrm{FARM}}$ in recovering the support of $\boldsymbol{\beta}^0$ when $\boldsymbol{\chi}_t$ is generated under (C0)–(C2) and $\boldsymbol{\xi}_t$ is generated under (E1) with varying n and p, averaged over 100 realisations. Vertical lines indicate where FPR = 0.05. Right: ROC curves for $\widehat{\boldsymbol{\Omega}}^{\mathrm{las}}$ and $\widehat{\boldsymbol{\Omega}}^{\mathrm{DS}}$ in recovering the support of $\boldsymbol{\Omega}$ when $\boldsymbol{\chi}_t$ is generated under (C1) and $\boldsymbol{\xi}_t$ is generated under (E1)–(E4) with varying n and p.

Overall, we observe that with increasing n, the performance of all estimators improve according to all metrics regardless of the data generating processes while increasing p has an adverse effect. Generally, whether the factor-driven component admits a static representation as in (C2) or not as in (C1), FNETS produces estimators of $\boldsymbol{\beta}^0$ that perform as well as those applied under the oracle setting of (C0) with $\boldsymbol{\chi}_t = \mathbf{0}$. Also both $\widehat{\boldsymbol{\beta}}^{\mathrm{las}}$ and $\widehat{\boldsymbol{\beta}}^{\mathrm{DS}}$ outperform FARM in all settings, and they perform very well in estimating the support of $\boldsymbol{\beta}^0$ (i.e. the edge set of $\mathcal{N}^{\mathrm{G}}$) without any thresholding in all scenarios. FARM tends to produce highly sparse estimators with low TPR (see the left panel of Figure 6.7; averaged ROC curves are not necessarily monotonic as it contains pointwise average TPR at given FPR). This may be attributed to the accumulation of errors in the estimates of $\boldsymbol{\xi}_t$, $1 \le t \le n$, possibly leading to low signal-to-noise ratio when estimating the VAR

parameters. When $\mathbf{\Gamma} = \mathbf{\Delta} = \mathbf{I}$ (as in (E1) and (E4)), FNETS performs similarly well in estimating $\mathbf{\Omega}$ regardless of the tail behaviour of $\boldsymbol{\varepsilon}_t$ and $\mathbf{u}_t$. When $\mathbf{\Delta} \neq \mathbf{I}$, it tends to incur larger errors in estimating $\mathbf{\Omega}$ compared to when $\mathbf{\Delta} = \mathbf{I}$, which is more noticeable in terms of support recovery (RHS of Figure 6.7). As shown in Barigozzi et al. (2023), the support of $\mathbf{\Omega}$ depends on the support of $\boldsymbol{\beta}^0$ and $\mathbf{\Delta}$ in a complex way, and generally it is greater than the union of the latter two which makes its estimation more challenging. This may be attributed to the difficulty in estimating non-diagonal $\mathbf{\Delta}$ carried forward to the estimator of $\mathbf{\Omega}$.

6.5.3 Forecasting

We assess the performance of FNETS in estimating the best linear predictors $\boldsymbol{\chi}_{n+1|n}$ by $\widehat{\boldsymbol{\chi}}^{\text{res}}_{n+1|n}$ ('restricted') and $\widehat{\boldsymbol{\chi}}^{\text{unr}}_{n+1|n}$ ('unrestricted'), $\boldsymbol{\xi}_{n+1|n}$ by $\widehat{\boldsymbol{\xi}}^{\text{las}}_{t+1|t}$ and $\widehat{\boldsymbol{\xi}}^{\text{DS}}_{t+1|t}$ (which denote the estimators with Lasso and DS estimators of $\boldsymbol{\beta}^0$, respectively) and finally, $\mathbf{X}_{n+1|n} = \boldsymbol{\chi}_{n+1|n} + \boldsymbol{\xi}_{n+1|n}$ by their combinations.

In Tables D.3.6 (under (E1)) and D.3.10 (under (E2)–(E4)), we report the estimation errors of a given forecast, say $\widehat{\boldsymbol{\Upsilon}}_{n+1|n}$, in estimating $\boldsymbol{\Upsilon}_{n+1|n}$ measured as

$$\frac{|\widehat{\boldsymbol{\Upsilon}}_{n+1|n} - \boldsymbol{\Upsilon}_{n+1|n}|_2^2}{|\boldsymbol{\Upsilon}_{n+1|n}|_2^2} \tag{6.24}$$

and additionally, report the in-sample estimation errors of $\widehat{\boldsymbol{\chi}}_t = \widehat{\boldsymbol{\chi}}^{\text{res}}_t$ and $\widehat{\boldsymbol{\chi}}_t = \widehat{\boldsymbol{\chi}}^{\text{unr}}_t$ measured as $\sum_t |\widehat{\boldsymbol{\chi}}_t - \boldsymbol{\chi}_t|_2^2 / (\sum_t |\boldsymbol{\chi}_t|_2^2)$. Tables D.3.7 (under (E1)) and Tables D.3.11 (under (E2)–(E4)) summarise the forecasting errors measured by

$$\frac{|\widehat{\boldsymbol{\Upsilon}}_{n+1|n} - \boldsymbol{\Upsilon}_{n+1|n}|_\infty}{|\boldsymbol{\Upsilon}_{n+1|n}|_\infty}. \tag{6.25}$$

We also report the forecasting errors measured as

$$\frac{|\widehat{\boldsymbol{\Upsilon}}_{n+1} - \boldsymbol{\Upsilon}_{n+1}|_2^2}{|\boldsymbol{\Upsilon}_{n+1}|_2}, \quad \text{and} \tag{6.26}$$

$$\frac{|\widehat{\boldsymbol{\Upsilon}}_{n+1} - \boldsymbol{\Upsilon}_{n+1}|_\infty^2}{|\boldsymbol{\Upsilon}_{n+1}|_\infty}, \tag{6.27}$$

see Tables D.3.8, D.3.9 (under (E1)), D.3.12 and D.3.13 (under (E2)–(E4)). Additionally, Table D.3.5 contains results in the above error measures obtained from the benchmark case when $\mathbf{X}_t = \boldsymbol{\xi}_t$ under (C0) when $\boldsymbol{\xi}_t$ is generated according to (E1).

Under (C2), occasionally the best linear predictor $\boldsymbol{\chi}_{n+1|n}$ has all its elements close to zero and the small value of $|\boldsymbol{\chi}_{n+1|n}|_2$ inflates the relative estimation error measured as in (6.24); this phenomenon is not observed from the forecasting errors measured with (6.26).

For FNETS, the forecasting performance improves as n increases regardless of the error measures. The estimation error for $\boldsymbol{\chi}_{n+1|n}$ decreases with p while it increases for $\boldsymbol{\xi}_{n+1|n}$, which is due to that the factor-adjustment step enjoys the blessing of dimensionality while VAR estimation tends to suffer from the increase of the dimensionality.

The forecasting method based on the unrestricted GDFM exhibits some instabilities stemming from the instability of the singular VAR equation system adopted for this purpose which, in turn, may be attributed to the possible over-specification of the VAR order (Hörmann and Nisol, 2021). As such, the in-sample estimation and forecasting method based on the restricted GDFM generally performs superior even when $\boldsymbol{\chi}_t$ does not admit a static representation (under (C1)), and the gap between the two forecasting methods gets wider when a static representation exists (under (C2)) as n and p increase. In general, we do not observe any systematic effect of the innovation distribution on the forecasting performance.

FARM performs reasonably well but suffers from lack of sample size when $n = 100$. When $n \geq 200$, it produces in-sample estimators of accuracy marginally better than $\widehat{\boldsymbol{\chi}}_t^{\text{res}}$, but the performance in estimating the VAR parameters carries forward to producing the forecast of $\boldsymbol{\xi}_{n+1}$, which results in worse forecasts of $\mathbf{X}_{n+1}$ overall.

6.5.4 Threshold selection

We assess the performance of the adaptive threshold. We generate $\boldsymbol{\chi}_t$ as in (C1) and fix $d = 1$ for generating $\boldsymbol{\xi}_t$ and further, treat d as known. We consider $(n, p) \in \{(200, 50), (200, 100), (500, 100), (500, 200)\}$. Then we estimate $\boldsymbol{\Omega}$ using the thresholded Lasso estimator of $\mathbf{A}_1$ (see (6.11) and (6.13)) with two choices of thresholds, $\mathrm{t} = \mathrm{t}_{\text{ada}}$ generated as described in Section 6.3.2 and $\mathrm{t} = 0$. To assess the performance of $\widehat{\boldsymbol{\Omega}} = [\widehat{\omega}_{ii'}]$ in recovering of the support of $\boldsymbol{\Omega} = [\omega_{ii'}]$, i.e. $\{(i, i') : \omega_{ii'} \neq 0\}$, we plot the receiver operating characteristic (ROC) curves of true positive rate (TPR) against false positive rate (FPR), where

$$\text{TPR} = \frac{|\{(i,i') : \widehat{\omega}_{ii'} \neq 0 \text{ and } \omega_{ii'} \neq 0\}|}{|\{(i,i') : \omega_{ii'} \neq 0\}|} \quad \text{and} \quad \text{FPR} = \frac{|\{(i,i') : \widehat{\omega}_{ii'} \neq 0 \text{ and } \omega_{ii'} = 0\}|}{|\{(i,i') : \omega_{ii'} = 0\}|}.$$

Tables 6.3 and 6.4 report the errors in estimating $\mathbf{A}_1$ and $\boldsymbol{\Omega}$ when the threshold $\mathrm{t} = \mathrm{t}_{\text{ada}}$ or $\mathrm{t} = 0$ is applied to the estimator of $\mathbf{A}_1$ obtained by either the Lasso (6.11) or the DS (6.12) estimators. With a matrix $\boldsymbol{\gamma}$ as an estimand we measure the estimation error of its estimator $\widehat{\boldsymbol{\gamma}}$ using the following (scaled) matrix norms:

$$L_F = \frac{\|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}\|_F}{\|\boldsymbol{\gamma}\|_F} \quad \text{and} \quad L_2 = \frac{\|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}\|}{\|\boldsymbol{\gamma}\|}.$$

Figure 6.8 plots the ROC curves averaged over 100 realisations when $\mathrm{t} = \mathrm{t}_{\text{ada}}$ and $\mathrm{t} = 0$. When $\boldsymbol{\Delta} = \mathbf{I}$ under (E1), we see little improvement from adopting t_{ada} as the support recovery performance is already good even without thresholding. However, when $\boldsymbol{\Delta} \neq \mathbf{I}$ under (E3), the adaptive threshold leads to improved support recovery especially when the sample size is large. Tables 6.3 and 6.4 additionally report the errors in estimating $\mathbf{A}_1$ and $\boldsymbol{\Omega}$ with and without thresholding, where we see little change is brought by thresholding. In summary, we conclude that the estimators already perform reasonably well without thresholding, and the adaptive threshold t_{ada} brings marginal improvement in support recovery which is of interest in network estimation.

Table 6.3: Errors in estimating $\mathbf{A}_1$ with $t \in \{0, t_{ada}\}$ in combination with the Lasso (6.11) and the DS (6.12) estimators, measured by L_F and L_2, averaged over 100 realisations (with standard errors reported in brackets). We also report the average TPR when FPR = 0.05 and the corresponding standard error.

			$t = 0$						$t = t_{ada}$					
			$\widehat{\boldsymbol{\beta}}^{las}$			$\widehat{\boldsymbol{\beta}}^{DS}$			$\widehat{\boldsymbol{\beta}}^{las}$			$\widehat{\boldsymbol{\beta}}^{DS}$		
Model	n	p	TPR	L_F	L_2	TPR	L_F	L_2	TPR	L_F	L_2	TPR	L_F	L_2
(E1)	200	50	0.9681	0.6234	0.7204	0.8991	0.4299	0.3747	0.9413	0.6226	0.7204	0.6932	0.4487	0.3960
			(0.050)	(0.081)	(0.118)	(0.096)	(0.280)	(0.225)	(0.112)	(0.088)	(0.121)	(0.216)	(0.256)	(0.206)
	200	100	0.9398	0.6696	0.8113	0.8810	0.5772	0.4362	0.8832	0.6710	0.8132	0.6491	0.6025	0.4642
			(0.091)	(0.096)	(0.096)	(0.094)	(0.449)	(0.271)	(0.182)	(0.108)	(0.100)	(0.246)	(0.418)	(0.250)
	500	100	0.9990	0.4648	0.6682	0.9304	0.2740	0.2604	0.9971	0.4608	0.6645	0.7237	0.2806	0.2699
			(0.003)	(0.054)	(0.094)	(0.065)	(0.158)	(0.138)	(0.010)	(0.056)	(0.095)	(0.199)	(0.133)	(0.111)
	500	200	0.9986	0.5068	0.7729	0.9167	0.3680	0.3882	0.9964	0.5023	0.7637	0.7095	0.3889	0.4014
			(0.003)	(0.058)	(0.081)	(0.076)	(0.196)	(0.134)	(0.006)	(0.061)	(0.082)	(0.256)	(0.187)	(0.126)
(E3)	200	50	0.9595	0.6375	0.7075	0.8828	0.4673	0.4280	0.9442	0.6356	0.7079	0.6720	0.4835	0.4433
			(0.053)	(0.077)	(0.094)	(0.107)	(0.324)	(0.255)	(0.064)	(0.079)	(0.096)	(0.212)	(0.303)	(0.241)
	200	100	0.9624	0.6200	0.6909	0.8093	0.4519	0.4090	0.9435	0.6175	0.6913	0.5903	0.4765	0.4324
			(0.072)	(0.079)	(0.089)	(0.100)	(0.385)	(0.251)	(0.093)	(0.082)	(0.090)	(0.182)	(0.371)	(0.243)
	500	100	0.9970	0.4657	0.5533	0.9304	0.3434	0.3621	0.9958	0.4638	0.5525	0.8384	0.3370	0.3634
			(0.006)	(0.056)	(0.076)	(0.089)	(0.158)	(0.153)	(0.008)	(0.058)	(0.077)	(0.182)	(0.140)	(0.144)
	500	200	0.9981	0.4702	0.5658	0.9205	0.3684	0.3740	0.9945	0.4686	0.5665	0.8154	0.3663	0.3803
			(0.003)	(0.065)	(0.091)	(0.088)	(0.182)	(0.162)	(0.014)	(0.068)	(0.093)	(0.205)	(0.159)	(0.145)

Table 6.4: Errors in estimating $\boldsymbol{\Omega}$ with $t \in \{0, t_{ada}\}$ applied to the estimator of $\mathbf{A}_1$ in combination with the Lasso (6.11) and the DS (6.12) estimators, measured by L_F and L_2, averaged over 100 realisations (with standard errors reported in brackets). We also report the average TPR when FPR = 0.05 and the corresponding standard error.

			$t = 0$						$t = t_{ada}$					
			$\widehat{\boldsymbol{\beta}}^{las}$			$\widehat{\boldsymbol{\beta}}^{DS}$			$\widehat{\boldsymbol{\beta}}^{las}$			$\widehat{\boldsymbol{\beta}}^{DS}$		
Model	n	p	TPR	L_F	L_2	TPR	L_F	L_2	TPR	L_F	L_2	TPR	L_F	L_2
(E1)	200	50	0.8714	0.4143	0.5553	0.8622	0.4217	0.5691	0.8685	0.4145	0.5559	0.8640	0.4217	0.5695
			(0.108)	(0.048)	(0.066)	(0.119)	(0.054)	(0.070)	(0.118)	(0.049)	(0.067)	(0.121)	(0.055)	(0.070)
	200	100	0.8827	0.4320	0.5890	0.8961	0.4379	0.5949	0.8684	0.4326	0.5892	0.8867	0.4386	0.5960
			(0.084)	(0.050)	(0.072)	(0.080)	(0.046)	(0.065)	(0.139)	(0.052)	(0.074)	(0.120)	(0.048)	(0.066)
	500	100	0.9909	0.3311	0.4916	0.9886	0.3391	0.4989	0.9928	0.3303	0.4901	0.9901	0.3380	0.4975
			(0.016)	(0.031)	(0.069)	(0.021)	(0.036)	(0.065)	(0.015)	(0.032)	(0.069)	(0.018)	(0.037)	(0.066)
	500	200	0.9942	0.3520	0.5287	0.9916	0.3511	0.5400	0.9954	0.3512	0.5273	0.9672	0.3528	0.5399
			(0.009)	(0.038)	(0.054)	(0.018)	(0.045)	(0.065)	(0.008)	(0.039)	(0.055)	(0.129)	(0.055)	(0.072)
(E3)	200	50	0.4074	0.7831	0.8353	0.4027	0.7942	0.8335	0.4063	0.7832	0.8353	0.4045	0.7943	0.8336
			(0.073)	(0.089)	(0.072)	(0.087)	(0.079)	(0.034)	(0.072)	(0.089)	(0.072)	(0.089)	(0.079)	(0.034)
	200	100	0.4178	0.8406	0.8690	0.3541	0.9119	0.8879	0.4486	0.8407	0.8690	0.4038	0.9120	0.8880
			(0.091)	(0.108)	(0.036)	(0.107)	(0.126)	(0.045)	(0.091)	(0.108)	(0.036)	(0.123)	(0.126)	(0.045)
	500	100	0.5405	0.8267	0.8118	0.5632	0.7910	0.7953	0.5406	0.8267	0.8117	0.5628	0.7910	0.7951
			(0.111)	(0.125)	(0.047)	(0.122)	(0.166)	(0.062)	(0.111)	(0.125)	(0.047)	(0.123)	(0.166)	(0.062)
	500	200	0.5951	0.8713	0.8519	0.6487	0.8184	0.8259	0.6918	0.8713	0.8519	0.7101	0.8184	0.8258
			(0.175)	(0.165)	(0.088)	(0.159)	(0.182)	(0.090)	(0.148)	(0.165)	(0.088)	(0.122)	(0.182)	(0.090)

6.5.5 VAR order selection

We compare the performance of the CV and eBIC methods proposed in Section 6.3.3 for selecting the order of the VAR process. Here, we consider the case when $\boldsymbol{\chi}_t = \mathbf{0}$ (setting (C0)) and when $\boldsymbol{\xi}_t$ is generated under (E1) with $d \in \{1, 3\}$. We set $(n, p) \in \{(200, 10), (200, 20), (500, 10), (500, 20)\}$ where the range of p is in line with the simulation studies conducted in the relevant literature (see e.g. Zheng (2022)). We consider $\{1, 2, 3, 4\}$ as the candidate VAR orders. Figure 6.9 and Table 6.5 show that CV works reasonably well regardless of $d \in \{1, 3\}$, with slightly better performance observed

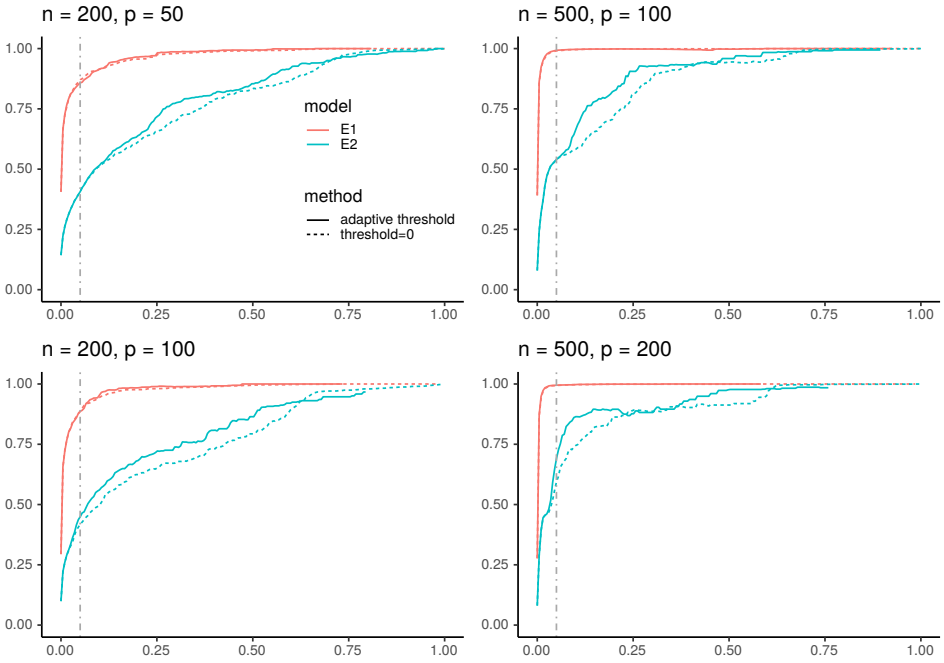


Figure 6.8: ROC curves of TPR against FPR for $\widetilde{\boldsymbol{\beta}}(\mathrm{t})$ (6.13) (with $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}^{\mathrm{las}}$) when $\mathrm{t} = \mathrm{t}_{ada}$ and $\mathrm{t} = 0$ in recovering the support of $\boldsymbol{\Omega}$, averaged over 100 realisations. Vertical lines indicate FPR = 0.05

together with the DS estimator. On the other hand, eBIC tends to over-estimate the VAR order when $d = 1$ while under-estimating it when $d = 3$, and hence is less reliable compared to the CV method.

Table 6.5: Distribution of $\widehat{d} - d$ over 100 realisations when the VAR order is selected by the CV and eBIC methods in combination with the Lasso (6.11) and the DS (6.12) estimators.

			CV								eBIC							
			$\widehat{\boldsymbol{\beta}}^{\mathrm{las}}$				$\widehat{\boldsymbol{\beta}}^{\mathrm{DS}}$				$\widehat{\boldsymbol{\beta}}^{\mathrm{las}}$				$\widehat{\boldsymbol{\beta}}^{\mathrm{DS}}$			
d	n	p	0	1	2	3	0	1	2	3	0	1	2	3	0	1	2	3
1	200	10	81	10	4	5	91	6	2	1	64	17	11	8	64	12	16	8
	200	20	94	6	0	0	94	5	1	0	68	10	9	13	75	10	7	8
	500	10	94	5	1	0	86	7	4	3	65	17	11	7	65	18	9	8
	500	20	97	2	0	1	98	1	1	0	70	15	8	7	64	14	10	12
			-2	-1	0	1	-2	-1	0	1	-2	-1	0	1	-2	-1	0	1
3	200	10	0	0	77	23	0	0	78	22	27	3	49	21	30	6	49	15
	200	20	0	0	97	3	0	0	85	15	32	1	48	19	31	2	58	9
	500	10	0	0	76	24	0	0	83	17	30	4	43	23	29	2	40	29
	500	20	0	0	74	26	0	0	97	3	29	3	45	23	25	4	53	18

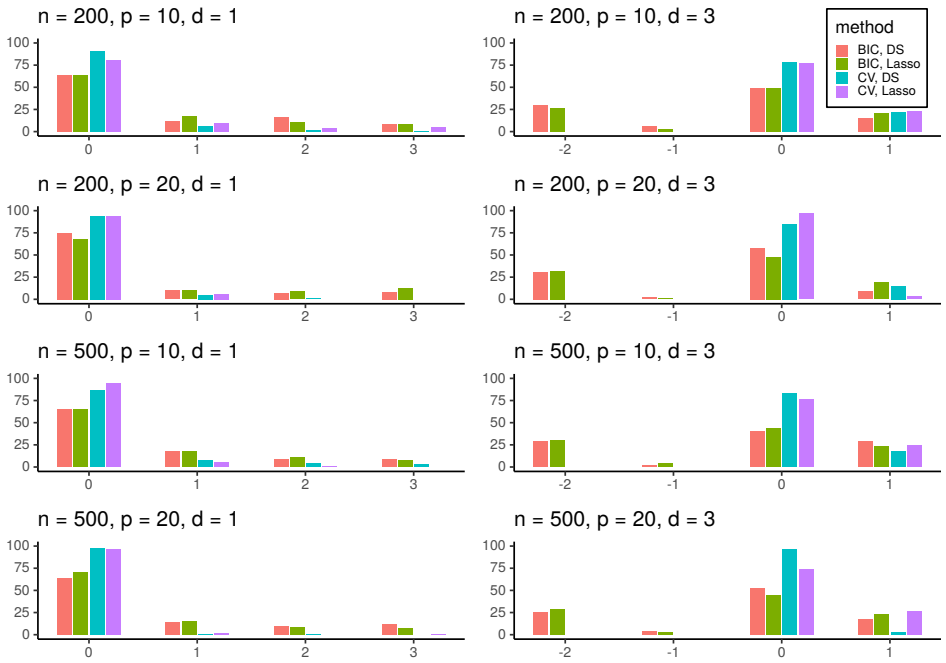


Figure 6.9: Box plots of $\widehat{d} - d$ over 100 realisations when the VAR order is selected by the CV and eBIC methods in combination with the Lasso (6.11) and the DS (6.12) estimators.

6.5.6 CLIME vs. ACLIME estimators

We compare the performance of the adaptive and non-adaptive estimators for the VAR innovation precision matrix $\mathbf{\Delta}$ and its impact on the estimation of $\mathbf{\Omega}$, the inverse of the long-run covariance matrix of the data (see Section 6.2.3.3). We generate $\boldsymbol{\chi}_t$ as in (C1), fix $d = 1$ and treat it as known and consider $(n, p) \in \{(200, 50), (200, 100), (500, 100), (500, 200)\}$.

In Tables 6.6 and 6.7, we report the errors of $\mathbf{\Delta}$ and $\mathbf{\Omega}$. We consider both the Lasso (6.11) and DS (6.12) estimators of VAR parameters, and CLIME and ACLIME estimators for $\mathbf{\Delta}$, which lead to four different estimators for $\mathbf{\Delta}$ and $\mathbf{\Omega}$, respectively. Overall, we observe that with increasing n, the performance of all estimators improve according to all metrics regardless of the scenarios (E1) or (E3), while increasing p has an adverse effect. The two methods perform similarly in setting (E1) when $\mathbf{\Delta} = \mathbf{I}$. There is marginal improvement for adopting the ACLIME estimator noticeable under (E3), particularly in TPR. Figures 6.10 and 6.11 shows the ROC curves for the support recovery of $\mathbf{\Delta}$ and $\mathbf{\Omega}$ when the Lasso estimator is used.

6.6 Data examples

We give demonstrations of the FNETS methodology for network estimation and forecasting on two real sets of data.

Table 6.6: Errors in estimating $\mathbf{\Delta}$ using CLIME and ACLIME estimators, measured by L_F and L_2, averaged over 100 realisations (with standard errors reported in brackets). We also report the average TPR when FPR = 0.05 and the corresponding standard errors.

			CLIME						ACLIME					
			$\widehat{\boldsymbol{\beta}}^{\text{las}}$			$\widehat{\boldsymbol{\beta}}^{\text{DS}}$			$\widehat{\boldsymbol{\beta}}^{\text{las}}$			$\widehat{\boldsymbol{\beta}}^{\text{DS}}$		
Model	n	p	TPR	L_F	L_2	TPR	L_F	L_2	TPR	L_F	L_2	TPR	L_F	L_2
(E1)	200	50	1.000	0.215	0.489	1.000	0.220	0.497	1.000	0.207	0.472	1.000	0.209	0.469
			(0.000)	(0.047)	(0.223)	(0.000)	(0.047)	(0.182)	(0.002)	(0.043)	(0.173)	(0.000)	(0.041)	(0.116)
	200	100	1.000	0.235	0.513	1.000	0.241	0.521	1.000	0.223	0.507	1.000	0.228	0.518
			(0.000)	(0.036)	(0.089)	(0.000)	(0.036)	(0.107)	(0.000)	(0.033)	(0.084)	(0.000)	(0.034)	(0.099)
	500	100	1.000	0.181	0.458	1.000	0.183	0.466	1.000	0.176	0.452	1.000	0.178	0.458
			(0.000)	(0.022)	(0.062)	(0.000)	(0.029)	(0.087)	(0.000)	(0.022)	(0.052)	(0.000)	(0.028)	(0.069)
	500	200	1.000	0.198	0.510	1.000	0.193	0.492	1.000	0.187	0.505	1.000	0.182	0.489
			(0.000)	(0.027)	(0.066)	(0.000)	(0.035)	(0.065)	(0.000)	(0.026)	(0.056)	(0.000)	(0.033)	(0.057)
(E3)	200	50	0.659	0.422	0.816	0.662	0.391	0.608	0.682	0.397	0.706	0.687	0.380	0.600
			(0.058)	(0.101)	(0.654)	(0.057)	(0.031)	(0.144)	(0.055)	(0.056)	(0.351)	(0.054)	(0.030)	(0.176)
	200	100	0.639	0.417	0.695	0.637	0.420	0.720	0.669	0.404	0.663	0.668	0.405	0.684
			(0.044)	(0.039)	(0.205)	(0.042)	(0.043)	(0.249)	(0.041)	(0.037)	(0.162)	(0.039)	(0.037)	(0.193)
	500	100	0.730	0.372	0.764	0.726	0.499	1.708	0.735	0.358	0.650	0.734	0.361	0.718
			(0.035)	(0.097)	(0.828)	(0.039)	(1.101)	(7.586)	(0.032)	(0.038)	(0.322)	(0.031)	(0.056)	(0.517)
	500	200	0.729	0.370	0.711	0.728	0.362	0.736	0.737	0.363	0.647	0.737	0.354	0.673
			(0.028)	(0.035)	(0.355)	(0.028)	(0.035)	(0.384)	(0.023)	(0.026)	(0.239)	(0.024)	(0.028)	(0.279)

Table 6.7: Errors in estimating $\mathbf{\Omega}$ using CLIME and ACLIME estimators of $\mathbf{\Delta}$, measured by L_F and L_2, averaged over 100 realisations (with standard errors reported in brackets). We also report the average TPR when FPR = 0.05 and the corresponding standard errors.

			CLIME						ACLIME					
			$\widehat{\boldsymbol{\beta}}^{\text{las}}$			$\widehat{\boldsymbol{\beta}}^{\text{DS}}$			$\widehat{\boldsymbol{\beta}}^{\text{las}}$			$\widehat{\boldsymbol{\beta}}^{\text{DS}}$		
Model	n	p	TPR	L_F	L_2	TPR	L_F	L_2	TPR	L_F	L_2	TPR	L_F	L_2
(E1)	200	50	0.871	0.415	0.557	0.862	0.422	0.571	0.867	0.411	0.558	0.856	0.417	0.570
			(0.108)	(0.050)	(0.070)	(0.119)	(0.055)	(0.080)	(0.106)	(0.051)	(0.088)	(0.114)	(0.053)	(0.083)
	200	100	0.883	0.432	0.589	0.896	0.438	0.595	0.868	0.423	0.583	0.883	0.429	0.587
			(0.084)	(0.050)	(0.072)	(0.080)	(0.046)	(0.065)	(0.088)	(0.048)	(0.077)	(0.085)	(0.045)	(0.061)
	500	100	0.991	0.331	0.492	0.989	0.339	0.499	0.991	0.328	0.490	0.989	0.337	0.498
			(0.016)	(0.031)	(0.069)	(0.021)	(0.036)	(0.065)	(0.015)	(0.033)	(0.070)	(0.019)	(0.036)	(0.067)
	500	200	0.994	0.352	0.529	0.992	0.351	0.540	0.994	0.344	0.525	0.990	0.342	0.537
			(0.009)	(0.038)	(0.054)	(0.018)	(0.045)	(0.065)	(0.009)	(0.038)	(0.056)	(0.014)	(0.044)	(0.068)
(E3)	200	50	0.509	0.532	0.724	0.510	0.514	0.664	0.504	0.518	0.679	0.507	0.506	0.658
			(0.078)	(0.071)	(0.243)	(0.068)	(0.043)	(0.137)	(0.071)	(0.055)	(0.162)	(0.063)	(0.043)	(0.141)
	200	100	0.511	0.541	0.683	0.513	0.542	0.695	0.509	0.531	0.674	0.504	0.531	0.679
			(0.059)	(0.047)	(0.082)	(0.065)	(0.051)	(0.093)	(0.062)	(0.045)	(0.084)	(0.061)	(0.046)	(0.084)
	500	100	0.640	0.450	0.655	0.624	0.544	1.099	0.642	0.441	0.597	0.637	0.440	0.617
			(0.066)	(0.072)	(0.402)	(0.079)	(0.866)	(3.714)	(0.059)	(0.036)	(0.118)	(0.060)	(0.047)	(0.204)
	500	200	0.670	0.461	0.630	0.658	0.450	0.630	0.677	0.456	0.612	0.661	0.445	0.605
			(0.045)	(0.041)	(0.116)	(0.043)	(0.040)	(0.117)	(0.041)	(0.036)	(0.075)	(0.037)	(0.037)	(0.082)

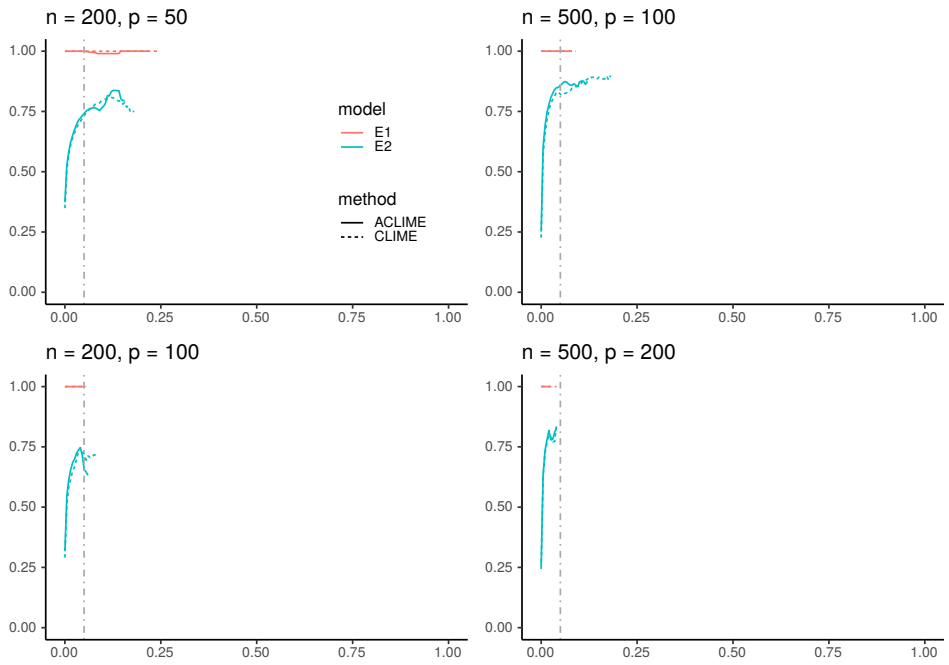


Figure 6.10: ROC curves of TPR against FPR for $\widehat{\mathbf{\Delta}}$ with CLIME and ACLIME estimators in recovering the support of $\mathbf{\Delta}$, averaged over 100 realisations. Vertical lines indicate FPR = 0.05.

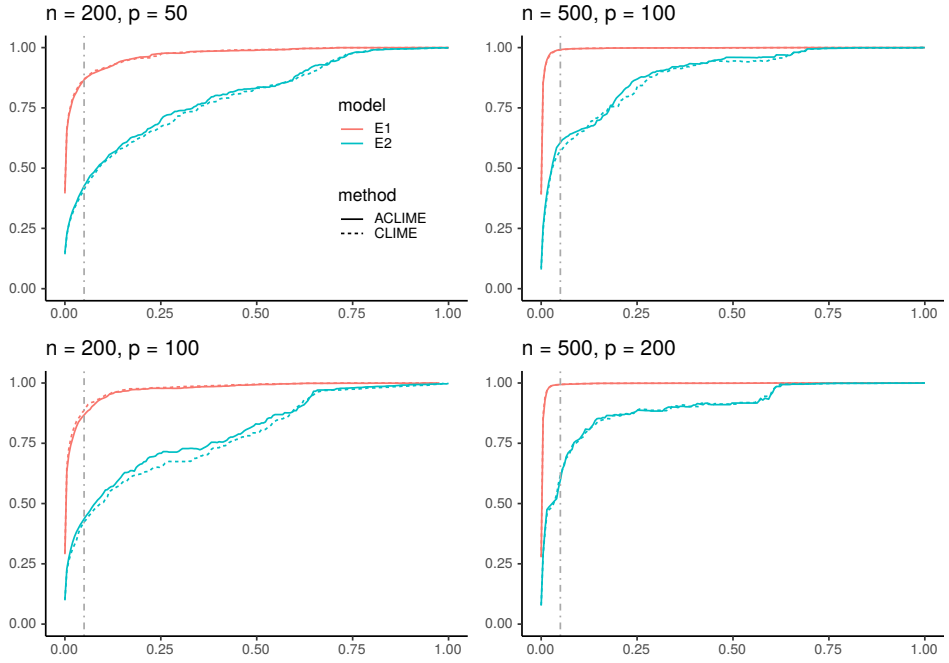


Figure 6.11: ROC curves of TPR against FPR for $\widehat{\mathbf{\Omega}}$ with CLIME and ACLIME estimators in recovering the support of $\mathbf{\Omega}$, averaged over 100 realisations. Vertical lines indicate FPR = 0.05.

6.6.1 Energy price data

Electricity is more difficult to store than physical commodities which results in high volatility and seasonality in spot prices (Han et al., 2022). Global market deregulation has increased the volume of electricity trading, which promotes the development of better forecasting and risk management methods. We analyse a dataset of node-specific prices in the PJM (Pennsylvania, New Jersey and Maryland) power pool area in the United States, accessed using `dataminer2.pjm.com`. There are four node types in the panel, which are Zone, Aggregate, Hub and Extra High Voltage (EHV); for their definitions, see Table D.2.1 and for the names and types of $p = 50$ nodes, see Table D.2.3, all found in Section D.2.1. The series we model is the sum of the real time congestion price and marginal loss price or, equivalently, the difference between the spot price at a given location and the overall system price, where the latter can be thought of as an observed factor in the local spot price. These are obtained as hourly prices and then averaged over each day as per Maciejowska and Weron (2013). We remove any short-term seasonality by subtracting a separate mean for each day of the week. Since the energy prices may take negative values, we adopt the inverse hyperbolic sine transformation as in Uniejewski et al. (2017) for variance stabilisation.

6.6.1.1 Network analysis

We analyse the data collected between 01/01/2021 and 19/07/2021 ($n = 200$). The information criterion in (6.21) returns a single factor ($\widehat{q} = 1$), and $\widehat{d} = 1$ is selected by CV. See Figure 6.14 for the heat maps visualising the three networks $\mathcal{N}^{\mathrm{G}}$, $\mathcal{N}^{\mathrm{C}}$ and $\mathcal{N}^{\mathrm{L}}$ described in Section 6.2.2, which are produced by **fnets**.

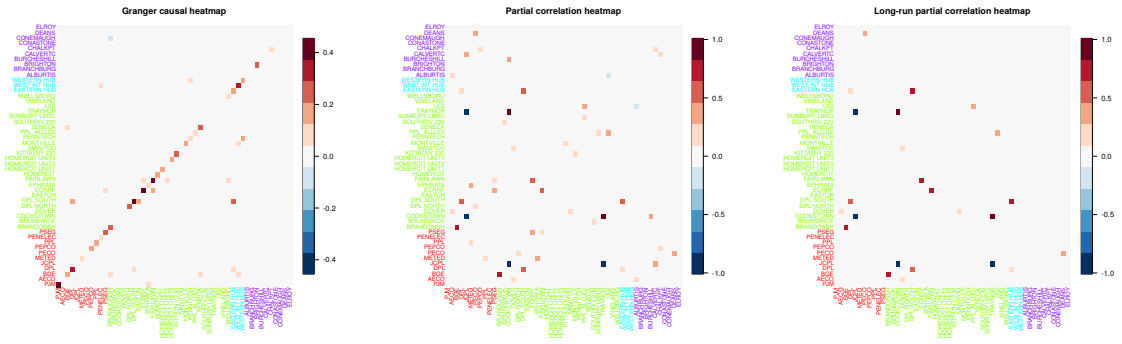


Figure 6.12: Heat maps of the three networks underlying the energy price data collected over the period 01/01/2021–19/07/2021. Left: $\mathcal{N}^{\mathrm{G}}$ obtained with the Lasso estimator (6.11) combined with the adaptive threshold $\mathrm{t}_{\mathrm{ada}}$. Middle: $\mathcal{N}^{\mathrm{C}}$ obtained with the ACLIME estimator of $\mathbf{\Delta}$. Right: $\mathcal{N}^{\mathrm{L}}$ obtained by combining the estimators of VAR parameters and $\mathbf{\Delta}$. In the axis labels, Zone-type nodes are coloured in red, Aggregate-types in green, Hub-types in blue and EHV-types in purple.

The non-zero entries of the VAR parameter matrix estimates tend to take positive values, indicating that high energy prices are persistent and spill over to other nodes. Considering the node types, Hub-type nodes (blue) tend to have out-going edges to nodes of different types, which

reflects the behaviour of the electrical transmission system. Some Zone-type nodes (red) have several in-coming edges from Aggregate-types (green) and Hub-types, while EHV-types (purple) have few edges in $\mathcal{N}^{\mathrm{G}}$, which carries forward to $\mathcal{N}^{\mathrm{L}}$ where we observe that those Zone-type nodes have strong long-run correlations with other nodes while EHV-types do not.

6.6.1.2 Forecasting

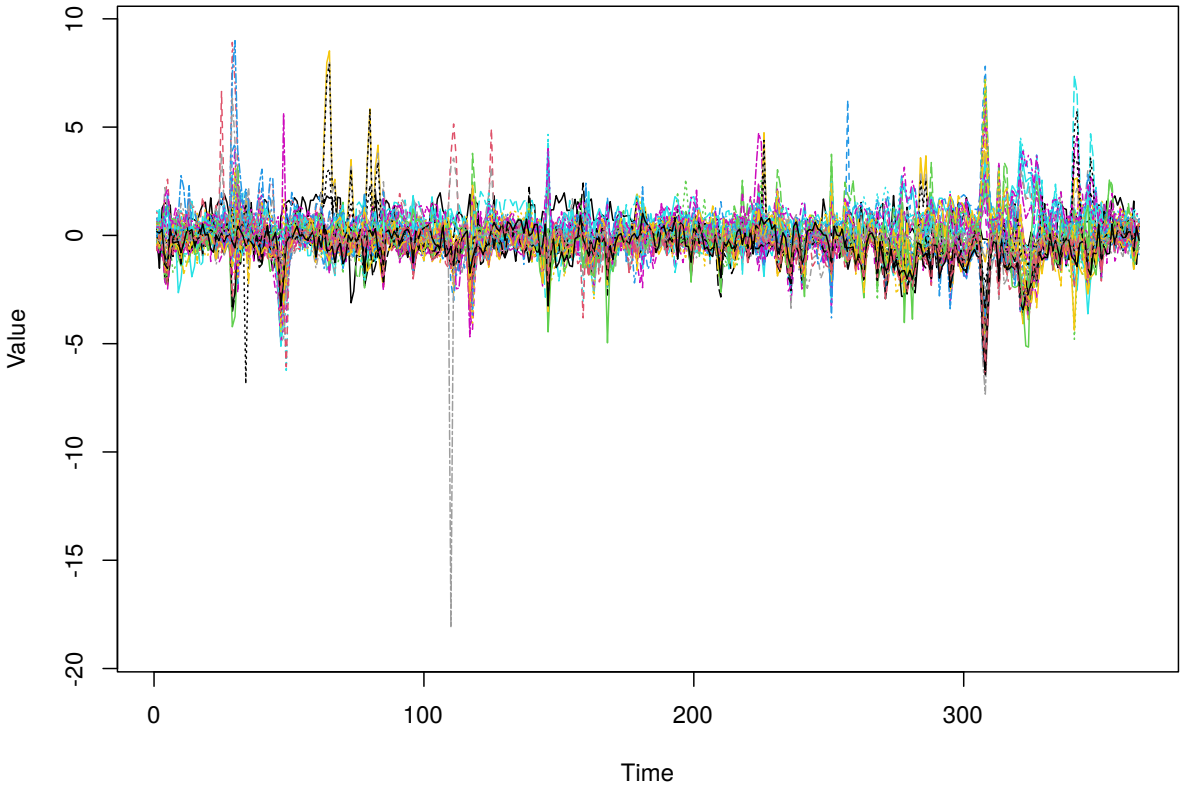


Figure 6.13: Time series plots of real-time congestion and marginal loss prices at 50 nodes in the PJM interchange, averaged daily over 2021. See Section 6.6.1 for a full description.

We perform a rolling window-based forecasting exercise. Starting from $T = 200$, we forecast $\mathbf{X}_{T+1}$ as $\widehat{\mathbf{X}}^{\mathrm{fnets}}_{T+1|T}(n) = \widehat{\boldsymbol{\chi}}_{T+1|T}(n) + \widehat{\boldsymbol{\xi}}_{T+1|T}(n)$, where $\widehat{\boldsymbol{\chi}}_{T+1|T}(n)$ (resp. $\widehat{\boldsymbol{\xi}}_{T+1|T}(n)$) denotes the forecast of $\boldsymbol{\chi}_{T+1}$ (resp. $\boldsymbol{\xi}_{T+1}$) using the preceding n data points $\{\mathbf{X}_t, T-n+1 \le t \le T\}$. We set $n = 200$. After the forecast $\widehat{\mathbf{X}}^{\mathrm{fnets}}_{T+1|T}(n)$ is generated, we update $T \leftarrow T+1$ and repeat the above procedure until $T = 364$ (the penultimate day of 2021). For $\widehat{\mathbf{X}}^{\mathrm{fnets}}_{T+1|T}(n)$, we consider six forecasts from the **fnets** methodology. The first two use $q = 0$ and fit a sparse VAR to the observed series, differing in their estimation methods: the first uses the Lasso, and the second the Dantzig selector. For the other four, we use the forecasting methods derived under the restricted factor model, denoted by $\widehat{\boldsymbol{\chi}}^{\mathrm{res}}_{T+1|T}(n)$, and unrestricted (by $\widehat{\boldsymbol{\chi}}^{\mathrm{unr}}_{T+1|T}(n)$) factor model. For $\widehat{\boldsymbol{\xi}}_{T+1|T}(n)$, we report forecasts from the Lasso ($\widehat{\boldsymbol{\xi}}^{\mathrm{las}}_{T+1|T}(n)$) and Dantzig selector ($\widehat{\boldsymbol{\xi}}^{\mathrm{DS}}_{T+1|T}(n)$) estimators of VAR transition matrices.

Selecting with the eBIC, we use the order $d = 1$ for both. For comparison, we report the forecast produced by a univariate AR model, denoted by $\widehat{\mathbf{X}}^{\mathrm{AR}}_{T+1|T}(n)$, where the order is selected by AIC. We evaluate the performance of $\widehat{\mathbf{X}}_{T+1|T}$ in forecasting $\mathbf{X}_{T+1}$ using scaled average and maximum absolute errors:

$$\mathrm{FE}^{\mathrm{avg}}_{T+1} = \frac{\left|\mathbf{X}_{T+1} - \widehat{\mathbf{X}}_{T+1|T}\right|_2^2}{|\mathbf{X}_{T+1}|_2^2} \quad \text{and} \quad \mathrm{FE}^{\mathrm{max}}_{T+1} = \frac{\left|\mathbf{X}_{T+1} - \widehat{\mathbf{X}}_{T+1|T}\right|_\infty}{|\mathbf{X}_{T+1}|_\infty},$$

See Table 6.9 for the summary of the forecasting results based on $\mathrm{FE}^{\mathrm{avg}}_{T+1}$ and $\mathrm{FE}^{\mathrm{max}}_{T+1}$.

Among the **fnets** forecasts, $\widehat{\boldsymbol{\xi}}^{\mathrm{las}}_{T+1|T}(n)$ with $q = 0$ performs the best. This outperforms $\widehat{\mathbf{X}}^{\mathrm{AR}}_{T+1|T}(n)$ in both metrics in terms of the mean and variance, but not in terms of the median, reflecting the induced bias from the Lasso. The forecasting method based on the unrestricted GDFM shows large instabilities and generally performs worse than the one based on the restricted GDFM, particularly when combined with the Dantzig selector estimator. Inspecting Figure 6.13, we suggest that this is due to structural instability in the data, noting that strong concerted movements are present towards the ends of the sample. This causes shrinkage in the common forecast, and the large residuals feed through to the idiosyncratic component, causing singularity. We can conclude that the the restricted version is fairly robust to outliers, while the unrestricted method is not. The resulting factor models perform similarly to the $q = 0$ case, giving some evidence that a single factor exists in the data.

Table 6.8: Energy data: Mean, median and standard errors of $\mathrm{FE}^{\mathrm{avg}}_{T+1}$ and $\mathrm{FE}^{\mathrm{max}}_{T+1}$ on days in 2021 for $\widehat{\mathbf{X}}^{\mathrm{fnets}}_{T+1|T}(n)$ (in the first four columns), in comparison with AR forecast with d selected by AIC; VAR orders are set to be $d = 1$ for $\widehat{\boldsymbol{\beta}}^{\mathrm{las}}$ and $\widehat{\boldsymbol{\beta}}^{\mathrm{DS}}$. Best performers for each metric are denoted in bold.

		FNETS						
		$q = 0$		Restricted		Unrestricted		
		$\widehat{\boldsymbol{\beta}}^{\mathrm{las}}$	$\widehat{\boldsymbol{\beta}}^{\mathrm{DS}}$	$\widehat{\boldsymbol{\beta}}^{\mathrm{las}}$	$\widehat{\boldsymbol{\beta}}^{\mathrm{DS}}$	$\widehat{\boldsymbol{\beta}}^{\mathrm{las}}$	$\widehat{\boldsymbol{\beta}}^{\mathrm{DS}}$	AR
$\mathrm{FE}^{\mathrm{avg}}$	Mean	**0.8462**	0.8655	0.8752	0.8914	5.1159	4.5259	0.8494
	Median	0.8100	0.8223	0.7931	0.8053	0.9527	0.9860	**0.6749**
	SE	**0.4211**	0.4850	0.6064	0.6101	13.3179	11.7680	0.6817
$\mathrm{FE}^{\mathrm{max}}$	Mean	**0.9284**	0.9394	0.9333	0.9474	1.4388	1.4766	0.9373
	Median	0.9206	0.9368	0.9068	0.9115	1.0225	1.0266	**0.8851**
	SE	**0.1878**	0.2118	0.2270	0.2436	1.1959	1.1504	0.3064

6.6.2 Equity volatility measures

We investigate the interconnectedness in a panel of volatility measures and evaluate its out-of-sample forecasting performance using FNETS. For this purpose, we consider a panel of $p = 46$ stock prices retrieved from the Wharton Research Data Service, of US companies which are

all classified as 'financials' according to the Global Industry Classification Standard; a list of company names and industry groups are found in Appendix D.2.2. The dataset spans the period between January 3, 2000 and December 31, 2012 (3267 trading days). Following Diebold and Yılmaz (2014), we measure the volatility using the high-low range as $\sigma_{it}^2 = 0.361(p_{it}^{\text{high}} - p_{it}^{\text{low}})^2$ where p_{it}^{high} and p_{it}^{low} denote, respectively, the maximum and the minimum log-price of stock i on day t, and set $X_{it} = \log(\sigma_{it}^2)$.

6.6.2.1 Network analysis

We focus on the period 03/2006–02/2010 corresponding to the Great Financial Crisis. We partition the data into four segments of length $n = 252$ each (corresponding to the number of trading days in a single year) and on each segment, we apply FNETS to estimate the three networks $\mathcal{N}^{\text{G}}$, $\mathcal{N}^{\text{C}}$ and $\mathcal{N}^{\text{L}}$ described in Section 6.2.2.

Each row of Figure 6.14 plots the heat maps of the three matrices underlying the three networks of interest. From all four segments, the CV-based approach returns $d = 1$ from the VAR orders $\{1, \ldots, 5\}$ when applied with the Lasso estimator of VAR parameters. Hence from left to right, they represent the Lasso estimator $\widehat{\mathbf{A}}_1 = (\widehat{\boldsymbol{\beta}}^{\text{las}})^\top$, partial correlations from the corresponding $\widehat{\boldsymbol{\Delta}}$ and long-run partial correlations from $\widehat{\boldsymbol{\Omega}}$ (with their diagonals set to be zero). The locations of the non-zero elements coincide with the edge sets of the corresponding networks, and the hues represent the (signed) edge weights.

Prior to March 2007, all networks exhibit a low degree of interconnectedness but the number of edges increases considerably in 03/2007–02/2008 due mainly to an overall increase in dynamic dependence and a prominent role of banks (blue group) not only in $\mathcal{N}^{\text{G}}$ but also in $\mathcal{N}^{\text{C}}$. In 03/2008–02/2009, the companies belonging to the insurance sector (red group) play a central role and in 03/2009–02/2010, the companies become highly interconnected with two particular firms having many outgoing edges in $\mathcal{N}^{\text{G}}$. Also, while most edges in $\mathcal{N}^{\text{L}}$, which captures the overall long-run dependence, have positive weights across time and companies, their weights become negative in this last segment.

6.6.2.2 Forecasting

We perform a rolling window-based forecasting exercise on the trading days in 2012. Starting from $T = 3016$ (the first trading day in 2012), we forecast $\mathbf{X}_{T+1}$ as $\widehat{\mathbf{X}}_{T+1|T}^{\text{fnets}}(n) = \widehat{\boldsymbol{\chi}}_{T+1|T}(n) + \widehat{\boldsymbol{\xi}}_{T+1|T}(n)$, where $\widehat{\boldsymbol{\chi}}_{T+1|T}(n)$ (resp. $\widehat{\boldsymbol{\xi}}_{T+1|T}(n)$) denotes the forecast of $\boldsymbol{\chi}_{T+1}$ (resp. $\boldsymbol{\xi}_{T+1}$) using the preceding n data points $\{\mathbf{X}_t, T - n + 1 \le t \le T\}$. We set $n = 252$. After the forecast $\widehat{\mathbf{X}}_{T+1|T}^{\text{fnets}}(n)$ is generated, we update $T \leftarrow T + 1$ and repeat the above procedure until $T = 3267$ (the last trading day in 2012).

For $\widehat{\boldsymbol{\chi}}_{T+1|T}(n)$, we consider the forecasting methods derived under restricted and unrestricted specifications, and for $\widehat{\boldsymbol{\xi}}_{T+1|T}(n)$, forecasts obtained with the Lasso ($\widehat{\boldsymbol{\xi}}_{T+1|T}^{\text{las}}(n)$) and DS ($\widehat{\boldsymbol{\xi}}_{T+1|T}^{\text{DS}}(n)$) estimators of VAR transition matrices. For Lasso estimation, fitting a VAR model of a large lag can lead to slow convergence, so we set the VAR order $d = 1$; for DS, we consider $d \in \{1, \ldots, 5\}$

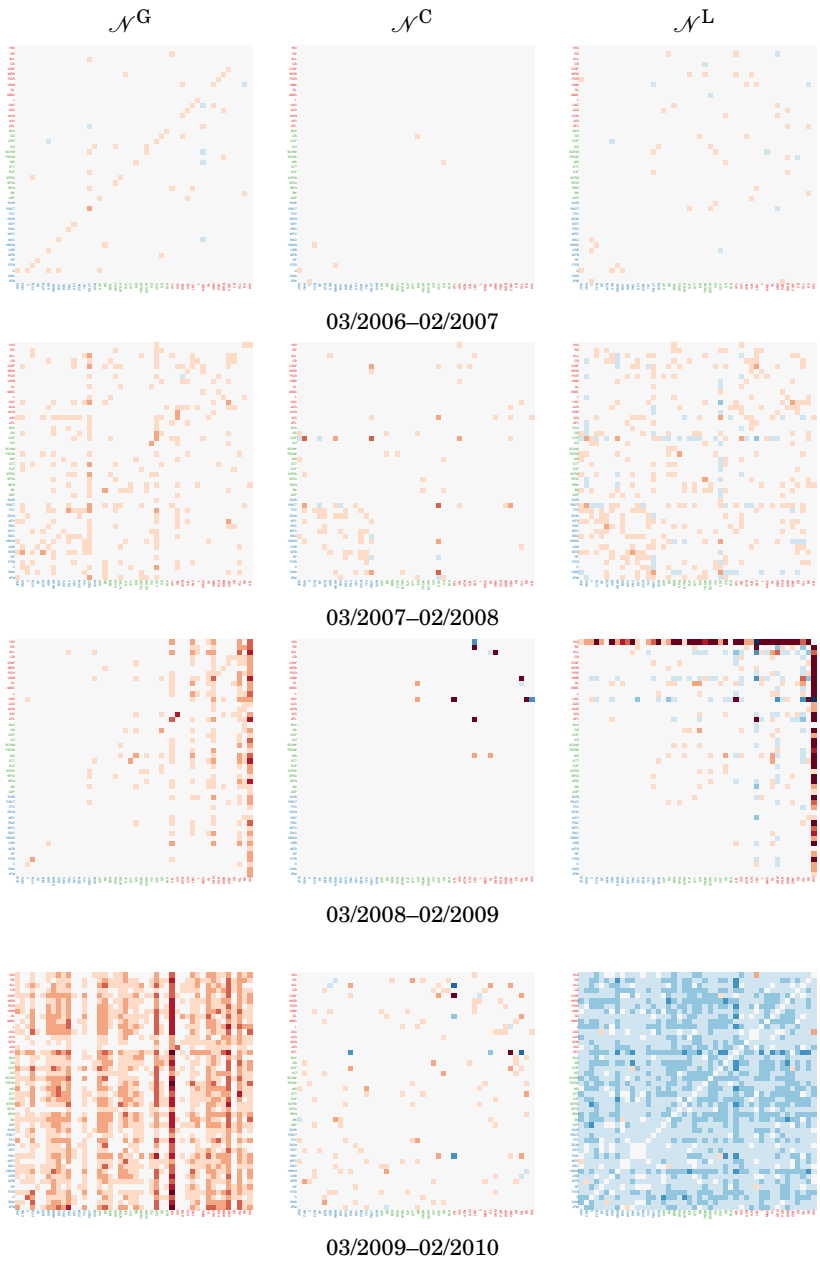


Figure 6.14: Heat maps of the estimators of the VAR transition matrices via Lasso, $\widehat{\boldsymbol{\beta}}^{\mathrm{las}}$, partial correlations from $\widehat{\boldsymbol{\Delta}}$ and long-run partial correlations from $\widehat{\boldsymbol{\Omega}}$ (left to right), which in turn estimate the networks $\mathcal{N}^{\mathrm{G}}$, $\mathcal{N}^{\mathrm{C}}$ and $\mathcal{N}^{\mathrm{L}}$, respectively, over three selected periods. The grouping of the companies according to their industry classifications are indicated by the axis label colours. The heat maps in the left column are in the scale of $[-0.81, 0.81]$ while the others are in the scale of $[-1, 1]$, with red hues denoting large positive values and blue hues large negative values.

where all lags lead to similar results. For comparison, we report the forecasting performance of FARM proposed in Fan et al. (2021). It first fits an AR model to each of the p series ('AR'), projects the residuals on their principal components as in a static factor model approach, and then

Table 6.9: Mean, median and standard errors of $\text{FE}^{\text{avg}}_{T+1}$ and $\text{FE}^{\text{max}}_{T+1}$ on the trading days in 2012 for $\widehat{\mathbf{X}}^{\text{fnets}}_{T+1|T}(n)$ (in the first four columns), in comparison with AR and FARM (Fan et al., 2021) forecasts; VAR orders are set to be $d = 1$ for $\widehat{\boldsymbol{\beta}}^{\text{las}}$ and $d = 5$ for $\widehat{\boldsymbol{\beta}}^{\text{DS}}$ and FARM. Best performers for each metric are denoted in bold.

		FNETS					
		Restricted		Unrestricted			
		$\widehat{\boldsymbol{\beta}}^{\text{las}}$	$\widehat{\boldsymbol{\beta}}^{\text{DS}}$	$\widehat{\boldsymbol{\beta}}^{\text{las}}$	$\widehat{\boldsymbol{\beta}}^{\text{DS}}$	AR	FARM
FE^{avg}	Mean	**0.7258**	0.7651	0.7466	0.9665	0.7572	0.7616
	Median	**0.6029**	0.6163	0.6412	0.6756	0.6511	0.6243
	SE	0.4929	0.5081	0.3748	1.088	**0.4162**	0.4946
FE^{max}	Mean	**0.8433**	0.8752	0.8729	0.9359	0.879	0.8745
	Median	**0.7925**	0.8217	0.8088	0.8708	0.8437	0.8259
	SE	0.2331	0.2406	0.2246	0.3246	**0.2169**	0.2337

fits VAR models to what remains via Lasso. Combining the three steps gives the final forecast $\widehat{\mathbf{X}}^{\text{Farm}}_{T+1|T}(n)$, where we set the first step AR order to be one and the third step VAR order to $d = 5$. The forecast produced by the first step univariate AR modelling, denoted by $\widehat{\mathbf{X}}^{\text{AR}}_{T+1|T}(n)$, is also included in the comparative study.

See Table 6.9 for the summary of the forecasting results based on $\text{FE}^{\text{avg}}_{T+1}$ and $\text{FE}^{\text{max}}_{T+1}$. Among the forecasts generated by FNETS, the combination of $\widehat{\boldsymbol{\chi}}^{\text{res}}_{T+1|T}(n)$ and $\widehat{\boldsymbol{\xi}}^{\text{las}}_{T+1|T}(n)$ performs the best in this exercise, which outperforms $\widehat{\mathbf{X}}^{\text{AR}}_{T+1|T}(n)$ and $\widehat{\mathbf{X}}^{\text{Farm}}_{T+1|T}(n)$ according to both FE^{avg} and FE^{max} on average. As seen in Section 6.5, the forecasting method based on the unrestricted GDFM shows instabilities and generally performs worse than the one based on the restricted GDFM, particularly in combination with the DS estimator.

6.7 Summary

We introduce the R package **fnets** which implements the FNETS methodology proposed by Barigozzi et al. (2023) for network estimation and forecasting of high-dimensional time series exhibiting strong correlations. It further implements data-driven methods for selecting tuning parameters, and provides tools for high-dimensional time series factor modelling under the GDFM which are of independent interest. The efficacy of our package is demonstrated on both real and simulated datasets.

CHAPTER 7

DISCUSSION

In this thesis, we have made methodological contributions to data segmentation and high-dimensional time series analysis. Here we review these contributions and discuss directions for future work.

In Chapter 3 we propose MOSEG, a high-dimensional data segmentation methodology for detecting multiple changes in the parameters under a linear regression model. This is a two-step procedure, which refines initial estimators obtained after scanning for differences in parameter estimates. An extension, MOSEG.MS, adapts to multiscale changes. We show consistency under serial dependence and heavy tails, and show (near-)minimax optimality under Gaussianity.

There are many ways this could be extended. We have considered perhaps the simplest high-dimensional regression model, and the algorithm here could be applied in more general settings, for example Generalised Linear models under structural assumptions, or indeed any model subject to penalised M-estimation.

In Chapter 4 we propose methods to detect and locate multiple change points in the second-order structure of multivariate time series approximated by a vector autoregressive model. We derive score- and Wald-type moving window procedures specific to this problem, and show these consistently estimate change points, extending the results of Reckrühm (2019) and Kirch and Reckrühm (2022). We also address challenges in terms of computation, detectability, and scale. We also use a projection for dimension reduction, which can be combined with a parametric bootstrap, to allow the procedure to be used with panels of larger dimension.

The extensions we have made to the MOSUM procedure are generic, so can be used for other models by specifying a different estimating function. An interesting direction of future research is to allow for penalised estimation procedures, particularly ℓ_2 regularisation through a ridge penalty, which would allow for the analysis of models with greater dimensions. It remains to

prove the consistency of the multiscale algorithm under a true multiscale assumption on the changes.

In Chapter 5 we propose a method for diffusion index forecasting under non-stationarity. This consists of two moving sum data segmentation methods for factor models: one for VAR dynamics, and the other for factor-augmented regression. We show that these consistently detect and locate changes. We highlight extensions to the methodology to reduce computational cost, and to address undetectable or multiscale changes. We then use a forecasting method with data-adaptive weights, which take the estimated change points into account. Through an application to a real macroeconomic dataset we show that our methodology can give superior forecast performance to the popularly-used rolling window method, while adding little computational burden.

Extending this method to the factor-augmented vector autoregression model would be straightforward. It would be of theoretical interest to derive the optimal localisation and detection rates in this setting, and to compare the rates of our method against these. The proposed weighted schemes could be used with other factor models, and these could be compared empirically. Further investigation of how the procedure works in practice, across different macroeconomic and financial datasets, would be of great interest. Finally, incorporating the estimated change points into Kalman filter-type estimation procedures would improve the potential for nowcasting.

In Chapter 6 we introduce the R package **fnets** which implements the FNETS methodology proposed by Barigozzi et al. (2023) for network estimation and forecasting of high-dimensional time series exhibiting strong correlations. It further implements data-driven methods for selecting tuning parameters, and provides tools for high-dimensional time series factor modelling under the GDFM which are of independent interest. The efficacy of the package, and the whole methodology, is demonstrated on both real and simulated datasets.

The question of how to produce forecasts from the model, given non-stationary behaviour, remains. Using the weighted estimation techniques from Chapter 5 is one route, though accounting for asynchronous breaks in the common and idiosyncratic components poses a challenge when the processes are latent.

Across Chapters 3, 4, and 5, consistency results are given for the proposed change point estimation procedures. Corollaries 3.3 and 3.5 provide finite-sample guarantees. The proofs establish that certain events hold with high probability, and then conditioning on these events, the estimation procedures are proven consistent. Theorems 4.2, B.2, 5.3, and 5.4 are shown to be asymptotically consistent, where the error shrinks as the sample size tends to infinity.

APPENDIX A

APPENDIX TO HIGH-DIMENSIONAL DATA SEGMENTATION IN REGRESSION SETTINGS PERMITTING HEAVY TAILS AND TEMPORAL DEPENDENCE

A.1 Proofs

In what follows, for any vector $\mathbf{a} \in \mathbb{R}^p$ and a set $\mathscr{A} \subset \{1,\ldots,p\}$, we denote by $\mathbf{a}(\mathscr{A}) = (a_i, i \in \mathscr{A})^\top$ the sub-vector of $\mathbf{a}$ supported on $\mathscr{A}$. We write the population counterpart of $T_k(G)$ with $\boldsymbol{\beta}^*_{s,e}$ defined in (3.5) as

$$T^*_k(G) = \sqrt{\frac{G}{2}} \left| \boldsymbol{\beta}^*_{k,k+G} - \boldsymbol{\beta}^*_{k-G,k} \right|_2 .$$

Further, we write $\mathscr{S}_{s,e} = \mathrm{supp}(\boldsymbol{\beta}^*_{s,e})$.

A.1.1 Proof of Theorem 3.1

A.1.1.1 Supporting lemmas

Lemma A.1.1. We have

$$T^*_k(G) = \begin{cases} \frac{1}{\sqrt{2G}}(G - |k-\theta_j|)\delta_j & \text{if } \{k-G+1,\ldots,k+G\} \cap \Theta = \{\theta_j\}, \\ 0 & \text{if } \{k-G+1,\ldots,k+G\} \cap \Theta = \emptyset \end{cases}$$

Lemma A.1.2. Define $\boldsymbol{\Delta}_{s,e} = \widehat{\boldsymbol{\beta}}_{s,e} - \boldsymbol{\beta}^*_{s,e}$. With $\lambda \geq 4C_{\mathrm{DEV}}\rho_{n,p}$, we have $\mathsf{P}(\mathscr{B}) \geq 1 - \mathsf{P}(\mathscr{R}^{(1)} \cap \mathscr{D}^{(2)})$ where

$$\mathscr{B} = \left\{ \left|\boldsymbol{\Delta}_{s,e}\right|_2 \leq \frac{12\sqrt{2\mathfrak{s}}\lambda}{\omega\sqrt{e-s}} \text{ and } \left|\boldsymbol{\Delta}_{s,e}(\mathscr{S}^c_{s,e})\right|_1 \leq 3\left|\boldsymbol{\Delta}_{s,e}(\mathscr{S}_{s,e})\right|_1 \text{ for all } 0 \leq s < e \leq n \right.$$
$$\left. \text{with } |\{s+1,\ldots,e\} \cap \Theta| \leq 1 \text{ and } e - s \geq C_0 \max\left[(\omega^{-1}\mathfrak{s}\log(p))^{\frac{1}{1-\tau}}, \rho^2_{n,p} \right] \right\}.$$

Proof. For given $0 \le s < e \le n$, we have

$$\sum_{t=s+1}^{e} \left(Y_t - \mathbf{x}_t^\top \widehat{\boldsymbol{\beta}}_{s,e}\right)^2 + \lambda\sqrt{e-s}\,|\widehat{\boldsymbol{\beta}}_{s,e}|_1 \le \sum_{t=s+1}^{e} \left(Y_t - \mathbf{x}_t^\top \boldsymbol{\beta}^*_{s,e}\right)^2 + \lambda\sqrt{e-s}\,|\boldsymbol{\beta}^*_{s,e}|_1,$$

from which it follows that

$$\begin{aligned}\lambda\sqrt{e-s}\left(|\boldsymbol{\beta}^*_{s,e}|_1 - |\widehat{\boldsymbol{\beta}}_{s,e}|_1\right) &\ge \sum_{t=s+1}^{e} \left[(\mathbf{x}_t^\top \widehat{\boldsymbol{\beta}}_{s,e})^2 - (\mathbf{x}_t^\top \boldsymbol{\beta}^*_{s,e})^2 - 2Y_t \mathbf{x}_t^\top (\widehat{\boldsymbol{\beta}}_{s,e} - \boldsymbol{\beta}^*_{s,e})\right] \\ &= \sum_{t=s+1}^{e} \left[\boldsymbol{\Delta}_{s,e}^\top \mathbf{x}_t \mathbf{x}_t^\top \boldsymbol{\Delta}_{s,e} - 2(Y_t - \mathbf{x}_t^\top \boldsymbol{\beta}^*_{s,e}) \mathbf{x}_t^\top \boldsymbol{\Delta}_{s,e}\right].\end{aligned}$$

Then, noting that $\boldsymbol{\beta}^*_{s,e}(\mathscr{S}^c_{s,e}) = \mathbf{0}$,

$$\begin{aligned}&\frac{1}{\sqrt{e-s}} \sum_{t=s+1}^{e} \left[\boldsymbol{\Delta}_{s,e}^\top \mathbf{x}_t \mathbf{x}_t^\top \boldsymbol{\Delta}_{s,e} - 2(Y_t - \mathbf{x}_t^\top \boldsymbol{\beta}^*_{s,e}) \mathbf{x}_t^\top \boldsymbol{\Delta}_{s,e}\right] + \lambda \left|\widehat{\boldsymbol{\beta}}_{s,e}(\mathscr{S}^c_{s,e})\right|_1 \\ &\qquad \le \lambda \left(\left|\boldsymbol{\beta}^*_{s,e}(\mathscr{S}_{s,e})\right|_1 - \left|\widehat{\boldsymbol{\beta}}_{s,e}(\mathscr{S}_{s,e})\right|_1\right) \le \lambda \left|\boldsymbol{\Delta}_{s,e}(\mathscr{S}_{s,e})\right|_1. \end{aligned} \tag{A.1.1}$$

Since $\lambda \ge 4C_{\mathrm{DEV}}\rho_{n,p}$, it follows from (A.1.1) that on $\mathscr{D}^{(2)}$,

$$\begin{aligned}&\frac{1}{\sqrt{e-s}} \sum_{t=s+1}^{e} \boldsymbol{\Delta}_{s,e}^\top \mathbf{x}_t \mathbf{x}_t^\top \boldsymbol{\Delta}_{s,e} - \frac{\lambda}{2}\left|\boldsymbol{\Delta}_{s,e}\right|_1 + \lambda \left|\boldsymbol{\Delta}_{s,e}(\mathscr{S}^c_{s,e})\right|_1 \le \lambda \left|\boldsymbol{\Delta}_{s,e}(\mathscr{S}_{s,e})\right|_1, \\ &\therefore 0 \le \frac{1}{\sqrt{e-s}} \sum_{t=s+1}^{e} \boldsymbol{\Delta}_{s,e}^\top \mathbf{x}_t \mathbf{x}_t^\top \boldsymbol{\Delta}_{s,e} \le \frac{\lambda}{2}\left(3\left|\boldsymbol{\Delta}_{s,e}(\mathscr{S}_{s,e})\right|_1 - \left|\boldsymbol{\Delta}_{s,e}(\mathscr{S}^c_{s,e})\right|_1\right),\end{aligned}$$

such that

$$\left|\boldsymbol{\Delta}_{s,e}(\mathscr{S}^c_{s,e})\right|_1 \le 3\left|\boldsymbol{\Delta}_{s,e}(\mathscr{S}_{s,e})\right|_1. \tag{A.1.2}$$

This in particular leads to

$$\left|\boldsymbol{\Delta}_{s,e}\right|_1 \le 4\left|\boldsymbol{\Delta}_{s,e}(\mathscr{S}_{s,e})\right|_1 \le 4\sqrt{2\mathfrak{s}}\left|\boldsymbol{\Delta}_{s,e}\right|_2$$

from the definition of $\mathfrak{s}$. Then on $\mathscr{R}^{(1)}$, we have

$$\begin{aligned}6\sqrt{2\mathfrak{s}}\lambda\left|\boldsymbol{\Delta}_{s,e}\right|_2 &\ge \frac{1}{\sqrt{e-s}} \sum_{t=s+1}^{e} \boldsymbol{\Delta}_{s,e}^\top \mathbf{x}_t \mathbf{x}_t^\top \boldsymbol{\Delta}_{s,e} \\ &\ge \omega\sqrt{e-s}\left|\boldsymbol{\Delta}_{s,e}\right|_2^2 - \frac{32C_{\mathrm{RSC}}\mathfrak{s}\log(p)(e-s)^\tau}{\sqrt{e-s}}\left|\boldsymbol{\Delta}_{s,e}\right|_2^2 \ge \frac{\omega}{2}\sqrt{e-s}\left|\boldsymbol{\Delta}_{s,e}\right|_2^2,\end{aligned}$$

where the last inequality follows for $(e-s)^{1-\tau} \ge 64C_{\mathrm{RSC}}\omega^{-1}\mathfrak{s}\log(p)$. In summary,

$$\left|\boldsymbol{\Delta}_{s,e}\right|_2 \le \frac{12\sqrt{2\mathfrak{s}}\lambda}{\omega\sqrt{e-s}}. \tag{A.1.3}$$

Combining (A.1.2) and (A.1.3), the proof is complete. ■

A.1.1.2 Proof of Theorem 3.1 (i)

Let $\mathscr{T}_j = \{\theta_j - \lfloor \eta G \rfloor + 1, \ldots, \theta_j + \lfloor \eta G \rfloor\} \cap \mathscr{T}$ for $1 \le j \le q$. Under Assumptions 3.4 and 3.5, we have $G \ge C_\delta^{-2} C_1 \max\{\omega^{-2}\mathfrak{s}\rho_{n,p}^2, (\omega^{-1}\mathfrak{s}\log(p))^{1/(1-\tau)}\}$ such that the lower bound on $(e-s)$ made in $\mathscr{B}$ (see Lemma A.1.2) is met by all $s = k$ and $e = k+G$, $k = 0, \ldots, n-G$. By Lemma A.1.2,

$$\max_{G \le k \le n-G} \left| T_k(G) - T_k^*(G) \right| \le$$
$$\max_{G \le k \le n-G} \sqrt{\frac{G}{2}} \left(\left| \widehat{\boldsymbol{\beta}}_{k-G,k} - \boldsymbol{\beta}^*_{k-G,k} \right|_2 + \left| \widehat{\boldsymbol{\beta}}_{k,k+G} - \boldsymbol{\beta}^*_{k,k+G} \right|_2 \right) \le \frac{24\sqrt{\mathfrak{s}}\lambda}{\omega}. \tag{A.1.4}$$

First, consider some k for which $\{k-G+2, \ldots, k+G-1\} \cap \Theta = \emptyset$. Then, we have $T_k^*(G) = 0$ from Lemma A.1.1 such that by (A.1.4),

$$\max_{k : \min_{1 \le j \le q} |k - \theta_j| \ge G} T_k(G) \le \max_{G \le \ell \le n-G} \left| T_\ell(G) - T_\ell^*(G) \right| \le \frac{24\sqrt{\mathfrak{s}}\lambda}{\omega} \le D. \tag{A.1.5}$$

This ensures that any $\widetilde{\theta} \in \widetilde{\Theta}$ satisfies $\min_{1 \le j \le q} |\widetilde{\theta} - \theta_j| < G$. Next, let θ_j^{L} and θ_j^{R} denote two points within $\mathscr{T}_j$ which are the closest to θ_j from the left and and the right of θ_j, respectively, with $\theta_j^{\mathrm{L}} = \theta_j^{\mathrm{R}}$ when $r = 1/G$. Then by construction of $\mathscr{T}$,

$$\max(\theta_j - \theta_j^{\mathrm{L}}, \theta_j^{\mathrm{R}} - \theta_j) \le \lfloor rG \rfloor \quad \text{and} \quad \min(\theta_j - \theta_j^{\mathrm{L}}, \theta_j^{\mathrm{R}} - \theta_j) \le \frac{\lfloor rG \rfloor}{2}, \tag{A.1.6}$$

such that from Lemma A.1.1,

$$\max\left(T^*_{\theta_j^{\mathrm{L}}}(G), T^*_{\theta_j^{\mathrm{R}}}(G) \right) \ge \frac{\delta_j (G - \lfloor rG \rfloor/2)}{\sqrt{2G}} \ge \sqrt{\frac{G}{2}}\, \delta_j (1 - r/2).$$

From this and by (A.1.4), at $\widetilde{\theta}_j = \arg\max_{k \in \mathscr{T}_j} T_k(G)$, we have

$$T_{\widetilde{\theta}_j}(G) \ge \max\left(T_{\theta_j^{\mathrm{L}}}(G), T_{\theta_j^{\mathrm{R}}}(G) \right) \ge \sqrt{\frac{G}{2}}\, \delta_j \left(1 - \frac{r}{2}\right) - \frac{24\sqrt{\mathfrak{s}}\lambda}{\omega} > \frac{1 - r/2}{2} \sqrt{\frac{G}{2}}\, \delta_j > D,$$

where the second last inequality follows from Assumption 3.5 (b), and the last one from (3.12). When $\eta = 1$, this and (A.1.5) indicates that such $\widetilde{\theta}_j$ satisfies (3.7). When $\eta < 1$, note that

$$\max\left(T_{\theta_j^{\mathrm{L}}}(G), T_{\theta^{\mathrm{R}}}(G) \right) - \max\left\{ T_k(G) : |k - \theta_j| > (1-\eta)G, k \in \mathscr{T} \right\}$$
$$\ge \sqrt{\frac{G}{2}}\, \delta_j \left(\eta - \frac{3r}{2} \right) - \frac{48\sqrt{\mathfrak{s}}\lambda}{\omega} \ge \frac{5\eta}{8\sqrt{2}} \min_{1 \le j \le q} \delta_j \sqrt{G} - \frac{48\sqrt{\mathfrak{s}}\lambda}{\omega} > 0$$

from (3.12). These arguments ensure that we detect at least one change point in $\mathscr{T}_j$ at $t = \widetilde{\theta}_j$ for each $j = 1, \ldots, q$. For such $\widetilde{\theta}_j$, suppose that $\theta_j^\circ = \arg\min_{k \in \{\theta_j^{\mathrm{L}}, \theta_j^{\mathrm{R}}\}} |\widetilde{\theta}_j - k|$. Then,

$$\frac{\delta_j}{\sqrt{2G}} (G - |\widetilde{\theta}_j - \theta_j|) + \frac{24\sqrt{\mathfrak{s}}\lambda}{\omega} \ge T_{\widetilde{\theta}_j}(G) \ge T_{\theta_j^\circ}(G) \ge \frac{\delta_j}{\sqrt{2G}} (G - |\theta_j^\circ - \theta_j|) - \frac{24\sqrt{\mathfrak{s}}\lambda}{\omega}$$

and re-arranging, we obtain

$$\frac{\delta_j}{\sqrt{2G}}\left(|\widetilde{\theta}_j-\theta_j|-|\theta_j^\circ-\theta_j|\right)\leq\frac{48\sqrt{\mathfrak{s}}\lambda}{\omega},\text{ such that }|\widetilde{\theta}_j-\theta_j|\leq\frac{48\sqrt{2\mathfrak{s}G}\lambda}{\omega\delta_j}+\lfloor rG\rfloor<\left\lfloor\frac{G}{2}\right\rfloor,$$

for large enough C_1 in Assumption 3.5 (b).

Finally, let $\mathscr{L}_{\mathscr{T}}(t)$ denote the largest time point $k'\in\mathscr{T}$ that satisfies $k'\leq t$, and define $\mathscr{R}_{\mathscr{T}}(t)$ analogously. Then, we establish that

$$T_{\mathscr{L}_{\mathscr{T}}(\theta_j-\frac{\eta G}{2}m)}(G)>\max\left\{T_k(G):\frac{\eta G}{2}(m+1)\leq\theta_j-k\leq\frac{\eta G}{2}(m+2),k\in\mathscr{T}\right\},\tag{A.1.7}$$

$$T_{\mathscr{R}_{\mathscr{T}}(\theta_j+\frac{\eta G}{2}m)}(G)>\max\left\{T_k(G):\frac{\eta G}{2}(m+1)\leq k-\theta_j\leq\frac{\eta G}{2}(m+2),k\in\mathscr{T}\right\},\tag{A.1.8}$$

for $m=0,\ldots,\lceil 2/\eta\rceil-2$. The inequality in (A.1.7) follows from noting that

$$\begin{aligned}&T_{\mathscr{L}_{\mathscr{T}}(\theta_j-\frac{\eta G}{2}m)}(G)-\max\left\{T_k(G):\frac{\eta G}{2}(m+1)\leq\theta_j-k\leq\frac{\eta G}{2}(m+2),k\in\mathscr{T}\right\}\\&\geq\sqrt{\frac{G}{2}}\delta_j\left(\frac{\eta}{2}-r\right)-\frac{48\sqrt{\mathfrak{s}}\lambda}{\omega}\geq\frac{\eta}{4\sqrt{2}}\min_{1\leq j\leq q}\delta_j\sqrt{G}-\frac{48\sqrt{\mathfrak{s}}\lambda}{\omega}>0\end{aligned}$$

under (3.12), and the inequality in (A.1.8) follows analogously. This ensures that $\widetilde{\theta}_j$ by its construction is the unique local maximiser of $T_k(G)$ within the interval $\{\theta_j-G+1,\ldots,\theta_j+G\}\cap\mathscr{T}$ satisfying (3.7) for each $j=1,\ldots,q$, which completes the proof.

A.1.1.3 Proof of Theorem 3.1 (ii)

Recalling (3.8), we write

$$Q_j(k)=\sum_{t=\widetilde{\theta}_j-G+1}^{k}(Y_t-\mathbf{x}_t^\top\widehat{\boldsymbol{\beta}}_j^{\mathrm{L}})^2+\sum_{t=k+1}^{\widetilde{\theta}_j+G}(Y_t-\mathbf{x}_t^\top\widehat{\boldsymbol{\beta}}_j^{\mathrm{R}})^2.$$

Theorem 3.1 (i) establishes that for each $j=1,\ldots,q$, we have $\widetilde{\theta}_j\in\widetilde{\Theta}$ that satisfies $|\widetilde{\theta}_j-\theta_j|<G/2$, and $\widetilde{\Theta}$ contains no other estimator. Then under Assumption 3.5 (a), we have the following statements satisfied for all j.

(i) Defining $\mathscr{I}(\widetilde{\theta}_j)=\{\widetilde{\theta}_j-G+1,\ldots,\widetilde{\theta}_j+G\}$, it fulfils $\mathscr{I}(\widetilde{\theta}_j)\cap\Theta=\{\theta_j\}$.

(ii) $\{\widetilde{\theta}_j^{\mathrm{L}}-G+1,\ldots,\widetilde{\theta}_j^{\mathrm{L}}\}\subset\{\theta_{j-1}+1,\ldots,\theta_j\}$ and $\{\widetilde{\theta}_j^{\mathrm{R}}+1,\ldots,\widetilde{\theta}_j^{\mathrm{R}}+G\}\subset\{\theta_j+1,\ldots,\theta_{j+1}\}$, such that denoting by $\boldsymbol{\Delta}_j^{\mathrm{L}}=\widehat{\boldsymbol{\beta}}_j^{\mathrm{L}}-\boldsymbol{\beta}_{j-1}$ and $\boldsymbol{\Delta}_j^{\mathrm{R}}=\widehat{\boldsymbol{\beta}}_j^{\mathrm{R}}-\boldsymbol{\beta}_j$, we have

$$\begin{aligned}&\max\left(\left|\boldsymbol{\Delta}_j^{\mathrm{L}}\right|_2,\left|\boldsymbol{\Delta}_j^{\mathrm{R}}\right|_2\right)\leq\frac{12\sqrt{2\mathfrak{s}}\lambda}{\omega\sqrt{G}},\\&\left|\boldsymbol{\Delta}_j^{\mathrm{L}}(\mathscr{S}_{j-1}^c)\right|_1\leq3\left|\boldsymbol{\Delta}_j^{\mathrm{L}}(\mathscr{S}_{j-1})\right|_1\text{ and }\left|\boldsymbol{\Delta}_j^{\mathrm{R}}(\mathscr{S}_j^c)\right|_1\leq3\left|\boldsymbol{\Delta}_j^{\mathrm{R}}(\mathscr{S}_j)\right|_1\end{aligned}\tag{A.1.9}$$

in $\mathscr{B}$, see Lemma A.1.2.

Then we show that for all $k \in \mathscr{I}(\widetilde{\theta}_j)$ satisfying $\delta_j^2|k-\theta_j| > v_{n,p}$ with

$$v_{n,p} = \max\left(\mathfrak{s}\rho_{n,p}^2, (\mathfrak{s}\log(p))^{\frac{1}{1-\tau}}\right) \cdot \max\left\{C_\delta^2 \max\left[\frac{9C_{\mathrm{RSC}}}{2\omega}, \frac{32C_{\mathrm{RSC}}}{\bar{\omega}}\right]^{\frac{1}{1-\tau}}, \left(\frac{96C_{\mathrm{DEV}}}{\omega}\right)^2\right\}, \tag{A.1.10}$$

we have $Q_j(k) - Q_j(\theta_j) > 0$, which completes the proof.

First, suppose that $k \geq \theta_j + 1$. Then,

$$\begin{aligned}
Q_j(k) - Q_j(\theta_j) &= \sum_{t=\theta_j+1}^{k} \left[(Y_t - \mathbf{x}_t^\top \widehat{\boldsymbol{\beta}}_j^{\mathrm{L}})^2 - (Y_t - \mathbf{x}_t^\top \widehat{\boldsymbol{\beta}}_j^{\mathrm{R}})^2\right] \\
&= \sum_{t=\theta_j+1}^{k} (\boldsymbol{\beta}_j - \widehat{\boldsymbol{\beta}}_j^{\mathrm{L}})^\top \mathbf{x}_t \mathbf{x}_t^\top (\boldsymbol{\beta}_j - \widehat{\boldsymbol{\beta}}_j^{\mathrm{L}}) - \sum_{t=\theta_j+1}^{k} (\widehat{\boldsymbol{\beta}}_j^{\mathrm{R}} - \boldsymbol{\beta}_j)^\top \mathbf{x}_t \mathbf{x}_t^\top (\widehat{\boldsymbol{\beta}}_j^{\mathrm{R}} - \boldsymbol{\beta}_j) \\
&\quad + 2\sum_{t=\theta_j+1}^{k} \varepsilon_t \mathbf{x}_t^\top \left[(\boldsymbol{\beta}_j - \boldsymbol{\beta}_{j-1}) + (\widehat{\boldsymbol{\beta}}_j^{\mathrm{R}} - \boldsymbol{\beta}_j) - (\widehat{\boldsymbol{\beta}}_j^{\mathrm{L}} - \boldsymbol{\beta}_{j-1})\right] = I_1 + I_2 + I_3.
\end{aligned}$$

From the definition of $\mathfrak{s}$ and the Cauchy-Schwarz inequality,

$$|\boldsymbol{\beta}_j - \boldsymbol{\beta}_{j-1}|_1 \leq \sqrt{2\mathfrak{s}}|\boldsymbol{\beta}_j - \boldsymbol{\beta}_{j-1}|_2 \tag{A.1.11}$$

and from (A.1.9), we have

$$|\boldsymbol{\Delta}_j^{\mathrm{L}}|_1 \leq 4|\boldsymbol{\Delta}_j^{\mathrm{L}}(\mathscr{S}_j)|_1 \leq 4\sqrt{2\mathfrak{s}}|\boldsymbol{\Delta}_j^{\mathrm{L}}|_2 \text{ and analogously, } |\boldsymbol{\Delta}_j^{\mathrm{R}}|_1 \leq 4\sqrt{2\mathfrak{s}}|\boldsymbol{\Delta}_j^{\mathrm{R}}|_2. \tag{A.1.12}$$

From (A.1.11)–(A.1.12), we derive

$$\begin{aligned}
&\left|\widehat{\boldsymbol{\beta}}_j^{\mathrm{L}} - \boldsymbol{\beta}_j\right|_2 \leq \delta_j\left(1 + \frac{12\sqrt{2\mathfrak{s}}\lambda}{\omega\delta_j\sqrt{G}}\right) \leq \frac{3\delta_j}{2} \text{ and similarly, } \left|\widehat{\boldsymbol{\beta}}_j^{\mathrm{L}} - \boldsymbol{\beta}_j\right|_2 \geq \frac{\delta_j}{2}, \\
&\left|\widehat{\boldsymbol{\beta}}_j^{\mathrm{L}} - \boldsymbol{\beta}_j\right|_1 \leq \sqrt{\mathfrak{s}}\delta_j\left(1 + \frac{96\sqrt{\mathfrak{s}}\lambda}{\omega\delta_j\sqrt{G}}\right) \leq \frac{3\sqrt{\mathfrak{s}}\delta_j}{2},
\end{aligned}$$

for a large enough C_1 in Assumption 3.5 (b). Then on $\mathscr{R}^{(1)}$, we have

$$I_1 \geq |k-\theta_j|\omega\delta_j^2\left(\frac{1}{4} - \frac{9C_{\mathrm{RSC}}\mathfrak{s}\log(p)}{4|k-\theta_j|^{1-\tau}\omega}\right) \geq \frac{\omega}{8}\delta_j^2|k-\theta_j| \tag{A.1.13}$$

from that $|k-\theta_j| > \delta_j^{-2}v_{n,p} \geq C_\delta^{-2}v_{n,p}$ (from Assumption 3.4) and (A.1.10). As for I_2, from Lemma A.1.2, (A.1.10) and (A.1.12) we have on $\mathscr{R}^{(2)}$,

$$|I_2| \leq \left|\boldsymbol{\Delta}_j^{\mathrm{R}}\right|_2^2\left[|k-\theta_j|\bar{\omega} + 32C_{\mathrm{RSC}}\mathfrak{s}\log(p)|k-\theta_j|^\tau\right] \leq 2|k-\theta_j|\bar{\omega}\left|\boldsymbol{\Delta}_j^{\mathrm{R}}\right|_2^2 \leq \frac{576\bar{\omega}\mathfrak{s}|k-\theta_j|\lambda^2}{\omega^2 G}. \tag{A.1.14}$$

Turning our attention to I_3, from (A.1.11)–(A.1.12),

$$\begin{aligned}
&\left|(\boldsymbol{\beta}_j - \boldsymbol{\beta}_{j-1}) + (\widehat{\boldsymbol{\beta}}_j^{\mathrm{R}} - \boldsymbol{\beta}_j) - (\widehat{\boldsymbol{\beta}}_j^{\mathrm{L}} - \boldsymbol{\beta}_{j-1})\right|_1 \leq \left|\boldsymbol{\beta}_j - \boldsymbol{\beta}_{j-1}\right|_1 + \left|\widehat{\boldsymbol{\beta}}_j^{\mathrm{R}} - \boldsymbol{\beta}_j\right|_1 + \left|\widehat{\boldsymbol{\beta}}_j^{\mathrm{L}} - \boldsymbol{\beta}_{j-1}\right|_1 \\
&\leq \sqrt{\mathfrak{s}}\delta_j\left(1 + \frac{192\sqrt{\mathfrak{s}}\lambda}{\omega\delta_j\sqrt{G}}\right) \leq 2\sqrt{\mathfrak{s}}\delta_j,
\end{aligned}$$

where the last inequality follows from Assumption 3.5 (b). Then on $\mathscr{D}^{(1)}$,

$$\frac{1}{2}|I_3| \le \left|\sum_{t=\theta_j+1}^{k} \varepsilon_t \mathbf{x}_t^\top\right|_\infty \left|(\boldsymbol{\beta}_j - \boldsymbol{\beta}_{j-1}) + (\widehat{\boldsymbol{\beta}}_j^{\mathrm{R}} - \boldsymbol{\beta}_j) - (\widehat{\boldsymbol{\beta}}_j^{\mathrm{L}} - \boldsymbol{\beta}_{j-1})\right|_1$$
$$\le 2C_{\mathrm{DEV}}\delta_j\sqrt{\mathfrak{s}(k-\theta_j)}\,\rho_{n,p}. \tag{A.1.15}$$

Then from (A.1.13), (A.1.14) and (A.1.15), we derive

$$\frac{|I_2|}{I_1} = \frac{4608\bar{\omega}\mathfrak{s}\lambda^2}{\omega^3\delta_j^2 G} \le \frac{1}{3} \quad \text{and} \quad \frac{|I_3|}{I_1} = \frac{32C_{\mathrm{DEV}}\sqrt{\mathfrak{s}}\,\rho_{n,p}}{\omega\delta_j\sqrt{k-\theta_j}} \le \frac{1}{3}$$

under Assumption 3.5 (b), for all $k \in \mathscr{I}_j$ satisfying $\delta_j^2|k-\theta_j| > v_{n,p}$ from (A.1.10). Analogous arguments apply when $k \le \theta_j$, and the above arguments are deterministic on $\mathscr{M}$. In summary, we have

$$\min_{1\le j\le q} \min_{\substack{k\in\mathscr{I}_j \\ \delta_j^2|k-\theta_j|>v_{n,p}}} \big(Q_j(k) - Q_j(\theta_j)\big) > \frac{\omega}{24}v_{n,p} > 0,$$

which concludes the proof.

A.1.2 Proof of Proposition 3.2

A.1.2.1 Supporting lemmas

Define $\mathbb{K}(b) = \mathbb{B}_0(b) \cap \mathbb{B}_2(1)$ with some $b \ge 1$, where $\mathbb{B}_d(r) = \{\mathbf{a} : |\mathbf{a}|_d \le r\}$ with the dimension of $\mathbf{a}$ determined within the context. Let $\mathbf{e}_i$ denote a vector that contains zeros except for its ith component set to be one. We denote the time-varying vector of parameters under (3.1) by $\boldsymbol{\beta}(t) = \sum_{j=1}^{q+1} \boldsymbol{\beta}_j \mathbb{1}_{\{\theta_{j-1}+1\le t\le\theta_j\}}$.

Denote by $\mathbf{Z}_t = (\mathbf{x}_t^\top, \varepsilon_t)^\top \in \mathbb{R}^{p+1}$ which admits $\mathbf{Z}_t = \sum_{\ell=0}^{\infty}\mathbf{D}_\ell\boldsymbol{\xi}_{t-\ell}$ under (3.13). For some $\mathbf{a},\mathbf{b} \in \mathbb{B}_2(1)$, define $U_t(\mathbf{a}) = \mathbf{a}^\top\mathbf{Z}_t$ and $W_t(\mathbf{a},\mathbf{b}) = \mathbf{a}^\top\mathbf{Z}_t\mathbf{Z}_t^\top\mathbf{b}$. Let $\boldsymbol{\xi}_t'$ denote an independent copy of $\boldsymbol{\xi}_t$, and define $\mathbf{Z}_{t,\{0\}} = \sum_{\ell=0,\ell\ne t}^{\infty}\mathbf{D}_\ell\boldsymbol{\xi}_{t-\ell} + \mathbf{D}_t\boldsymbol{\xi}_0'$. We denote the functional dependence measure and the dependence-adjusted norm for $U_t(\mathbf{a})$ as defined in Zhang and Wu (2017), by

$$\delta_{t,\nu}(\mathbf{a}) = \left\|\mathbf{a}^\top\mathbf{Z}_t - \mathbf{a}^\top\mathbf{Z}_{t,\{0\}}\right\|_\nu \quad \text{and} \quad |||U_\cdot(\mathbf{a})|||_\nu = \sum_{t=0}^{\infty}\delta_{t,\nu}(\mathbf{a}),$$

respectively. Analogously, we define

$$\delta_{t,\nu}(\mathbf{a},\mathbf{b}) = \left\|\mathbf{a}^\top\mathbf{Z}_t\mathbf{Z}_t^\top\mathbf{b} - \mathbf{a}^\top\mathbf{Z}_{t,\{0\}}\mathbf{Z}_{t,\{0\}}^\top\mathbf{b}\right\|_\nu \quad \text{and} \quad |||W_\cdot(\mathbf{a},\mathbf{b})|||_\nu = \sum_{t=0}^{\infty}\delta_{t,\nu}(\mathbf{a},\mathbf{b})$$

for $W_t(\mathbf{a},\mathbf{b})$. Finally, for some $\kappa \ge 0$, we denote the dependence adjusted sub-exponential norm of $W_t(\mathbf{a},\mathbf{b})$ by $\|W_\cdot(\mathbf{a},\mathbf{b})\|_{\psi_\kappa} = \sup_{\nu\ge2}\nu^{-\kappa}|||W_\cdot(\mathbf{a},\mathbf{b})|||_\nu$. In what follows, we denote by C_Π with $\Pi \subset \{\gamma,\nu,\Xi,\varsigma\}$ a constant that depends on the parameters included in Π which may vary from one occasion to another.

Lemma A.1.3. Suppose that Condition 1 holds.

(i) Under Condition 1 (a), we have $\sup_{\mathbf{a},\mathbf{b}\in\mathbb{B}_2(1)}\|W_\cdot(\mathbf{a},\mathbf{b})\|_{\psi_\kappa}\le C_{\gamma,\Xi,\varsigma}C_\xi^2<\infty$ with $\kappa=2\gamma+1$.

(ii) Under Condition 1 (b), we have $\sup_{\mathbf{a}\in\mathbb{B}_2(1)}|||U_\cdot(\mathbf{a})|||_2\le C_{\Xi,\varsigma}$.

Proof. In what follows, we denote by $\mu_\nu=\|\xi_{it}\|_\nu$. For given $\nu>1$, we have

$$\sup_{\mathbf{a}\in\mathbb{B}_2(1)}\delta_{t,\nu}(\mathbf{a})=\left\|\mathbf{a}^\top\mathbf{D}_t(\boldsymbol{\xi}_0-\boldsymbol{\xi}_0')\right\|_\nu\le C_\nu\mu_\nu\sqrt{2\sup_{\mathbf{a}\in\mathbb{B}_2(1)}\left|\mathbf{a}^\top\mathbf{D}_t\right|_2^2}\le C_\nu\mu_\nu\Xi(1+t)^{-\varsigma}\tag{A.1.16}$$

with $C_\nu=\max(1/(\nu-1),\sqrt{\nu-1})$, where the first inequality follows from Lemma 2 of Chen et al. (2021) (Burkholder's inequality) and Minkowski's inequality, and the second inequality from Condition 1 and from that $\|\mathbf{D}_t\|_2\le\sqrt{\|\mathbf{D}_t\|_1\|\mathbf{D}_t\|_\infty}$ (with $\|\cdot\|_a$ denoting the induced matrix norms). Therefore, under Condition 1 (b),

$$\sup_{\mathbf{a}\in\mathbb{B}_2(1)}|||U_\cdot(\mathbf{a})|||_2\le\Xi\sum_{t=0}^\infty(1+t)^{-\varsigma}\le C_{\Xi,\varsigma},$$

which proves (ii). Note that by Hölder and Minkowski's inequalities,

$$\begin{aligned}\delta_{t,\nu}(\mathbf{a},\mathbf{b})\le&\left\|\sum_{\ell=0}^\infty\mathbf{a}^\top\mathbf{D}_\ell\boldsymbol{\xi}_{t-\ell}\right\|_{2\nu}\left\|\mathbf{b}^\top\mathbf{D}_t(\boldsymbol{\xi}_0-\boldsymbol{\xi}_0')\right\|_{2\nu}\\&+\left\|\sum_{\ell=0,\ell\neq t}^\infty\mathbf{b}^\top\mathbf{D}_\ell\boldsymbol{\xi}_{t-\ell}+\mathbf{b}^\top\mathbf{D}_t\boldsymbol{\xi}_0'\right\|_{2\nu}\left\|\mathbf{a}^\top\mathbf{D}_t(\boldsymbol{\xi}_0-\boldsymbol{\xi}_0')\right\|_{2\nu}.\end{aligned}$$

For given $\nu>2$, similarly as in (A.1.16), we can show that

$$\begin{aligned}&\sup_{\mathbf{a}\in\mathbb{B}_2(1)}\left\|\sum_{\ell=0}^\infty\mathbf{a}^\top\mathbf{D}_\ell\boldsymbol{\xi}_{t-\ell}\right\|_{2\nu}\le\sum_{\ell=0}^\infty\sup_{\mathbf{a}\in\mathbb{B}_2(1)}\left\|\mathbf{a}^\top\mathbf{D}_\ell\boldsymbol{\xi}_{t-\ell}\right\|_{2\nu}\\&\le C_{2\nu}\mu_{2\nu}\sum_{\ell=0}^\infty\sqrt{\sup_{\mathbf{a}\in\mathbb{B}_2(1)}\left|\mathbf{a}^\top\mathbf{D}_\ell\right|_2^2}\le C_{2\nu}\mu_{2\nu}\sum_{\ell=0}^\infty\Xi(1+\ell)^{-\varsigma}\le C_{\gamma,\Xi,\varsigma}C_\xi\nu^{\gamma+1/2}\end{aligned}\tag{A.1.17}$$

under Condition 1 (a). Then, (A.1.16)–(A.1.17) lead to

$$\begin{aligned}&\sup_{\mathbf{a},\mathbf{b}\in\mathbb{B}_2(1)}\delta_{t,\nu}(\mathbf{a},\mathbf{b})\le C_{\gamma,\Xi,\varsigma}C_\xi^2\nu^{2\gamma+1}(1+t)^{-\varsigma},\quad\text{and}\\&\sup_{\mathbf{a},\mathbf{b}\in\mathbb{B}_2(1)}|||W_\cdot(\mathbf{a},\mathbf{b})|||_\nu\le C_{\gamma,\Xi,\varsigma}C_\xi^2\nu^{2\gamma+1}\sum_{t=0}^\infty(1+t)^{-\varsigma}\le C_{\gamma,\Xi,\varsigma}C_\xi^2\nu^{2\gamma+1},\end{aligned}$$

such that we have $\sup_{\mathbf{a},\mathbf{b}\in\mathbb{B}_2(1)}\|W_\cdot(\mathbf{a},\mathbf{b})\|_{\psi_\kappa}\le C_{\gamma,\Xi,\varsigma}C_\xi^2$ with $\kappa=2\gamma+1$, which proves (i). ■

Lemma A.1.4. Under Condition 1 (a), there exist fixed constants $C',C''>0$ such that for all $0\le s<e\le n$ and $z>0$, we have

$$\sup_{\mathbf{a},\mathbf{b}\in\mathbb{B}_2(1)}\mathsf{P}\left(\frac{1}{\sqrt{e-s}}\left|\sum_{t=s+1}^e\mathbf{a}^\top\mathbf{Z}_t\mathbf{Z}_t^\top\mathbf{b}-\mathbb{E}\left(\sum_{t=s+1}^e\mathbf{a}^\top\mathbf{Z}_t\mathbf{Z}_t^\top\mathbf{b}\right)\right|\ge z\right)\le C'\exp\left(-C''z^{\frac{2}{4\gamma+3}}\right).$$

Proof. By Lemma A.1.3 (i) and Lemma C.4 of Zhang and Wu (2017), there exist constants $C', C'' > 0$ that depend on γ, Ξ, ς and C_ξ, such that for all $z > 0$,

$$\sup_{\mathbf{a},\mathbf{b}\in\mathbb{B}_2(1)} \mathsf{P}\left(\frac{1}{\sqrt{e-s}}\left|\sum_{t=s+1}^{e}\mathbf{a}^\top\mathbf{Z}_t\mathbf{Z}_t^\top\mathbf{b} - \mathbb{E}\left(\sum_{t=s+1}^{e}\mathbf{a}^\top\mathbf{Z}_t\mathbf{Z}_t^\top\mathbf{b}\right)\right| \geq z\right)$$
$$\leq C'\exp\left(-\frac{(4\gamma+3)z^{\frac{2}{4\gamma+3}}}{4e(C_{\gamma,\Xi,\varsigma}C_\xi^2)^{\frac{2}{4\gamma+3}}}\right) \leq C'\exp\left(-C''z^{\frac{2}{4\gamma+3}}\right).$$

■

Lemma A.1.5. Under Condition 1 (b), there exists a fixed constants $C''' > 0$ such that for all $0 \leq s < e \leq n$ and $0 < z < C_{\Xi,\varsigma}^2\sqrt{e-s}$, we have

$$\sup_{\mathbf{a},\mathbf{b}\in\mathbb{B}_2(1)} \mathsf{P}\left(\frac{1}{\sqrt{e-s}}\left|\sum_{t=s+1}^{e}\mathbf{a}^\top\mathbf{Z}_t\mathbf{Z}_t^\top\mathbf{b} - \mathbb{E}\left(\sum_{t=s+1}^{e}\mathbf{a}^\top\mathbf{Z}_t\mathbf{Z}_t^\top\mathbf{b}\right)\right| \geq z\right) \leq 6\exp(-C'''z^2).$$

Proof. By Lemma A.1.3 (ii) and Theorem 6.6 of Zhang and Wu (2021), there exists an absolute constant $C > 0$ such that for all $0 < z < C_{\Xi,\varsigma}^2\sqrt{e-s}$,

$$\sup_{\mathbf{a}\in\mathbb{B}_2(1)} \mathsf{P}\left(\frac{1}{\sqrt{e-s}}\left|\sum_{t=s+1}^{e}\mathbf{a}^\top\mathbf{Z}_t\mathbf{Z}_t^\top\mathbf{a} - \mathbb{E}\left(\sum_{t=s+1}^{e}\mathbf{a}^\top\mathbf{Z}_t\mathbf{Z}_t^\top\mathbf{a}\right)\right| \geq z\right)$$
$$\leq 2\exp\left[-C\min\left(\frac{z^2}{C_{\Xi,\varsigma}^4}, \frac{z\sqrt{e-s}}{C_{\Xi,\varsigma}^2}\right)\right] \leq 2\exp(-CC_{\Xi,\varsigma}^{-4}z^2).$$

Then noting that

$$\sup_{\mathbf{a},\mathbf{b}\in\mathbb{B}_2(1)} \mathsf{P}\left(\frac{2}{\sqrt{e-s}}\left|\sum_{t=s+1}^{e}\mathbf{a}^\top\mathbf{Z}_t\mathbf{Z}_t^\top\mathbf{b} - \mathbb{E}\left(\sum_{t=s+1}^{e}\mathbf{a}^\top\mathbf{Z}_t\mathbf{Z}_t^\top\mathbf{b}\right)\right| \geq z\right) \leq$$
$$\sup_{\mathbf{a},\mathbf{b}\in\mathbb{B}_2(1)} \mathsf{P}\left(\frac{1}{\sqrt{e-s}}\left|\sum_{t=s+1}^{e}(\mathbf{a}+\mathbf{b})^\top\mathbf{Z}_t\mathbf{Z}_t^\top(\mathbf{a}+\mathbf{b}) - \mathbb{E}\left(\sum_{t=s+1}^{e}(\mathbf{a}+\mathbf{b})^\top\mathbf{Z}_t\mathbf{Z}_t^\top(\mathbf{a}+\mathbf{b})\right)\right| \geq \frac{z}{3}\right)$$
$$+2\sup_{\mathbf{a}\in\mathbb{B}_2(1)} \mathsf{P}\left(\frac{1}{\sqrt{e-s}}\left|\sum_{t=s+1}^{e}\mathbf{a}^\top\mathbf{Z}_t\mathbf{Z}_t^\top\mathbf{a} - \mathbb{E}\left(\sum_{t=s+1}^{e}\mathbf{a}^\top\mathbf{Z}_t\mathbf{Z}_t^\top\mathbf{a}\right)\right| \geq \frac{z}{3}\right) \leq 6\exp\left(-\frac{Cz^2}{9C_{\Xi,\varsigma}^4}\right),$$

we can find C''' that depends on Ξ and ς. ■

A.1.2.2 Proof of Proposition 3.2 (i)

Recalling C' from Lemma A.1.4, we set $c_1 = 3C'$.

Verification of Assumption 3.2:

By assumption, we have $\mathbb{E}(\mathbf{x}_t\varepsilon_t) = \mathbf{0}$. Then setting $\mathbf{a} = \mathbf{e}_i$, $i = 1,\ldots,p$, $\mathbf{b} = \mathbf{e}_{p+1}$ and $z = C_{\mathrm{DEV}}\log^{2\gamma+3/2}(p\vee n)$ in Lemma A.1.4,

$$\mathsf{P}(\mathscr{D}^{(1)}) \geq 1 - C'pn^2\exp\left(-C''C_{\mathrm{DEV}}^{\frac{2}{4\gamma+3}}\log(p\vee n)\right). \tag{A.1.18}$$

Next, by construction,

$$\sum_{t=s+1}^{e}(\boldsymbol{\beta}(t)-\boldsymbol{\beta}^*_{s,e})=\mathbf{0} \quad \text{and} \quad \max_{\substack{0\le s<e\le n\\ |\{s+1,\ldots,e\}\cap\Theta|\le 1}} \max_{s<t\le e}\left|\boldsymbol{\beta}(t)-\boldsymbol{\beta}^*_{s,e}\right|_2 \le C_\delta \tag{A.1.19}$$

under Assumption 3.4, and

$$\mathbb{E}\left[\sum_{t=s+1}^{e}\mathbf{x}_t\mathbf{x}_t^\top(\boldsymbol{\beta}(t)-\boldsymbol{\beta}^*_{s,e})\right]=\boldsymbol{\Sigma}_x\sum_{t=s+1}^{e}(\boldsymbol{\beta}(t)-\boldsymbol{\beta}^*_{s,e})=\mathbf{0} \tag{A.1.20}$$

under Assumption 3.1. Then setting $\mathbf{a}=\mathbf{e}_i$, $i=1,\ldots,p$, $\mathbf{b}=\boldsymbol{\beta}(t)-\boldsymbol{\beta}^*_{s,e}$ for given s,e and $t\in\{s+1,\ldots,e\}$ and $z=C_{\mathrm{DEV}}C_\delta\log^{2\gamma+3/2}(p\vee n)$ in Lemma A.1.4,

$$\mathsf{P}(\mathscr{D}^{(2)})\ge 1-C'pn^3\exp\left(-C''(C_{\mathrm{DEV}}C_\delta)^{\frac{2}{4\gamma+3}}\log(p\vee n)\right), \tag{A.1.21}$$

from (A.1.19) and (A.1.20). Combining (A.1.18) and (A.1.21), we can find large enough C_{DEV} that depends only on C'', γ, C_δ and c_2 such that $\mathsf{P}(\mathscr{D}^{(1)}\cap\mathscr{D}^{(2)})\ge 1-2c_1(p\vee n)^{-c_2}/3$.

Verification of Assumption 3.3:

Let $b_{s,e}$ denote an integer that depends on $(e-s)$ for some $0\le s<e\le n$, and define

$$\mathscr{R}=\left\{\sup_{\mathbf{a}\in\mathbb{K}(2b_{s,e})}\frac{1}{e-s}\left|\sum_{t=s+1}^{e}\mathbf{a}^\top\left(\mathbf{x}_t\mathbf{x}_t^\top-\boldsymbol{\Sigma}_x\right)\mathbf{a}\right|\ge\frac{\Lambda_{\min}(\boldsymbol{\Sigma}_x)}{54}\text{ for all }0\le s<e\le n\right.$$
$$\left.\text{with } e-s\ge C_0\log^{4\gamma+3}(p\vee n)\right\}.$$

By Lemma A.1.4 and Lemma F.2 of Basu and Michailidis (2015), we have

$$\begin{aligned}\mathsf{P}\left(\mathscr{R}^c\right)&\le\sum_{\substack{0\le s<e\le n\\ e-s\ge C_0\log^{4\gamma+3}(p\vee n)}}C'\exp\left[-C''\left(\frac{\sqrt{e-s}\,\Lambda_{\min}(\boldsymbol{\Sigma}_x)}{54}\right)^{\frac{2}{4\gamma+3}}+2b_{s,e}\log(p)\right]\\&\le C'n^2\exp\left[-\frac{C''}{2}\left(\frac{C_0^{1/2}\Lambda_{\min}(\boldsymbol{\Sigma}_x)}{54}\right)^{\frac{2}{4\gamma+3}}\log(p\vee n)\right],\end{aligned}$$

where the last inequality follows with

$$b_{s,e}=\left\lfloor\frac{C''}{4\log(p)}\left(\frac{\sqrt{e-s}\,\Lambda_{\min}(\boldsymbol{\Sigma}_x)}{54}\right)^{\frac{2}{4\gamma+3}}\right\rfloor,$$

which satisfies $b_{s,e}\ge 1$ for large enough C_0. Further, we can find C_0 that depends only on C'', $\Lambda_{\min}(\boldsymbol{\Sigma}_x)$, γ and c_2 which leads to $\mathsf{P}(\mathscr{R})\ge 1-c_1(p\vee n)^{-c_2}/3$. Then, by Lemma 12 of Loh and Wainwright (2012), on $\mathscr{R}$, we have

$$\begin{aligned}\sum_{t=s+1}^{e}\mathbf{a}^\top\mathbf{x}_t\mathbf{x}_t^\top\mathbf{a}\ge&\Lambda_{\min}(\boldsymbol{\Sigma}_x)(e-s)|\mathbf{a}|_2^2\\&-\frac{\Lambda_{\min}(\boldsymbol{\Sigma}_x)}{2}(e-s)\left(|\mathbf{a}|_2^2+\frac{4\log(p)}{C''}\left(\frac{54}{\sqrt{e-s}\,\Lambda_{\min}(\boldsymbol{\Sigma}_x)}\right)^{\frac{2}{4\gamma+3}}|\mathbf{a}|_1^2\right)\\\ge&\omega(e-s)|\mathbf{a}|_2^2-C_{\mathrm{RSC}}\log(p)(e-s)^{\frac{4\gamma+2}{4\gamma+3}}|\mathbf{a}|_1^2\end{aligned}$$

for all $\mathbf{a} \in \mathbb{R}^p$, with $\omega = \Lambda_{\min}(\mathbf{\Sigma}_x)/2$ and C_{RSC} depending only on C'', γ and $\Lambda_{\min}(\mathbf{\Sigma}_x)$. Analogously we have on $\mathscr{R}$,

$$\sum_{t=s+1}^{e} \mathbf{a}^\top \mathbf{x}_t \mathbf{x}_t^\top \mathbf{a} \le \bar{\omega}(e-s)|\mathbf{a}|_2^2 + C_{\mathrm{RSC}} \log(p)(e-s)^{\frac{4\gamma+2}{4\gamma+3}} |\mathbf{a}|_1^2$$

for all $\mathbf{a} \in \mathbb{R}^p$, with $\bar{\omega} = 3\Lambda_{\max}(\mathbf{\Sigma}_x)/2$.

Combining the arguments above, we have $\mathsf{P}(\mathscr{D}^{(1)} \cap \mathscr{D}^{(2)} \cap \mathscr{R}^{(1)} \cap \mathscr{R}^{(2)}) \ge 1 - c_1(p \vee n)^{-c_2}$, with $\tau = (4\gamma+2)/(4\gamma+3)$ and $\rho_{n,p} = \log^{2\gamma+3/2}(p \vee n)$.

A.1.2.3 Proof of Proposition 3.2 (ii)

We set $c_1 = 18$.

Verification of Assumption 3.2:

By assumption, we have $\mathbb{E}(\mathbf{x}_t \varepsilon_t) = \mathbf{0}$. Then setting $\mathbf{a} = \mathbf{e}_i$, $i = 1, \ldots, p$, $\mathbf{b} = \mathbf{e}_{p+1}$ and $z = C_{\mathrm{DEV}} \sqrt{\log(p \vee n)}$ in Lemma A.1.5,

$$\mathsf{P}(\mathscr{D}^{(1)}) \ge 1 - 6pn^2 \exp\left(-C''' C_{\mathrm{DEV}}^2 \log(p \vee n)\right), \tag{A.1.22}$$

provided that $C_0 > C_{\Xi,\varsigma}^{-4} C_{\mathrm{DEV}}^2$. Also, setting $\mathbf{a} = \mathbf{e}_i$, $i = 1, \ldots, p$, $\mathbf{b} = \boldsymbol{\beta}(t) - \boldsymbol{\beta}_{s,e}^*$ for given s, e and $t \in \{s+1, \ldots, e\}$ and $z = C_{\mathrm{DEV}} C_\delta \sqrt{\log(p \vee n)}$ in Lemma A.1.5,

$$\mathsf{P}(\mathscr{D}^{(2)}) \ge 1 - 6pn^3 \exp\left(-C''' C_{\mathrm{DEV}}^2 C_\delta^2 \log(p \vee n)\right), \tag{A.1.23}$$

from (A.1.19) and (A.1.20). Combining (A.1.22) and (A.1.23), we can find large enough C_{DEV} that depends only on C''', C_δ and c_2 such that $\mathsf{P}(\mathscr{D}^{(1)} \cap \mathscr{D}^{(2)}) \ge 1 - 2c_1(p \vee n)^{-c_2}/3$.

Verification of Assumption 3.3:

Let $b_{s,e}$ denote an integer that depends on $(e-s)$ for some $0 \le s < e \le n$, and define

$$\mathscr{R} = \left\{ \sup_{\mathbf{a} \in \mathbb{K}(2b_{s,e})} \frac{1}{e-s} \left| \sum_{t=s+1}^{e} \mathbf{a}^\top \left(\mathbf{x}_t \mathbf{x}_t^\top - \mathbf{\Sigma}_x \right) \mathbf{a} \right| \ge \frac{\Lambda_{\min}(\mathbf{\Sigma}_x)}{54} \text{ for all } 0 \le s < e \le n \right.$$
$$\left. \text{with } e - s \ge C_0 \log(p \vee n) \right\}.$$

Then by Lemma A.1.5 and Lemma F.2 of Basu and Michailidis (2015), we have

$$\begin{aligned} \mathsf{P}\left(\mathscr{R}^c\right) &\le \sum_{\substack{0 \le s < e \le n \\ e - s \ge C_0 \log(p \vee n)}} 6 \exp\left[-C'''(e-s) \left(\frac{\Lambda_{\min}(\mathbf{\Sigma}_x)}{54} \right)^2 + 2b_{s,e} \log(p) \right] \\ &\le 6n^2 \exp\left[-\frac{C''' C_0}{2} \left(\frac{\Lambda_{\min}(\mathbf{\Sigma}_x)}{54} \right)^2 \log(p \vee n) \right], \end{aligned}$$

where the last inequality follows with

$$b_{s,e} = \left\lfloor \frac{C'''(e-s)}{4 \log(p)} \left(\frac{\Lambda_{\min}(\mathbf{\Sigma}_x)}{54} \right)^2 \right\rfloor,$$

which satisfies $b_{s,e} \geq 1$ for large enough C_0. Further, we can find C_0 that depends only on C''', $\Lambda_{\min}(\mathbf{\Sigma}_x)$ and c_2 which leads to $\mathsf{P}(\mathscr{R}) \geq 1 - c_1(p \vee n)^{-c_2}/3$. Then, by Lemma 12 of Loh and Wainwright (2012), on $\mathscr{R}$, we have

$$\begin{aligned}\sum_{t=s+1}^{e} \mathbf{a}^\top \mathbf{x}_t \mathbf{x}_t^\top \mathbf{a} \geq & \Lambda_{\min}(\mathbf{\Sigma}_x)(e-s)|\mathbf{a}|_2^2 \\ & - \frac{\Lambda_{\min}(\mathbf{\Sigma}_x)}{2}(e-s)\left(|\mathbf{a}|_2^2 + \frac{4\log(p)}{C'''(e-s)}\left(\frac{54}{\Lambda_{\min}(\mathbf{\Sigma}_x)}\right)^2 |\mathbf{a}|_1^2\right) \\ \geq & \omega(e-s)|\mathbf{a}|_2^2 - C_{\mathrm{RSC}}\log(p)|\mathbf{a}|_1^2\end{aligned}$$

for all $\mathbf{a} \in \mathbb{R}^p$, with $\omega = \Lambda_{\min}(\mathbf{\Sigma}_x)/2$ and C_{RSC} depending only on C''' and $\Lambda_{\min}(\mathbf{\Sigma}_x)$. Analogously we have on $\mathscr{R}$,

$$\sum_{t=s+1}^{e} \mathbf{a}^\top \mathbf{x}_t \mathbf{x}_t^\top \mathbf{a} \leq \bar{\omega}(e-s)|\mathbf{a}|_2^2 + C_{\mathrm{RSC}}\log(p)|\mathbf{a}|_1^2$$

for all $\mathbf{a} \in \mathbb{R}^p$, with $\bar{\omega} = 3\Lambda_{\max}(\mathbf{\Sigma}_x)/2$.

Combining the arguments above, we have $\mathsf{P}(\mathscr{D}^{(1)} \cap \mathscr{D}^{(2)} \cap \mathscr{R}^{(1)} \cap \mathscr{R}^{(2)}) \geq 1 - c_1(p \vee n)^{-c_2}$, with $\tau = 0$ and $\rho_{n,p} = \sqrt{\log(p \vee n)}$.

A.1.3 Proof of Theorem 3.4

In what follows, we operate on $\mathscr{M} = \mathscr{D}^{(1)} \cap \mathscr{D}^{(2)} \cap \mathscr{R}^{(1)} \cap \mathscr{R}^{(2)} \cap \mathscr{B}$. Under Assumption 3.5′, we have all $G \in \mathscr{G}$ satisfy $G \geq C_0 \max\{\rho_{n,p}^2, (\omega^{-1}\mathfrak{s}\log(p))^{1/(1-\tau)}\}$ such that the lower bound on $(e-s)$ made in $\mathscr{B}$ (see Lemma A.1.2) is met by all $s = k$ and $e = k+G$, $k = 0, \ldots, n-G$.

For some k and $G \in \mathscr{G}$, we write $\mathscr{I}(k,G) = \{k-G+1, \ldots, k+G\}$. Recall that for each pre-estimator $\widetilde{\theta} \in \widetilde{\Theta}(G)$, we denote by $\mathscr{I}(\widetilde{\theta}) = \mathscr{I}(\widetilde{\theta}, G)$ its detection interval. By the same arguments adopted in (A.1.4) and Lemmas A.1.1 and A.1.2, we have

$$\max_{G \in \mathscr{G}} \max_{\substack{G \leq k \leq n-G \\ |\mathscr{I}(k,G) \cap \Theta| \leq 1}} \left|T_k(G) - T_k^*(G)\right| \leq \frac{24\sqrt{\mathfrak{s}}\lambda}{\omega} \quad \text{and} \quad T_k^*(G) = 0 \ \text{ if } \ \mathscr{I}(k,G) \cap \Theta = \emptyset. \tag{A.1.24}$$

Then, we make the following observations.

(i) From (A.1.24) and the requirement on D in (3.19), we have $\mathscr{I}(\widetilde{\theta}) \cap \Theta \neq \emptyset$ for all $\widetilde{\theta} \in \widetilde{\Theta}(\mathscr{G})$, i.e. each pre-estimator in $\widetilde{\Theta}(\mathscr{G})$ has (at least) one change point in its detection interval.

(ii) Under Assumption 3.5′, for each θ_j, $j = 1, \ldots, q$, there exists one pre-estimator $\widetilde{\theta} \in \widetilde{\Theta}(G_{(j)})$ such that $\mathscr{I}(\widetilde{\theta}) \cap \Theta = \{\theta_j\}$ and $|\widetilde{\theta} - \theta_j| < \lfloor G_{(j)}/2 \rfloor$, by the arguments used in the proof of Theorem 3.1 (i).

Thanks to (ii), there exists an anchor estimator $\widetilde{\theta}^A \in \widetilde{\Theta}^A$ for each θ_j, in the sense that $\theta_j \in \mathscr{I}(\widetilde{\theta}^A)$ and further, this anchor estimator $\widetilde{\theta}^A$ is detected with some bandwidth $G \leq G_{(j)}$. At the same time, there is at most a single anchor estimator $\widetilde{\theta}^A$ fulfilling $\theta_j \in \mathscr{I}(\widetilde{\theta}^A)$ by its construction

in (3.15), and (i) ensures that all anchor estimators contain one change point in its detection interval. Therefore, we have $\widehat{q} = |\widetilde{\Theta}^A| = q$ and we may write $\widetilde{\Theta}^A = \{\widetilde{\theta}_j^A, 1 \le j \le q : \widetilde{\theta}_1^A < \ldots < \widetilde{\theta}_q^A\}$.

Next, by (ii), there exists some $\widetilde{\theta} \in \widetilde{\Theta}(G_{(j)})$ fulfilling (3.16) for each $j = 1, \ldots, q$. To see this, note that if $\widetilde{\theta} \in \widetilde{\Theta}(G_{(j)})$ detects θ_j in the sense that $\theta_j \in \mathscr{I}(\widetilde{\theta})$,

$$\begin{aligned}
&\left\{\widetilde{\theta} - G_{(j)} - \left\lfloor \frac{G_{(j)}}{2} \right\rfloor + 1, \ldots, \widetilde{\theta} + G_{(j)} + \left\lfloor \frac{G_{(j)}}{2} \right\rfloor \right\} \subset \{\theta_j - 2G_{(j)} + 1, \theta_j + 2G_{(j)}\}, \text{ while} \\
&\mathscr{I}(\widetilde{\theta}_{j-1}^A) \subset \{\theta_{j-1} - 2G_{(j-1)} + 1, \ldots, \theta_{j-1} + 2G_{(j-1)}\} \text{ and} \\
&\mathscr{I}(\widetilde{\theta}_{j+1}^A) \subset \{\theta_{j+1} - 2G_{(j+1)} + 1, \ldots, \theta_{j+1} + 2G_{(j+1)}\},
\end{aligned}$$

and the sets on RHS do not overlap under Assumption 3.5′ (a). This in turn implies that we have $|\mathscr{C}_j| \ge 1$. Also for $\widetilde{\theta}_j^M \in \mathscr{C}_j$, we have that its detection bandwidth G_j^M satisfies

$$\frac{3}{2} G_j^M \le \min(\theta_{j+1} - \theta_j, \theta_j - \theta_{j-1}) \quad \text{and} \quad G_j^M \ge G_{(j)}$$

by the construction of $\mathscr{C}_j$. Also, the bandwidths generated as in Remark 3.2 satisfy

$$G_{\ell-1} + \frac{1}{2} G_{\ell-1} \le G_{\ell-1} + G_{\ell-2} = G_\ell \le 2G_{\ell-1}, \quad \text{such that} \quad \frac{1}{2} G_\ell \le G_{\ell-1} \le \frac{2}{3} G_\ell \text{ for } \ell \ge 2,$$

and therefore

$$\frac{1}{4} G_{(j)} \le G_j^* \quad \text{and} \quad G_j^* \le \left(\frac{3}{4} \cdot \frac{2}{3} + \frac{1}{4}\right) G_j^M \le \frac{1}{2} \min(\theta_{j+1} - \theta_j, \theta_j - \theta_{j-1}). \tag{A.1.25}$$

Further, by that $|\widetilde{\theta}_j^m - \theta_j| < G_j^m$ (see (i)) and

$$2G_j^m + G_j^* = \frac{11}{4} G_j^m + \frac{1}{4} G_j^M \le \frac{11}{4} G_{(j)} + \frac{1}{4} G_j^M \le \frac{41}{48} \min(\theta_{j+1} - \theta_j, \theta_j - \theta_{j-1}),$$

we have

$$\{\widetilde{\theta}_j^m - G_j^m - G_j^* + 1, \ldots, \widetilde{\theta}_j^m - G_j^m\} \cap \{\widetilde{\theta}_j^m + G_j^m + 1, \ldots, \widetilde{\theta}_j^m + G_j^m + G_j^*\} \cap \Theta = \emptyset. \tag{A.1.26}$$

From (A.1.25) and Assumption 3.5′ (b), we have

$$\delta_j^2 G_j^* \ge C_1 \max\left\{\omega^{-2} \mathfrak{s} \rho_{n,p}^2, \left(\omega^{-1} \mathfrak{s} \log(p)\right)^{1/(1-\tau)}\right\}$$

and from (A.1.26) and Lemma A.1.2, we have $\boldsymbol{\Delta}_j^{\mathrm{L}} = \widehat{\boldsymbol{\beta}}_j^{\mathrm{L}} - \boldsymbol{\beta}_{j-1}$ and $\boldsymbol{\Delta}_j^{\mathrm{R}} = \widehat{\boldsymbol{\beta}}_j^{\mathrm{R}} - \boldsymbol{\beta}_j$ satisfy

$$\begin{aligned}
&\max\left(\left|\boldsymbol{\Delta}_j^{\mathrm{L}}\right|_2, \left|\boldsymbol{\Delta}_j^{\mathrm{R}}\right|_2\right) \le \frac{12\sqrt{2\mathfrak{s}}\lambda}{\omega\sqrt{G_j^*}} \le \frac{24\sqrt{2\mathfrak{s}}\lambda}{\omega\sqrt{G_{(j)}}}, \\
&\left|\boldsymbol{\Delta}_j^{\mathrm{L}}(\mathscr{S}_{j-1}^c)\right|_1 \le 3\left|\boldsymbol{\Delta}_j^{\mathrm{L}}(\mathscr{S}_{j-1})\right|_1 \text{ and } \left|\boldsymbol{\Delta}_j^{\mathrm{R}}(\mathscr{S}_j^c)\right|_1 \le 3\left|\boldsymbol{\Delta}_j^{\mathrm{R}}(\mathscr{S}_j)\right|_1,
\end{aligned}$$

such that the arguments analogous to those employed in the proof of Theorem 3.1 (ii) are applicable to establish the localisation rate of $\check{\theta}_j$, which completes the proof.

A.2 Further information on the real dataset

Table A.2.1 lists the covariates included in the dataset analysed in Section 6.6.2.

Table A.2.1: Covariates contained in the equity premium dataset analysed in Section 6.6.2 (cf. Koo et al. (2020), Table 3)

Name	Description
d/p	Dividend price ratio: difference between the log of dividends and the log of prices
d/y	Dividend yield: difference between the log of dividends and the log of lagged prices
e/p	Earnings price ratio: difference between the log of earnings and the log of prices
d/e	Dividend payout ratio: difference between the log of dividends and the log of earnings
b/m	Book-to-market ratio: ratio of book value to market value for the Dow Jones Industrial Average
ntis	Net equity expansion: ratio of 12-month moving sums of net issues by NYSE listed stocks over the total end-of-year market capitalization of NYSE stocks
tbl	Treasury bill rates: 3-month Treasury bill rates
lty	Long-term yield: long-term government bond yield
tms	Term spread: difference between the long term bond yield and the Treasury bill rate
dfy	Default yield spread: difference between Moody's BAA and AAA-rated corporate bond yields
dfr	Default return spread: difference between the returns of long-term corporate and government bonds
svar	Log of stock varianceobtained as the sum of squared daily returns on S&P500 index
infl	Inflation: CPI inflation for all urban consumers
ltr	Long-term return: return of long term government bonds

APPENDIX B

APPENDIX TO MOVING SUM DATA SEGMENTATION FOR VECTOR AUTOREGRESSIVE TIME SERIES

B.1 MOSUM Wald procedure

Here we propose an alternative methodology using the Wald detector

$$\widehat{W}(G) = \max_{G \le k \le n-G} \widehat{W}_k(G), \quad \widehat{W}_k(G) = \sqrt{\frac{G}{2}} \left\| \widehat{\mathbf{\Gamma}}_k^{-1/2} \left(\widehat{\boldsymbol{a}}_{k+1,k+G} - \widehat{\boldsymbol{a}}_{k-G+1,k} \right) \right\|, \tag{B.1.1}$$

which has the population counterpart

$$W(G) = \max_{G \le k \le n-G} W_k(G), \quad W_k(G) = \sqrt{\frac{G}{2}} \left\| \mathbf{\Gamma}_k^{-1/2} \left(\widehat{\boldsymbol{a}}_{k+1,k+G} - \widehat{\boldsymbol{a}}_{k-G+1,k} \right) \right\|.$$

The parameter vector $\widehat{\boldsymbol{a}}_{s,e}$ solves (4.6). The matrix $\widehat{\mathbf{\Gamma}}_k$ estimates

$$\mathbf{\Gamma}_k = \boldsymbol{V}_{(j)}^{-1} \mathbf{\Sigma}_{(j)} \boldsymbol{V}_{(j)}^{-\top}, \text{ for } k_{j-1} + 1 \le k \le k_j, \tag{B.1.2}$$

where $\mathbf{\Sigma}_{(j)} = \mathbf{\Sigma}_{(j)}(\boldsymbol{a}_j)$ and $\boldsymbol{V}_{(j)} = \mathbb{E}(\nabla \boldsymbol{H}(\boldsymbol{X}_t^{(j)}, \mathbb{X}_{t-1}^{(j)}, \boldsymbol{a}_j))$ is the expectation of the gradient of the estimating function. In contrast to the score detector in (4.3), the Wald detector in (B.1.1) compares the difference in the least squares parameter estimates from each G window either side of a candidate k. This is expensive relative to simply evaluating predictive scores from a single inspection parameter, and requires a greater window size to ensure reasonable estimation, but all changes are detectable in the sense of Remark 4.1 and this can achieve more accurate localisation in practice.

As with the MOSUM score procedure, to locate changes we identify all pairs of indices (v_j, w_j) such that for some $\epsilon \in (0, 1/2)$,

$$\begin{aligned} &\widehat{W}_k(G) > D(G, \alpha) \text{ for } v_j \le k \le w_j, \text{ and} \\ &\widehat{W}_k(G) \le D(G, \alpha) \text{ for } k = v_j - 1, w_j + 1, \end{aligned} \tag{B.1.3}$$

where $w_j - v_j \geq \epsilon G$, and $D(G,\alpha)$ is as in (4.11). We take the number of these pairs as an estimator for the number of changes:

$$\widehat{q} = \text{ number of pairs } \big(v_j, w_j\big),$$

and for each $j = 1,\ldots,\widehat{q}$, we use the local maximum between v_j and w_j, as a location estimator for k_j, i.e.

$$\widehat{k}_j = \arg\max_{v_j \leq k \leq w_j} \widehat{W}_k(G).$$

In practice we recommend to use the η-criterion discussed in Appendix B.4.

B.1.1 Estimation of $\boldsymbol{\Gamma}_k$

Next we propose an estimators for the MOSUM Wald procedure of the form $\widehat{\boldsymbol{\Gamma}}_k = \widehat{\boldsymbol{V}}_k^{-1}\widehat{\boldsymbol{\Sigma}}_k\widehat{\boldsymbol{V}}_k^{-\top}$. For $\boldsymbol{V}_k = \boldsymbol{V}_{(j)}, k_{j-1}+1 \leq k \leq k_j$, we propose

$$\widehat{\boldsymbol{V}}_k = \text{BlockDiagonal}_p(\widehat{\boldsymbol{C}}_{k-G+1,k+G}) = \begin{pmatrix} \widehat{\boldsymbol{C}}_{k-G+1,k+G} & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \boldsymbol{0} & \widehat{\boldsymbol{C}}_{k-G+1,k+G} & & \\ \vdots & & \ddots & \\ \boldsymbol{0} & \cdots & & \widehat{\boldsymbol{C}}_{k-G+1,k+G} \end{pmatrix}. \quad \text{(B.1.4)}$$

For $\boldsymbol{\Sigma}_k$, we propose the estimator

$$\widehat{\boldsymbol{\Sigma}}_k^{(3)} = \widehat{\boldsymbol{S}}_k \otimes \widehat{\boldsymbol{C}}_{k-G+1,k+G}, \quad \text{(B.1.5)}$$

where we use the estimator for $\boldsymbol{S}$

$$\widehat{\boldsymbol{S}}_k^{(W)} = \frac{1}{2G}\Bigg(\sum_{t=k-G+1}^{k} \widehat{\boldsymbol{\varepsilon}}_t(\widehat{\boldsymbol{a}}_{k-G+1,k})\widehat{\boldsymbol{\varepsilon}}_t^{\top}(\widehat{\boldsymbol{a}}_{k-G+1,k}) + \sum_{t=k+1}^{k+G} \widehat{\boldsymbol{\varepsilon}}_t(\widehat{\boldsymbol{a}}_{k+1,k+G})\widehat{\boldsymbol{\varepsilon}}_t^{\top}(\widehat{\boldsymbol{a}}_{k+1,k+G})\Bigg). \quad \text{(B.1.6)}$$

This uses residuals from both halves of the Wald detector window, averaging a covariance estimate on either side of the candidate k. By localising we prevent contamination between regimes, and as with (4.8) we gain detection power by overestimating the covariance around true change points.

As with (4.9) we can use the estimating function to define an estimator, the only difference here being that $\widehat{\boldsymbol{a}}_{s,e}$ is used for the regression parameters. We define

$$\widehat{\boldsymbol{\Sigma}}_k^{(4)} = \frac{1}{2G}\Bigg[\Bigg(\sum_{t=k-G+1}^{k} \big(\boldsymbol{H}_t(\widehat{\boldsymbol{a}}_{k-G+1,k}) - \bar{\boldsymbol{H}}_{k-G+1,k}(\widehat{\boldsymbol{a}}_{k-G+1,k})\big)\big(\boldsymbol{H}(\widehat{\boldsymbol{a}}_{k-G+1,k}) - \bar{\boldsymbol{H}}_{k-G+1,k}(\widehat{\boldsymbol{a}}_{k-G+1,k})\big)^{\top} + \sum_{t=k+1}^{k+G} \big(\boldsymbol{H}_t(\widehat{\boldsymbol{a}}_{k+1,k+G}) - \bar{\boldsymbol{H}}_{k+1,k+G}(\widehat{\boldsymbol{a}}_{k+1,k+G})\big)\big(\boldsymbol{H}_t(\widehat{\boldsymbol{a}}_{k+1,k+G}) - \bar{\boldsymbol{H}}_{k+1,k+G}(\widehat{\boldsymbol{a}}_{k+1,k+G})\big)^{\top}\Bigg)\Bigg]. \quad \text{(B.1.7)}$$

B.1.2 Extensions

B.1.2.1 Grid-based procedure

Over the grid (4.14) we define the grid-based Wald detector

$$\widehat{W}_{\mathcal{T}}(G) = \max_{k \in \mathcal{T}} \widehat{W}_k(G), \tag{B.1.8}$$

where $\widehat{W}_k(G)$ is as defined in (B.1.1). After scanning over the grid, intervals with detectors exceeding the threshold are filled in with score statistics, where the inspection parameter is calculated over each contiguous significant interval. To locate changes, we can use either of the ϵ- or η-criteria; our theoretical results are derived with the former but will hold with the latter by the same arguments as Lemma B.28.

MOSUM Wald grid-based Procedure

1. Identify all $k \in \mathcal{T}$ such that $\widehat{W}_k(G) \geq D(G, \alpha)$

2. Collect these k into contiguous intervals such that $k = s_i, s_i + 1, \ldots, e_i, i = 1 \ldots, Q$ and extend these outwards to define the sets

$$\mathcal{T}_i = \{s_i - G, s_i - G + 1, \ldots, e_i + G\}.$$

3. Compute the estimator $\widehat{\boldsymbol{a}}_{\mathcal{T}_i}$ as in (4.6) with $s = s_i - G$ and $e = e_i + G$.

4. For each i, calculate the statistic $\widehat{T}_k(G, \widehat{\boldsymbol{a}}_{\mathcal{T}_i})$ for each k in the set $\mathcal{T}_i$. For each k, assign T_k to be the pointwise maximum of $\widehat{T}_k(G, \widehat{\boldsymbol{a}}_{\mathcal{T}_i})$ over i.

5. Locate changes with the ϵ-criterion as in (4.5).

B.1.2.2 Multiscale method

We can use $\widehat{W}_k$ in place of $\widehat{T}_k$ in Algorithm 8 to obtain a multiscale extension to the MOSUM Wald procedure.

B.1.2.3 Threshold bootstrap

In the bootstrap procedure of Section 4.5.3, we can use the detector $\widehat{W}^m(G)$ as per (B.1.1), evaluated over the simulated dataset $\{\boldsymbol{X}_t^m\}_{t=1}^n$ as in step 2. Taking a quantile of these maxima over $m = 1, \ldots, M$ will give a threshold with better size control than $D(G, \alpha)$ as in (4.11).

B.1.3 Theoretical properties

We provide theoretical results supporting the MOSUM Wald procedure. Analogously to Assumption 4.7, Assumption B.1 requires estimators for $\mathbf{\Gamma}_k$ to be consistent away from changes, and the spectral norm to be finite near to changes.

Assumption B.1. The estimator $\widehat{\mathbf{\Gamma}}_k$ of the covariance matrix $\mathbf{\Gamma}_k$ is positive definite and satisfies

(a)
$$\max_{k_{j-1}+G\leq k\leq k_j-G}\left\|\widehat{\mathbf{\Gamma}}_k^{-1/2}-\mathbf{\Gamma}_k^{-1/2}\right\|_F=o_P\left(\log(n/G)^{-1}\right)$$
for any $j=1,\ldots,q+1$.

(b) For any $j=1,\ldots,q$ it holds that
$$\max_{k:|k-k_j|<(1-\varepsilon)G}\left\|\widehat{\mathbf{\Gamma}}_k^{1/2}\right\|_F<\infty,\text{ and }\max_{k:|k-k_j|<(1-\varepsilon)G}\left\|\widehat{\mathbf{\Gamma}}_k^{-1/2}\right\|_F<\infty.$$

We show in Proposition B.1 that under the null hypothesis, a transformation of the Wald statistic converges to a Gumbel distribution.

Proposition B.1. Let Assumptions 4.1–4.4 hold. Then, under H_0,

(a)
$$a(n/G)W(G)-b(n/G)\xrightarrow{\mathscr{D}}G_2,$$
where G_2, $a(x)$, and $b(x)$ are as in Proposition 4.1.

(b) The result of (a) holds with an estimator $\widehat{\mathbf{\Gamma}}_k$, with $\widehat{\mathbf{\Sigma}}_k$ as in (B.1.5) or (B.1.7), in place of the true $\mathbf{\Gamma}_k$.

With the critical value c_α, we identify the asymptotic distribution of the transformed statistics under H_0.

As a result of Proposition B.1, we have a testing procedure for $W(G)$ with asymptotic level α, where we reject H_0 if $W(G)$ exceeds $D(G,\alpha)$ as in (4.11).

Assumption 4.5′ translates Assumption 4.5 to the Wald setting, where the jump size does not depend on an inspection parameter. When $\boldsymbol{a}_{j+1}\neq\boldsymbol{a}_j$, we necessarily have that $\delta_j>0$.

Assumption 4.5′. For $j=1,\ldots,q$, define the jump size $\delta_j=\|\boldsymbol{d}_j\|$, where
$$\boldsymbol{d}_j=\mathbb{E}\left(\boldsymbol{H}(\boldsymbol{X}_t^{(j+1)},\mathbb{X}_{t-1}^{(j+1)},\boldsymbol{a}_{j+1})\right)-\mathbb{E}\left(\boldsymbol{H}(\boldsymbol{X}_t^{(j)},\mathbb{X}_{t-1}^{(j)},\boldsymbol{a}_j)\right).$$
As $n\to\infty$, we let $\min_{1\leq j\leq q}\delta_j\geq c_{\delta,n}>0$, where $c_{\delta,n}\cdot\sqrt{\frac{G}{\log(n/G)}}\to\infty$.

Theorem B.2 establishes that the MOSUM Wald procedure consistently estimates both the number and locations of detectable changes.

Theorem B.2 (MOSUM Wald procedure consistency)**.** Let Assumptions 4.1-4.4, 4.5′, and 4.6 hold. Using the MOSUM Wald procedure, we have that

(a)
$$\mathsf{P}\left(\widehat{q} = q; \max_{1\le j\le q} |\widehat{k}_j - k_j| < G\right) \to 1$$
as $n \to \infty$.

(b) Letting $w_n \to \infty$ and $0 < w_n < G \cdot \min_{j=1,\dots,q} \delta_j^2$, there exists some $\gamma > 2$ such that
$$\mathsf{P}\left(\max_{1\le j\le q} \left|\widehat{k}_j - k_j\right| \delta_j^2 > w_n\right) = O(w_n^{-\gamma/2}) + o(1).$$

(c) The results hold using an estimator $\widehat{\mathbf{\Gamma}}_k$, with $\widehat{\mathbf{\Sigma}}_k$ as in (B.1.5) or (B.1.7), in place of the true $\mathbf{\Gamma}_k$.

Theorem B.3 supports the use of the grid-based procedure with the Wald detector.

Theorem B.3. The results of Theorem B.2 holds for the output $\{\widehat{k}_j : 1 \le j \le \widehat{q}\}$ from the Wald grid-based procedure.

B.2 Verifying conditions

Here we outline the conditions used in deriving asymptotic results in Reckrühm (2019) and Kirch and Reckrühm (2022) for a more general process, and we show these hold for the piecewise stationary VAR model.

In Assumption B.2 we specify a general invariance principle.

Assumption B.2 (Invariance principle)**.** Let $\{\boldsymbol{Y}_t\}_{t\in\mathbb{Z}}$ be a stochastic process where $\boldsymbol{Y}_t \in \mathbb{R}^m$, $\mathbb{E}(\boldsymbol{Y}_1) = \mathbf{0}$ and $\mathsf{Cov}(\boldsymbol{Y}_1) = \mathbf{\Sigma}$. The partial sum process $S(k) = \sum_{t=1}^{k} \boldsymbol{Y}_t$ satisfies a strong invariance principle such that (possibly after changing the probability space) there exists a m-dimensional standard Wiener process $\{\boldsymbol{W}(k)\}_{k\ge 0}$ with covariance $\boldsymbol{I}_m$ and $\widetilde{\nu} > 0$ such that
$$\left\| \mathbf{\Sigma}^{-1/2}(S(k) - \mathbb{E}(S(k))) - \boldsymbol{W}(k) \right\| = O\left(k^{1/(2+\widetilde{\nu})}\right) \text{ a.s.}$$
as k goes to infinity.

In Lemma B.4, we show that under the piecewise stationary VAR model, the generating process satisfies the regularity conditions stated in Reckrühm (2019).

Lemma B.4. Let Assumptions 4.1–4.4 hold. The model (4.1) satisfies the following:

(a) For regimes $j = 1, \dots, q+1$, the components of $\left\{\mathbb{X}_t^{(j)}(\mathbb{X}_t^{(j)})^\top - \boldsymbol{C}_{(j)}\right\}_{t=1}^{n}$ satisfy a strong invariance principle as in Assumption B.2.

(b) For regimes $j = 1,\dots,q+1$,

$$\left\{\left((\mathbb{X}_{t-1}^{(j)})^\top \varepsilon_{1t}, (\mathbb{X}_{t-1}^{(j)})^\top \varepsilon_{2t}, \dots, (\mathbb{X}_{t-1}^{(j)})^\top \varepsilon_{pt}\right)^\top\right\}_{t=1}^{n}$$

satisfies a strong invariance principle as in Assumption B.2.

(c) For $j = 1,\dots,q$, The matrix $\delta \boldsymbol{C}_{(j)} + (1-\delta)\boldsymbol{C}_{(j+1)}$ is positive definite for all $\delta \in [0,1]$, and $\sup_{\delta\in[0,1]} \left\| \left(\delta \boldsymbol{C}_{(j)} + (1-\delta)\boldsymbol{C}_{(j+1)}\right)^{-1} \right\|_F < \infty$.

Proof. The components of each sequence $\left\{\mathbb{X}_t^{(j)}(\mathbb{X}_t^{(j)})^\top - \boldsymbol{C}_{(j)} : t \geq 1\right\}$, $j = 1,\dots,q+1$ have expectation 0 and and at least $4+\widetilde{\nu}$ finite moments by Assumption 4.1, so Assumption B.2 follows from Theorem 4.1 of Kirch and Reckrühm (2022), and (a) is satisfied. Part (b) follows similarly. By definition, $\boldsymbol{C}_{(j)}$ is positive definite iff $\boldsymbol{z}^\top \boldsymbol{C}_{(j)} \boldsymbol{z} > 0$ for all conformable $\boldsymbol{z}$. Also, $\boldsymbol{C}_{(j)}$ is positive definite for all j by Assumption 4.1. Hence it follows that $\boldsymbol{z}^\top \left(\delta \boldsymbol{C}_{(j)} + (1-\delta)\boldsymbol{C}_{(j+1)}\right)\boldsymbol{z} > 0$ for any $0 \leq \delta \leq 1$, so (c) holds. ■

In Lemma B.5 we give the form of the expected Hessian of the estimating function.

Lemma B.5. The matrix $\boldsymbol{V}_{(j)}(\widetilde{\boldsymbol{a}}) = \mathbb{E}\left(\nabla \boldsymbol{H}(\boldsymbol{X}_t^{(j)}, \mathbb{X}_{t-1}^{(j)}, \widetilde{\boldsymbol{a}})\right)$ has the following form:

$$\boldsymbol{V}_{(j)}(\widetilde{\boldsymbol{a}}) = \begin{pmatrix} \boldsymbol{C}_{(j)} & \boldsymbol{O} & \dots & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{C}_{(j)} & & \vdots \\ \vdots & & \ddots & \\ \boldsymbol{O} & \dots & & \boldsymbol{C}_{(j)} \end{pmatrix}. \tag{B.2.1}$$

Hence, by Assumption 4.1,

$$\boldsymbol{V}_{(j)}^{-1}(\widetilde{\boldsymbol{a}}) = \begin{pmatrix} \boldsymbol{C}_{(j)}^{-1} & \boldsymbol{O} & \dots & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{C}_{(j)}^{-1} & & \vdots \\ \vdots & & \ddots & \\ \boldsymbol{O} & \dots & & \boldsymbol{C}_{(j)}^{-1} \end{pmatrix}. \tag{B.2.2}$$

Proof. For $i = 1,\dots,p$, we have that

$$\nabla \boldsymbol{H}\left(X_{it}^{(j)}, \mathbb{X}_{t-1}^{(j)}, \widetilde{\boldsymbol{a}}(i)\right) = \frac{\partial \boldsymbol{H}\left(X_{it}^{(j)}, \mathbb{X}_{t-1}^{(j)}, \widetilde{\boldsymbol{a}}(i)\right)}{\partial \widetilde{\boldsymbol{a}}(i)} = \mathbb{X}_{t-1}^{(j)}(\mathbb{X}_{t-1}^{(j)})^\top, \tag{B.2.3}$$

implying that $\mathbb{E}\left(\nabla \boldsymbol{H}(X_{it}^{(j)}, \mathbb{X}_{t-1}^{(j)}, \widetilde{\boldsymbol{a}}(i))\right) = \mathbb{E}(\mathbb{X}_{t-1}^{(j)}(\mathbb{X}_{t-1}^{(j)})^\top) = \boldsymbol{C}_{(j)}$. Moreover, for $i \neq i'$,

$$\frac{\partial \boldsymbol{H}\left(X_{it}^{(j)}, \mathbb{X}_{t-1}^{(j)}, \widetilde{\boldsymbol{a}}(i)\right)}{\partial \widetilde{\boldsymbol{a}}(i')} = \boldsymbol{O}.$$

Using the blockwise inverse of matrices and that each principal submatrix is positive definite, we have (B.2.2). ■

Lemma B.6 controls the moments of the series, the gradient, and the Hessian in each regime.

Lemma B.6. Let Assumptions 4.1–4.4 hold. There exists a $\nu > 0$ such that the model (4.1) satisfies the following for $j = 1, \dots, q+1$

(a) $\mathbb{E}\left(\sup_{\widetilde{\boldsymbol{a}} \in \boldsymbol{\Theta}} \left\| \boldsymbol{H}\left(\boldsymbol{X}_t^{(j)}, \mathbb{X}_{t-1}^{(j)}, \widetilde{\boldsymbol{a}}\right)\right\|^{2+\nu}\right) < \infty.$

(b) $\mathbb{E}\left(\sup_{\widetilde{\boldsymbol{a}} \in \boldsymbol{\Theta}} \left\| \nabla \boldsymbol{H}\left(\boldsymbol{X}_t^{(j)}, \mathbb{X}_{t-1}^{(j)}, \widetilde{\boldsymbol{a}}\right)\right\|_F^{2+\nu}\right) < \infty.$

(c) For $i = 1, \dots, p$, $\mathbb{E}\left(\sup_{\widetilde{\boldsymbol{a}} \in \boldsymbol{\Theta}} \left\| \nabla^2 \boldsymbol{H}\left(X_{it}^{(j)}, \mathbb{X}_{t-1}^{(j)}, \widetilde{\boldsymbol{a}}(i)\right)\right\|_F^{2+\nu}\right) < \infty.$

Proof. (a) Noting that

$$-\boldsymbol{H}(X_{it}^{(j)}, \mathbb{X}_{t-1}^{(j)}, \widetilde{\boldsymbol{a}}(i)) = \mathbb{X}_{t-1}^{(j)}(\mathbb{X}_{t-1}^{(j)})^\top(\widetilde{\boldsymbol{a}}(i) - \boldsymbol{a}_j(i)) + \mathbb{X}_{t-1}^{(j)}\varepsilon_{it},$$

for $i = 1, \dots, p$ we have that

$$\begin{aligned}
&\mathbb{E}(\sup_{\widetilde{\boldsymbol{a}} \in \boldsymbol{\Theta}} \|\boldsymbol{H}(X_{it}^{(j)}, \mathbb{X}_{t-1}^{(j)}, \widetilde{\boldsymbol{a}}(i)\|^{2+\nu}) \\
&\leq \mathbb{E}(\sup_{\widetilde{\boldsymbol{a}} \in \boldsymbol{\Theta}} \|\mathbb{X}_{t-1}^{(j)}(\mathbb{X}_{t-1}^{(j)})^\top(\widetilde{\boldsymbol{a}}(i) - \boldsymbol{a}_j(i))\|^{2+\nu}) + \mathbb{E}(\|\mathbb{X}_{t-1}^{(j)}\varepsilon_{it}\|^{2+\nu}) \\
&\leq \mathbb{E}(\|\mathbb{X}_{t-1}^{(j)}\|^{4+2\nu} \sup_{\widetilde{\boldsymbol{a}} \in \boldsymbol{\Theta}} \|(\widetilde{\boldsymbol{a}}(i) - \boldsymbol{a}_j(i))\|^{2+\nu}) + \mathbb{E}(\|\mathbb{X}_{t-1}^{(j)}\|^{2+\nu}\|\varepsilon_{it}\|^{2+\nu}) < \infty,
\end{aligned}$$

where the last inequality follows by the moment conditions in Assumptions 4.1 and 4.2 with $\widetilde{\nu} = 2\nu$, and the conditional independence of processes in Assumption 4.3. Each subvector of $\boldsymbol{H}\left(\boldsymbol{X}_t^{(j)}, \mathbb{X}_{t-1}^{(j)}, \widetilde{\boldsymbol{a}}\right)$ is thusly bounded, so the statement holds.

(b) Using (B.2.3), this holds by Assumption 4.1.

(c) $\nabla^2 \boldsymbol{H}\left(X_{it}^{(j)}, \mathbb{X}_{t-1}^{(j)}, \widetilde{\boldsymbol{a}}(i)\right) = \boldsymbol{O}$ for all $\widetilde{\boldsymbol{a}} \in \boldsymbol{\Theta}$, so this holds.

■

B.2.1 MOSUM score procedure

Here we verify conditions specific to the MOSUM score procedure. Recall $\boldsymbol{a}_{s,e}$, the unique solution of (4.12), and $\mathbb{K}_{s,e}$ in (4.13). Define $s = \lfloor \gamma_s n \rfloor$ and $e = \lfloor \gamma_e n \rfloor$ for $0 \leq \gamma_s < \gamma_e \leq 1$.

Lemma B.7 states that each series of estimating functions obeys a forwards inequality.

Lemma B.7. Define $\boldsymbol{H}_0(\boldsymbol{X}_t, \mathbb{X}_{t-1}, \widetilde{\boldsymbol{a}}) = \boldsymbol{H}(\boldsymbol{X}_t, \mathbb{X}_{t-1}, \widetilde{\boldsymbol{a}}) - \mathbb{E}(\boldsymbol{H}(\boldsymbol{X}_t, \mathbb{X}_{t-1}, \widetilde{\boldsymbol{a}}))$. Let Assumptions 4.1–4.4 hold.

(a) The following hold for any $\xi_n \to \infty$ and any j such that $k_j \in \mathbb{K}_{s,e}$:

(i) For $m \in \{k_{j-1}+G, k_j-G, k_j\}$, it holds

$$\max_{\xi_n \le k \le G} \frac{1}{k} \left\| \sum_{t=m-k+1}^{m} \boldsymbol{H}_0\left(\boldsymbol{X}_t^{(j)}, \mathbb{X}_{t-1}^{(j)}, \boldsymbol{a}_{s,e}\right) \right\| = o_P(1),$$

(ii) For $m \in \{k_{j-1}, k_{j-1}+G, k_j-G\}$, it holds

$$\max_{\xi_n \le k \le G} \frac{1}{k} \left\| \sum_{t=m+1}^{m+k} \boldsymbol{H}_0\left(\boldsymbol{X}_t^{(j)}, \mathbb{X}_{t-1}^{(j)}, \boldsymbol{a}_{s,e}\right) \right\| = o_P(1).$$

(b) The following backward law of large numbers holds for any $\xi_n \to \infty$ and any j such that $k_j \in \mathbb{K}_{s,e}$:

$$\max_{\xi_n \le k \le G} \frac{1}{k} \left\| \sum_{t=G-k+1}^{G} \boldsymbol{H}_0\left(\boldsymbol{X}_t^{(j)}, \mathbb{X}_{t-1}^{(j)}, \boldsymbol{a}_{s,e}\right) \right\| = o_P(1).$$

Proof. By Lemma B.6 and Kirch and Reckrühm (2022) Theorem 4.4, part (a) holds. Part (b) holds by Lavielle and Moulines (2000), Theorem 1. ■

Lemma B.8 verifies that the estimating function evaluated on each stationary segment fulfils an invariance principle.

Lemma B.8. Let Assumptions 4.1–4.4 hold. Let $\widetilde{\boldsymbol{a}}$ be a fixed inspection parameter. Each series $\{\boldsymbol{H}\left(\boldsymbol{X}_t^{(j)}, \mathbb{X}_{t-1}^{(j)}, \widetilde{\boldsymbol{a}}\right)\}_{t=1}^{n}$ for j such that $k_j \in \mathbb{K}_{s,e}$ satisfies Assumption B.2.

Proof. For each $j = 1, \ldots, q+1$, define $S(j,k,\widetilde{\boldsymbol{a}}) = \sum_{t=1}^{k} \boldsymbol{H}\left(\boldsymbol{X}_t^{(j)}, \mathbb{X}_{t-1}^{(j)}, \widetilde{\boldsymbol{a}}\right)$. For $i = 1, \ldots, p$, $\boldsymbol{H}(\boldsymbol{X}_t, \mathbb{X}_{t-1}, \widetilde{\boldsymbol{a}})$ consists of sub-vectors

$$-\mathbb{X}_{t-1}^{(j)}(\mathbb{X}_{t-1}^{(j)})^\top(\widetilde{\boldsymbol{a}}(i) - \boldsymbol{a}_j(i)) - \mathbb{X}_{t-1}\varepsilon_{it}.$$

The sequences $\left\{\mathbb{X}_{t-1}^{(j)}(\mathbb{X}_{t-1}^{(j)})^\top : t \ge 1\right\}$ and $\left\{\mathbb{X}_{t-1}^{(j)}\varepsilon_{it} : t \ge 1\right\}$, $i = 1, \ldots, p, j = 1, \ldots, q+1$, each satisfy invariance by Lemma B.4, so the statement holds. ■

In Lemma B.9, we verify that the least squares estimator is $\sqrt{n}-$consistent for the best-approximating global parameter, even when that $\mathbb{X}_{t-1}$ may contain regressors from multiple regimes. Hence by Lemma B.10, the divergence of the difference vector $\boldsymbol{m}_k(G, \widehat{\boldsymbol{a}}_{s,e})$ from $\boldsymbol{m}_k(G, \boldsymbol{a}_{s,e})$ must be bounded in probability.

Lemma B.9. Let Assumptions 4.1–4.4 hold. Then

$$\sqrt{n}\left(\widehat{\boldsymbol{a}}_{s,e} - \boldsymbol{a}_{s,e}\right) = O_P(1).$$

Proof. Lemma B.6 holds, so the result follows by Kirch and Reckrühm (2022) Theorem 4.3. ■

Lemma B.10. Let Assumptions 4.1–4.4 hold. For any j such that $k_j \in \mathbb{K}_{s,e}$, the estimator $\widehat{\boldsymbol{a}}_{s,e}$ fulfils

$$(i) \max_{k_{j-1}+G \leq k \leq k_j - G} \frac{1}{\sqrt{2G}} \|\boldsymbol{m}_k(G, \widehat{\boldsymbol{a}}_{s,e}) - \boldsymbol{m}_k(G, \boldsymbol{a}_{s,e})\| = o_P\left((\log(n/G))^{-1/2}\right)$$
$$(ii) \max_{k:|k-k_j|<G} \frac{1}{\sqrt{2G}} \|\boldsymbol{m}_k(G, \widehat{\boldsymbol{a}}_{s,e}) - \boldsymbol{m}_k(G, \boldsymbol{a}_{s,e})\| = o_P\left((\log(n/G))^{1/2}\right).$$

Proof. By Assumption 4.2, the process $\boldsymbol{X}_t$ is stable, with white noise errors. By Lemma B.9, $\widehat{\boldsymbol{a}}_{s,e}$ is $\sqrt{n}$-consistent for $\boldsymbol{a}_{s,e}$. By Lemma B.6 and Kirch and Reckrühm (2022), Theorem 4.4, the statement holds. ■

Lemma B.11 states that for $\boldsymbol{a}_{s,e}$, at least one k_j in the interval of interest causes a change in the expectation of the estimating function.

Lemma B.11 (Lemma 3.3, Kirch and Reckrühm (2022))**.** For at least one j such that $k_j \in \mathbb{K}_{s,e}$ it holds that

$$\mathbb{E}\left(\boldsymbol{H}\left(\boldsymbol{X}_t^{(j)}, \mathbb{X}_{t-1}^{(j)}, \boldsymbol{a}_{s,e}\right)\right) \neq \mathbb{E}\left(\boldsymbol{H}\left(\boldsymbol{X}_t^{(j+1)}, \mathbb{X}_{t-1}^{(j+1)}, \boldsymbol{a}_{s,e}\right)\right).$$

B.2.2 MOSUM Wald procedure

In Lemma B.12 we verify the conditions underlying the MOSUM Wald procedure.

Lemma B.12. Let Assumptions 4.1–4.3 hold. Then, the following hold:

(a) $\boldsymbol{V}_{(j)}(\widetilde{\boldsymbol{a}}) = \mathbb{E}\left(\nabla \boldsymbol{H}\left(\boldsymbol{X}_t^{(j)}, \mathbb{X}_{t-1}^{(j)}, \widetilde{\boldsymbol{a}}\right)\right)^\top, j = 1, \ldots, q+1$, is a regular matrix for all $\widetilde{\boldsymbol{a}} \in \boldsymbol{\Theta}$ and

$$\sup_{\widetilde{\boldsymbol{a}} \in \boldsymbol{\Theta}} \left\| \boldsymbol{V}_{(j)}^{-1}(\widetilde{\boldsymbol{a}}) \right\|_F < \infty.$$

(b) Let $\delta \boldsymbol{V}_{(j)}(\widetilde{\boldsymbol{a}}) + (1-\delta)\boldsymbol{V}_{(j+1)}(\widetilde{\boldsymbol{a}})$ be a regular matrix for all $\widetilde{\boldsymbol{a}} \in \boldsymbol{\Theta}$ and all $\delta \in [0,1]$ and let

$$\sup_{\delta \in [0,1]} \sup_{\widetilde{\boldsymbol{a}} \in \boldsymbol{\Theta}} \left\| \left(\delta \boldsymbol{V}_{(j)}(\widetilde{\boldsymbol{a}}) + (1-\delta)\boldsymbol{V}_{(j+1)}(\widetilde{\boldsymbol{a}})\right)^{-1} \right\|_F < \infty, \quad j = 1, \ldots, q.$$

Proof. Part (a) holds by the Cauchy-Schwartz inequality and that $\boldsymbol{V}_{(j)}^{-1}(\widetilde{\boldsymbol{a}})\boldsymbol{V}_{(j)}(\widetilde{\boldsymbol{a}}) = \boldsymbol{I}$, for any $\widetilde{\boldsymbol{a}} \in \boldsymbol{\Theta}$, we have that

$$\left\| \boldsymbol{V}^{-1}(\widetilde{\boldsymbol{a}}) \right\|_F \leq \|\boldsymbol{V}(\widetilde{\boldsymbol{a}})\|_F^{-1} < \infty.$$

For part (b), for any $\widetilde{\boldsymbol{a}} \in \boldsymbol{\Theta}$ and $\delta \in [0,1]$, by part (a) we have

$$\begin{aligned} \left\| \left(\delta \boldsymbol{V}_{(j)}(\widetilde{\boldsymbol{a}}) + (1-\delta)\boldsymbol{V}_{(j+1)}(\widetilde{\boldsymbol{a}})\right)^{-1} \right\|_F &\leq \left\| \left(\delta \boldsymbol{V}_{(j)}(\widetilde{\boldsymbol{a}}) + (1-\delta)\boldsymbol{V}_{(j+1)}(\widetilde{\boldsymbol{a}})\right) \right\|_F^{-1} \\ &\leq \max\{\left\| \boldsymbol{V}_{(j)}(\widetilde{\boldsymbol{a}}) \right\|^{-1}, \left\| \boldsymbol{V}_{(j+1)}(\widetilde{\boldsymbol{a}}) \right\|^{-1}\} \\ &< \infty. \end{aligned}$$

■

Lemma B.13 bounds the scaled estimation error in the MOSUM Wald procedure.

Lemma B.13. Denote by $\mathbf{1}_m = (1,\dots,1)^\top \in \mathbb{R}^m$. Let Assumptions 4.1–4.4 hold. For $j = 1,\dots,q+1$ it holds that

$$\max_{k_{j-1}+1\le k\le k_j-G} \left\| \sqrt{\frac{G}{2}} \left(\widehat{\boldsymbol{a}}_{k+1,k+G} - \boldsymbol{a}_j\right) \boldsymbol{C}_{(j)} \otimes \mathbf{1}_{p(dp+1)} - \frac{1}{\sqrt{2G}} \sum_{t=k+1}^{k+G} \boldsymbol{H}\left(\boldsymbol{X}_t^{(j)}, \mathbb{X}_{t-1}^{(j)}, \boldsymbol{a}_j\right) \right\| = o_P\left((\log(n/G))^{-1/2}\right).$$

Proof. By the definition of $\widehat{\boldsymbol{a}}_{k+1,k+G}$ in (4.6) we have

$$\sum_{t=k+1}^{k+G} \varepsilon_{it} \mathbb{X}_{t-1}^{(j)} = \sum_{t=k+1}^{k+G} (\widehat{\boldsymbol{a}}_{k+1,k+G}(i) - \boldsymbol{a}_j(i))^\top \mathbb{X}_{t-1}^{(j)} (\mathbb{X}_{t-1}^{(j)})^\top,$$

hence

$$\begin{aligned} &\frac{1}{\sqrt{2G}} \sum_{t=k+1}^{k+G} \varepsilon_{it} \mathbb{X}_{t-1}^{(j)} - \sqrt{\frac{G}{2}} (\widehat{\boldsymbol{a}}_{k+1,k+G}(i) - \boldsymbol{a}_j(i))^\top \boldsymbol{C}_{(j)} \\ &= \frac{1}{G} \sum_{t=k+1}^{k+G} \sqrt{\frac{G}{2}} (\widehat{\boldsymbol{a}}_{k+1,k+G}(i) - \boldsymbol{a}_j(i))^\top (\mathbb{X}_{t-1}^{(j)} (\mathbb{X}_{t-1}^{(j)})^\top - \boldsymbol{C}_{(j)}). \end{aligned}$$

By Lemmas B.4, B.9, and B.15,

$$\begin{aligned} &\max_{k_{j-1}+1\le k\le k_j-G} \left\| \sqrt{\frac{G}{2}} \left(\widehat{\boldsymbol{a}}_{k+1,k+G} - \boldsymbol{a}_j\right) \boldsymbol{C}_{(j)} - \frac{1}{\sqrt{2G}} \sum_{t=k+1}^{k+G} \varepsilon_{it} \mathbb{X}_{t-1}^{(j)} \right\| \\ &\le \max_{k_{j-1}+1\le k\le k_j-G} \sqrt{\frac{G}{2}} \|\widehat{\boldsymbol{a}}_{k+1,k+G}(i) - \boldsymbol{a}_j(i)\| \max_{k_{j-1}+1\le k\le k_j-G} \frac{1}{G} \| \sum_{k+1}^{k+G} (\mathbb{X}_{t-1} \mathbb{X}_{t-1}^\top - \boldsymbol{C}_{(j)}) \|_F \\ &= O_P\left(\frac{\log(n/G)}{\sqrt{G}}\right) = o_P\left((\log(n/G))^{-1/2}\right). \end{aligned}$$

The last step follows by Assumption 4.4 on the bandwidth G. This holds for $i = 1,\dots,p$ so the statement holds by the union bound. ■

B.3 Proofs and supporting results

B.3.1 MOSUM score procedure

Proposition 4.1

Proof. Lemmas B.7 and B.8 hold. The estimator $\widehat{\boldsymbol{a}}_{1,n}$ obeys the bounds of Lemma B.10, while $\widehat{\boldsymbol{\Sigma}}_k(\widetilde{\boldsymbol{a}})$ as in (4.7) or (4.9) meets Assumption 4.7 (a) by Lemmas B.22 and B.24. The statement then follows by Kirch and Reckrühm (2022), Theorem 2.1. ■

Theorem 4.2

Proof. By Lemma B.8, the estimating function series obeys an invariance principle as in Assumption B.2. When $q \geq 1$, all changes are detectable by Assumption 4.5. The bandwidth meets Assumption 4.4. The estimator $\widehat{\boldsymbol{a}}_{1,n}$ obeys the bounds of Lemma B.10, while $\widehat{\boldsymbol{\Sigma}}_k(\widetilde{\boldsymbol{a}})$ as in (4.7) or (4.9) meets Assumption 4.7 by Lemmas B.22 and B.24. Hence by Kirch and Reckrühm (2022) Theorem 3.5, part (a) holds.

For part (b), we want to show that for each k_j, and each $\epsilon > 0$, there exists $w_n \in (0, G \cdot \min_{j=1,\ldots,q} \delta_j^2(\widetilde{\boldsymbol{a}}))$ such that $w_n \to \infty$ and

$$\mathsf{P}(\max_{1\leq j\leq q} |\widehat{k}_j - k_j| \delta_j^2(\widetilde{\boldsymbol{a}}) > w_n) \leq \epsilon.$$

We condition on the event $\{\widehat{q} = q\}$. By Assumption 4.5, following the proof of Kirch and Reckrühm (2022) Theorem 3.6 we have that

$$\mathsf{P}\left(\widehat{k}_j - k_j < \frac{-w_n}{\delta_j^2(\widetilde{\boldsymbol{a}})}\right) = O\left(w_n^{-\gamma/2}\right) + o(1),$$

and

$$\mathsf{P}\left(\widehat{k}_j - k_j > \frac{w_n}{\delta_j^2(\widetilde{\boldsymbol{a}})}\right) = O\left(w_n^{-\gamma/2}\right) + o(1).$$

Hence, for each $1 \leq j \leq q$,

$$\mathsf{P}\left(\left|\widehat{k}_j - k_j\right| \delta_j^2(\widetilde{\boldsymbol{a}}) > w_n\right) = O\left(w_n^{-\gamma/2}\right) + o(1),$$

so by the union bound over $1 \leq j \leq q$, since q is fixed, we conclude

$$\mathsf{P}\left(\max_{1\leq j\leq q} \left|\widehat{k}_l - k_j\right| \delta_j^2(\widetilde{\boldsymbol{a}}) > w_n\right) = O(q w_n^{-\gamma/2}) + o(1) = O(w_n^{-\gamma/2}) + o(1).$$

■

B.3.2 MOSUM Wald procedure

Proposition B.1

Proof. Lemmas B.7, B.8, and B.13 hold. The statement then follows by Kirch and Reckrühm (2022), Theorem 2.1. ■

Theorem B.2

Proof. By Lemma B.8, the estimating function series obeys an invariance principle as in Assumption B.2. The bandwidth meets Assumption 4.4, and the estimator $\widehat{\boldsymbol{\Gamma}}_k$ meets Assumption B.1 by Remark B.1 and Lemmas B.26 and B.27. The error bound in Lemma B.13 holds. We note that Proposition 3.1 of Kirch and Reckrühm (2022) holds in our setting in consideration of the

boundary effect between regimes. At $t = k_j + 1, \ldots, k_j + d$ the regressor $\mathbb{X}_{t-1}$ can contain at most d observations drawn from the previous regime, e.g. $\boldsymbol{X}_{t-1}^{(j)}$. The estimator $\widehat{\boldsymbol{a}}_{k-G+1,k}$ consistently estimates

$$\frac{G-d}{G}\boldsymbol{a}_j + \frac{d}{G}\boldsymbol{a}^\dagger$$

in the sense of Lemma B.16, where $\boldsymbol{a}^\dagger$ is some fixed parameter. Since d is fixed and $G \to \infty$ with n, $\widehat{\boldsymbol{a}}_{k-G+1,k}$ is consistent for $\boldsymbol{a}_j$. Hence by Theorem 3.2 of Kirch and Reckrühm (2022), part (a) holds. Part (b) holds by the same arguments as the proof of Theorem 4.2. ■

B.3.3 Recursive segmentation

Theorem 4.3

Proof. By Theorem 4.2, with probability tending to 1, any change point k_j meeting Assumption 4.5 will have a corresponding estimator $\widehat{k}_j$ such that $|\widehat{k}_j - k_j| < \delta_j^{-2}(\boldsymbol{a}_{1,n}) \cdot w_n$ for some w_n; we denote this event $\mathscr{R}_0$. If all changes meet Assumption 4.5 with $\widetilde{\boldsymbol{a}} = \boldsymbol{a}_{1,n}$, then we are done. Otherwise, we set $i \leftarrow 0$, $\widetilde{s}_{old} = 1, \widetilde{e}_{old} = n$ and argue recursively.

We condition on the event $\mathscr{R}_i$. Suppose the existence of k_j, where $j \in \{1, \ldots, q\}$, which was not detected. By Lemma B.11, we at least have that $\delta_j(\boldsymbol{a}_{\widetilde{s},\widetilde{e}}) > 0$, where $\widetilde{s}$ is the largest element of $\{\widehat{k}_{j'} + 1 : j' \leq j-1\}$, and $\widetilde{e}$ is the smallest element of $\{\widehat{k}_{j''} : j'' \geq j+1\}$, with the conventions that $\widehat{k}_0 = 0$ and $\widehat{k}_{q+1} = n$. First, we suppose k_j has a jump size such that $\delta_j(\boldsymbol{a}_{\widetilde{s}_{old},\widetilde{e}_{old}}) < c_{\delta,n} \leq \delta_j(\boldsymbol{a}_{\widetilde{s},\widetilde{e}})$. Letting $s = k_{j'} + 1$ and $e = k_{j''}$, we have (for $j', j'' \in \{1, \ldots, q\}$)

$$|\widetilde{s} - s| \leq \delta_{j'}^{-2}(\boldsymbol{a}_{\widetilde{s}_{old},\widetilde{e}_{old}}) \cdot w_n, \text{ and } |\widetilde{e} - e| \leq \delta_{j''}^{-2}(\boldsymbol{a}_{\widetilde{s}_{old},\widetilde{e}_{old}}) \cdot w_n, \tag{B.3.1}$$

due to $\mathscr{R}_i$. We apply the MOSUM score procedure to the resulting segmentation $\widetilde{s}, \ldots, \widetilde{e}$ without loss of generality. We have that $\min_{1 \leq j \leq q+1} k_j - k_{j-1} > 2G > \lfloor c_{\delta,n}^{-2} \cdot w_n \rfloor$ by Assumptions 4.4 and 4.8 (a), so

$$1 \leq s + \lfloor c_{\delta,n}^{-2} \cdot w_n \rfloor < k_j < e - \lfloor c_{\delta,n}^{-2} \cdot w_n \rfloor \leq n,$$

and hence by (B.3.1)

$$1 \leq \widetilde{s} + \lfloor c_{\delta,n}^{-2} \cdot w_n' \rfloor < k_j < \widetilde{e} - \lfloor c_{\delta,n}^{-2} \cdot w_n' \rfloor \leq n,$$

where $w_n' = 2w_n$, so k_j must meet Assumption 4.8 (a) with $\widetilde{s}$ and $\widetilde{e}$.

If $s < \widetilde{s}$ and $\widetilde{e} < e$, then by Theorem 4.2, k_j will be detected and estimated with error less than $c_{\delta,n}^{-2} \cdot w_n$. Otherwise, suppose that $\widetilde{s} < s$; the case for $e < \widetilde{e}$ is analogous. Then for each $k \in \left[\widetilde{s} + G + 1, \widetilde{s} + G + \lfloor \delta_{j'}^{-2}(\boldsymbol{a}_{\widetilde{s}_{old},\widetilde{e}_{old}}) \cdot w_n' \rfloor\right]$, i.e. those in the interval such that the corresponding

detector may draw from two regimes, we have that

$$\begin{aligned}\|\mathbb{E}\boldsymbol{m}_k(G,\boldsymbol{a}_{\widetilde{s},\widetilde{e}})\| &= \left\| \sum_{t=k+1}^{k+G} \mathbb{E}\boldsymbol{H}_t^{(j')}(\boldsymbol{a}_{\widetilde{s},\widetilde{e}}) - \sum_{t=k-G+1}^{k_{j'}} \mathbb{E}\boldsymbol{H}_t^{(j'-1)}(\boldsymbol{a}_{\widetilde{s},\widetilde{e}}) - \sum_{t=k_{j'}+1}^{k} \mathbb{E}\boldsymbol{H}_t^{(j')}(\boldsymbol{a}_{\widetilde{s},\widetilde{e}}) \right\| \\ &\leq (k_{j'} - k + G - 1)\|\mathbb{E}\boldsymbol{H}_t^{(j')}(\boldsymbol{a}_{\widetilde{s},\widetilde{e}}) - \mathbb{E}\boldsymbol{H}_t^{(j'-1)}(\boldsymbol{a}_{\widetilde{s},\widetilde{e}})\| \\ &\leq \delta_{j'}^{-2}(\boldsymbol{a}_{\widetilde{s}_{old},\widetilde{e}_{old}}) \cdot w_n' \cdot \delta_{j'}^2(\boldsymbol{a}_{\widetilde{s},\widetilde{e}}) \\ &= O(1),\end{aligned}$$

where the third line holds by Assumption 4.8 (b). Hence by Lemma B.8,

$$\max_{\widetilde{s}+G+1\leq k\leq \widetilde{s}+G+\lfloor \delta_{j'}^{-2}(\boldsymbol{a}_{\widetilde{s}_{old},\widetilde{e}_{old}})\cdot w_n' \rfloor} T_k(\boldsymbol{a}_{\widetilde{s},\widetilde{e}}) = O(1) + o_P(\sqrt{\log(n/G)}) = o_P(D(G,\alpha)),$$

so the detector will not exceed the threshold with high probability, and we will not re-detect $k_{j'}$ (likewise, $k_{j''}$). The estimator $\widehat{\boldsymbol{a}}_{\widetilde{s},\widetilde{e}}$ is $\sqrt{n}$-consistent for $\boldsymbol{a}_{\widetilde{s},\widetilde{e}}$ by Lemma B.9, so the bounds of Lemma B.10 hold, while $\widehat{\boldsymbol{\Sigma}}_k(\widetilde{\boldsymbol{a}})$ as in (4.7) or (4.9) meets Assumption 4.7 by Lemmas B.22 and B.24. Hence,

$$\begin{aligned}\max_{\widetilde{s}+G+1\leq k\leq \widetilde{s}+G+\lfloor \delta_{j'}^{-2}(\boldsymbol{a}_{\widetilde{s}_{old},\widetilde{e}_{old}})\cdot w_n' \rfloor} \widehat{T}_k(\widehat{\boldsymbol{a}}_{\widetilde{s},\widetilde{e}}) &= \max_{\widetilde{s}+G+1\leq k\leq \widetilde{s}+G+\lfloor \delta_{j'}^{-2}(\boldsymbol{a}_{\widetilde{s}_{old},\widetilde{e}_{old}})\cdot w_n' \rfloor} \widehat{T}_k(\boldsymbol{a}_{\widetilde{s},\widetilde{e}}) + o_P((a(n/G))^{-1}) \\ &= o_P(D(G,\alpha)),\end{aligned}$$

so the result holds. By Theorem 4.2, k_j will be detected and estimated with error at most $c_{\delta,n}^{-2}\cdot w_n$. Define $\mathscr{R}_{i+1} = \mathscr{R}_i \cap \mathscr{B}_{\widetilde{s},\widetilde{e}}$, where $\mathscr{B}_{\widetilde{s},\widetilde{e}}$ denotes the event that we detect, and estimate with error less than $c_{\delta,n}^{-2}\cdot w_n$, all change points meeting Assumption 4.8 with $\boldsymbol{a}_{\widetilde{s},\widetilde{e}}$. We set $i \leftarrow i+1, \widetilde{s}_{old} = \widetilde{s}, \widetilde{e}_{old} = \widetilde{e}$ and reiterate the argument until all change points are detected.

Next, we consider change points with small jump sizes, i.e. those in the set

$$\mathbb{Q}_{\widetilde{s},\widetilde{e}} = \{k_j : \widetilde{s} < k_j < \widetilde{e}, \text{ and } \delta_j(\boldsymbol{a}_{\widetilde{s},\widetilde{e}}) < c_{\delta,n}\}.$$

It remains to show that the procedure will not detect these, and so is consistent under Assumption 4.8. It holds for $\left|k-k_j\right| \leq G$ with $j \in \mathbb{Q}_{\widetilde{s},\widetilde{e}}$ that $\|\mathbb{E}\boldsymbol{m}_k(G,\boldsymbol{a}_{\widetilde{s},\widetilde{e}})\| < c_{\delta,n}$, so

$$\max_{|k-k_j|\leq G} \frac{1}{\sqrt{2G}} \|\mathbb{E}\boldsymbol{m}_k(G,\boldsymbol{a}_{\widetilde{s},\widetilde{e}})\| = o(D(G,\alpha)) \tag{B.3.2}$$

when $c_{\delta,n}\cdot\sqrt{\frac{G}{\log(n/G)}} \to \infty$. By Assumption 4.4 it holds for such $j \in \mathbb{Q}_{\widetilde{s},\widetilde{e}}$ that

$$\begin{aligned}\max_{k_j\leq k<k_j+G} T_k(G,\boldsymbol{a}_{\widetilde{s},\widetilde{e}}) &= \frac{1}{\sqrt{2G}} \left\| (\boldsymbol{\Sigma}_k(\widetilde{\boldsymbol{a}}_{\widetilde{s},\widetilde{e}}))^{-1/2}\boldsymbol{m}_k(G,\widetilde{\boldsymbol{a}}_{\widetilde{s},\widetilde{e}}) \right\| \\ &= o(D(G,\alpha)) + O_P\left(\max_{k_j\leq k<k_j+G} \frac{1}{\sqrt{G}} \left\| \boldsymbol{W}(k+G) - 2\boldsymbol{W}(k) + \boldsymbol{W}\left(k_j\right) \right\| \right) \\ &+ O_P\left(\max_{k_j\leq k<k_j+G} \frac{1}{\sqrt{G}} \|\boldsymbol{\Sigma}_{(j+1)}^{-1/2}\boldsymbol{\Sigma}_{(j)}^{1/2}(\boldsymbol{W}(k_j) - \boldsymbol{W}(k-G)))\| \right) \\ &= o(D(G,\alpha)) + O_P(G^{-1/2}k^{1/(2+\widetilde{\nu})}) = o_P(D(G,\alpha)),\end{aligned}$$

where the second line follows by (B.3.2) and Lemma B.8 with Assumption B.2, and the last line follows by the self-similarity of Wiener processes, the stationarity of its increments and the continuous sample paths. A similar result holds for $k_j - G \leq k < k_j$, and so

$$P\left(\max_{j\in\mathbb{Q}_{\tilde{s},\tilde{e}}}\max_{|k-k_j|<G} T_k(G,\boldsymbol{a}_{\tilde{s},\tilde{e}}) \geq D(G,\alpha)\right) \to 0.$$

Since q is fixed, the algorithm terminates in a finite number of steps, and $\mathscr{R}_i$ for $i \leq q$ has probability tending to 1.

The statement holds for data-driven inspection parameters, because by Lemma B.10 it holds for $j \in \mathbb{Q}_{\tilde{s},\tilde{e}}$

$$\begin{aligned}\max_{|k-k_j|<G} T_k\left(G,\widehat{\boldsymbol{a}}_{\tilde{s},\tilde{e}}\right) &= \max_{|k-k_j|<G} T_k(G,\boldsymbol{a}_{\tilde{s},\tilde{e}}) + o_P(\sqrt{\log(n/G)}) \\ &= o_P(D(G,\alpha)).\end{aligned}$$

The estimators $\widehat{\boldsymbol{\Sigma}}_k(\widetilde{\boldsymbol{a}})$ in (4.7) or (4.9) meets Assumption 4.7 by Lemmas B.22 and B.24, so the result holds using these also. ■

B.3.4 Grid-based procedures

We give theoretical results for the grid-based procedures, giving the same asymptotic guarantees as for the MOSUM score and Wald procedures. We define the grid-based detectors

$$\widehat{T}_{\mathscr{T}}(G) = \max_{k\in\mathscr{T}} \widehat{T}_k(G,\widetilde{\boldsymbol{a}}), \text{ and } \widehat{W}_{\mathscr{T}}(G) = \max_{k\in\mathscr{T}} \widehat{W}_k(G). \tag{B.3.3}$$

where $\widehat{T}_k(G,\widetilde{\boldsymbol{a}})$ is as defined in (4.3). Define the grid-based detectors

$$T_{\mathscr{T}}(G,\widetilde{\boldsymbol{a}}) = \max_{k\in\mathscr{T}} T_k(G,\widetilde{\boldsymbol{a}}), \text{ and } W_{\mathscr{T}}(G) = \max_{k\in\mathscr{T}} W_k(G).$$

B.3.4.1 Null results

In Lemma B.14 we show that the grid-based procedure inherits size control under the null hypothesis.

Lemma B.14. Let the conditions of Proposition 4.1 hold. The grid-based procedure using the score detector (B.3.3) has asymptotic level at most α. Under the conditions of Proposition B.1, the same holds for the Wald detector (B.1.8).

Proof. Since $\mathscr{T} \subseteq \{G,\ldots,n-G\}$, we have $T_{\mathscr{T}}(G,\widetilde{\boldsymbol{a}}) \leq T(G,\widetilde{\boldsymbol{a}})$ and $W_{\mathscr{T}}(G) \leq W(G)$, so as $n \to \infty$,

$$\mathsf{P}(a(n/G)T_{\mathscr{T}}(G,\widetilde{\boldsymbol{a}}) - b(n/G) > c_\alpha) \leq \alpha$$

and

$$\mathsf{P}(a(n/G)W_{\mathscr{T}}(G) - b(n/G) > c_\alpha) \leq \alpha.$$

The statement follows. ■

B.3.4.2 Alternative results

First, we provide a useful result for bounding matrix norms.

Lemma B.15. Let $\boldsymbol{A} \in \mathbb{R}^{p\times p}$ and $\boldsymbol{x} \in \mathbb{R}^p$. Then, $\|\boldsymbol{Ax}\| \leq \|\boldsymbol{A}\|_F \|\boldsymbol{x}\|$.

Proof. On noting that $\boldsymbol{Ax} \in \mathbb{R}^p$, that the Frobenius norm and Euclidean norm coincide for a real-valued vector, and the submultiplicativity of the Frobenius norm yield

$$\|\boldsymbol{Ax}\| = \|\boldsymbol{Ax}\|_F \leq \|\boldsymbol{A}\|_F \|\boldsymbol{x}\|_F = \|\boldsymbol{A}\|_F \|\boldsymbol{x}\|.$$

■

We now define some sets used in proofs. Let A_G be the set of points which are at least G away from any change point:

$$A_G = \bigcup_{j=1,\ldots,q+1} A_{j,G} \text{ where } A_{j,G} = \{k : k_{j-1} + G + 1 \leq k \leq k_j - G\}. \tag{B.3.4}$$

The sets

$$B_G^{(1)} = \bigcup_{j=1,\ldots,q+1} B_{j,G}^{(1)}, \text{ where } B_{j,G}^{(1)} = \{k : k + 1 \leq k_j \leq k + G\}, \text{ and} \tag{B.3.5}$$

$$B_G^{(2)} = \bigcup_{j=1,\ldots,q+1} B_{j,G}^{(2)}, \text{ where } B_{j,G}^{(2)} = \{k : k - G + 1 \leq k_j \leq k\} \tag{B.3.6}$$

contain points k such that a change point is within distance G to the right and to the left respectively. We define $B_G = B_G^{(1)} \cup B_G^{(2)}$. For each $j = 1,\ldots,q$, we define the point $k_{j,\mathcal{T}}$ to be the smallest k such that $k \in \mathcal{T}$ and $|k - k_j|$ is minimised, i.e.

$$k_{j,\mathcal{T}} = \underset{k\in\mathcal{T}}{\arg\min} |k - k_j|. \tag{B.3.7}$$

By the definition of $\mathcal{T}$, we have the bound $|k_j - k_{j,\mathcal{T}}| \leq R/2 \leq G/2$.

Lemma B.16 shows that parameter estimates under the grid-based method are $\sqrt{G}$-consistent.

Lemma B.16. Let Assumptions 4.1–4.4 hold. Then, $\sqrt{G}\left\|\widehat{\boldsymbol{a}}_{k_{j,\mathcal{T}}+1,k_{j,\mathcal{T}}+G} - \boldsymbol{a}_{j+1}\right\| = O_P(G^{-1/2})$ and $\sqrt{G}\left\|\widehat{\boldsymbol{a}}_{k_{j,\mathcal{T}}-G+1,k_{j,\mathcal{T}}} - \boldsymbol{a}_j\right\| = O_P(G^{-1/2})$ for all $k_{j,\mathcal{T}}, j = 1,\ldots,q$.

Proof. Lemma B.6 holds, so for each $j = 1,\ldots,q$, we have $k_{j,\mathcal{T}} \in B_{j,G}^{(1)} \cup B_{j,G}^{(2)}$. The statement follows by Reckrühm (2019), Lemma 3.1.11 parts (b) and (c). ■

When $q \geq 1$, we show in Theorems B.17 and B.18 that grid-based procedures have asymptotic power 1.

Theorem B.17. Let H_1 be true, so that $q \geq 1$. Let Assumptions 4.1–4.5 hold. Then, using the score grid-based procedure,

(a)

$$\forall z \in R, \lim_{n\to\infty} \mathsf{P}(a(n/G)T_{\mathscr{T}}(G, \widetilde{\boldsymbol{a}}) - b(n/G) \geq z) = 1$$

(b) The same holds with the estimator $\widehat{\boldsymbol{a}}_{1,n}$ in (4.6):

$$\forall z \in R, \lim_{n\to\infty} \mathsf{P}\left(a(n/G)T_{\mathscr{T}}(G, \widehat{\boldsymbol{a}}_{1,n}) - b(n/G) \geq z\right) = 1$$

(c) We can replace $\boldsymbol{\Sigma}_k(\widetilde{\boldsymbol{a}})$ with an estimator $\widehat{\boldsymbol{\Sigma}}_k(\widetilde{\boldsymbol{a}})$ as in (4.7) or (4.9) and the results of part (a) and (b) hold.

Proof. By Lemma B.8 the invariance principle holds for the estimating function series. We have Assumption 4.4 on the bandwidth. The estimator $\widehat{\boldsymbol{a}}_{1,n}$ obeys the bounds of Lemma B.10, while $\widehat{\boldsymbol{\Sigma}}_k(\widetilde{\boldsymbol{a}})$ as in (4.7) or (4.9) meets Assumption 4.7 (a) by Lemmas B.22 and B.24. By Assumption 4.5, all changes are detectable. The statement then follows similarly to the proof of Reckrühm (2019), Theorem 2.1.5. ■

Theorem B.18. Let H_1 be true, so that $q \geq 1$. Let Assumptions 4.1–4.4 and 4.5′ hold. Then,

(a)

$$\forall z \in \mathbb{R} \lim_{n\to\infty} \mathsf{P}(a(n/G)W_{\mathscr{T}}(G) - b(n/G) \geq z) = 1.$$

(b) The result in (a) holds when evaluating $W_k(G)$ with estimators $\widehat{\boldsymbol{\Gamma}}_k$, with $\widehat{\boldsymbol{\Sigma}}_k$ as in (B.1.5) or (B.1.7), in place of $\boldsymbol{\Gamma}_k$.

Proof. Lemma B.12 holds on the moments of the estimating function, and we have Assumption 4.4 on the bandwidth. The estimator $\widehat{\boldsymbol{\Gamma}}_k$ meets Assumption B.1 by Remark B.1 and Lemmas B.26 and B.27. The proof then follows similarly to that of Reckrühm (2019) Theorem 3.1.12.

(a) Similar to the proof of Reckrühm (2019) Theorem 2.1.5 it is sufficient to show that $W_{\mathscr{T}}(G) - \frac{z+b(n/G)}{a(n/G)} \xrightarrow{P} \infty$, since the inequality $a(n/G)W_{\mathscr{T}}(G) - b(n/G) \geq z$ is equivalent to

$$W_{\mathscr{T}}(G) - \frac{z + b(n/G)}{a(n/G)} \geq 0.$$

We consider $W_{k_j}(G)$ and $W_{k_{j,\mathscr{T}}}(G)$, assuming $k_j \geq k_{j,\mathscr{T}}$ and $k_{j,\mathscr{T}} \in B_G^{(1)}$. This is done without loss of generality by symmetry of the argument around k_j, and we consider $k_{j,\mathscr{T}} \in B_G^{(2)}$ in the second case. Then $\boldsymbol{a}^* = (1-w)\boldsymbol{a}_j + w\boldsymbol{a}_{j+1}$ where $w = |k_j - k_{j,\mathscr{T}}|/G \leq R/(2G) \leq 1/2$.

First we use $W_{\mathcal{F}}(G) \geq W_{k_{j,\mathcal{F}}}(G)$, before we split the statistic $W_{k_{j,\mathcal{F}}}(G)$ into noise and signal. We can apply Lemmas B.9, B.15, and B.16 to show

$$\begin{aligned}
W_{\mathcal{F}}(G) = \max_{k\in\mathcal{F}} W_k(G) \geq W_{k_{j,\mathcal{F}}}(G) &= \frac{\sqrt{G}}{\sqrt{2}} \left\| \boldsymbol{\Gamma}_{k_{j,\mathcal{F}}}^{-1/2} \left(\widehat{\boldsymbol{a}}_{k_{j,\mathcal{F}}+1,k_{j,\mathcal{F}}+G} - \widehat{\boldsymbol{a}}_{k_{j,\mathcal{F}}-G+1,k_{j,\mathcal{F}}} \right) \right\| \\
&\geq \frac{\sqrt{G}}{\sqrt{2}} \left(\left\| \boldsymbol{\Gamma}_{k_{j,\mathcal{F}}}^{-1/2} \left(\boldsymbol{a}_{j+1} - \boldsymbol{a}^* \right) \right\| - \left\| \boldsymbol{\Gamma}_{k_{j,\mathcal{F}}}^{-1/2} \left(\widehat{\boldsymbol{a}}_{k_{j,\mathcal{F}}+1,k_{j,\mathcal{F}}+G} - \boldsymbol{a}_{j+1} - (\widehat{\boldsymbol{a}}_{k_{j,\mathcal{F}}-G+1,k_{j,\mathcal{F}}} - \boldsymbol{a}^*) \right) \right\| \right) \\
&\geq \frac{\sqrt{G}}{\sqrt{2}} \| \boldsymbol{\Gamma}_{k_{j,\mathcal{F}}}^{-1/2} \left(\boldsymbol{a}_{j+1} - \boldsymbol{a}^* \right) \| \\
&- \frac{\sqrt{G}}{\sqrt{2}} \left\| \boldsymbol{\Gamma}_{k_{j,\mathcal{F}}}^{-1/2} \right\|_F \left(\| \widehat{\boldsymbol{a}}_{k_{j,\mathcal{F}}+1,k_{j,\mathcal{F}}+G} - \boldsymbol{a}_{j+1} \| + \| \widehat{\boldsymbol{a}}_{k_{j,\mathcal{F}}-G+1,k_{j,\mathcal{F}}} - \boldsymbol{a}^* \| \right) \\
&= \frac{\sqrt{G}}{\sqrt{2}} \left\| \boldsymbol{\Gamma}_{k_{j,\mathcal{F}}}^{-1/2} \left(\boldsymbol{a}_{j+1} - \boldsymbol{a}^* \right) \right\| + O_P(G^{-1}) \\
&= \frac{\sqrt{G}}{\sqrt{2}} \left\| \boldsymbol{\Gamma}_{k_{j,\mathcal{F}}}^{-1/2} \left(\boldsymbol{a}_{j+1} - \boldsymbol{a}^* \right) \right\| + O_P(1),
\end{aligned}$$

since $\left\| \boldsymbol{\Gamma}_{k_j}^{-1/2} \right\|_F = O(1)$ and $G^{-1} = O(1)$ for $G \geq 1$. Then, since $\boldsymbol{\Sigma}_{(j)}$ and $\boldsymbol{V}_{(j)}\left(\boldsymbol{a}_j\right)$ are positive definite, we have that

$$\boldsymbol{\Gamma}_{k_{j,\mathcal{F}}} = \boldsymbol{V}_{(j)}\left(\boldsymbol{a}_j\right)^{-1} \boldsymbol{\Sigma}_{(j)} \left(\boldsymbol{V}_{(j)}\left(\boldsymbol{a}_j\right)^{-1} \right)^\top$$

is a positive definite matrix. This implies that $\boldsymbol{\Gamma}_{k_j}^{-1}$ is positive definite also. Hence, on noting that $\boldsymbol{a}_{j+1} \neq \boldsymbol{a}^*$, we have that

$$\begin{aligned}
\left\| \boldsymbol{\Gamma}_{k_{j,\mathcal{F}}}^{-1/2} \left(\boldsymbol{a}_{j+1} - \boldsymbol{a}^* \right) \right\| &= \sqrt{\left(\boldsymbol{a}_{j+1} - \boldsymbol{a}^* \right) \boldsymbol{\Gamma}_{k_{j,\mathcal{F}}}^{-1} \left(\boldsymbol{a}_{j+1} - \boldsymbol{a}^* \right)^\top} \\
&= \sqrt{\left(\boldsymbol{a}_{j+1} - \boldsymbol{a}^* \right) \left(\boldsymbol{V}_{(j)}\left(\boldsymbol{a}_j\right)^{-1} \boldsymbol{\Sigma}_{(j)} \left(\boldsymbol{V}_{(j)}\left(\boldsymbol{a}_j\right)^{-1} \right)^\top \right)^{-1} \left(\boldsymbol{a}_{j+1} - \boldsymbol{a}^* \right)^\top} \\
&\geq (1-w) c_{\delta,n}.
\end{aligned}$$

Since $\frac{z+b(n/G)}{a(n/G)} = o(\sqrt{G})$ by Assumption 4.4, using Assumption 4.5 we can conclude that

$$W_{\mathcal{F}}(G) - \frac{z+b(n/G)}{a(n/G)} \geq \frac{\sqrt{G(1-w)c_{\delta,n}}}{\sqrt{2}} + O_P(1) - \frac{z+b(n/G)}{a(n/G)} = \sqrt{G} \left(\frac{(1-w)c_{\delta,n}}{\sqrt{2}} + o_P(1) \right) \xrightarrow{P} \infty,$$

which implies part (a).

(b) By Assumption B.1 on the estimator sequence, we have $\left\| \widehat{\boldsymbol{\Gamma}}_{k_{j,\mathcal{F}}}^{-1/2} \right\|_F < \infty$, so the same arguments as in part (a) can be applied here to obtain

$$\widehat{W}_{\mathcal{F}}(G) \geq \frac{\sqrt{G}}{\sqrt{2}} \left\| \widehat{\boldsymbol{\Gamma}}_{k_{j,\mathcal{F}}}^{-1/2} \left(\boldsymbol{a}_{j+1} - \boldsymbol{a}^* \right) \right\| + O_P(1),$$

and hence

$$\left\| \widehat{\boldsymbol{\Gamma}}_{k_{j,\mathcal{F}}}^{-1/2} \left(\boldsymbol{a}_{j+1} - \boldsymbol{a}^* \right) \right\| \geq (1-w) c_{\delta,n}.$$

Finally, similar to (a) we can conclude that

$$\widehat{W}_{\mathcal{T}}(G) - \frac{z + b(n/G)}{a(n/G)} \geq \sqrt{G}\left((1-w)c_{\delta,n} + o_P(1)\right) \to \infty,$$

which yields the statement of (b). ■

We show in Theorems B.19 and B.20 that the grid-based procedures estimate q correctly, and in Corollaries B.19.1 and B.20.1 that the procedures give consistent location estimators.

Theorem B.19. Let $q \geq 1$. Let Assumptions 4.1–4.4, 4.5′, and 4.6 hold. Then, using the Wald grid-based procedure,

(a) $\mathsf{P}(\widehat{q} = q) = 1$ as $n \to \infty$.

(b) The result holds using an estimator $\widehat{\mathbf{\Gamma}}_k$, with $\widehat{\mathbf{\Sigma}}_k$ as in (B.1.5) or (B.1.7), in place of the true $\mathbf{\Gamma}_k$.

Proof. Here, we adapt the proof of Reckrühm (2019), Theorem 3.1.15 to the grid setting. The forwards/backwards inequalities of Lemma B.7 hold. The estimator $\widehat{\mathbf{\Gamma}}_k$ meets Assumption B.1 by Remark B.1 and Lemmas B.26 and B.27. Define $A_{G,\mathcal{T}} = A_G \cap \mathcal{T}$ as the set corresponding to A_G in (B.3.4) for the grid procedure. We also use $B^{(1)}_{j,(1-\epsilon)G}$ and $B^{(2)}_{j,(1-\epsilon)G}$ as per (B.3.5) and (B.3.6), and ϵ as in (4.5). By the localisation criterion, we only need to consider k in $A_{G,\mathcal{T}}$ or in the set $B_{(1-\epsilon)G,\mathcal{T}} = B_{(1-\epsilon)G} \cap \mathcal{T}$. For the MOSUM Wald procedure, we have that

$$\begin{aligned}\mathsf{P}(\widehat{q} = q) &\geq \mathsf{P}\left(\max_{k \in A_{G,\mathcal{T}}} W_k(G) < D(G,\alpha), \min_{k \in B_{(1-\epsilon)G,\mathcal{T}}} W_k(G) \geq D(G,\alpha)\right) \\ &\geq \mathsf{P}\left(\max_{k \in A_{G,\mathcal{T}}} W_k(G) < D(G,\alpha)\right) + \mathsf{P}\left(\min_{k \in B_{(1-\epsilon)G,\mathcal{T}}} W_k(G) \geq D(G,\alpha)\right) - 1.\end{aligned}$$

Thus, it is sufficient to show that

1. $\mathsf{P}\left(\max_{k \in A_{G,\mathcal{T}}} W_k(G) < D(G,\alpha)\right) \to 1$, and
2. $\mathsf{P}\left(\min_{k \in B_{(1-\epsilon)G,\mathcal{T}}} W_k(G) \geq D(G,\alpha)\right) \to 1$.

Part (1) is simple to see, since $\mathsf{P}\left(\max_{k \in A_G} W_k(G) < D(G,\alpha)\right) \to 1$ is shown to be true in Reckrühm (2019), Theorem 3.1.15, and $A_{G,\mathcal{T}} \subset A_G$, so $\max_{k \in A_{G,\mathcal{T}}} W_k(G) < \max_{k \in A_G} W_k(G)$. Also, part (2) is shown to hold with $r = 1$ without loss of generality, since $\min_{k \in B_{(1-\epsilon)G,\mathcal{T}}} W_k(G) \geq \min_{k \in B_{(1-\epsilon)G}} W_k(G)$. ■

Corollary B.19.1. Let the conditions of Theorem B.19 hold. Using the Wald grid-based procedure,

$$\mathsf{P}(\max_{1 \leq j \leq q} |\widehat{k}_j - k_j| < G) \to 1,$$

and in particular part (b) of Theorem B.2 holds.

Proof. We have that

$$\left\{\min_{k\in B_{(1-\epsilon)G,\mathcal{T}}} W_k(G) \geq D(G,\alpha); \widehat{q} = q\right\} \subset \left\{\max_{1\leq j\leq \hat{q}} \left|\widehat{k}_j - k_j\right| < G; \widehat{q} = q\right\},$$

so by Theorems B.18 and B.19,

$$\mathsf{P}\left(\max_{1\leq j\leq \hat{q}} \left|\widehat{k}_j - k_j\right| < G; \widehat{q} = q\right) \geq \mathsf{P}\left(\min_{k\in B_{(1-\epsilon)G,\mathcal{T}}} W_k(G) \geq D(G,\alpha); \widehat{q} = q\right) \to 1.$$

■

Theorem B.20. Let $q \geq 1$. Let Assumptions 4.1–4.6 hold. Then, using the score grid-based procedure,

(a) $\mathsf{P}(\widehat{q}(\widetilde{\boldsymbol{a}}) = q) = 1$ as $n \to \infty$.

(b) The results hold using $\widetilde{\boldsymbol{a}} = \widehat{\boldsymbol{a}}_{1,n}$.

(c) The results hold using an estimator $\widehat{\boldsymbol{\Sigma}}_k(\widetilde{\boldsymbol{a}})$ as in (4.7) or (4.9) in place of the true $\boldsymbol{\Sigma}_k(\widetilde{\boldsymbol{a}})$.

Proof. For the MOSUM score procedure, we make similar arguments to the proof of Theorem B.19 adapting Reckrühm (2019), Theorem 2.1.8. ■

Corollary B.20.1. Let the conditions of Theorem B.20 hold. Using the score grid-based procedure,

$$\mathsf{P}(\max_{1\leq j\leq q} |\widehat{k}_j(\widetilde{\boldsymbol{a}}) - k_j| < G) \to 1,$$

and in particular part (b) of Theorem 4.2 holds.

Proof. The result holds similarly to Corollary B.19.1. ■

Proof of Theorem 4.4

Proof. Lemma B.14, Theorem B.17, Theorem B.20 and Corollary B.20.1 hold, so the statement holds similarly to the proof of Theorem 4.2. ■

Proof of Theorem B.3

Proof. Lemma B.14, Theorem B.18, Theorem B.19 and Corollary B.19.1 hold, so the statement holds similarly to the proof of Theorem B.2. ■

B.3.5 Estimators

Here we verify that the estimators proposed in Sections 4.3.2 and B.1.1 are consistent in the sense of Assumptions 4.7.

By the definition of the piecewise stationary VAR model, for each $j = 1, \ldots, q+1$ the covariance matrix $\boldsymbol{\Sigma}_{(j)}$ is equal to the Kronecker product of $\boldsymbol{S}$, partitioned into scalar components $s_{ii'}$, and $\boldsymbol{C}_{(j)}$:

$$\boldsymbol{\Sigma}_{(j)} = \boldsymbol{S} \otimes \boldsymbol{C}_{(j)} = \begin{pmatrix} s_{11}\boldsymbol{C}_{(j)} & s_{12}\boldsymbol{C}_{(j)} & \ldots & s_{1p}\boldsymbol{C}_{(j)} \\ s_{21}\boldsymbol{C}_{(j)} & s_{22}\boldsymbol{C}_{(j)} & & \vdots \\ \vdots & & \ddots & \\ s_{p1}\boldsymbol{C}_{(j)} & \ldots & & s_{pp}\boldsymbol{C}_{(j)} \end{pmatrix}.$$

We start by stating a uniform law of large numbers which applies to matrix estimators.

Lemma B.21. (Kirch and Reckrühm, 2022, Lemma D.1 (b)) Let $\nu' > 0$. Let $\{\boldsymbol{F}(\boldsymbol{X}_t, \widetilde{\boldsymbol{a}})\}_{t=1}^n$ be a stationary and ergodic random sequence with values in $\mathbb{R}^{a \times b}$ satisfying

$$\text{(a) } \mathbb{E}\left(\sup_{\widetilde{\boldsymbol{a}} \in \boldsymbol{\Theta}} \|\boldsymbol{F}(\boldsymbol{X}_t, \widetilde{\boldsymbol{a}})\|_F^{2+\nu'}\right) < \infty, \text{ and (b) } \mathbb{E}\left(\sup_{\widetilde{\boldsymbol{a}} \in \boldsymbol{\Theta}} \|\nabla \operatorname{vech}(\boldsymbol{F}(\boldsymbol{X}_t, \widetilde{\boldsymbol{a}}))\|_F^{2+\nu'}\right) < \infty,$$

where $\operatorname{vech} : \mathbb{R}^{a \times b} \to \mathbb{R}^{ab}$. Then we have

$$\sup_{\widetilde{\boldsymbol{a}} \in \boldsymbol{\Theta}} \max_{0 \le k \le n-G} \frac{1}{G} \left\| \sum_{t=k+1}^{k+G} \{\boldsymbol{F}(\boldsymbol{X}_t, \widetilde{\boldsymbol{a}}) - \mathbb{E}(\boldsymbol{F}(\boldsymbol{X}_t, \widetilde{\boldsymbol{a}}))\} \right\|_F = o_P(1).$$

B.3.5.1 Score estimators

Lemma B.22. Let Assumptions 4.1–4.4 hold. $\widehat{\boldsymbol{\Sigma}}_k^{(2)}$ in (4.9) meets Assumption 4.7.

Proof. Part (b) holds with probability 1 when $G \ge d(dp+1)$. We show that part (a) holds by verifying the conditions of Lemma B.21, firstly under H_0. Define the function

$$\boldsymbol{F}(\boldsymbol{X}_t, \widetilde{\boldsymbol{a}}, s, e) = \big(\boldsymbol{H}_t(\widetilde{\boldsymbol{a}}) - \bar{\boldsymbol{H}}_{s,e}(\widetilde{\boldsymbol{a}})\big)\big(\boldsymbol{H}_t(\widetilde{\boldsymbol{a}}) - \bar{\boldsymbol{H}}_{s,e}(\widetilde{\boldsymbol{a}})\big)^\top,$$

with inspection parameter $\widetilde{\boldsymbol{a}}$. We can then express the estimator as

$$\widehat{\boldsymbol{\Sigma}}_k^{(2)} = \frac{1}{2G} \sum_{t=k+1}^{k+G} \boldsymbol{F}(\boldsymbol{X}_t, \widetilde{\boldsymbol{a}}, k+1, k+G) + \frac{1}{2G} \sum_{t=k-G+1}^{k} \boldsymbol{F}(\boldsymbol{X}_t, \widetilde{\boldsymbol{a}}, k-G+1, k).$$

We hence have that, for any $s \le t \le e$,

$$\begin{aligned} &\mathbb{E}\left(\sup_{\widetilde{\boldsymbol{a}} \in \boldsymbol{\Theta}} \|\boldsymbol{F}(\boldsymbol{X}_t, \widetilde{\boldsymbol{a}}, s, e)\|_F^{2+\widetilde{\nu}/2}\right) \\ &\le \mathbb{E} \sup_{\widetilde{\boldsymbol{a}} \in \boldsymbol{\Theta}} \left(\|\boldsymbol{H}_t(\widetilde{\boldsymbol{a}})\|^{2(2+\widetilde{\nu}/2)} + \frac{2}{e-s+1} \|\boldsymbol{H}_t(\widetilde{\boldsymbol{a}}) \bar{\boldsymbol{H}}_{s,e}^\top(\widetilde{\boldsymbol{a}})\|^{2+\widetilde{\nu}/2} + \|\bar{\boldsymbol{H}}_{s,e}(\widetilde{\boldsymbol{a}})\|^{2(2+\widetilde{\nu}/2)} \right) \\ &< \infty, \end{aligned}$$

and condition (a) of Lemma B.21 holds with $\nu' = \tilde{\nu}/2$. Condition (b) of Lemma B.21 holds by application of Lemma B.6 and the product rule. The series $\{\boldsymbol{F}(\boldsymbol{X}_t, \tilde{\boldsymbol{a}}, s, e)\}_{t=1}^n$ is measurable, so the ergodic and stationary properties of $\{\boldsymbol{X}_t\}_{t=1}^n$ carry over.

By Lemma B.21 we have

$$\sup_{\tilde{\boldsymbol{a}} \in \boldsymbol{\Theta}} \max_{0 \le k \le n-G} \frac{1}{G} \left\| \sum_{t=k+1}^{k+G} \{\boldsymbol{F}(\boldsymbol{X}_t, \tilde{\boldsymbol{a}}, k+1, k+G) - \mathbb{E}(\boldsymbol{F}(\boldsymbol{X}_t, \tilde{\boldsymbol{a}}, k+1, k+G))\} \right\|_F = o_P(1).$$

We have $\mathbb{E}(\boldsymbol{F}(\boldsymbol{X}_t, \tilde{\boldsymbol{a}}, s, e)) = \boldsymbol{\Sigma}_{(1)}(\tilde{\boldsymbol{a}})$ for all pairs s, e. Hence, $\left\| \widehat{\boldsymbol{\Sigma}}_k(\tilde{\boldsymbol{a}}) - \boldsymbol{\Sigma}_{(1)}(\tilde{\boldsymbol{a}}) \right\|_F = o_P(1)$ for all k. It holds that $\boldsymbol{\Sigma}_{(1)}(\tilde{\boldsymbol{a}})$ is invertible as a result of Assumptions 4.1–4.2, and $\widehat{\boldsymbol{\Sigma}}_k(\tilde{\boldsymbol{a}})$ is also invertible for sufficiently large G (with probability 1), so by continuity of the matrix root operation and the Continuous Mapping Theorem we have

$$\max_{G \le k \le n-G} \left\| (\widehat{\boldsymbol{\Sigma}}_k(\tilde{\boldsymbol{a}}))^{-1/2} - (\boldsymbol{\Sigma}_{(1)}(\tilde{\boldsymbol{a}}))^{-1/2} \right\|_F = o_P(1),$$

so part (a) holds.

Now letting $q \ge 1$, we verify Assumption 4.7 (a). That

$$\max_{k_{j-1}+G \le k \le k_j - G} \left\| (\widehat{\boldsymbol{\Sigma}}_k(\tilde{\boldsymbol{a}}))^{-1/2} - (\boldsymbol{\Sigma}_k(\tilde{\boldsymbol{a}}))^{-1/2} \right\|_F = o_P(1)$$

follows similarly to the above verification of part (a) under the null. We must consider the effect of estimating at the boundary of regimes, however, since for $t = k_j + 1, \ldots, k_j + d$ the regressor $\mathbb{X}_{t-1}$ can contain at most d observations drawn from the previous regime. Since d is fixed and $G \to \infty$ with n, for each $j = 1, \ldots, q$, and for all k such that $\min(|k - k_j|, |k + G - k_j|) \ge G$, it holds that

$$\sup_{\tilde{\boldsymbol{a}} \in \boldsymbol{\Theta}} \frac{1}{G} \left\| \sum_{t=k+1}^{k+G} \{\boldsymbol{F}(\boldsymbol{X}_t, \tilde{\boldsymbol{a}}, k+1, k+G) - \boldsymbol{\Sigma}_k(\tilde{\boldsymbol{a}})\} \right\| = o_P(1).$$

Again by the Continuous Mapping Theorem, the result holds. ■

We show in Lemma B.23 that the error covariance estimator (4.8) satisfies the conditions for Lemma B.21, and is hence consistent.

Lemma B.23. Let Assumptions 4.1–4.4 hold. Recall $\widehat{\boldsymbol{S}}_k(\tilde{\boldsymbol{a}})$ in (4.8). Then, for all $G \le k \le n - G$,

$$\sup_{\tilde{\boldsymbol{a}} \in \boldsymbol{\Theta}} \left\| \widehat{\boldsymbol{S}}_k(\tilde{\boldsymbol{a}}) - \boldsymbol{S}(\tilde{\boldsymbol{a}}) \right\|_F = o_P(1).$$

Proof. We verify that the conditions of Lemma B.21 hold. Consider the function $\boldsymbol{F}_t(\tilde{\boldsymbol{a}}) = \widehat{\boldsymbol{\varepsilon}}_t(\tilde{\boldsymbol{a}})\widehat{\boldsymbol{\varepsilon}}_t^\top(\tilde{\boldsymbol{a}})$. Over a compact parameter space $\boldsymbol{\Theta}$, we have that $\sup_{\tilde{\boldsymbol{a}} \in \boldsymbol{\Theta}} \|\tilde{\boldsymbol{a}}\|_F < \infty$. We can decompose $\boldsymbol{F}_t(\tilde{\boldsymbol{a}})$ as follows:

$$\begin{aligned} \boldsymbol{F}_t(\tilde{\boldsymbol{a}}) &= (\boldsymbol{X}_t - \tilde{\boldsymbol{a}}\mathbb{X}_{t-1})(\boldsymbol{X}_t - \tilde{\boldsymbol{a}}\mathbb{X}_{t-1})^\top \\ &= \boldsymbol{X}_t\boldsymbol{X}_t^\top - \boldsymbol{X}_t\mathbb{X}_{t-1}^\top\tilde{\boldsymbol{a}}^\top - \tilde{\boldsymbol{a}}\mathbb{X}_{t-1}\boldsymbol{X}_t^\top + \tilde{\boldsymbol{a}}\mathbb{X}_{t-1}\mathbb{X}_{t-1}^\top\tilde{\boldsymbol{a}}^\top, \end{aligned}$$

and this allows us to bound the expectation of the supremum over $\boldsymbol{\Theta}$:

$$\begin{aligned}
&\mathbb{E}\left(\sup_{\widetilde{\boldsymbol{a}}\in\boldsymbol{\Theta}}\|\boldsymbol{F}_t(\widetilde{\boldsymbol{a}})\|_F^{2+\nu'}\right)\\
&=\mathbb{E}\left(\sup_{\widetilde{\boldsymbol{a}}\in\boldsymbol{\Theta}}\left\|\boldsymbol{X}_t\boldsymbol{X}_t^\top-\boldsymbol{X}_t\mathbb{X}_{t-1}^\top\widetilde{\boldsymbol{a}}^\top-\widetilde{\boldsymbol{a}}\mathbb{X}_{t-1}\boldsymbol{X}_t^\top+\widetilde{\boldsymbol{a}}\mathbb{X}_{t-1}\mathbb{X}_{t-1}^\top\widetilde{\boldsymbol{a}}^\top\right\|_F^{2+\nu'}\right)\\
&\leq\mathbb{E}\sup_{\widetilde{\boldsymbol{a}}\in\boldsymbol{\Theta}}\left(\left\|\boldsymbol{X}_t\boldsymbol{X}_t^\top\right\|_F^{2+\nu'}+2\|\boldsymbol{X}_t\mathbb{X}_{t-1}^\top\widetilde{\boldsymbol{a}}^\top\|_F^{2+\nu'}+\|\widetilde{\boldsymbol{a}}\mathbb{X}_{t-1}\mathbb{X}_{t-1}^\top\widetilde{\boldsymbol{a}}^\top\right\|_F^{2+\nu'}\right)\\
&\leq\mathbb{E}\left(\|\boldsymbol{X}_t\boldsymbol{X}_t^\top\|_F^{2+\nu'}\right)+2\,\mathbb{E}\left(\sup_{\widetilde{\boldsymbol{a}}\in\boldsymbol{\Theta}}\|\boldsymbol{X}_t\mathbb{X}_{t-1}^\top\widetilde{\boldsymbol{a}}^\top\|_F^{2+\nu'}\right)+\mathbb{E}\left(\sup_{\widetilde{\boldsymbol{a}}\in\boldsymbol{\Theta}}\|\widetilde{\boldsymbol{a}}\mathbb{X}_{t-1}\mathbb{X}_{t-1}^\top\widetilde{\boldsymbol{a}}^\top\|_F^{2+\nu'}\right)\\
&\leq\mathbb{E}\left(\|\boldsymbol{X}_t\boldsymbol{X}_t^\top\|_F^{2+\nu'}\right)+2\sup_{\widetilde{\boldsymbol{a}}\in\boldsymbol{\Theta}}\|\widetilde{\boldsymbol{a}}\|_F^{2+\nu'}\,\mathbb{E}\left(\|\boldsymbol{X}_t\mathbb{X}_{t-1}^\top\|_F^{2+\nu'}\right)+\sup_{\widetilde{\boldsymbol{a}}\in\boldsymbol{\Theta}}\|\widetilde{\boldsymbol{a}}\|_F^{4+2\nu'}\,\mathbb{E}\left(\|\mathbb{X}_{t-1}\mathbb{X}_{t-1}^\top\|_F^{2+\nu'}\right)\\
&<\infty,
\end{aligned}\tag{B.3.8}$$

using the Cauchy-Schwartz inequality, and Assumption 4.1 with $\widetilde{\nu}=2\nu'$. Then we consider the function

$$\begin{aligned}
\boldsymbol{G}_t(\widetilde{\boldsymbol{a}},s,e)&=\left(\widehat{\boldsymbol{\varepsilon}}_t(\widetilde{\boldsymbol{a}})-\bar{\boldsymbol{\varepsilon}}_{s,e}(\widetilde{\boldsymbol{a}})\right)\left(\widehat{\boldsymbol{\varepsilon}}_t(\widetilde{\boldsymbol{a}})-\bar{\boldsymbol{\varepsilon}}_{s,e}(\widetilde{\boldsymbol{a}})\right)^\top\\
&=\widehat{\boldsymbol{\varepsilon}}_t(\widetilde{\boldsymbol{a}})\widehat{\boldsymbol{\varepsilon}}_t(\widetilde{\boldsymbol{a}})^\top-\widehat{\boldsymbol{\varepsilon}}_t(\widetilde{\boldsymbol{a}})\bar{\boldsymbol{\varepsilon}}_{s,e}^\top(\widetilde{\boldsymbol{a}})-\bar{\boldsymbol{\varepsilon}}_{s,e}(\widetilde{\boldsymbol{a}})\widehat{\boldsymbol{\varepsilon}}_t(\widetilde{\boldsymbol{a}})+\bar{\boldsymbol{\varepsilon}}_{s,e}(\widetilde{\boldsymbol{a}})\bar{\boldsymbol{\varepsilon}}_{s,e}^\top(\widetilde{\boldsymbol{a}}),
\end{aligned}$$

which induces the estimator

$$\widehat{\boldsymbol{S}}_k=\frac{1}{2G}\left(\sum_{t=k+1}^{k+G}\boldsymbol{G}_t(\widetilde{\boldsymbol{a}},k+1,k+G)+\sum_{t=k-G+1}^{k}\boldsymbol{G}_t(\widetilde{\boldsymbol{a}},k-G+1,k)\right).$$

We have that, for $\nu'>0$,

$$\mathbb{E}\left(\sup_{\widetilde{\boldsymbol{a}}\in\boldsymbol{\Theta}}\|\widehat{\boldsymbol{\varepsilon}}_t(\widetilde{\boldsymbol{a}})\bar{\boldsymbol{\varepsilon}}_{s,e}(\widetilde{\boldsymbol{a}})^\top\|_F^{2+\nu'}\right)\leq\mathbb{E}\left(\sup_{\widetilde{\boldsymbol{a}}\in\boldsymbol{\Theta}}\|\widehat{\boldsymbol{\varepsilon}}_t(\widetilde{\boldsymbol{a}})\|^{2+\nu'}\|\bar{\boldsymbol{\varepsilon}}_{s,e}(\widetilde{\boldsymbol{a}})^\top\|^{2+\nu'}\right)\leq\mathbb{E}\left(\sup_{\widetilde{\boldsymbol{a}}\in\boldsymbol{\Theta}}\|\widehat{\boldsymbol{\varepsilon}}_t(\widetilde{\boldsymbol{a}})\|^{2(2+\nu')}\right).\tag{B.3.9}$$

We verify the $2+\nu'$-th power is finite:

$$\begin{aligned}
\mathbb{E}\left(\sup_{\widetilde{\boldsymbol{a}}\in\boldsymbol{\Theta}}\|\boldsymbol{G}_t(\widetilde{\boldsymbol{a}},s,e)\|_F^{2+\nu'}\right)&\leq\mathbb{E}\left(\sup_{\widetilde{\boldsymbol{a}}\in\boldsymbol{\Theta}}\|\boldsymbol{F}_t(\widetilde{\boldsymbol{a}})\|_F^{2+\nu'}+2\|\widehat{\boldsymbol{\varepsilon}}_t(\widetilde{\boldsymbol{a}})\bar{\boldsymbol{\varepsilon}}_{s,e}(\widetilde{\boldsymbol{a}})^\top\|_F^{2+\nu'}+\|\bar{\boldsymbol{F}}_{s,e}(\widetilde{\boldsymbol{a}})\|_F^{2+\nu'}\right)\\
&\leq\mathbb{E}\left(\sup_{\widetilde{\boldsymbol{a}}\in\boldsymbol{\Theta}}\|\boldsymbol{F}_t(\widetilde{\boldsymbol{a}})\|_F^{2+\nu'}\right)+2\,\mathbb{E}\left(\sup_{\widetilde{\boldsymbol{a}}\in\boldsymbol{\Theta}}\|\widehat{\boldsymbol{\varepsilon}}_t(\widetilde{\boldsymbol{a}})\bar{\boldsymbol{\varepsilon}}_{s,e}(\widetilde{\boldsymbol{a}})^\top\|_F^{2+\nu'}\right)+\mathbb{E}\left(\sup_{\boldsymbol{a}\in\boldsymbol{\Theta}}\|\bar{\boldsymbol{F}}_{s,e}(\widetilde{\boldsymbol{a}})\|_F^{2+\nu'}\right)\\
&<\infty,
\end{aligned}$$

which follows by (B.3.8) and (B.3.9), so condition (a) of Lemma B.21 holds. Condition (b) of Lemma B.21 follows similarly, using the product rule.

This has expectation $\mathbb{E}(\boldsymbol{G}(\widetilde{\boldsymbol{a}},s,e))=(1-(e-s+1)^{-1})\,\mathbb{E}(\boldsymbol{F}_t(\widetilde{\boldsymbol{a}}))=(1-(e-s+1)^{-1})\frac{n-dp-1}{n}\boldsymbol{S}(\widetilde{\boldsymbol{a}})$. Moreover, since both functions are measurable, the ergodic and stationary properties of $\boldsymbol{X}_t$ carry over to the function sequences. Hence, by Lemma B.21, we have that

$$\sup_{\widetilde{\boldsymbol{a}}\in\boldsymbol{\Theta}}\max_{0\leq k\leq n-G}\frac{1}{G}\left\|\sum_{t=k+1}^{k+G}\{\boldsymbol{F}_t(\widetilde{\boldsymbol{a}})-\mathbb{E}(\boldsymbol{F}_t(\widetilde{\boldsymbol{a}}))\}\right\|_F=o_P(1),\tag{B.3.10}$$

and

$$\sup_{\widetilde{\boldsymbol{a}}\in\boldsymbol{\Theta}}\max_{0\le k\le n-G}\frac{1}{G}\left\|\sum_{t=k+1}^{k+G}\{\boldsymbol{G}_t(\widetilde{\boldsymbol{a}},k+1,k+G)-\mathbb{E}(\boldsymbol{G}_k(\widetilde{\boldsymbol{a}},k+1,k+G))\}\right\|_F=o_P(1).$$

For any k, we have

$$\begin{aligned}\left\|\widehat{\boldsymbol{S}}_k^{(2)}(\widetilde{\boldsymbol{a}})-\boldsymbol{S}(\widetilde{\boldsymbol{a}})\right\|_F&=\left\|\frac{1}{2G}\sum_{t=k+1}^{k+G}\boldsymbol{G}_t(\widetilde{\boldsymbol{a}},k+1,k+G)+\frac{1}{2G}\sum_{t=k-G+1}^{k}\boldsymbol{G}_t(\widetilde{\boldsymbol{a}},k-G+1,k)-\boldsymbol{S}(\widetilde{\boldsymbol{a}})\right\|_F\\&\le\left\|\frac{1}{2G}\sum_{t=k+1}^{k+G}\boldsymbol{G}_t(\widetilde{\boldsymbol{a}},k+1,k+G)-\frac{1}{2}\mathbb{E}(\boldsymbol{G}_k(\widetilde{\boldsymbol{a}},k+1,k+G))\right\|_F\\&+\left\|\frac{1}{2}\mathbb{E}(\boldsymbol{G}_k(\widetilde{\boldsymbol{a}},k+1,k+G))-\frac{1}{2}\boldsymbol{S}(\widetilde{\boldsymbol{a}})\right\|_F\\&+\left\|\frac{1}{2G}\sum_{t=k-G+1}^{k}\boldsymbol{G}_t(\widetilde{\boldsymbol{a}},k-G+1,k)-\frac{1}{2}\mathbb{E}(\boldsymbol{G}_k(\widetilde{\boldsymbol{a}},k-G+1,k))\right\|_F\\&+\left\|\frac{1}{2}\mathbb{E}(\boldsymbol{G}_k(\widetilde{\boldsymbol{a}},k-G+1,k))-\frac{1}{2}\boldsymbol{S}(\widetilde{\boldsymbol{a}})\right\|_F=o_P(1),\end{aligned}$$

and we are done. ■

In Lemma B.24 we use Lemma B.23 to show that the covariance estimator $\widehat{\boldsymbol{\Sigma}}_k(\widetilde{\boldsymbol{a}})$ defined by (4.8) is consistent in the sense of Assumption 4.7.

Lemma B.24. Let Assumptions 4.1–4.4 hold. $\widehat{\boldsymbol{\Sigma}}_k^{(1)}(\widetilde{\boldsymbol{a}})$ in (4.7) with the error covariance estimator (4.8) meets Assumption 4.7.

Proof. Let $\widehat{\boldsymbol{\Sigma}}_k(\widetilde{\boldsymbol{a}})=\widehat{\boldsymbol{\Sigma}}_k^{(1)}(\widetilde{\boldsymbol{a}})$. We begin by showing this meets Assumption 4.7 (b). By the construction of (4.7), for each k we have the inequality

$$\left\|(\widehat{\boldsymbol{\Sigma}}_k(\widetilde{\boldsymbol{a}}))^{-1/2}\right\|_F=\left\|(\widehat{\boldsymbol{S}}_k(\widetilde{\boldsymbol{a}}))^{-1/2}\otimes\widehat{\boldsymbol{C}}_{k-G+1,k+G}^{-1/2}\right\|_F\le\left\|(\widehat{\boldsymbol{S}}_k(\widetilde{\boldsymbol{a}}))^{-1/2}\right\|_F\left\|\widehat{\boldsymbol{C}}_{k-G+1,k+G}^{-1/2}\right\|_F.$$

For sufficiently large n and G, both $\widehat{\boldsymbol{S}}_k(\widetilde{\boldsymbol{a}})$ and $\widehat{\boldsymbol{C}}_{k-G+1,k+G}$ are positive definite with probability 1. As a result, the inverse square roots of both exist, so $\max_{G\le k\le n-G}\left\|(\widehat{\boldsymbol{\Sigma}}_k(\widetilde{\boldsymbol{a}}))^{-1/2}\right\|_F<\infty$.

Next we verify Assumption 4.7 (a). By Lemma B.23, the Continuous Mapping Theorem, and continuity of the matrix root operator, we have for each k

$$\|(\widehat{\boldsymbol{S}}_k(\widetilde{\boldsymbol{a}}))^{-1/2}-(\boldsymbol{S}(\widetilde{\boldsymbol{a}}))^{-1/2}\|_F=o_P(1).$$

By Kirch and Reckrühm (2022), Lemma D.1 (a) we have that

$$\begin{aligned}&\max_{k_{j-1}+G\le k\le k_j-G}\left\|\widehat{\boldsymbol{C}}_{k-G+1,k+G}-\boldsymbol{C}_{(j)}\right\|_F\\&=\max_{k_{j-1}+G\le k\le k_j-G}\frac{1}{\sqrt{2G}}\left\|\sum_{t=k-G}^{k+G-1}(\mathbb{X}_t\mathbb{X}_t^\top-\boldsymbol{C}_{(j)})\right\|_F\\&=O_P\left(\frac{\sqrt{\log(n/G)}}{\sqrt{G}}\right)=o_P(1)\end{aligned}$$

for each $j = 1, \ldots, q+1$. Then, since $\widehat{\boldsymbol{C}}_{k,k+G}$ is positive definite with probability 1 for sufficiently large G, and by continuity of the matrix inverse root operator, we have $\left\|\widehat{\boldsymbol{C}}_{k,k+G}^{-1/2}\right\|_F = o_P(1)$ uniformly in k. Using these facts we find, for each $j = 1, \ldots q+1$,

$$
\begin{aligned}
&\max_{k_{j-1}+G \le k \le k_j - G} \left\| (\widehat{\boldsymbol{\Sigma}}_k(\widetilde{\boldsymbol{a}}))^{-1/2} - (\boldsymbol{\Sigma}_{(j)}(\widetilde{\boldsymbol{a}}))^{-1/2} \right\|_F \\
&= \max_{k_{j-1}+G \le k \le k_j - G} \left\| (\widehat{\boldsymbol{S}}_k(\widetilde{\boldsymbol{a}}) \otimes \widehat{\boldsymbol{C}}_{k-G+1,k+G})^{-1/2} - (\boldsymbol{S}_k(\widetilde{\boldsymbol{a}}) \otimes \boldsymbol{C}_{(j)})^{-1/2} \right\|_F \\
&= \max_{k_{j-1}+G \le k \le k_j - G} \left\| (\widehat{\boldsymbol{S}}_k(\widetilde{\boldsymbol{a}}))^{-1/2} \otimes \widehat{\boldsymbol{C}}_{k-G+1,k+G}^{-1/2} - (\boldsymbol{S}_k(\widetilde{\boldsymbol{a}}))^{-1/2} \otimes \boldsymbol{C}_{(j)}^{-1/2} \right\|_F \\
&= \max_{k_{j-1}+G \le k \le k_j - G} \left\| (\boldsymbol{S}_k(\widetilde{\boldsymbol{a}}) + o_P(1))^{-1/2} \otimes (\boldsymbol{C}_{(j)} + o_P(1))^{-1/2} - (\boldsymbol{S}_k(\widetilde{\boldsymbol{a}}))^{-1/2} \otimes \boldsymbol{C}_{(j)}^{-1/2} \right\|_F \\
&= o_P(1).
\end{aligned}
$$

■

B.3.5.2 Wald estimators

Next we verify conditions for the MOSUM Wald procedure. In finding an estimator for $\boldsymbol{\Gamma}_k$, we observe Remark B.1.

Remark B.1. Assumption B.1 on $\widehat{\boldsymbol{\Gamma}}_k$ is fulfilled if

$$
\widehat{\boldsymbol{\Gamma}}_k = \widehat{\boldsymbol{V}}_k^{-1} \widehat{\boldsymbol{\Sigma}}_k \left(\widehat{\boldsymbol{V}}_k^{-1} \right)^\top, \tag{B.3.11}
$$

where $\widehat{\boldsymbol{\Sigma}}_k$ satisfies Assumption 4.7, and $\widehat{\boldsymbol{V}}_k$ satisfies the following:

(a)
$$
\max_{k_{j-1}+G \le k \le k_j - G} \left\| \widehat{\boldsymbol{V}}_k^{-1/2} - \boldsymbol{V}_k^{-1/2} \right\|_F = o_P\left(\log(n/G)^{-1} \right)
$$
for any $j = 1, \ldots, q+1$.

(b) For any $j = 1, \ldots, q$ it holds that
$$
\max_{k:|k-k_j|<G} \left\| \widehat{\boldsymbol{V}}_k^{1/2} \right\|_F < \infty, \text{ and } \max_{k:|k-k_j|<G} \left\| \widehat{\boldsymbol{V}}_k^{-1/2} \right\|_F < \infty.
$$

This is a direct result of Remarks 3.1.9. and 3.1.13. of Reckrühm (2019).

Lemma B.25. Let Assumptions 4.1–4.4 hold. $\widehat{\boldsymbol{V}}_k$ in (B.1.4) meets the conditions of Remark B.1.

Proof. This follows by Lemmas B.4–B.5. The boundary issue is handled similarly to Lemma B.10. ■

We show in Lemma B.26 that the estimator in (B.1.7) meets the null and alternative consistency conditions.

Lemma B.26. Let Assumptions 4.1–4.4 hold. The estimator $\widehat{\boldsymbol{\Sigma}}_k^{(4)}$ in (B.1.7) meets Assumption 4.7.

Proof. The necessary arguments here are the same as for Lemma B.22. ■

Finally, we show in Lemma B.27 that the estimator in (B.1.5) is consistent under the null and the alternative.

Lemma B.27. Let Assumptions 4.1–4.4 hold. The estimator $\widehat{\boldsymbol{\Sigma}}_k^{(W)}$ in (B.1.5) with error variance estimator $\widehat{\boldsymbol{S}}^{(W)}$ in (B.1.6) meets Assumption 4.7.

Proof. Using $\widehat{\boldsymbol{\varepsilon}}_t(\widehat{\boldsymbol{a}}_{k-G+1,k})$ and $\widehat{\boldsymbol{\varepsilon}}_t(\widehat{\boldsymbol{a}}_{k+1,k+G})$, the necessary arguments here are the same as for Lemmas B.23 and B.24. ■

B.4 η-criterion

We define an alternative location procedure to the ϵ-criterion of (4.5) and (B.1.3). We refer to this as the η-criterion as per Meier et al. (2021), also called the 'maximum-check' in Eichinger and Kirch (2018). For some $\eta \in (0,1)$, the point $\widehat{k}$ is a change point estimator if and only if it exceeds the critical value D and the detector at $\widehat{k}$ is the maximal point within its own ηG window, i.e. for the score detector

$$\widehat{k}_j = \underset{\widehat{k}_j - \eta G \leq k \leq \widehat{k}_j + \eta G}{\operatorname{argmax}} \widehat{T}_k(G, \widetilde{\boldsymbol{a}}), \text{ and } \widehat{T}_{\widehat{k}_j}(G, \widetilde{\boldsymbol{a}}) \geq D(G, \alpha), \tag{B.4.1}$$

and for the Wald detector,

$$\widehat{k}_j = \underset{\widehat{k}_j - \eta G \leq k \leq \widehat{k}_j + \eta G}{\operatorname{argmax}} \widehat{W}_k(G), \text{ and } \widehat{W}_{\widehat{k}_j}(G) \geq D(G, \alpha). \tag{B.4.2}$$

Then $\widehat{q}$ is the number of these estimators. This is a more suitable procedure in cases where the detector does not drop below the threshold between change points, as may be the case with large bandwidths.

As a result of Lemma B.28, Theorems 4.2 and B.2 hold when changes are localised using the η-criterion.

Lemma B.28. Fix $\eta \in (0,1)$. Using the localisation procedures defined in (B.4.1) and (B.4.2), the results of Theorems 4.2 and B.2 hold.

Proof. The result follows by similar arguments to the proof of Lemma D.1 of Cho and Kirch (2021b).

■

B.5 Computational considerations

B.5.1 Sequential estimation

Here we describe how to use sequential estimation for reducing computational costs. The computation of the detector $\widehat{T}_k(G, \widetilde{\boldsymbol{a}})$ in (4.3) relies on the difference vector $\boldsymbol{m}_k(G, \widetilde{\boldsymbol{a}})$, which we can calculate with rolling sums. This also requires access to $(\widehat{\boldsymbol{\Sigma}}_{k+1}(\widetilde{\boldsymbol{a}}))^{-1}$ for each $k = G+d, \ldots, n-G$. For the estimators (4.7) and (4.9), rather than directly calculating $(n-2G)$-many matrices, we can use the recursive relationship between $(\widehat{\boldsymbol{\Sigma}}_k(\widetilde{\boldsymbol{a}}))^{-1}$ and $(\widehat{\boldsymbol{\Sigma}}_{k+1}(\widetilde{\boldsymbol{a}}))^{-1}$, via the Woodbury identity (Woodbury, 1950), which reduces the operations required in the finite sample. For instance, in the score estimator in (4.9), we have that

$$\widehat{\boldsymbol{\Sigma}}_{k+1}(\widetilde{\boldsymbol{a}}) = \widehat{\boldsymbol{\Sigma}}_k(\widetilde{\boldsymbol{a}}) + \boldsymbol{H}_{k+1+G}(\widetilde{\boldsymbol{a}})\boldsymbol{H}_{k+1+G}(\widetilde{\boldsymbol{a}})^\top - \boldsymbol{H}_{k-G}(\widetilde{\boldsymbol{a}})\boldsymbol{H}_{k-G}(\widetilde{\boldsymbol{a}})^\top - 2\boldsymbol{H}_k(\widetilde{\boldsymbol{a}})\boldsymbol{H}_k(\widetilde{\boldsymbol{a}})^\top.$$

Denoting

$$\boldsymbol{U} = (\boldsymbol{H}_{k+1+G}(\widetilde{\boldsymbol{a}}), -\boldsymbol{H}_{k-G}(\widetilde{\boldsymbol{a}}), 2\boldsymbol{H}_k(\widetilde{\boldsymbol{a}})), \quad \boldsymbol{V} = (\boldsymbol{H}_{k+1+G}(\widetilde{\boldsymbol{a}}), \boldsymbol{H}_{k-G}(\widetilde{\boldsymbol{a}}), \boldsymbol{H}_k(\widetilde{\boldsymbol{a}}))^\top,$$

we can express this update as $\widehat{\boldsymbol{\Sigma}}_{k+1}(\widetilde{\boldsymbol{a}}) = \widehat{\boldsymbol{\Sigma}}_k(\widetilde{\boldsymbol{a}}) + \boldsymbol{U}\boldsymbol{V}$, so that

$$(\widehat{\boldsymbol{\Sigma}}_{k+1}(\widetilde{\boldsymbol{a}}))^{-1} = (\widehat{\boldsymbol{\Sigma}}_k(\widetilde{\boldsymbol{a}}))^{-1} - (\widehat{\boldsymbol{\Sigma}}_k(\widetilde{\boldsymbol{a}}))^{-1}\boldsymbol{U}(\boldsymbol{I}_3 + \boldsymbol{V}(\widehat{\boldsymbol{\Sigma}}_k(\widetilde{\boldsymbol{a}}))^{-1}\boldsymbol{U})^{-1}\boldsymbol{V}(\widehat{\boldsymbol{\Sigma}}_k(\widetilde{\boldsymbol{a}}))^{-1},$$

and the iteration begins from $(\widehat{\boldsymbol{\Sigma}}_{G+d}(\widetilde{\boldsymbol{a}}))^{-1}$. Similar arguments apply to $\widehat{\boldsymbol{C}}_k$ and $\widehat{\boldsymbol{S}}_k(\widetilde{\boldsymbol{a}})$ for the estimator in (4.7).

Making similar arguments for the Wald estimators (B.1.5) and (B.1.7) is non-trivial, since each evaluation of $\widehat{\boldsymbol{\Sigma}}_k$ depends on new parameters $\widehat{\boldsymbol{a}}_{k+1,k+G}$ and $\widehat{\boldsymbol{a}}_{k-G+1,k}$. For the former, we can calculate the $\widehat{\boldsymbol{C}}_k$ sequentially, but not the error covariance estimator $\widehat{\boldsymbol{S}}_k$. Procedural updating of $\widehat{\boldsymbol{a}}_{s,e}$ is possible, so we leave as further work to find a recursive formula relating the covariance estimators.

B.5.2 Parallelisation

Both procedures are easy to parallelise. The evaluation of a statistic at each time point can be sent to a different worker and then merged, so we can divide the overall complexity of evaluation (see Table 5.2) by the number of workers L. When not using sequential estimation, this can be split in any order, while when using updating estimators, the sample can be split into L many contiguous ordered subamples.

APPENDIX C

APPENDIX TO SEGMENTING AND FORECASTING NONSTATIONARY FACTOR-AUGMENTED REGRESSION MODELS

C.1 Estimators

C.1.1 Factor VAR

To evaluate the score detector, we require data-driven choices for the inspection parameter $\widetilde{\boldsymbol{a}}$ and the covariance matrix $\boldsymbol{\Sigma}_k(\widetilde{\boldsymbol{a}})$. As mentioned in Section 5.3.1, for $\widetilde{\boldsymbol{a}}$ we use the global least-squares solution $\widehat{\boldsymbol{a}}_{1,n}$ solving (5.8) with $w_t = 1$.

For $\boldsymbol{\Sigma}_k(\widetilde{\boldsymbol{a}})$, we need an estimator which meets Assumption 5.11. We propose two estimators, which combine an estimator for $\boldsymbol{C}_{(j)} = \mathrm{Cov}(\mathbb{F}_t^{(j)})$ and a choice for the error covariance $\boldsymbol{S}(\widetilde{\boldsymbol{a}}) = \mathrm{Cov}(\widehat{\boldsymbol{\eta}}_t(\widetilde{\boldsymbol{a}}))$ where $\widehat{\boldsymbol{\eta}}_t(\widetilde{\boldsymbol{a}}) = \widehat{\boldsymbol{F}}_t - \widetilde{\boldsymbol{a}}\widehat{\mathbb{F}}_{t-1}$. For matrices $\boldsymbol{A},\boldsymbol{B}$ such that $\boldsymbol{A} \in \mathbb{R}^{n\times m}$, let

$$\boldsymbol{A}\otimes\boldsymbol{B} = \begin{pmatrix} a_{11}\boldsymbol{B} & a_{12}\boldsymbol{B} & \dots & a_{1m}\boldsymbol{B} \\ a_{21}\boldsymbol{B} & a_{22}\boldsymbol{B} & & \vdots \\ \vdots & & \ddots & \\ a_{n1}\boldsymbol{B} & \dots & & a_{nm}\boldsymbol{B} \end{pmatrix}$$

be the Kronecker product. We have the estimator

$$\widehat{\boldsymbol{\Sigma}}_k(\widetilde{\boldsymbol{a}}) = \widehat{\boldsymbol{S}}_k(\widetilde{\boldsymbol{a}}) \otimes \widehat{\boldsymbol{C}}_{k-G+1,k+G}, \tag{C.1.1}$$

where

$$\widehat{\boldsymbol{C}}_{s,e} = \frac{1}{e-s+1}\sum_{t=s}^{e} \widehat{\mathbb{F}}_{t-1}\widehat{\mathbb{F}}_{t-1}^{\top}.$$

The matrix $\widehat{\boldsymbol{\Sigma}}_k(\widetilde{\boldsymbol{a}})$ has submatrices of the form $\widehat{s}_{ij}(\widetilde{\boldsymbol{a}})\widehat{\boldsymbol{C}}_{k-G+1,k+G}$. For the error covariance $\boldsymbol{S}(\widetilde{\boldsymbol{a}})$ we have the estimator

$$\begin{aligned}\widehat{\boldsymbol{S}}_k(\widetilde{\boldsymbol{a}}) = \frac{1}{2G}\Bigg(&\sum_{t=k-G+1}^{k} \big(\widehat{\boldsymbol{\eta}}_t(\widetilde{\boldsymbol{a}}) - \bar{\boldsymbol{\eta}}_{k-G+1,k}(\widetilde{\boldsymbol{a}})\big)\big(\widehat{\boldsymbol{\eta}}_t(\widetilde{\boldsymbol{a}}) - \bar{\boldsymbol{\eta}}_{k-G+1,k}(\widetilde{\boldsymbol{a}})\big)^\top \\ &+ \sum_{t=k+1}^{k+G} \big(\widehat{\boldsymbol{\eta}}_t(\widetilde{\boldsymbol{a}}) - \bar{\boldsymbol{\eta}}_{k+1,k+G}(\widetilde{\boldsymbol{a}})\big)\big(\widehat{\boldsymbol{\eta}}_t(\widetilde{\boldsymbol{a}}) - \bar{\boldsymbol{\eta}}_{k+1,k+G}(\widetilde{\boldsymbol{a}})\big)^\top\Bigg), \qquad \text{(C.1.2)}\\ \text{where } \bar{\boldsymbol{\eta}}_{s,e}(\widetilde{\boldsymbol{a}}) &= \frac{1}{e-s+1}\sum_{t=s}^{e}\widehat{\boldsymbol{\eta}}_t(\widetilde{\boldsymbol{a}}).\end{aligned}$$

C.1.1.1 Consistency

As $\boldsymbol{F}_t^{(j)}, j = 1,\dots,q+1$ is only identifiable up to the rotation $\widetilde{\boldsymbol{F}}_t^{(j)} = \boldsymbol{R}^\top \boldsymbol{F}_t^{(j)}$, the representation of (5.3) as

$$\boldsymbol{F}_t^{(j)} = \sum_{l=1}^{d} \boldsymbol{A}_l^{(j)} \boldsymbol{F}_{t-l}^{(j)} + \boldsymbol{\eta}_t$$

with transfer matrices $\boldsymbol{A}_l^{(j)}, l = 1,\dots,d$ implies

$$\widetilde{\boldsymbol{F}}_t^{(j)} = \boldsymbol{R}^\top \boldsymbol{F}_t^{(j)} = \sum_{l=1}^{d} \boldsymbol{R}^\top \boldsymbol{A}_l^{(j)} \boldsymbol{R} \widetilde{\boldsymbol{F}}_{t-l}^{(j)} + \boldsymbol{R}^\top \boldsymbol{\eta}_t.$$

As such, we are interested in the parameter vectors $\bar{\boldsymbol{a}}_j$ corresponding to the matrices $\bar{\boldsymbol{A}}_l^{(j)} = \boldsymbol{R}^\top \boldsymbol{A}_l^{(j)} \boldsymbol{R}, l = 1,\dots,d$. Define the mixture

$$\boldsymbol{a} = \frac{1}{n}\sum_{j=1}^{q+1} (k_j - k_{j-1})\bar{\boldsymbol{a}}_j \qquad \text{(C.1.3)}$$

Lemma C.1 states that the mixture (C.1.3) is consistently estimated by the least squares estimator, and Lemma C.2 states that the divergence of the difference vector $\boldsymbol{m}_k(G, \widehat{\boldsymbol{a}}_{1,n})$ or $\widehat{\boldsymbol{m}}_k(G, \widehat{\boldsymbol{a}}_{1,n})$ from $\boldsymbol{m}_k(G, \boldsymbol{a})$ must be bounded in probability.

Let $\lambda_{\min}, \lambda_{\max}$ respectively denote the minimum and maximum eigenvalue operators. Denote the p-dimensional normalised eigenvectors corresponding to the j-th largest eigenvalues of $\widehat{\boldsymbol{\Gamma}}_x$ and $\boldsymbol{\Gamma}_\chi$, by $\widehat{\mathbf{w}}_{x,j}$ and $\mathbf{w}_{\chi,j}$, respectively. We further define the $p \times r$ matrices $\widehat{\mathbf{W}}_x = \big[\widehat{\mathbf{w}}_{x,1},\dots,\widehat{\mathbf{w}}_{x,r}\big]$ and $\mathbf{W}_\chi = \big[\mathbf{w}_{\chi,1},\dots,\mathbf{w}_{\chi,r}\big]$. Let $\widehat{\boldsymbol{M}}_x = \operatorname{diag}(\widehat{\mu}_{x,1},\dots,\widehat{\mu}_{x,r})$ and $\boldsymbol{M}_\chi = \operatorname{diag}(\mu_{\chi,1},\dots,\mu_{\chi,r})$.

Lemma C.1. Let Assumptions 5.1–5.5 hold. Then,

$$\|\widehat{\boldsymbol{a}}_{1,n} - \boldsymbol{a}\| = O_P\left(\sqrt{\frac{\log p}{n}} \vee \frac{1}{p}\right).$$

Consequently, Assumption 5.13 (a) holds for $\widetilde{\boldsymbol{a}} = \widehat{\boldsymbol{a}}_{1,n}$ with high probability.

Proof. Define

$$\boldsymbol{\Gamma}_\chi^{acv} = \begin{pmatrix} \boldsymbol{\Gamma}_\chi(0) & \boldsymbol{\Gamma}_\chi(1) & \dots & \boldsymbol{\Gamma}_\chi(d-1) \\ \boldsymbol{\Gamma}_\chi(1) & \boldsymbol{\Gamma}_\chi(0) & & \vdots \\ \vdots & & \ddots & \\ \boldsymbol{\Gamma}_\chi(d-1) & \dots & & \boldsymbol{\Gamma}_\chi(0) \end{pmatrix}.$$

and

$$\boldsymbol{\gamma}_\chi^{acv} = \begin{pmatrix} \boldsymbol{\Gamma}_\chi(1) \\ \boldsymbol{\Gamma}_\chi(2) \\ \vdots \\ \boldsymbol{\Gamma}_\chi(d) \end{pmatrix}.$$

with the sample equivalents $\widehat{\boldsymbol{\Gamma}}_x^{acv}$ and $\widehat{\boldsymbol{\gamma}}_x^{acv}$. Note that $\widehat{\boldsymbol{\Gamma}}_x^{acv}$ and $\widehat{\boldsymbol{C}}_{d+1,n}$ are similar but averages are taken over different intervals. We define $\boldsymbol{\Gamma}_F^{acv}$ and $\boldsymbol{\gamma}_F^{acv}$ with the block matrix entries $\boldsymbol{\Gamma}_F(\ell)$, where

$$\boldsymbol{\Gamma}_F(0) = \frac{1}{p}\boldsymbol{M}_\chi,$$

and $\boldsymbol{\Gamma}_F(\ell)$ for $\ell = 1,\dots,d$ are defined similarly using a singular value decomposition. These have sample equivalents $\widehat{\boldsymbol{\Gamma}}_F^{acv}$ and $\widehat{\boldsymbol{\gamma}}_F^{acv}$ defined with block matrix entries $\widehat{\boldsymbol{\Gamma}}_F(\ell)$, where

$$\widehat{\boldsymbol{\Gamma}}_F(0) = \frac{1}{p}\widehat{\boldsymbol{M}}_x,$$

and $\widehat{\boldsymbol{\Gamma}}_F(\ell)$ for $\ell = 1,\dots,d$ are defined similarly. Recall that the PCA estimator is only consistent up to the rotation $\boldsymbol{R}$, so we define $\widetilde{\boldsymbol{\Gamma}}_F(\ell) = \boldsymbol{R}^\top \boldsymbol{\Gamma}_F(\ell)\boldsymbol{R}$ as the blockwise entries of $\widetilde{\boldsymbol{\Gamma}}_F^{acv}$ and $\widetilde{\boldsymbol{\gamma}}_F^{acv}$. Then we define the OLS estimator as $\widehat{\boldsymbol{a}}_{1,n} = (\widehat{\boldsymbol{\Gamma}}_F^{acv})^{-1}\widehat{\boldsymbol{\gamma}}_F^{acv}$, while $\boldsymbol{a} = (\widetilde{\boldsymbol{\Gamma}}_F^{acv})^{-1}\widetilde{\boldsymbol{\gamma}}_F^{acv}$.

As a result of (C.2.1) and Weyl's inequality, for all $1 \le j \le r$ we have

$$\frac{1}{p}|\widehat{\mu}_{\chi,j} - \mu_{\chi,j}| \le \frac{1}{p}\|\widehat{\boldsymbol{\Gamma}}_x - \boldsymbol{\Gamma}_\chi\| = O_P\left(\sqrt{\frac{\log p}{n}} \vee \frac{1}{p}\right),$$

and from Remark 5.4 (i) we have $\mu_{\chi,r}/p > \underline{\gamma}$ for large enough n, so

$$p^{-1}\widehat{\mu}_{\chi,r} > \underline{\gamma}(1 + o_P(1)).$$

Hence, $\widehat{\boldsymbol{\Gamma}}_F(0)$ has its minimum eigenvalue bounded below with high probability, and $\widehat{\boldsymbol{\Gamma}}_F^{acv}$ has the same eigenvalues by the properties of block matrices, so $\widehat{\boldsymbol{\Gamma}}_F^{acv}$ is invertible for large enough n and p. Then

$$\|(\widehat{\boldsymbol{\Gamma}}_F^{acv})^{-1}\| = \frac{p}{\widehat{\mu}_{\chi,r}} < \frac{1}{\underline{\gamma}(1 + o_P(1))} = O_P(1). \tag{C.1.4}$$

By (C.2.1), and that $\|\boldsymbol{R}\| = 1$, we have

$$\|\widehat{\boldsymbol{\Gamma}}_F(0) - \widetilde{\boldsymbol{\Gamma}}_F(0)\| = \|\frac{1}{p}\widehat{\boldsymbol{M}}_x - \frac{1}{p}\boldsymbol{R}^\top \boldsymbol{M}_\chi \boldsymbol{R}\| = O_P\left(\sqrt{\frac{\log p}{n}} \vee \frac{1}{p}\right).$$

This gives a bound for $\|\widehat{\mathbf{\Gamma}}_F(\ell) - \widetilde{\mathbf{\Gamma}}_F(\ell)\|$ with $\ell = 0$, and the bounds for $\ell = 1, \ldots, d$ follow similarly since $\mathbf{\Gamma}_\varepsilon(\ell) = \mathbb{E}(\boldsymbol{\varepsilon}_t \boldsymbol{\varepsilon}_{t-\ell}^\top) = \boldsymbol{O}$ under the white noise condition on $\boldsymbol{\varepsilon}_t$. Hence,

$$\|\widehat{\mathbf{\Gamma}}_F^{acv} - \widetilde{\mathbf{\Gamma}}_F^{acv}\| = O_P\left(\sqrt{\frac{\log p}{n}} \vee \frac{1}{p}\right).$$

Then by part 2 of Lemma A.1 of Fan et al. (2011),

$$\|(\widehat{\mathbf{\Gamma}}_F^{acv})^{-1} - (\mathbf{\Gamma}_F^{acv})^{-1}\| = O_P\left(\sqrt{\frac{\log p}{n}} \vee \frac{1}{p}\right). \tag{C.1.5}$$

Similarly,

$$\|\widehat{\boldsymbol{\gamma}}_F^{acv} - \widetilde{\boldsymbol{\gamma}}_F^{acv}\| = O_P\left(\sqrt{\frac{\log p}{n}} \vee \frac{1}{p}\right). \tag{C.1.6}$$

Also by Remark 5.4 (i),

$$\|\widetilde{\boldsymbol{\gamma}}_F^{acv}\| = O(1). \tag{C.1.7}$$

Then, using (C.1.4), (C.1.5), (C.1.6), and (C.1.7),

$$\begin{aligned}
\|\widehat{\boldsymbol{a}}_{1,n} - \boldsymbol{a}\| &= \|(\widehat{\mathbf{\Gamma}}_F^{acv})^{-1}\widehat{\boldsymbol{\gamma}}_F^{acv} - (\widetilde{\mathbf{\Gamma}}_F^{acv})^{-1}\widetilde{\boldsymbol{\gamma}}_F^{acv}\| \\
&\leq \|(\widehat{\mathbf{\Gamma}}_F^{acv})^{-1}\|\|\widehat{\boldsymbol{\gamma}}_F^{acv} - \widetilde{\boldsymbol{\gamma}}_F^{acv}\| + \|\widetilde{\boldsymbol{\gamma}}_F^{acv}\|\|(\widehat{\mathbf{\Gamma}}_F^{acv})^{-1} - (\widetilde{\mathbf{\Gamma}}_F^{acv})^{-1}\| \\
&= O_P(1)O_P\left(\sqrt{\frac{\log p}{n}} \vee \frac{1}{p}\right) + O(1)O_P\left(\sqrt{\frac{\log p}{n}} \vee \frac{1}{p}\right) \\
&= O_P\left(\sqrt{\frac{\log p}{n}} \vee \frac{1}{p}\right).
\end{aligned}$$

Hence using Assumption 5.4, $\|\widehat{\boldsymbol{a}}_{1,n}\| \leq \|\widehat{\boldsymbol{a}}_{1,n} - \boldsymbol{a}\| + \|\boldsymbol{a}\| = O_P(1)$. ■

Lemma C.2. Let the conditions of Lemma C.1 hold. The estimator $\widehat{\boldsymbol{a}}_{1,n}$ fulfils

$$\begin{aligned}
&(i) \max_{k_{j-1}+G \leq k \leq k_j - G} \frac{1}{\sqrt{2G}} \|\boldsymbol{m}_k(G, \widehat{\boldsymbol{a}}_{1,n}) - \boldsymbol{m}_k(G, \boldsymbol{a})\| = o_P\left((\log(n/G))^{-1/2}\right), j = 1, \ldots, q+1. \\
&(ii) \max_{k:|k-k_j|<G} \frac{1}{\sqrt{2G}} \|\boldsymbol{m}_k(G, \widehat{\boldsymbol{a}}_{1,n}) - \boldsymbol{m}_k(G, \boldsymbol{a})\| = o_P\left((\log(n/G))^{1/2}\right), j = 1, \ldots, q. \\
&(iii) \max_{G \leq k \leq n-G} \frac{1}{\sqrt{2G}} \|\widehat{\boldsymbol{m}}_k(G, \widehat{\boldsymbol{a}}_{1,n}) - \boldsymbol{m}_k(G, \boldsymbol{a})\| = O_P\left(\log^{2v}(n)\right).
\end{aligned}$$

Proof. (i) and (ii) hold by Lemma C.1 and Kirch and Reckrühm (2022), Theorem 4.4, since $O_P(G)O_P\left(\sqrt{\frac{\log p}{n}} \vee \frac{1}{p}\right) = o_P(\sqrt{\frac{G}{\log(n/G)}})$, using $n = O(p^2)$ from Assumption 5.4. We have

$$\max_{G \leq k \leq n-G} \frac{1}{\sqrt{2G}} \|\widehat{\boldsymbol{m}}_k(G, \widehat{\boldsymbol{a}}_{1,n}) - \boldsymbol{m}_k(G, \widehat{\boldsymbol{a}}_{1,n})\| = O_P\left(\log^{2v}(n)\right)$$

by arguments similar to Lemma 5.1, so using (i), (ii), and the triangle inequality, (iii) holds. ■

Lemma C.3 validates Assumption 5.11.

Lemma C.3. Let the conditions of Lemma 5.1 hold. $\widehat{\boldsymbol{\Sigma}}_k(\widetilde{\boldsymbol{a}})$ in (C.1.1) meets Assumption 5.11.

Proof. By similar arguments to Lemma 5.1 we have

$$\max_{G\leq k\leq n-G}\|\widehat{\boldsymbol{C}}^{1/2}_{k-G+1,k+G}\| = \max_{G\leq k\leq n-G}\|\widehat{\boldsymbol{C}}_{k-G+1,k+G}\|^{1/2} = O_P(\log^v(n)),$$

and

$$\max_{G\leq k\leq n-G}\|(\widehat{\boldsymbol{S}}_k(\widetilde{\boldsymbol{a}}))^{1/2}\| = \max_{G\leq k\leq n-G}\|\widehat{\boldsymbol{S}}_k(\widetilde{\boldsymbol{a}})\|^{1/2} = O_P(\log^v(n)).$$

Combining these, we have

$$\max_{G\leq k\leq n-G}\|(\widehat{\boldsymbol{\Sigma}}_k(\widetilde{\boldsymbol{a}}))^{1/2}\| = O_P(\log^{2v}(n)).$$

Since the estimator is positive definite with probability 1, we also have

$$\|(\widehat{\boldsymbol{\Sigma}}_k(\widetilde{\boldsymbol{a}}))^{-1/2}\| = \lambda^{1/2}_{\min}(\widehat{\boldsymbol{\Sigma}}_k(\widetilde{\boldsymbol{a}})) \leq \lambda^{1/2}_{\max}(\widehat{\boldsymbol{\Sigma}}_k(\widetilde{\boldsymbol{a}})),$$

so

$$\max_{G\leq k\leq n-G}\|(\widehat{\boldsymbol{\Sigma}}_k(\widetilde{\boldsymbol{a}}))^{-1/2}\| = O_P(\log^{2v}(n)),$$

and we are done.

■

C.1.2 Factor-augmented regression

Similarly to the factor VAR case, for the factor-augmented regression we require choices for $\widetilde{\boldsymbol{\beta}}$ and $\boldsymbol{\Sigma}^y_k(\widetilde{\boldsymbol{\beta}})$. As mentioned in 5.3.1, for $\widetilde{\boldsymbol{\beta}}$ we use the global least-squares solution $\widehat{\boldsymbol{\beta}}_{1,n}$, which solves (5.9) with $w_t = 1$. For $\boldsymbol{\Sigma}^y_k(\widetilde{\boldsymbol{\beta}})$, we use

$$\widehat{\boldsymbol{\Sigma}}^y_k(\widetilde{\boldsymbol{\beta}}) = \widehat{\sigma}^2_k(\widetilde{\boldsymbol{\beta}})\widehat{\boldsymbol{\Gamma}}_F, \tag{C.1.8}$$

where we denote the estimator of $\boldsymbol{\Gamma}_F$ as

$$\widehat{\boldsymbol{\Gamma}}_F = \frac{1}{n}\sum_{t=1}^{n}\widehat{\boldsymbol{F}}_t\widehat{\boldsymbol{F}}_t^\top,$$

and

$$\widehat{\sigma}^2_k(\widetilde{\boldsymbol{\beta}}) = \frac{1}{2G}\left(\sum_{t=k-G+1}^{k}\left(\widehat{\varepsilon}^y_t(\widetilde{\boldsymbol{\beta}}) - \bar{\varepsilon}^y_{k-G+1,k}(\widetilde{\boldsymbol{\beta}})\right)^2 + \sum_{t=k+1}^{k+G}\left(\widehat{\varepsilon}^y_t(\widetilde{\boldsymbol{\beta}}) - \bar{\varepsilon}^y_{k+1,k+G}(\widetilde{\boldsymbol{\beta}})\right)^2\right),$$

where $\widehat{\varepsilon}^y_t(\widetilde{\boldsymbol{\beta}}) = y_t - \widehat{\boldsymbol{z}}^\top\widetilde{\boldsymbol{\beta}}$.

C.1.2.1 Consistency

As in Appendix C.1.1.1, $\boldsymbol{F}_t^{(j)}, j = 1, \ldots, q+1$ is only identifiable up to the rotation $\widetilde{\boldsymbol{F}}_t^{(j)} = \boldsymbol{R}^\top \boldsymbol{F}_t^{(j)}$, so we are interested in the parameter vectors $\bar{\boldsymbol{\beta}}_j = [\boldsymbol{R}, \boldsymbol{I}]\boldsymbol{\beta}_j$. Define the mixture $\boldsymbol{\beta} = \frac{1}{n}\sum_{j=1}^{q^y+1}(k_j^y - k_{j-1}^y)\bar{\boldsymbol{\beta}}_j$. Lemma C.4 states that this is consistently estimated by the least squares estimator, and Lemma C.5 states that the divergence of the difference vector $\boldsymbol{m}_k(G, \widehat{\boldsymbol{\beta}}_{1,n})$ or $\widehat{\boldsymbol{m}}_k(G, \widehat{\boldsymbol{\beta}}_{1,n})$ from $\boldsymbol{m}_k(G, \boldsymbol{\beta})$ must be bounded in probability.

Lemma C.4. Let Assumptions 5.1–5.6 hold. Then,

$$\|\widehat{\boldsymbol{\beta}}_{1,n} - \boldsymbol{\beta}\| = O_P\left(\sqrt{\frac{\log p}{n}} \vee \frac{1}{p}\right).$$

Consequently, Assumption 5.13 (b) holds for $\widetilde{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}_{1,n}$ with high probability.

Proof. The proof follows as a simpler case of the proof of Lemma C.1. ■

Lemma C.5. Let the conditions of Lemma C.4 hold. The estimator $\widehat{\boldsymbol{\beta}}_{1,n}$ fulfils

$$(i) \max_{k_{j-1}^y + G \le k \le k_j^y - G} \frac{1}{\sqrt{2G}} \|\boldsymbol{m}_k^y(G, \widehat{\boldsymbol{\beta}}_{1,n}) - \boldsymbol{m}_k^y(G, \boldsymbol{\beta})\| = o_P\left((\log(n/G))^{-1/2}\right), j = 1, \ldots, q^y + 1.$$

$$(ii) \max_{k:|k-k_j^y|<G} \frac{1}{\sqrt{2G}} \|\boldsymbol{m}_k^y(G, \widehat{\boldsymbol{\beta}}_{1,n}) - \boldsymbol{m}_k^y(G, \boldsymbol{\beta})\| = o_P\left((\log(n/G))^{1/2}\right), j = 1, \ldots, q^y.$$

$$(iii) \max_{G \le k \le n-G} \frac{1}{\sqrt{2G}} \|\widehat{\boldsymbol{m}}_k^y(G, \widehat{\boldsymbol{\beta}}_{1,n}) - \boldsymbol{m}_k^y(G, \boldsymbol{\beta})\| = O_P\left(\log^{2v}(n)\right).$$

Proof. (i) and (ii) hold by Lemma C.4 and Kirch and Reckrühm (2022), Theorem 4.4. (iii) holds similarly to Lemma C.2 (iii). ■

Lemma C.6 validates Assumption 5.12.

Lemma C.6. Let the conditions of Lemma 5.1 hold. $\widehat{\boldsymbol{\Sigma}}_k^y(\widetilde{\boldsymbol{\beta}})$ as in (C.1.8) meets Assumption 5.12.

Proof. This follows by similar arguments to the proof of Lemma C.3. ■

C.2 Proofs

C.2.1 Factor consistency

In this subsection, we give results on our ability to recover the factor series through PCA estimation.

Lemma C.7. There exists an $r \times r$-orthogonal matrix $\boldsymbol{R}$, such that

$$\left\|\widehat{\mathbf{W}}_x - \mathbf{W}_\chi \boldsymbol{R}\right\| = O_P\left(\sqrt{\frac{\log p}{n}} \vee \frac{1}{p}\right).$$

Proof. Under Assumption 5.1 (iv) (via Remark 5.3 (ii)) Assumption 5.3 (ii), and Assumption 5.5, Lemmas A.3 and B.1 (ii) of Fan et al. (2011) show that

$$\max_{1\le i,i'\le p}\left|\frac{1}{n}\sum_{t=1}^{n}X_{it}X_{i't}-\mathbb{E}\left(\frac{1}{n}\sum_{t=1}^{n}X_{it}X_{i't}\right)\right|\le \max_{1\le j,j'\le r} r^2\bar{\lambda}^2\left|\frac{1}{n}\sum_{t=1}^{n}F_{jt}F_{j't}-\mathbb{E}\left(\frac{1}{n}\sum_{t=1}^{n}F_{jt}F_{j't}\right)\right|$$
$$+\max_{1\le i,i'\le p}\left|\frac{1}{n}\sum_{t=1}^{n}\varepsilon_{it}\varepsilon_{i't}-\mathbb{E}\left(\frac{1}{n}\sum_{t=1}^{n}\varepsilon_{it}\varepsilon_{i't}\right)\right|+2\max_{\substack{1\le j\le r\\ 1\le i\le p}} r\bar{\lambda}\left|\frac{1}{n}\sum_{t=1}^{n}F_{jt}\varepsilon_{it}\right|=O_P\left(\sqrt{\frac{\log p}{n}}\right).$$

Therefore,

$$\frac{1}{p}\left\|\widehat{\mathbf{\Gamma}}_x-\mathbf{\Gamma}_\chi\right\|\le\frac{1}{p}\left\|\widehat{\mathbf{\Gamma}}_x-\mathbf{\Gamma}_x\right\|_F+\frac{1}{p}\left\|\mathbf{\Gamma}_\varepsilon\right\|=O_P\left(\sqrt{\frac{\log p}{n}}\vee\frac{1}{p}\right) \tag{C.2.1}$$

under Assumption 5.3 (i) (by Remark 5.4 (ii)). From Theorem 2 in Yu et al. (2015), we have

$$\left\|\widehat{\mathbf{W}}_x-\mathbf{W}_\chi\boldsymbol{R}\right\|\le\frac{2^{3/2}\sqrt{r}\left\|\widehat{\mathbf{\Gamma}}_x-\mathbf{\Gamma}_\chi\right\|}{\min\left(\mu_{\chi,0}-\mu_{\chi,1},\mu_{\chi,r}-\mu_{\chi,r+1}\right)}, \tag{C.2.2}$$

where $\mu_{\chi,0}=\infty$ and $\mu_{\chi,r+1}=0$. The denominator of (C.2.2) is bounded from the below by $\underline{\gamma}p$ by Remark 5.4 (i), so the statement of the Lemma follows by (C.2.2). ■

Lemma C.8. For a fixed $v\ge 1+\left(1/b_f\vee 1/b_\varepsilon\right)$ and $w_t\in\{0,1\}$, which may be stochastic or deterministic, we have

$$\max_{1\le k\le n-G+1}G^{-1/2}\left\|\sum_{t=k}^{k+G-1}w_t\boldsymbol{\varepsilon}_t\right\|=O_P\left(\sqrt{p}\log^v(n)\right), \tag{C.2.3}$$

$$\max_{1\le k\le n-G+1}G^{-1/2}\left\|\sum_{t=k}^{k+G-1}w_t\mathbf{X}_t\right\|=O_P\left(\sqrt{p}\log^v(n)\right). \tag{C.2.4}$$

Proof. The proof of (C.2.3) and (C.2.4) when all $w_t=1$ is given in (Barigozzi et al., 2018, Lemma 4). We note that for such w_t, the exponentially decaying tail behaviour of F_{jt} and ε_{it} in Remark 5.3 and Assumption 5.3, and the mixing property in Assumption 5.5, carries over to w_tF_{jt} and $w_t\varepsilon_{it}$, and hence the identical arguments are applicable to the proof of (C.2.4) for any w_t. ■

In Lemma C.9, we derive a uniform bound on the partial sums of the differences between the true and estimated factors.

Lemma C.9. With $v>0$ as in Lemma C.8, we have

$$\max_{1\le k\le n-G+1}G^{-1/2}\left\|\sum_{t=k}^{k+G-1}\left(\widehat{\mathbf{F}}_t-\boldsymbol{R}^\top\mathbf{F}_t\right)\right\|=O_P\left(\log^v(n)\right).$$

Proof. As the factor space is identified up to rotation, we set $\mathbf{F}_t = \mathbf{W}_\chi^\top \boldsymbol{\chi}_t/\sqrt{p} = \mathbf{W}_\chi^\top(\mathbf{X}_t - \boldsymbol{\varepsilon}_t)/\sqrt{p}$. Also recall that $\widehat{\mathbf{F}}_t = \widehat{\mathbf{W}}_x^\top \mathbf{X}_t/\sqrt{p}$. Then,

$$G^{-1/2}\left\|\sum_{t=k}^{k+G-1}\left(\widehat{\mathbf{F}}_t - \boldsymbol{R}^\top \mathbf{F}_t\right)\right\| \le (pG)^{-1/2}\left\|\sum_{t=k}^{k+G-1}\left(\widehat{\mathbf{W}}_x - \mathbf{W}_\chi \boldsymbol{R}\right)^\top \mathbf{X}_t\right\| + (pG)^{-1/2}\left\|\sum_{t=k}^{k+G-1}\mathbf{W}_\chi^\top \boldsymbol{\varepsilon}_t\right\|$$

and

$$\begin{aligned}\max_{1\le k\le n-G+1}(pG)^{-1/2}\left\|\sum_{t=k}^{k+G-1}\left(\widehat{\mathbf{W}}_x - \mathbf{W}_\chi \boldsymbol{R}\right)^\top \mathbf{X}_t\right\| &\le \left\|\widehat{\mathbf{W}}_x - \mathbf{W}_\chi \boldsymbol{R}\right\| \max_{1\le k\le n-G+1}(pG)^{-1/2}\left\|\sum_{t=k}^{k+G-1}\mathbf{X}_t\right\| \\ &= O_P\left\{\left(\sqrt{\frac{\log p}{n}} \vee \frac{1}{p}\right)\log^v(n)\right\},\end{aligned}$$

from Lemmas C.7–C.8 (for the latter, set $w_t = 1$ for all t in (C.2.4)). Also, due to (C.2.3),

$$\max_{1\le k\le n-G+1}(pG)^{-1/2}\left\|\sum_{t=k}^{k+G-1}\mathbf{W}_\chi^\top \boldsymbol{\varepsilon}_t\right\| \le \left\|\mathbf{W}_\chi\right\| \max_{1\le k\le n-G+1}(pG)^{-1/2}\left\|\sum_{t=k}^{k+G-1}\boldsymbol{\varepsilon}_t\right\| = O_P\left(\log^v(n)\right),$$

which completes the proof. ■

Lemma C.10 provides a supporting result for Proposition 5.1.

Lemma C.10. For a fixed $k > r$, let $\widehat{\mathbf{V}} = \left[\widehat{\mathbf{w}}_{x,r+1}, \ldots, \widehat{\mathbf{w}}_{x,k}\right]$. Then

$$\left\|\widehat{\mathbf{V}}^\top \boldsymbol{\Lambda}\right\| = O_P\left(\sqrt{\frac{p\log p}{n}} \vee \frac{1}{\sqrt{p}}\right).$$

Proof. Since $\widehat{\mathbf{V}}^\top \widehat{\mathbf{W}}_x = \mathbf{O}_{(k-r)\times r}$ and $\|\widehat{\mathbf{V}}\| = 1$, we have

$$\left\|\widehat{\mathbf{V}}^\top \mathbf{W}_\chi\right\| = \left\|\widehat{\mathbf{V}}^\top \mathbf{W}_\chi \mathbf{R}\right\| = \left\|\widehat{\mathbf{V}}^\top\left(\widehat{\mathbf{W}}_x - \mathbf{W}_\chi \mathbf{R}\right)\right\| \le \left\|\widehat{\mathbf{W}}_x - \mathbf{W}_\chi \mathbf{R}\right\| = O_P\left(\sqrt{\frac{\log p}{n}} \vee \frac{1}{p}\right) \tag{C.2.5}$$

from Lemma C.7. Recall $\mathbf{M}_\chi = \operatorname{diag}\left(\mu_{\chi,j}, 1 \le j \le r\right)$ and $\boldsymbol{\Gamma}_F = n^{-1}\sum_{t=1}^n \mathbb{E}\left(\mathbf{F}_t \mathbf{F}_t^\top\right)$, where the latter is positive definite with its operator norm bounded by Remark 5.4. Using that $\boldsymbol{\Gamma}_\chi = \mathbf{W}_\chi \mathbf{M}_\chi \mathbf{W}_\chi^\top = \boldsymbol{\Lambda}\boldsymbol{\Gamma}_F \boldsymbol{\Lambda}^\top$, we have

$$\begin{aligned}\left\|\widehat{\mathbf{V}}^\top \boldsymbol{\Lambda}\right\| &= \left\|\widehat{\mathbf{V}}^\top \mathbf{W}_\chi \mathbf{M}_\chi \mathbf{W}_\chi^\top \boldsymbol{\Lambda}\left(\boldsymbol{\Lambda}^\top\boldsymbol{\Lambda}\right)^{-1}\boldsymbol{\Gamma}_F^{-1}\right\| \le \left\|\widehat{\mathbf{V}}^\top \mathbf{W}_\chi\right\|\left\|\mathbf{M}_\chi\right\|\left\|\boldsymbol{\Lambda}\left(\boldsymbol{\Lambda}^\top\boldsymbol{\Lambda}\right)^{-1}\right\|\left\|\boldsymbol{\Gamma}_F^{-1}\right\| \\ &= O_P\left\{\left(\sqrt{\frac{\log p}{n}} \vee \frac{1}{p}\right)\cdot p \cdot \frac{1}{\sqrt{p}}\right\} = O_P\left(\sqrt{\frac{p\log p}{n}} \vee \frac{1}{\sqrt{p}}\right).\end{aligned}$$

using (C.2.5) and Assumption 5.2. ■

Proof of Proposition 5.1

Proof. Define $\widehat{\varepsilon}_{it}^k = X_{it} - \widehat{\mathbf{\Lambda}}_i \widehat{\boldsymbol{F}}_t^k$, where $\widehat{\boldsymbol{F}}_t^k \in \mathbb{R}^k$ is the k-dimensional factor estimate. We first show that $V(k)$ attains its minimum when $k = r$, where

$$V(k) = \frac{1}{np}\sum_{i=1}^{p}\sum_{t=1}^{n}\left(\widehat{\varepsilon}_{it}^k\right)^2 + kg(n,p) = \frac{1}{np}\sum_{t=1}^{n}\left\|\widehat{\mathbf{W}}_{1:k}\widehat{\mathbf{W}}_{1:k}^\top \mathbf{X}_t\right\|^2 + kg(n,p),$$

where $\widehat{\mathbf{W}}_{a:b} = \left[\widehat{\mathbf{w}}_{x,a},\ldots,\widehat{\mathbf{w}}_{x,b}\right]$ for $1 \le a \le b \le n$. Firstly, let $k > r$. Then, since $\widehat{\mathbf{w}}_{x,j}, j = 1,2,\ldots$ are orthonormal,

$$V(k) - V(r) = \frac{1}{np}\sum_{t=1}^{n}\left\|\widehat{\mathbf{V}}\widehat{\mathbf{V}}^\top \mathbf{X}_t\right\|^2 + (k-r)g(n,p)$$

where $\widehat{\mathbf{V}} = \widehat{\mathbf{W}}_{(r+1):k}$. Note that

$$\frac{1}{np}\sum_{t=1}^{n}\left\|\widehat{\mathbf{V}}\widehat{\mathbf{V}}^\top \mathbf{X}_t\right\|^2 \le \frac{2}{np}\sum_{t=1}^{n}\left\|\widehat{\mathbf{V}}\widehat{\mathbf{V}}^\top \boldsymbol{\chi}_t\right\|^2 + \frac{2}{np}\sum_{t=1}^{n}\left\|\widehat{\mathbf{V}}\widehat{\mathbf{V}}^\top \boldsymbol{\varepsilon}_t\right\|^2 = I + II.$$

Then, using Lemma C.10,

$$I \le \frac{2}{np}\sum_{t=1}^{n}\|\widehat{\mathbf{V}}\|^2\left\|\widehat{\mathbf{V}}^\top \mathbf{\Lambda}\right\|^2\|\mathbf{F}_t\|^2 = \frac{2}{p}O_P\left(\frac{p\log p}{n}\vee\frac{1}{p}\right)O_P\left(\log^{2/b_f} n\right) = O_P\left\{\left(\frac{\log p}{n}\vee\frac{1}{p^2}\right)\log^{2/b_f} n\right\}$$

under Remark 5.3 (ii) and Assumption 5.4. Also,

$$\begin{aligned} II &= \frac{2}{p}\operatorname{trace}\left[\widehat{\mathbf{V}}\widehat{\mathbf{V}}^\top\frac{1}{n}\sum_{t=1}^{n}\left\{\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}_t^\top - \mathbb{E}\left(\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}_t^\top\right)\right\}\right] + \frac{2}{p}\operatorname{trace}\left\{\widehat{\mathbf{V}}\widehat{\mathbf{V}}^\top\frac{1}{n}\sum_{t=1}^{n}\mathbb{E}\left(\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}_t^\top\right)\right\} \\ &\le \frac{2(k-r)}{np}\left\|\sum_{t=1}^{n}\left\{\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}_t^\top - \mathbb{E}\left(\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}_t^\top\right)\right\}\right\| + \frac{2(k-r)}{np}\left\|\sum_{t=1}^{n}\mathbb{E}\left(\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}_t^\top\right)\right\| = O_P\left(\sqrt{\frac{\log p}{n}}\right) \end{aligned}$$

using our Assumptions 5.3 and 5.5, and Lemma A.3 of Fan et al. (2011). Hence, under the conditions imposed on $g(n,p)$, we conclude that $V(k) > V(r)$ for any fixed $k > r$ with probability tending to one as $n,p \to \infty$. Now, let $k < r$. Again,

$$V(k) - V(r) = \frac{1}{np}\sum_{t=1}^{T}\left\|\widehat{\mathbf{V}}\widehat{\mathbf{V}}^\top \mathbf{X}_t\right\|^2 + (r-k)g(n,p)$$

where $\widehat{\mathbf{V}} = \widehat{\mathbf{W}}_{(k+1):r}$. Further,

$$\begin{aligned} \frac{1}{np}\sum_{t=1}^{n}\left\|\widehat{\mathbf{V}}\widehat{\mathbf{V}}^\top \mathbf{X}_t\right\|^2 &= \frac{1}{np}\sum_{t=1}^{n}\left\|\widehat{\mathbf{V}}\widehat{\mathbf{V}}^\top \boldsymbol{\chi}_t\right\|^2 + \frac{2}{np}\sum_{t=1}^{n}\boldsymbol{\chi}_t^\top\widehat{\mathbf{V}}\widehat{\mathbf{V}}^\top \boldsymbol{\varepsilon}_t + \frac{1}{np}\sum_{t=1}^{n}\left\|\widehat{\mathbf{V}}\widehat{\mathbf{V}}^\top \boldsymbol{\varepsilon}_t\right\|^2 \\ &= III + IV + V. \end{aligned}$$

Then, we can bound $V = O_P(\sqrt{\log p/n})$ as with II. From Lemma C.7, there exists an $r \times (r-k)$ matrix $\tilde{\mathbf{Z}}$ with orthonormal columns so that

$$\left\|\widehat{\mathbf{V}} - \mathbf{W}_\chi\tilde{\mathbf{Z}}\right\| = O_p\left(\sqrt{\frac{\log p}{n}}\vee\frac{1}{p}\right),$$

and hence

$$\left\|\widehat{\mathbf{V}}\widehat{\mathbf{V}}^\top - \mathbf{W}_\chi \tilde{\mathbf{Z}}\tilde{\mathbf{Z}}^\top \mathbf{W}_\chi^\top\right\| \le \left\|\widehat{\mathbf{V}}\left(\widehat{\mathbf{V}} - \mathbf{W}_\chi \tilde{\mathbf{Z}}\right)^\top\right\| + \left\|\left(\widehat{\mathbf{V}} - \mathbf{W}_\chi \tilde{\mathbf{Z}}\right)\mathbf{W}_\chi \tilde{\mathbf{Z}}^\top\right\| = O_P\left(\sqrt{\frac{\log p}{n}} \vee \frac{1}{p}\right). \tag{C.2.6}$$

Note that

$$III \ge \overbrace{\frac{1}{np}\sum_{t=1}^n \left\|\mathbf{W}_\chi \tilde{\mathbf{Z}}^\top \tilde{\mathbf{W}}_\chi^\top \boldsymbol{\chi}_t\right\|^2}^{VI} - \overbrace{\frac{2}{np}\sum_{t=1}^n \left\|\mathbf{W}_\chi \tilde{\mathbf{Z}}\tilde{\mathbf{Z}}^\top \mathbf{W}_\chi^\top \boldsymbol{\chi}_t\right\| \left\|\left(\widehat{\mathbf{V}}^\top - \mathbf{W}_\chi \tilde{\mathbf{Z}}\tilde{\mathbf{Z}}^\top \mathbf{W}_\chi^\top\right)\boldsymbol{\chi}_t\right\|}^{VII}$$
$$+ \underbrace{\frac{1}{np}\sum_{t=1}^n \left\|\left(\widehat{\mathbf{V}}\widehat{\mathbf{V}}^\top - \mathbf{W}_\chi \tilde{\mathbf{Z}}\tilde{\mathbf{Z}}^\top \mathbf{W}_\chi^\top\right)\boldsymbol{\chi}_t\right\|^2}_{VIII}.$$

Letting $\widehat{\boldsymbol{\Gamma}}_\chi = n^{-1}\sum_{t=1}^n \boldsymbol{\chi}_t \boldsymbol{\chi}_t^\top$,

$$\begin{aligned}
VI &= \frac{1}{np}\sum_{t=1}^n \boldsymbol{\chi}_t^\top \mathbf{W}_\chi \tilde{\mathbf{Z}}\tilde{\mathbf{Z}}^\top \left(\mathbf{W}_\chi\right)^\top \boldsymbol{\chi}_t = \frac{1}{p}\operatorname{trace}\left(\mathbf{W}_\chi \tilde{\mathbf{Z}}\tilde{\mathbf{Z}}^\top \left(\mathbf{W}_\chi\right)^\top \widehat{\boldsymbol{\Gamma}}_\chi\right) \\
&= \frac{1}{p}\operatorname{trace}\left(\mathbf{W}_\chi \tilde{\mathbf{Z}}\tilde{\mathbf{Z}}^\top \left(\mathbf{W}_\chi\right)^\top \boldsymbol{\Gamma}_\chi\right) + \frac{1}{p}\operatorname{trace}\left\{\mathbf{W}_\chi \tilde{\mathbf{Z}}\tilde{\mathbf{Z}}^\top \left(\mathbf{W}_\chi\right)^\top \left(\widehat{\Gamma}_\chi - \boldsymbol{\Gamma}_\chi\right)\right\} \\
&\le \frac{1}{p}\operatorname{trace}\left\{\mathbf{W}_\chi \tilde{\mathbf{Z}}\tilde{\mathbf{Z}}^\top \left(\mathbf{W}_\chi\right)^\top \boldsymbol{\Gamma}_\chi\right\} + \left\|\mathbf{W}_\chi \tilde{\mathbf{Z}}\tilde{\mathbf{Z}}^\top \left(\mathbf{W}_\chi\right)^\top\right\|_F \cdot \frac{1}{p}\left\|\widehat{\boldsymbol{\Gamma}}_\chi - \boldsymbol{\Gamma}_\chi\right\|_F \\
&= \frac{1}{p}\operatorname{trace}\left(\mathbf{W}_\chi \tilde{\mathbf{Z}}\tilde{\mathbf{Z}}^\top \left(\mathbf{W}_\chi\right)^\top \boldsymbol{\Gamma}_\chi\right) + O_P\left(\sqrt{\frac{\log p}{n}}\right) = \frac{1}{p}\operatorname{trace}\left(\tilde{\mathbf{Z}}\tilde{\mathbf{Z}}^\top \mathbf{M}_\chi^b\right) + O_P\left(\sqrt{\frac{\log p}{n}}\right),
\end{aligned}$$

where $p^{-1}\left\|\widehat{\boldsymbol{\Gamma}}_\chi - \boldsymbol{\Gamma}_\chi\right\|_F = O_P(\sqrt{\log p/n})$ is analogously shown as in Lemma C.7 under Remark 5.3 (ii) and Assumption 5.4. The form of VI involves the trace of the projection of $\mathbf{M}_\chi$ with respect to a rank $r-k$ projection matrix, and hence $VI > 0$. Also, using (C.2.6) and Remark 5.3 (ii),

$$\left\|\left(\widehat{\mathbf{V}}\widehat{\mathbf{V}}^\top - \mathbf{W}_\chi \tilde{\mathbf{Z}}\tilde{\mathbf{Z}}^\top \mathbf{W}_\chi^\top\right)\boldsymbol{\chi}_t\right\| = O_P\left\{\left(\sqrt{\frac{p\log p}{n}} \vee \frac{1}{\sqrt{p}}\right)\log^{1/b_f} n\right\}$$

and

$$\left\|\mathbf{W}_\chi \tilde{\mathbf{Z}}\tilde{\mathbf{Z}}^\top \mathbf{W}_\chi^\top \boldsymbol{\chi}_t\right\| = O_P\left(\sqrt{p}\log^{1/b_f} n\right)$$

uniformly in t, and therefore $VII = O_P\left\{(\sqrt{\log p/n} \vee 1/p)\log^{2/b_f} n\right\}$. Also,

$$VIII \le \frac{1}{np}\sum_{t=1}^n \left\|\widehat{\mathbf{V}}\widehat{\mathbf{V}}^\top - \mathbf{W}_\chi \tilde{\mathbf{Z}}\tilde{\mathbf{Z}}^\top \mathbf{W}_\chi^\top\right\|^2 \left\|\boldsymbol{\chi}_t\right\|^2 = O_P\left\{\left(\frac{\log n}{n} \vee \frac{1}{p^2}\right)\log^{2/b_f} n\right\}.$$

Combining the bounds on VI, VII and $VIII$, we conclude that III is bounded away from zero with probability tending to one. Finally, under Remark 5.3 (ii) and Assumptions 5.3 and 5.4, Lemma B.1 (ii) of Fan et al. (2011) leads to

$$IV = \frac{2}{np}\operatorname{trace}\left(\widehat{\mathbf{V}}\widehat{\mathbf{V}}^\top \sum_{t=1}^p \chi_t \boldsymbol{\varepsilon}_t^\top\right) \le \frac{2(r-k)}{np}\left\|\sum_{t=1}^n \chi_t \boldsymbol{\varepsilon}_t^\top\right\| = O_P\left(\sqrt{\frac{\log p}{n}}\right),$$

which leads to $V(k) > V(r)$ with probability converging to one. Having shown that $V(k)$ is minimised at r, the proof follows similarly to Corollary 1 in Bai and Ng (2002).

■

C.2.2 VAR segmentation consistency

Lemma C.11. For each $j = 1, \dots, q+1$, $\mathrm{Cov}(\mathbb{F}_1^{(j)}) = \boldsymbol{C}_{(j)}$ is a positive definite covariance matrix.

Proof. This is a consequence of Remark 5.3 (i) and Assumption 5.1. ■

As a result of Lemma C.11 and Assumption 5.1, each series $\{\boldsymbol{H}(\boldsymbol{F}_t^{(j)}, \mathbb{F}_{t-1}^{(j)}, \widetilde{\boldsymbol{a}})\}_{t=1}^n$ has a positive definite covariance matrix $\boldsymbol{\Sigma}_{(j)}(\widetilde{\boldsymbol{a}})$.

As the factors are only identifiable up to rotation, we are interested in estimating $\widetilde{\boldsymbol{F}}_t = \boldsymbol{R}^\top \boldsymbol{F}_t$, with $\widetilde{\mathbb{F}}_{t-1}$ collecting the respective past observations. Define the population estimating function

$$\boldsymbol{H}\left(\widetilde{\boldsymbol{F}}_t, \widetilde{\mathbb{F}}_{t-1}, \widetilde{\boldsymbol{a}}\right) = -(\widetilde{\boldsymbol{F}}_t - \widetilde{\boldsymbol{a}}\widetilde{\mathbb{F}}_{t-1}) \otimes \widetilde{\mathbb{F}}_{t-1}.$$

We monitor these for changes in expectation using the score detector

$$T(G, \widetilde{\boldsymbol{a}}) = \max_{G \le k \le n-G} T_k(G, \widetilde{\boldsymbol{a}}), \quad T_k(G, \widetilde{\boldsymbol{a}}) = \frac{1}{\sqrt{2G}} \left\| (\boldsymbol{\Sigma}_k(\widetilde{\boldsymbol{a}}))^{-1/2} \boldsymbol{m}_k(G, \widetilde{\boldsymbol{a}}) \right\|,$$

where the population difference vector at time k, evaluated with inspection parameter $\widetilde{\boldsymbol{a}}$, is

$$\boldsymbol{m}_k(G, \widetilde{\boldsymbol{a}}) = \sum_{t=k+1}^{k+G} \boldsymbol{H}\left(\widetilde{\boldsymbol{F}}_t, \widetilde{\mathbb{F}}_{t-1}, \widetilde{\boldsymbol{a}}\right) - \sum_{t=k-G+1}^{k} \boldsymbol{H}\left(\widetilde{\boldsymbol{F}}_t, \widetilde{\mathbb{F}}_{t-1}, \widetilde{\boldsymbol{a}}\right).$$

Proof of Lemma 5.1

Proof. Denote

$$\boldsymbol{B}_t(\widetilde{\boldsymbol{a}}) = (\widetilde{\boldsymbol{F}}_t - \widetilde{\boldsymbol{a}}\widetilde{\mathbb{F}}_{t-1}), \text{ and } \widehat{\boldsymbol{B}}_t(\widetilde{\boldsymbol{a}}) = (\widehat{\boldsymbol{F}}_t - \widetilde{\boldsymbol{a}}\widehat{\mathbb{F}}_{t-1}).$$

We have

$$\begin{aligned}
&\max_{G \le k \le n-G} \frac{1}{\sqrt{2G}} \|\widehat{\boldsymbol{m}}_k(G, \widetilde{\boldsymbol{a}}) - \boldsymbol{m}_k(G, \widetilde{\boldsymbol{a}})\| \\
&\le \max_{0 \le k \le n-G} \frac{2}{\sqrt{2G}} \| \sum_{t=k+1}^{k+G} \boldsymbol{H}\left(\widehat{\boldsymbol{F}}_t, \widehat{\mathbb{F}}_{t-1}, \widetilde{\boldsymbol{a}}\right) - \boldsymbol{H}\left(\widetilde{\boldsymbol{F}}_t, \widetilde{\mathbb{F}}_{t-1}, \widetilde{\boldsymbol{a}}\right) \| \\
&\le \max_{0 \le k \le n-G} \frac{2}{\sqrt{2G}} \| \sum_{t=k+1}^{k+G} \widehat{\boldsymbol{B}}_t(\widetilde{\boldsymbol{a}}) \otimes \widehat{\mathbb{F}}_{t-1} - \boldsymbol{B}_t(\widetilde{\boldsymbol{a}}) \otimes \widetilde{\mathbb{F}}_{t-1} \| \\
&\le \max_{0 \le k \le n-G} \frac{2}{\sqrt{2G}} \| \sum_{t=k+1}^{k+G} (\widehat{\boldsymbol{B}}_t(\widetilde{\boldsymbol{a}}) - \boldsymbol{B}_t(\widetilde{\boldsymbol{a}})) \otimes \widehat{\mathbb{F}}_{t-1} \| + \max_{0 \le k \le n-G} \frac{2}{\sqrt{2G}} \| \sum_{t=k+1}^{k+G} \boldsymbol{B}_t(\widetilde{\boldsymbol{a}}) \otimes (\widehat{\mathbb{F}}_{t-1} - \widetilde{\mathbb{F}}_{t-1}) \| \\
&= I + II.
\end{aligned}$$

We start by bounding I. We have that

$$\max_{0\le k\le n-G}\frac{2}{\sqrt{2G}}\|\sum_{t=k+1}^{k+G}(\widehat{\boldsymbol{B}}_t(\widetilde{\boldsymbol{a}})-\boldsymbol{B}_t(\widetilde{\boldsymbol{a}}))\otimes\widehat{\mathbb{F}}_{t-1}\|$$
$$\le\sqrt{r^2d}\max_{1\le i\le r}\max_{1\le i'\le rd}\max_{0\le k\le n-G}\frac{2}{\sqrt{2G}}|\sum_{t=k+1}^{k+G}(\widehat{B}_{it}-B_{it})\widehat{\mathbb{F}}_{i',t-1}|. \tag{C.2.7}$$

Letting $w_{it}=\mathbb{I}(\widehat{B}_{it}-B_{it}\ge 0)$,

$$\max_{1\le i\le r}\max_{1\le i'\le rd}\max_{0\le k\le n-G}\frac{2}{\sqrt{2G}}|\sum_{t=k+1}^{k+G}(\widehat{B}_{it}-B_{it})\widehat{\mathbb{F}}_{i',t-1}|$$
$$\le\max_{1\le i\le r}\max_{0\le k\le n-G}\frac{2}{\sqrt{2G}}|\sum_{t=k+1}^{k+G}w_{it}(\widehat{B}_{it}-B_{it})|\max_{1\le i'\le rd}\max_{1\le t\le n}|\widehat{\mathbb{F}}_{i',t-1}|$$
$$+\max_{1\le i\le r}\max_{0\le k\le n-G}\frac{2}{\sqrt{2G}}|\sum_{t=k+1}^{k+G}(1-w_{it})(\widehat{B}_{it}-B_{it})|\max_{1\le i'\le rd}\max_{1\le t\le n}|\widehat{\mathbb{F}}_{i',t-1}|=III+IV.$$

Modifying the proof of Lemma C.9, and since $\|\widetilde{\boldsymbol{a}}\|=O(1)$ by Assumption 5.13, we have

$$\max_{1\le i\le r}\max_{0\le k\le n-G}\frac{1}{\sqrt{G}}\|\sum_{t=k+1}^{k+G}w_{it}(\widehat{\boldsymbol{B}}_t-\boldsymbol{B}_t)\|$$
$$\le\max_{1\le i\le r}\max_{0\le k\le n-G}\frac{1}{\sqrt{G}}\|\sum_{t=k+1}^{k+G}w_{it}(\widehat{\boldsymbol{F}}_t-\widetilde{\boldsymbol{F}}_t)\|+\max_{1\le i\le r}\max_{0\le k\le n-G}\frac{1}{\sqrt{G}}\|\widetilde{\boldsymbol{a}}\|\|\sum_{t=k+1}^{k+G}w_{it}(\widehat{\mathbb{F}}_{t-1}-\widetilde{\mathbb{F}}_{t-1})\|$$
$$\le(1+\sqrt{d}\,\|\widetilde{\boldsymbol{a}}\|)\max_{1\le i\le r}\max_{0\le k\le n-G}\frac{1}{\sqrt{G}}\|\sum_{t=k+1}^{k+G}w_{it}(\widehat{\boldsymbol{F}}_t-\widetilde{\boldsymbol{F}}_t)\|=O_P(\log^v(n)).$$

Hence,

$$\max_{1\le i\le r}\max_{0\le k\le n-G}\frac{1}{\sqrt{G}}|\sum_{t=k+1}^{k+G}w_{it}(\widehat{F}_{it}-\widetilde{F}_{it})|=O_P(\log^v(n)). \tag{C.2.8}$$

By Remark 5.3 (ii), using Lemma C.9,

$$\max_{1\le i'\le rd}\max_{1\le t\le n}|\widehat{\mathbb{F}}_{i',t-1}|\le\sqrt{d}\max_{1\le i\le r}\max_{1\le t\le n}|\widehat{F}_{it}|$$
$$\le\sqrt{d}\max_{1\le i\le r}\max_{1\le t\le n}|\widehat{F}_{it}-\widetilde{F}_{it}|+\sqrt{d}\max_{1\le i\le r}\max_{1\le t\le n}|\widetilde{F}_{it}|=O_P(\log^v(n)). \tag{C.2.9}$$

Combining (C.2.8)–(C.2.9), we have $III=O_P(\log^{2v}n)$, and by similar arguments, $IV=O_P(\log^{2v}n)$. Using these results in (C.2.7), we bound $I=O_P(\log^{2v}n)$. Using Assumption 5.3 (iv), $II=$

$O_P(\log^{2v} n)$ follows similarly. Then,

$$\begin{aligned}
&\max_{G\le k\le n-G} \frac{1}{\sqrt{2G}} \|(\widehat{\boldsymbol{\Sigma}}_k(\widetilde{\boldsymbol{a}}))^{-1/2}\widehat{\boldsymbol{m}}_k(G,\widetilde{\boldsymbol{a}}) - (\boldsymbol{\Sigma}_k(\widetilde{\boldsymbol{a}}))^{-1/2}\boldsymbol{m}_k(G,\widetilde{\boldsymbol{a}})\| \\
&\le \max_{G\le k\le n-G} \frac{1}{\sqrt{2G}} \|(\widehat{\boldsymbol{\Sigma}}_k(\widetilde{\boldsymbol{a}}))^{-1/2}\widehat{\boldsymbol{m}}_k(G,\widetilde{\boldsymbol{a}}) - (\widehat{\boldsymbol{\Sigma}}_k(\widetilde{\boldsymbol{a}}))^{-1/2}\boldsymbol{m}_k(G,\widetilde{\boldsymbol{a}}) \\
&+ (\widehat{\boldsymbol{\Sigma}}_k(\widetilde{\boldsymbol{a}}))^{-1/2}\boldsymbol{m}_k(G,\widetilde{\boldsymbol{a}}) - (\boldsymbol{\Sigma}_k(\widetilde{\boldsymbol{a}}))^{-1/2}\boldsymbol{m}_k(G,\widetilde{\boldsymbol{a}})\| \\
&\le \max_{G\le k\le n-G} \frac{1}{\sqrt{2G}} \|(\widehat{\boldsymbol{\Sigma}}_k(\widetilde{\boldsymbol{a}}))^{-1/2}\widehat{\boldsymbol{m}}_k(G,\widetilde{\boldsymbol{a}}) - (\widehat{\boldsymbol{\Sigma}}_k(\widetilde{\boldsymbol{a}}))^{-1/2}\boldsymbol{m}_k(G,\widetilde{\boldsymbol{a}})\| \\
&+ \|(\widehat{\boldsymbol{\Sigma}}_k(\widetilde{\boldsymbol{a}}))^{-1/2}\boldsymbol{m}_k(G,\widetilde{\boldsymbol{a}}) - (\boldsymbol{\Sigma}_k(\widetilde{\boldsymbol{a}}))^{-1/2}\boldsymbol{m}_k(G,\widetilde{\boldsymbol{a}})\| \\
&\le \max_{G\le k\le n-G} \frac{1}{\sqrt{2G}} \|(\widehat{\boldsymbol{\Sigma}}_k(\widetilde{\boldsymbol{a}}))^{-1/2}\| \|\widehat{\boldsymbol{m}}_k(G,\widetilde{\boldsymbol{a}}) - \boldsymbol{m}_k(G,\widetilde{\boldsymbol{a}})\| + \|(\widehat{\boldsymbol{\Sigma}}_k(\widetilde{\boldsymbol{a}}))^{-1/2} - (\boldsymbol{\Sigma}_k(\widetilde{\boldsymbol{a}}))^{-1/2}\| \|\boldsymbol{m}_k(G,\widetilde{\boldsymbol{a}})\| \\
&= O_P(\log^{2v}(n))O_P(\log^{2v}(n)) + O_P(\log^{2v}(n)),
\end{aligned}$$

where the last line follows by Assumption 5.11 and Reckrühm (2019) Lemma 2.1.4. ■

In the rest of this section, we demonstrate that the data segmentation procedure for the piecewise stationary VAR model is consistent when working with the factor series. These results are inherited from the results for observed series given in Chapter 4. We condition on the event $\{\widehat{r} = r\}$, which holds with probability tending to 1 by Proposition 5.1.

Proof of Theorem 5.3

Proof. By Lemma 5.1, we have $\widehat{T}_k(G,\widetilde{\boldsymbol{a}}) = T_k(G,\widetilde{\boldsymbol{a}}) + O_P(\log^{2v}(n))$ for all $k = G,\dots,n-G$. Similarly to the proof of Kirch and Reckrühm (2022) Proposition 3.4, we show that

$$\begin{aligned}
&\mathsf{P}\left(\max_{1\le j\le q+1} \max_{k_{j-1}+G\le k\le k_j-G} \widehat{T}_k(G,\widetilde{\boldsymbol{a}}) \ge D\right) \\
&\le \sum_{j=1}^{q+1} \mathsf{P}\left(a(n/G) \max_{k_{j-1}+G\le k\le k_j-G} \widehat{T}_k(G,\widetilde{\boldsymbol{a}}) - b(n/G) \ge c_\alpha\right) \\
&\le \sum_{j=1}^{q+1} \mathsf{P}\left(a(n/G) \max_{k_{j-1}+G\le k\le k_j-G} (T_k(G,\widetilde{\boldsymbol{a}}) + C\log^{2v}(n)) - b(n/G) \ge c_\alpha\right) \\
&\le \sum_{j=1}^{q+1} (\alpha + o(1)) \to 0.
\end{aligned}$$

Hence,

$$\begin{aligned}
&\mathsf{P}\left(\max_{1\le j\le q+1} \max_{k_{j-1}+G\le k\le k_j-G} \widehat{T}_k(G,\widetilde{\boldsymbol{a}}) < D\right) \to 1, \text{ and similarly when } q \ge 1, \\
&\mathsf{P}\left(\min_{1\le j\le q} \min_{k:|k-k_j|\le(1-\epsilon)G} \widehat{T}_k(G,\widetilde{\boldsymbol{a}}) \ge D\right) \to 1
\end{aligned}$$

as $n \to \infty$, so

$$\mathsf{P}(\widehat{q}(\widetilde{\boldsymbol{a}}) = q) \to 1.$$

Likewise by Kirch and Reckrühm (2022) Theorem 3.5,

$$\mathrm{P}\left(\widehat{q}(\widetilde{\boldsymbol{a}}) = q; \max_{1\le j\le q} \left|\widehat{k}_j - k_j\right| < G\right) \to 1.$$

Using $\widetilde{\boldsymbol{a}} = \widehat{\boldsymbol{a}}_{1,n}$, Assumption 5.13 (a) holds by Lemma C.1, and we have $\widehat{T}_k(G, \widehat{\boldsymbol{a}}_{1,n}) = T_k(G, \boldsymbol{a}) + O_P(\log^{4v}(n))$ by Lemma C.2, so the result still holds. ■

C.2.3 Regression segmentation consistency

We demonstrate that the data segmentation procedure for the regression model is consistent when working with the factor series. These results are inherited from the results for observed series given in Reckrühm (2019). The population estimating function is

$$\boldsymbol{H}^y\left(y_t, \boldsymbol{z}_t, \widetilde{\boldsymbol{\beta}}\right) = -(y_t - \boldsymbol{z}_t^\top \widetilde{\boldsymbol{\beta}})\boldsymbol{z}_t.$$

Over moving windows of length G, we compare these for changes in expectation using the score detector

$$T^y(G, \widetilde{\boldsymbol{\beta}}) = \max_{G\le k\le n-G} T_k^y(G, \widetilde{\boldsymbol{\beta}}), \quad T_k^y(G, \widetilde{\boldsymbol{\beta}}) = \frac{1}{\sqrt{2G}} \left\| (\boldsymbol{\Sigma}_k^y(\widetilde{\boldsymbol{\beta}}))^{-1/2} \boldsymbol{m}_k^y(G, \widetilde{\boldsymbol{\beta}}) \right\|,$$

and the difference vector at time k, evaluated with inspection parameter $\widetilde{\boldsymbol{\beta}}$, is

$$\boldsymbol{m}_k^y(G, \widetilde{\boldsymbol{\beta}}) = \sum_{t=k+1}^{k+G} \boldsymbol{H}^y\left(y_t, \boldsymbol{z}_t, \widetilde{\boldsymbol{\beta}}\right) - \sum_{t=k-G+1}^{k} \boldsymbol{H}^y\left(y_t, \boldsymbol{z}_t, \widetilde{\boldsymbol{\beta}}\right).$$

Proof of Lemma 5.2

Proof. The proof is similar to that of Lemma 5.1, and hence omitted. ■

Proof of Theorem 5.4

Proof. The proof is similar to that of Theorem 5.3, using Lemmas 5.2, C.4 and C.5. ■

C.3 Further simulations

Settings We adopt the following two designs from Cho et al. (2022), where the common component is drawn from a Generalised Dynamic Factor Model (GDFM). Let L denote the backshift operator, and $[\cdot]$ denote the rounding function. Both examples use $n = 2000$, and we vary $p = 50, 100, 150$. The common component is generated such that $\boldsymbol{\chi} = \boldsymbol{\chi}_t^{[j]}$ for $k_{j-1} + 1 \le t \le k_j$.

(GDFM1) $\boldsymbol{\chi}_t^{[j]}$ admits a static factor model representation, as

$$\chi_{it}^{[j]} = \sum_{j=1}^{r} (B_{0,ij}^j + B_{1,ii'}^j L + B_{2,ii'}^j L^2) u_{i't}, \quad 1 \le j \le q+1,$$

where $u_{i't} \sim_{iid} N(0,\sigma_j^2)$ with $(\sigma_1,\sigma_2) = (1,0.5)$, and the MA coefficients are generated as $(B^j_{0,ii'},B^j_{1,ii'},B^j_{2,ii'}) \sim_{iid} N(\mathbf{0},\mathbf{I}_3)$ for all $1 \le i \le p$ and $1 \le i' \le r$ when $q = 0$. Then sequentially for $j = 1,\dots,q$, we draw $\Pi^k_\chi \subset \{1,\dots,p\}$ with $|\Pi^k_\chi| = [0.5p]$ such that for all i', $(B^j_{0,ii'},B^j_{1,ii'},B^j_{2,ii'}) \sim_{iid} N(\mathbf{0},\mathbf{I}_3)$ when $i \in \Pi^j_\chi$ while $(B^j_{0,ii'},B^j_{1,ii'},B^j_{2,ii'}) = (B^{j-1}_{0,ii'},B^{j-1}_{1,ii'},B^{j-1}_{2,ii'})$ when $i \notin \Pi^j_\chi$. We set $(k_1,k_2,k_3) = (500,1000,1500)$.

(GDFM2) $\boldsymbol{\chi}^{[j]}_t$ does not admit a static factor model representation, as

$$\chi^{[j]}_{it} = \sum_{i'=1}^{r} \{a_{ii'}(1-\alpha^r_{ii'}L)^{-1}\}u_{i't}, \quad 1 \le j \le q+1,$$

where $u_{i't} \sim_{iid} N(0,1)$ and the coefficients a_{ij} are drawn uniformly as $a_{ij} \sim_{iid} U[-1,1]$ with $U[a,b]$ denoting a uniform distribution. The AR coefficients are generated as $\alpha^j_{ij} \sim_{iid} U[-0.8,0.8]$ when $q = 0$ and then sequentially for $j = 1,\dots,q$, we draw $\Pi^j_\chi \subset \{1,\dots,p\}$ with $|\Pi^j_\chi| = [0.5p]$ such that for all i', we have $\alpha^{[j]}_{ii'} = -\alpha^{[j-1]}_{ii'}$ when $i \in \Pi^j_\chi$ and $\alpha^j_{ii'} = \alpha^{[j-1]}_{ii'}$ when $i \notin \Pi^j_\chi$. We set $(k_1,k_2) = (666,1333)$.

Results We report the results in Table C.3.1. We also report results localising with $\epsilon = 0.3$. As we should expect, `fvarseg` outperforms `mosumfvar` in these settings. `mosumfvar` estimates the locations of changes reasonably under Setting (GDFM1), but lacks detection power under (GDFM2). As per Remark 5.1, the current `mosumfvar` algorithm is not designed to detect changes to the contemporaneous structure, so allowing for changes in $\boldsymbol{S}$ may improve the results seen here.

Table C.3.1: (GDFM1)–(GDFM2): Distributions of $\widehat{q} - q$ and the covering metric $\mathscr{C}(\widehat{\mathscr{P}}, \mathscr{P})$ of the estimated segmentations when $q \geq 1$, and the empirical size when $q = 0$ returned by `mosumfvar` with automatic parameter selection, using $\eta = 0.3$ and $\epsilon = 0.3$, and `fvarseg`. The best performer for each metric is given in bold.

			$\widehat{q}$–q								
Model	Method	p	-3	-2	-1	0	1	2	3	CM	Size
(GDFM1)	`mosumfvar`	50	51	14	7	14	7	1	6	0.4796	0.99
	(η)	100	44	14	5	21	10	4	2	0.5537	1
		150	31	39	3	19	4	3	1	0.5432	0.99
	`mosumfvar`	50	49	4	5	20	8	13	1	0.5501	1
	(ϵ)	100	41	4	4	24	17	10	0	0.6082	1
		150	31	4	1	51	11	1	1	0.7155	1
	`fvarseg`	50	2	3	15	**80**	0	0	0	**0.9186**	**0**
		100	1	3	11	**85**	0	0	0	**0.9390**	**0**
		150	3	4	8	**85**	0	0	0	**0.9261**	**0.01**
(GDFM2)	`mosumfvar`	50	0	86	6	1	1	1	5	0.3782	0.99
	(η)	100	0	0	97	3	0	0	0	0.5788	1
		150	0	0	100	0	0	0	0	0.5631	1
	`mosumfvar`	50	0	71	11	3	4	4	7	0.4093	0.99
	(ϵ)	100	0	0	0	0	2	4	94	0.5208	1
		150	0	0	0	1	0	4	95	0.5045	1
	`fvarseg`	50	0	0	0	**99**	1	0	0	**0.9888**	**0.01**
		100	0	0	0	**100**	0	0	0	**0.9915**	**0.01**
		150	0	0	0	**100**	0	0	0	**0.9915**	**0.01**

APPENDIX D

APPENDIX TO FACTOR-ADJUSTED NETWORK ESTIMATION AND FORECASTING FOR HIGH-DIMENSIONAL TIME SERIES

D.1 Information criteria for factor number selection

Here we list information criteria for factor number estimation which are implemented in **fnets** and accessible by the functions `fnets`, `fnets.factor.model` and `factor.number` by setting the argument `ic.op` at an integer belonging to $\{1,\dots,6\}$. When `fm.restricted = FALSE`, we have

IC_1: $\left(\frac{1}{p}\sum_{j=b+1}^{p}\frac{1}{2m+1}\sum_{k=-m}^{m}\widehat{\mu}_{x,j}(\omega_k)\right) + b\cdot c\cdot(m^{-2}+\sqrt{m/n}+p^{-1})\cdot\log(\min(p,m^2,\sqrt{n/m}))$,

IC_2: $\left(\frac{1}{p}\sum_{j=b+1}^{p}\frac{1}{2m+1}\sum_{k=-m}^{m}\widehat{\mu}_{x,j}(\omega_k)\right) + b\cdot c\cdot(\min(p,m^2,\sqrt{n/m}))^{-1/2}$,

IC_3: $\left(\frac{1}{p}\sum_{j=b+1}^{p}\frac{1}{2m+1}\sum_{k=-m}^{m}\widehat{\mu}_{x,j}(\omega_k)\right) + b\cdot c\cdot(\min(p,m^2,\sqrt{n/m}))^{-1}\cdot\log(\min(p,m^2,\sqrt{n/m}))$,

IC_4: $\log\left(\frac{1}{p}\sum_{j=b+1}^{p}\frac{1}{2m+1}\sum_{k=-m}^{m}\widehat{\mu}_{x,j}(\omega_k)\right) + b\cdot c\cdot(m^{-2}+\sqrt{m/n}+p^{-1})\cdot\log(\min(p,m^2,\sqrt{n/m}))$,

IC_5: $\log\left(\frac{1}{p}\sum_{j=b+1}^{p}\frac{1}{2m+1}\sum_{k=-m}^{m}\widehat{\mu}_{x,j}(\omega_k)\right) + b\cdot c\cdot(\min(p,m^2,\sqrt{n/m}))^{-1/2}$,

IC_6: $\log\left(\frac{1}{p}\sum_{j=b+1}^{p}\frac{1}{2m+1}\sum_{k=-m}^{m}\widehat{\mu}_{x,j}(\omega_k)\right) + b\cdot c\cdot(\min(p,m^2,\sqrt{n/m}))^{-1}\cdot\log(\min(p,m^2,\sqrt{n/m}))$.

When `fm.restricted = TRUE`, we use one of

IC_1: $\left(\frac{1}{p}\sum_{j=b+1}^{p}\widehat{\mu}_{x,j}\right) + b\cdot c\cdot(n+p)/(np)\cdot\log(np/(n+p))$,

IC_2: $\left(\frac{1}{p}\sum_{j=b+1}^{p}\widehat{\mu}_{x,j}\right) + b\cdot c\cdot(n+p)/(np)\cdot\log(np/(n+p))$,

IC_3: $\left(\frac{1}{p}\sum_{j=b+1}^{p}\widehat{\mu}_{x,j}\right) + b\cdot c\cdot\log(\min(n,p))/(\min(n,p))$,

IC_4: $\log\left(\frac{1}{p}\sum_{j=b+1}^{p}\widehat{\mu}_{x,j}\right) + b\cdot c\cdot(n+p)/(np)\cdot\log(np/(n+p))$,

IC_5: $\log\left(\frac{1}{p}\sum_{j=b+1}^{p}\widehat{\mu}_{x,j}\right) + b\cdot c\cdot(n+p)/(np)\cdot\log(np/(n+p))$,

IC_6: $\log\left(\frac{1}{p}\sum_{j=b+1}^{p}\widehat{\mu}_{x,j}\right) + b\cdot c\cdot\log(\min(n,p))/(\min(n,p))$.

Whether `fm.restricted = FALSE` or not, the default choice is `ic.op = 5`.

D.2 Dataset information

D.2.1 Energy price data

Table D.2.1 defines the four node types in the panel. Table D.2.3 describes the dataset analysed in Data example.

Table D.2.1: Node type definitions for energy price data.

Name	Definition
Zone	A transmission owner's area within the PJM Region.
Aggregate	A group of more than one individual bus into a pricing node (pnode) that is considered as a whole in the Energy Market and other various systems and Markets within PJM.
Hub	A group of more than one individual bus into a regional pricing node (pnode) developed to produce a stable price signal in the Energy Market and other various systems and Markets within PJM.
Extra High Voltage (EHV)	Nodes at 345kV and above on the PJM system.

D.2.2 Equity volatility measures

Table D.2.3 provides the list of the 46 companies included in the application presented in Section 6.6.2 along with their tickers and industry and sub-industry classifications according to Global Industry Classification Standard.

D.3 Complete simulation results

Here we report the full set of simulation results for Sections 6.5.2–6.5.3, from Figure D.3.1 to Table D.3.13.

D.3.1 Estimation

In Tables D.3.1–D.3.3, we report the estimation errors of $\widehat{\boldsymbol{\beta}}^{\text{las}}$ and $\widehat{\boldsymbol{\beta}}^{\text{DS}}$ in estimating $\boldsymbol{\beta}^0$, and in Tables D.3.2–D.3.4, those of $\widehat{\boldsymbol{\Omega}}^{\text{las}}$ and $\widehat{\boldsymbol{\Omega}}^{\text{DS}}$ ($\widehat{\boldsymbol{\Omega}}$ obtained with $\widehat{\boldsymbol{\beta}}^{\text{las}}$ and $\widehat{\boldsymbol{\beta}}^{\text{DS}}$, respectively) in estimating $\boldsymbol{\Omega}$ averaged over 100 realisations, and the corresponding standard errors. In Figures D.3.1–D.3.2 and Figure 6.7 we report FPR and TPR values for support recovery. We additionally report the results of the TPR value when FPR is set at 0.05, with and without thresholding the estimators as described in Section 6.3.

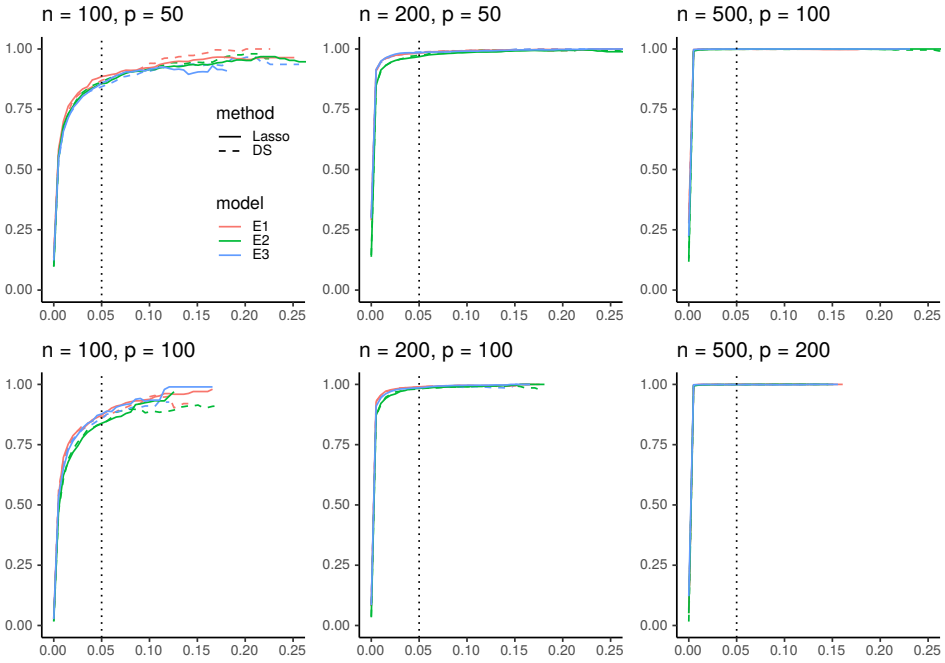


Figure D.3.1: ROC curves of TPR against FPR for $\widehat{\boldsymbol{\beta}}^{\text{las}}$, $\widehat{\boldsymbol{\beta}}^{\text{DS}}$ and $\widehat{\boldsymbol{\beta}}^{\text{FARM}}$ in recovering the support of $\boldsymbol{\beta}^0$ when $\boldsymbol{\chi}_t$ is generated under (C1) and $\boldsymbol{\xi}_t$ is generated under (E2)–(E4) with varying n and p, averaged over 100 realisations. Vertical lines indicate FPR = 0.05. For comparison, we also plot the corresponding curves (from $\widehat{\boldsymbol{\beta}}^{\text{las}}$ and $\widehat{\boldsymbol{\beta}}^{\text{DS}}$) obtained under (C0) i.e. when $\boldsymbol{\chi}_t = \mathbf{0}$.

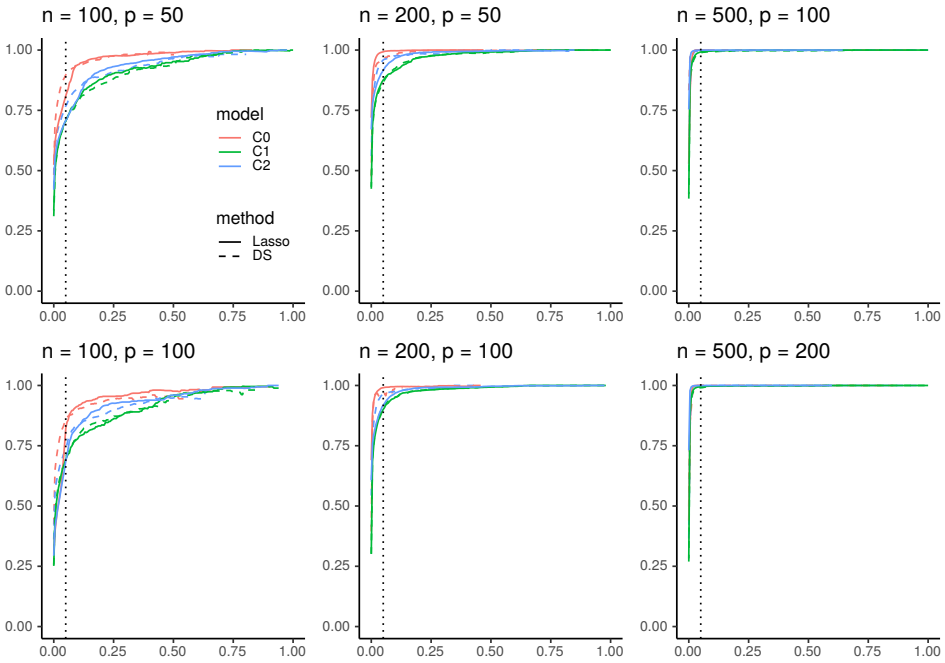


Figure D.3.2: ROC curves of TPR against FPR for $\widehat{\boldsymbol{\Omega}}^{\text{las}}$ and $\widehat{\boldsymbol{\Omega}}^{\text{DS}}$ in recovering the support of $\boldsymbol{\Omega}$ when $\boldsymbol{\chi}_t$ is generated under (C1)–(C2) and $\boldsymbol{\xi}_t$ is generated under (E1) with varying n and p, averaged over 100 realisations. Vertical lines indicate FPR = 0.05. For comparison, we also plot the corresponding curves obtained under (C0) i.e. when $\boldsymbol{\chi}_t = \mathbf{0}$.

Table D.2.2: Names, IDs and Types for the 50 power nodes in the energy price dataset.

Name	Node ID	Node Type
PJM	1	ZONE
AECO	51291	ZONE
BGE	51292	ZONE
DPL	51293	ZONE
JCPL	51295	ZONE
METED	51296	ZONE
PECO	51297	ZONE
PEPCO	51298	ZONE
PPL	51299	ZONE
PENELEC	51300	ZONE
PSEG	51301	ZONE
BRANDONSH	51205	AGGREGATE
BRUNSWICK	51206	AGGREGATE
COOKSTOWN	51211	AGGREGATE
DOVER	51214	AGGREGATE
DPL NORTH	51215	AGGREGATE
DPL SOUTH	51216	AGGREGATE
EASTON	51218	AGGREGATE
ECRRF	51219	AGGREGATE
EPHRATA	51220	AGGREGATE
FAIRLAWN	51221	AGGREGATE
HOMERCIT	51229	AGGREGATE
HOMERCIT UNIT1	51230	AGGREGATE
HOMERCIT UNIT2	51231	AGGREGATE
HOMERCIT UNIT3	51232	AGGREGATE
KITTATNY 230	51238	AGGREGATE
MANITOU	51239	AGGREGATE
MONTVILLE	51241	AGGREGATE
PENNTECH	51246	AGGREGATE
PPL_ALLUGI	51252	AGGREGATE
SENECA	51255	AGGREGATE
SOUTHRIV 230	51261	AGGREGATE
SUNBURY LBRG	51270	AGGREGATE
TRAYNOR	51277	AGGREGATE
UGI	51279	AGGREGATE
VINELAND	51280	AGGREGATE
WELLSBORO	51285	AGGREGATE
EASTERN HUB	51217	HUB
WEST INT HUB	51287	HUB
WESTERN HUB	51288	HUB
ALBURTIS	52443	EHV
BRANCHBURG	52444	EHV
BRIGHTON	52445	EHV
BURCHESHILL	52446	EHV
CALVERTC	52447	EHV
CHALKPT	52448	EHV
CONASTONE	52449	EHV
CONEMAUGH	52450	EHV
DEANS	52451	EHV
ELROY	52452	EHV

Table D.2.3: Tickers, industry and sub-industry classifications of the 46 companies.

Name	Ticker	Industry	Sub-industry
JPMORGAN CHASE & CO	JPM	Banks	Diversified banks
COMERICA INC	CMA	Banks	Regional banks
CITIGROUP INC	C	Banks	Diversified banks
FIFTH THIRD BANCORP	FITB	Banks	Regional banks
REGIONS FINANCIAL CORP	RF	Banks	Regional banks
M & T BANK CORP	MTB	Banks	Regional banks
U S BANCORP	USB	Banks	Diversified banks
HUNTINGTON BANCSHARES	HBAN	Banks	Regional banks
BANK OF AMERICA CORP	BAC	Banks	Diversified banks
WELLS FARGO & CO	WFC	Banks	Diversified banks
PNC FINANCIAL SVCS GROUP INC	PNC	Banks	Regional banks
KEYCORP	KEY	Banks	Regional banks
ZIONS BANCORPORATION NA	ZION	Banks	Regional banks
TRUIST FINANCIAL CORP	TFC	Banks	Regional banks
PEOPLE'S UNITED FINL INC	PBCT	Banks	Regional banks
SVB FINANCIAL GROUP	SIVB	Banks	Regional banks
AMERICAN EXPRESS CO	AXP	Diversified Financials	Consumer finance
BANK OF NEW YORK MELLON CORP	BK	Diversified Financials	Asset Management & Custody Banks
FRANKLIN RESOURCES INC	BEN	Diversified Financials	Asset Management & Custody Banks
S&P GLOBAL INC	SPGI	Diversified Financials	Financial Exchanges & Data
NORTHERN TRUST CORP	NTRS	Diversified Financials	Asset Management & Custody Banks
RAYMOND JAMES FINANCIAL CORP	RJF	Diversified Financials	Investment Banking & Brokerage
STATE STREET CORP	STT	Diversified Financials	Asset Management & Custody Banks
MORGAN STANLEY	MS	Diversified Financials	Investment Banking & Brokerage
PRICE (T. ROWE) GROUP	TROW	Diversified Financials	Asset Management & Custody Banks
SCHWAB (CHARLES) CORP	SCHW	Diversified Financials	Investment Banking & Brokerage
INVESCO LTD	IVZ	Diversified Financials	Asset Management & Custody Banks
CAPITAL ONE FINANCIAL CORP	COF	Diversified Financials	Consumer finance
GOLDMAN SACHS GROUP INC	GS	Diversified Financials	Investment Banking & Brokerage
BLACKROCK INC	BLK	Diversified Financials	Asset Management & Custody Banks
AFLAC INC	AFL	Insurance	Life & Health Insurance
AMERICAN INTERNATIONAL GROUP	AIG	Insurance	Multi-line Insurance
AON PLC	AON	Insurance	Insurance Brokers
ARTHUR J GALLAGHER & CO	AJG	Insurance	Insurance Brokers
LINCOLN NATIONAL CORP	LNC	Insurance	Life & Health Insurance
LOEWS CORP	L	Insurance	Property & Casualty Insurance
MARSH & MCLENNAN COS	MMC	Insurance	Insurance Brokers
GLOBE LIFE INC	GL	Insurance	Life & Health Insurance
UNUM GROUP	UNM	Insurance	Life & Health Insurance
PROGRESSIVE CORP-OHIO	PGR	Insurance	Property & Casualty Insurance
BERKLEY (W R) CORP	WRB	Insurance	Property & Casualty Insurance
CINCINNATI FINANCIAL CORP	CINF	Insurance	Property & Casualty Insurance
CHUBB LTD	CB	Insurance	Property & Casualty Insurance
ALLSTATE CORP	ALL	Insurance	Property & Casualty Insurance
EVEREST RE GROUP LTD	RE	Insurance	Reinsurance
HARTFORD FINANCIAL SERVICES	HIG	Insurance	Multi-line Insurance

Table D.3.1: Errors of $\widehat{\boldsymbol{\beta}}^{\text{las}}$, $\widehat{\boldsymbol{\beta}}^{\text{DS}}$ and $\widehat{\boldsymbol{\beta}}^{\text{FARM}}$ in estimating $\boldsymbol{\beta}^0$ measured by L_F and L_2 averaged over 100 realisations (also reported are the standard errors) under the model (E1) for the generation of $\boldsymbol{\xi}_t$ and (C0)–(C2) for $\boldsymbol{\chi}_t$ with varying n and p. We also report the TPR when FPR = 0.05 without and with thresholding for $\widehat{\boldsymbol{\beta}}^{\text{las}}$ and $\widehat{\boldsymbol{\beta}}^{\text{DS}}$.

								TPR (5%)			
				L_F		L_2		Without		With	
	n	p	Method	Mean	SD	Mean	SD	Mean	SD	Mean	SD
(C0)	100	50	Lasso	0.633	0.177	0.68	0.163	0.836	0.243	0.815	0.262
			DS	0.613	0.072	0.679	0.082	0.94	0.105	0.893	0.125
	100	100	Lasso	0.854	0.142	0.88	0.129	0.517	0.27	0.466	0.283
			DS	0.669	0.076	0.738	0.079	0.892	0.16	0.839	0.159
	200	50	Lasso	0.421	0.048	0.473	0.059	0.999	0.004	0.997	0.009
			DS	0.532	0.071	0.589	0.084	0.989	0.019	0.982	0.024
	200	100	Lasso	0.454	0.034	0.52	0.056	0.999	0.004	0.996	0.008
			DS	0.58	0.044	0.654	0.066	0.982	0.018	0.966	0.031
	500	100	Lasso	0.402	0.034	0.441	0.041	1	0	0.987	0.020
			DS	0.29	0.074	0.308	0.085	1	0.001	0.998	0.008
	500	200	Lasso	0.425	0.034	0.47	0.048	1	0.001	0.986	0.024
			DS	0.46	0.128	0.493	0.15	0.999	0.002	0.98	0.021
(C1)	100	50	Lasso	0.805	0.094	0.875	0.111	0.757	0.216	0.681	0.252
			DS	0.815	0.084	0.883	0.107	0.748	0.19	0.684	0.209
			FARM	0.914	0.047	0.954	0.088	0.404	0.127	-	-
	100	100	Lasso	0.863	0.077	0.925	0.098	0.66	0.228	0.561	0.257
			DS	0.848	0.071	0.924	0.09	0.701	0.209	0.608	0.223
			FARM	0.927	0.026	0.96	0.086	0.361	0.086	-	-
	200	50	Lasso	0.613	0.075	0.708	0.111	0.973	0.038	0.951	0.089
			DS	0.617	0.083	0.715	0.119	0.969	0.052	0.951	0.070
			FARM	0.804	0.057	0.871	0.135	0.726	0.106	-	-
	200	100	Lasso	0.647	0.08	0.794	0.094	0.963	0.062	0.936	0.094
			DS	0.643	0.072	0.776	0.102	0.971	0.039	0.941	0.079
			FARM	0.794	0.045	0.841	0.095	0.733	0.098	-	-
	500	100	Lasso	0.461	0.054	0.657	0.094	0.999	0.003	0.996	0.015
			DS	0.48	0.057	0.665	0.107	0.999	0.004	0.998	0.006
			FARM	0.625	0.037	0.725	0.124	0.961	0.03	-	-
	500	200	Lasso	0.501	0.058	0.763	0.083	0.999	0.003	0.996	0.008
			DS	0.518	0.066	0.813	0.107	0.999	0.003	0.961	0.176
			FARM	0.611	0.035	0.704	0.122	0.969	0.021	-	-
(C2)	100	50	Lasso	0.721	0.118	0.756	0.116	0.819	0.236	0.805	0.246
			DS	0.704	0.057	0.759	0.074	0.888	0.08	0.837	0.094
			FARM	0.857	0.046	0.888	0.071	0.534	0.137	-	-
	100	100	Lasso	0.868	0.084	0.886	0.089	0.572	0.251	0.517	0.274
			DS	0.749	0.08	0.786	0.077	0.826	0.216	0.766	0.214
			FARM	0.882	0.031	0.894	0.071	0.483	0.085	-	-
	200	50	Lasso	0.503	0.04	0.551	0.065	0.996	0.01	0.994	0.012
			DS	0.575	0.051	0.635	0.077	0.988	0.02	0.971	0.034
			FARM	0.737	0.057	0.774	0.093	0.821	0.089	-	-
	200	100	Lasso	0.53	0.05	0.559	0.057	0.995	0.019	0.99	0.026
			DS	0.568	0.042	0.625	0.062	0.987	0.015	0.973	0.023
			FARM	0.726	0.046	0.722	0.064	0.83	0.078	-	-
	500	100	Lasso	0.374	0.026	0.417	0.042	1	0	1	0.000
			DS	0.448	0.05	0.494	0.064	1	0.001	0.994	0.016
			FARM	0.551	0.043	0.566	0.089	0.99	0.018	-	-
	500	200	Lasso	0.383	0.023	0.425	0.035	1	0	1	0.000
			DS	0.478	0.033	0.528	0.045	1	0.001	0.995	0.018
			FARM	0.559	0.033	0.551	0.046	0.988	0.012	-	-

Table D.3.2: Errors of $\widehat{\mathbf{\Omega}}^{\text{las}}$ and $\widehat{\mathbf{\Omega}}^{\text{DS}}$ in estimating $\mathbf{\Omega}$ measured by L_F and L_2, averaged over 100 realisations (also reported are the standard errors) under the model (E1) for the generation of $\boldsymbol{\xi}_t$ and (C0)–(C2) for $\boldsymbol{\chi}_t$ with varying n and p. We also report the TPR when FPR = 0.05 without and with thresholding.

								TPR (5%)			
				L_F		L_2		Without		With	
	n	p	Method	Mean	SD	Mean	SD	Mean	SD	Mean	SD
(C0)	100	50	Lasso	0.452	0.087	0.587	0.105	0.789	0.18	0.704	0.236
			DS	0.456	0.042	0.593	0.048	0.896	0.092	0.721	0.111
	100	100	Lasso	0.587	0.095	0.738	0.1	0.583	0.17	0.411	0.195
			DS	0.467	0.054	0.624	0.056	0.846	0.113	0.658	0.106
	200	50	Lasso	0.373	0.026	0.488	0.038	0.993	0.016	0.854	0.090
			DS	0.423	0.043	0.553	0.058	0.979	0.042	0.755	0.134
	200	100	Lasso	0.376	0.022	0.507	0.03	0.991	0.016	0.839	0.064
			DS	0.444	0.024	0.593	0.03	0.98	0.021	0.727	0.069
	500	100	Lasso	0.327	0.02	0.453	0.029	1	0.001	0.784	0.040
			DS	0.236	0.045	0.328	0.064	1	0.002	0.975	0.067
	500	200	Lasso	0.328	0.017	0.466	0.029	1	0.001	0.77	0.029
			DS	0.336	0.059	0.473	0.09	0.999	0.003	0.845	0.123
(C1)	100	50	Lasso	0.486	0.057	0.652	0.154	0.697	0.148	0.578	0.169
			DS	0.488	0.064	0.662	0.18	0.691	0.129	0.545	0.162
	100	100	Lasso	0.515	0.069	0.696	0.099	0.641	0.127	0.475	0.153
			DS	0.503	0.062	0.687	0.137	0.662	0.118	0.498	0.155
	200	50	Lasso	0.474	0.723	0.812	2.66	0.876	0.106	0.769	0.145
			DS	0.403	0.052	0.563	0.123	0.872	0.089	0.769	0.131
	200	100	Lasso	0.416	0.046	0.573	0.071	0.898	0.071	0.728	0.149
			DS	0.417	0.048	0.572	0.066	0.91	0.059	0.737	0.130
	500	100	Lasso	0.33	0.033	0.488	0.068	0.992	0.014	0.881	0.089
			DS	0.337	0.037	0.495	0.065	0.989	0.019	0.864	0.096
	500	200	Lasso	0.348	0.04	0.523	0.055	0.995	0.008	0.841	0.088
			DS	0.35	0.046	0.535	0.062	0.992	0.018	0.828	0.103
(C2)	100	50	Lasso	0.433	0.067	0.576	0.093	0.696	0.159	0.666	0.195
			DS	0.433	0.033	0.584	0.044	0.768	0.098	0.668	0.108
	100	100	Lasso	0.526	0.084	0.68	0.102	0.595	0.133	0.446	0.199
			DS	0.458	0.06	0.617	0.065	0.727	0.138	0.617	0.130
	200	50	Lasso	0.349	0.03	0.48	0.046	0.915	0.068	0.843	0.077
			DS	0.399	0.034	0.541	0.043	0.96	0.047	0.769	0.097
	200	100	Lasso	0.334	0.027	0.471	0.042	0.917	0.054	0.838	0.078
			DS	0.391	0.03	0.544	0.038	0.966	0.03	0.76	0.066
	500	100	Lasso	0.287	0.019	0.413	0.032	1	0.001	0.884	0.080
			DS	0.321	0.043	0.456	0.058	0.998	0.008	0.819	0.086
	500	200	Lasso	0.292	0.022	0.428	0.028	1	0.001	0.889	0.067
			DS	0.34	0.021	0.491	0.034	1	0.002	0.775	0.050

Table D.3.3: Errors of $\widehat{\boldsymbol{\beta}}^{\text{las}}$, $\widehat{\boldsymbol{\beta}}^{\text{DS}}$ and $\widehat{\boldsymbol{\beta}}^{\text{FARM}}$ in estimating $\boldsymbol{\beta}^0$ measured by L_F and L_2 averaged over 100 realisations (also reported are the standard errors) under the models (E2)–(E4) for the generation of $\boldsymbol{\xi}_t$ and (C1) for $\boldsymbol{\chi}_t$ with varying n and p. We also report the TPR when FPR = 0.05 without and with thresholding.

								TPR (5%)			
				L_F		L_2		Without		With	
	n	p	Method	Mean	SD	Mean	SD	Mean	SD	Mean	SD
(E2)	100	50	Lasso	0.814	0.088	0.898	0.229	0.787	0.161	0.727	0.188
			DS	0.82	0.092	0.908	0.199	0.753	0.185	0.677	0.230
	100	100	Lasso	0.883	0.063	0.963	0.099	0.636	0.216	0.536	0.234
			DS	0.889	0.074	0.983	0.143	0.662	0.228	0.552	0.244
	200	50	Lasso	0.655	0.068	0.753	0.105	0.959	0.053	0.941	0.079
			DS	0.651	0.082	0.743	0.111	0.959	0.052	0.932	0.086
	200	100	Lasso	0.694	0.07	0.842	0.086	0.948	0.07	0.904	0.116
			DS	0.697	0.081	0.849	0.108	0.941	0.08	0.893	0.143
	500	100	Lasso	0.519	0.062	0.73	0.109	0.998	0.004	0.996	0.009
			DS	0.524	0.06	0.755	0.111	0.999	0.004	0.997	0.007
	500	200	Lasso	0.549	0.055	0.83	0.088	0.997	0.004	0.993	0.010
			DS	0.557	0.05	0.907	0.092	0.997	0.006	0.993	0.014
(E4)	100	50	Lasso	0.813	0.092	0.867	0.111	0.745	0.183	0.676	0.218
			DS	0.829	0.09	0.893	0.111	0.709	0.225	0.649	0.247
	100	100	Lasso	0.857	0.078	0.936	0.104	0.654	0.201	0.558	0.223
			DS	0.864	0.08	0.946	0.126	0.635	0.234	0.538	0.270
	200	50	Lasso	0.617	0.07	0.701	0.095	0.972	0.048	0.95	0.086
			DS	0.617	0.075	0.699	0.094	0.97	0.037	0.949	0.060
	200	100	Lasso	0.668	0.078	0.808	0.1	0.948	0.066	0.909	0.122
			DS	0.655	0.087	0.796	0.11	0.953	0.07	0.918	0.122
	500	100	Lasso	0.474	0.055	0.648	0.095	0.999	0.004	0.998	0.007
			DS	0.474	0.062	0.653	0.118	0.999	0.003	0.998	0.006
	500	200	Lasso	0.489	0.055	0.766	0.085	0.999	0.002	0.998	0.005
			DS	0.516	0.053	0.811	0.11	0.999	0.003	0.988	0.082

Table D.3.4: Errors of $\widehat{\mathbf{\Omega}}^{\text{las}}$ and $\widehat{\mathbf{\Omega}}^{\text{DS}}$ in estimating $\mathbf{\Omega}$ measured by L_F and L_2, averaged over 100 realisations (also reported are the standard errors) under the models (E2)–(E4) for the generation of $\boldsymbol{\xi}_t$ and (C1) for $\boldsymbol{\chi}_t$ with varying n and p. We also report the TPR when FPR = 0.05 without and with thresholding.

								TPR (5%)			
				L_F		L_2		Without		With	
	n	p	Method	Mean	SD	Mean	SD	Mean	SD	Mean	SD
(E2)	100	50	Lasso	0.582	0.055	0.732	0.155	0.407	0.071	0.329	0.086
			DS	0.582	0.054	0.726	0.105	0.396	0.076	0.323	0.099
	100	100	Lasso	0.615	0.061	0.756	0.059	0.352	0.065	0.25	0.071
			DS	0.621	0.059	0.765	0.064	0.35	0.066	0.243	0.078
	200	50	Lasso	0.504	0.066	0.649	0.157	0.513	0.072	0.454	0.080
			DS	0.502	0.038	0.638	0.06	0.514	0.065	0.452	0.087
	200	100	Lasso	0.522	0.047	0.658	0.067	0.515	0.06	0.392	0.080
			DS	0.525	0.048	0.669	0.064	0.518	0.063	0.392	0.089
	500	100	Lasso	0.442	0.042	0.61	0.149	0.646	0.06	0.524	0.079
			DS	0.436	0.042	0.594	0.135	0.635	0.058	0.53	0.085
	500	200	Lasso	0.457	0.039	0.608	0.066	0.674	0.043	0.484	0.059
			DS	0.447	0.038	0.598	0.063	0.659	0.041	0.493	0.064
(E4)	100	50	Lasso	0.495	0.057	0.652	0.113	0.684	0.125	0.546	0.171
			DS	0.502	0.056	0.655	0.097	0.661	0.147	0.525	0.168
	100	100	Lasso	0.519	0.055	0.696	0.082	0.645	0.125	0.46	0.148
			DS	0.521	0.058	0.696	0.081	0.628	0.145	0.47	0.161
	200	50	Lasso	0.4	0.048	0.541	0.073	0.882	0.071	0.763	0.122
			DS	0.403	0.045	0.547	0.067	0.885	0.066	0.763	0.132
	200	100	Lasso	0.429	0.05	0.586	0.074	0.893	0.073	0.69	0.160
			DS	0.423	0.049	0.571	0.071	0.895	0.074	0.72	0.154
	500	100	Lasso	0.338	0.039	0.499	0.064	0.992	0.016	0.856	0.109
			DS	0.336	0.039	0.499	0.067	0.989	0.019	0.867	0.094
	500	200	Lasso	0.349	0.04	0.522	0.056	0.996	0.009	0.849	0.087
			DS	0.357	0.042	0.542	0.061	0.995	0.009	0.825	0.082

D.3.2 Forecasting

See Tables D.3.6–D.3.13 reporting the errors in estimating the best linear predictors according to different measures as well as forecasting errors, along with Table D.3.5 containing the benchmark results obtained in the oracle setting of (C0). The estimators $\widehat{\boldsymbol{\xi}}^{\mathrm{las}}_{t+1|t}$ and $\widehat{\boldsymbol{\xi}}^{\mathrm{DS}}_{t+1|t}$ depend on the choice of the in-sample estimator of $\boldsymbol{\chi}_t$ (which automatically yields the in-sample estimator of $\boldsymbol{\xi}_t$) but we suppress this dependence in their notations.

Table D.3.5: Errors in forecasting $\mathbf{X}_{n+1}$ by the FNETS measured by (6.24)–(6.27) averaged over 100 realisations (also reported are the standard errors) under the model (E1) for the generation of $\boldsymbol{\xi}_t$ and (C0) for $\boldsymbol{\chi}_t$ with varying n and p, which serve as a benchmark.

	n	p	Method	Mean	SD		n	p	Method	Mean	SD
(6.24)	100	50	Lasso	0.437	0.253	(6.26)	100	50	Lasso	0.898	0.090
			DS	0.378	0.122				DS	0.896	0.077
	100	100	Lasso	0.748	0.208		100	100	Lasso	0.958	0.044
			DS	0.463	0.133				DS	0.913	0.045
	200	50	Lasso	0.176	0.056		200	50	Lasso	0.853	0.078
			DS	0.277	0.092				DS	0.87	0.069
	200	100	Lasso	0.207	0.036		200	100	Lasso	0.873	0.054
			DS	0.326	0.054				DS	0.891	0.046
	500	100	Lasso	0.161	0.035		500	100	Lasso	0.876	0.054
			DS	0.095	0.058				DS	0.868	0.065
	500	200	Lasso	0.179	0.032		500	200	Lasso	0.879	0.035
			DS	0.225	0.108				DS	0.884	0.041
(6.25)	100	50	Lasso	0.651	0.221	(6.27)	100	50	Lasso	0.935	0.117
			DS	0.638	0.146				DS	0.94	0.109
	100	100	Lasso	0.873	0.143		100	100	Lasso	0.974	0.051
			DS	0.724	0.149				DS	0.951	0.071
	200	50	Lasso	0.447	0.116		200	50	Lasso	0.93	0.102
			DS	0.556	0.14				DS	0.939	0.091
	200	100	Lasso	0.486	0.102		200	100	Lasso	0.929	0.092
			DS	0.599	0.107				DS	0.936	0.078
	500	100	Lasso	0.418	0.073		500	100	Lasso	0.913	0.103
			DS	0.305	0.106				DS	0.906	0.113
	500	200	Lasso	0.468	0.079		500	200	Lasso	0.919	0.081
			DS	0.49	0.146				DS	0.919	0.083

Table D.3.6: Forecasting errors of FNETS and FARM measured by (6.24) averaged over 100 realisations (also reported are the standard errors) under the model (E1) for the generation of $\boldsymbol{\xi}_t$ and (C1)–(C2) for $\boldsymbol{\chi}_t$ with varying n and p. We also report the errors of restricted and unrestricted in-sample estimators of $\boldsymbol{\chi}_t$, $1 \le t \le n$.

			Method		In-sample		$\boldsymbol{\chi}_{n+1\|n}$		$\boldsymbol{\xi}_{n+1\|n}$		$\mathbf{x}_{n+1\|n}$	
	n	p	$\boldsymbol{\chi}_t$	$\boldsymbol{\xi}_t$	Mean	SD	Mean	SD	Mean	SD	Mean	SD
(C1)	100	50	Restricted	Lasso	0.355	0.095	0.517	0.358	0.732	0.181	0.491	0.206
			Unrestricted		0.37	0.104	0.543	0.309	0.814	0.326	0.491	0.198
			Restricted	DS	-	-	-	-	0.746	0.177	0.493	0.209
			Unrestricted		-	-	-	-	0.845	0.33	0.493	0.189
				FARM	0.381	0.143	2.88	7.88	1.07	0.482	1.76	2.460
	100	100	Restricted	Lasso	0.279	0.091	0.455	0.28	0.797	0.137	0.462	0.186
			Unrestricted		0.319	0.094	0.495	0.233	0.924	0.374	0.462	0.159
			Restricted	DS	-	-	-	-	0.78	0.128	0.453	0.174
			Unrestricted		-	-	-	-	0.899	0.36	0.452	0.155
				FARM	0.288	0.119	4.39	9.37	1.08	0.511	3.2	6.150
	200	50	Restricted	Lasso	0.287	0.07	0.517	0.383	0.52	0.186	0.41	0.205
			Unrestricted		0.391	0.116	0.662	0.543	0.628	0.375	0.456	0.201
			Restricted	DS	-	-	-	-	0.526	0.172	0.409	0.198
			Unrestricted		-	-	-	-	0.646	0.456	0.461	0.212
				FARM	0.291	0.084	0.574	0.418	0.989	1.01	0.568	0.310
	200	100	Restricted	Lasso	0.215	0.065	0.379	0.307	0.522	0.137	0.342	0.156
			Unrestricted		0.259	0.086	0.473	0.382	0.651	0.323	0.35	0.138
			Restricted	DS	-	-	-	-	0.509	0.119	0.342	0.156
			Unrestricted		-	-	-	-	0.637	0.35	0.354	0.151
				FARM	0.216	0.072	0.539	0.696	0.873	0.481	0.516	0.452
	500	100	Restricted	Lasso	0.148	0.022	0.224	0.173	0.276	0.08	0.204	0.100
			Unrestricted		0.274	0.091	0.386	0.231	0.429	0.235	0.262	0.120
			Restricted	DS	-	-	-	-	0.291	0.085	0.207	0.103
			Unrestricted		-	-	-	-	0.452	0.259	0.262	0.113
				FARM	0.148	0.022	0.251	0.231	0.698	0.477	0.318	0.162
	500	200	Restricted	Lasso	0.1	0.016	0.183	0.115	0.278	0.067	0.185	0.079
			Unrestricted		0.21	0.099	0.306	0.159	0.452	0.201	0.216	0.091
			Restricted	DS	-	-	-	-	0.297	0.076	0.188	0.077
			Unrestricted		-	-	-	-	0.474	0.218	0.22	0.095
				FARM	0.1	0.016	0.207	0.16	0.673	0.373	0.291	0.154
(C2)	100	50	Restricted	Lasso	0.165	0.029	0.874	2.77	0.596	0.19	0.444	0.318
			Unrestricted		0.379	0.145	0.639	0.585	0.673	0.226	0.58	0.327
			Restricted	DS	-	-	-	-	0.565	0.116	0.431	0.296
			Unrestricted		-	-	-	-	0.607	0.161	0.569	0.331
				FARM	0.167	0.031	3.04	18.1	0.846	0.197	0.798	0.992
	100	100	Restricted	Lasso	0.118	0.03	0.573	1.26	0.779	0.132	0.454	0.302
			Unrestricted		0.333	0.165	0.57	0.628	0.803	0.126	0.611	0.278
			Restricted	DS	-	-	-	-	0.619	0.133	0.39	0.257
			Unrestricted		-	-	-	-	0.674	0.131	0.558	0.253
				FARM	0.118	0.03	4.82	37.9	0.9	0.185	0.917	1.820
	200	50	Restricted	Lasso	0.125	0.016	0.77	2.43	0.343	0.09	0.265	0.233
			Unrestricted		0.509	0.142	0.667	0.439	0.451	0.218	0.522	0.293
			Restricted	DS	-	-	-	-	0.405	0.087	0.287	0.244
			Unrestricted		-	-	-	-	0.464	0.177	0.541	0.273
				FARM	0.136	0.029	0.75	2.42	0.668	0.202	0.397	0.357
	200	100	Restricted	Lasso	0.081	0.017	0.333	0.878	0.352	0.082	0.229	0.186
			Unrestricted		0.271	0.128	0.449	0.42	0.427	0.133	0.385	0.213
			Restricted	DS	-	-	-	-	0.378	0.071	0.242	0.198
			Unrestricted		-	-	-	-	0.424	0.098	0.392	0.209
				FARM	0.081	0.017	0.248	0.507	0.716	0.179	0.326	0.235
	500	100	Restricted	Lasso	0.057	0.007	0.289	0.903	0.189	0.045	0.135	0.107
			Unrestricted		0.37	0.132	0.536	0.553	0.284	0.142	0.361	0.189
			Restricted	DS	-	-	-	-	0.244	0.055	0.154	0.116
			Unrestricted		-	-	-	-	0.329	0.153	0.373	0.182
				FARM	0.06	0.013	0.323	1.07	0.553	0.239	0.241	0.172
	500	200	Restricted	Lasso	0.036	0.006	0.214	0.517	0.176	0.03	0.108	0.087
			Unrestricted		0.214	0.118	0.392	0.485	0.243	0.116	0.262	0.148
			Restricted	DS	-	-	-	-	0.25	0.038	0.132	0.100
			Unrestricted		-	-	-	-	0.297	0.091	0.286	0.153
				FARM	0.036	0.006	0.202	0.493	0.529	0.182	0.21	0.141

Table D.3.7: Forecasting errors of FNETS and FARM measured by (6.25) averaged over 100 realisations (also reported are the standard errors) under the model (E1) for the generation of $\boldsymbol{\xi}_t$ and (C1)–(C2) for $\boldsymbol{\chi}_t$ with varying n and p.

			Method		$\boldsymbol{\chi}_{n+1\|n}$		$\boldsymbol{\xi}_{n+1\|n}$		$\mathbf{X}_{n+1\|n}$	
	n	p	$\boldsymbol{\chi}_t$	$\boldsymbol{\xi}_t$	Mean	SD	Mean	SD	Mean	SD
(C1)	100	50	Restricted	Lasso	0.682	0.277	0.87	0.158	0.654	0.208
			Unrestricted		0.731	0.28	0.929	0.314	0.646	0.208
			Restricted	DS	-	-	0.883	0.201	0.653	0.210
			Unrestricted		-	-	0.971	0.354	0.652	0.199
				FARM	1.05	1.03	1.08	0.388	0.984	0.641
	100	100	Restricted	Lasso	0.636	0.288	0.925	0.159	0.594	0.223
			Unrestricted		0.71	0.247	1.04	0.413	0.607	0.199
			Restricted	DS	-	-	0.928	0.154	0.602	0.212
			Unrestricted		-	-	1.04	0.412	0.605	0.203
				FARM	1.12	0.872	1.1	0.413	1.09	0.768
	200	50	Restricted	Lasso	0.704	0.357	0.742	0.208	0.633	0.266
			Unrestricted		0.825	0.414	0.866	0.435	0.669	0.246
			Restricted	DS	-	-	0.755	0.211	0.635	0.258
			Unrestricted		-	-	0.881	0.456	0.674	0.252
				FARM	0.677	0.318	1.08	0.702	0.725	0.305
	200	100	Restricted	Lasso	0.636	0.355	0.733	0.19	0.557	0.211
			Unrestricted		0.746	0.386	0.917	0.443	0.562	0.205
			Restricted	DS	-	-	0.722	0.164	0.554	0.210
			Unrestricted		-	-	0.91	0.452	0.568	0.210
				FARM	0.614	0.359	0.969	0.46	0.619	0.287
	500	100	Restricted	Lasso	0.461	0.22	0.536	0.123	0.449	0.187
			Unrestricted		0.671	0.315	0.8	0.413	0.507	0.199
			Restricted	DS	-	-	0.545	0.125	0.447	0.178
			Unrestricted		-	-	0.797	0.427	0.497	0.180
				FARM	0.409	0.22	0.941	0.616	0.548	0.272
	500	200	Restricted	Lasso	0.44	0.201	0.549	0.126	0.432	0.171
			Unrestricted		0.585	0.194	0.863	0.382	0.441	0.157
			Restricted	DS	-	-	0.561	0.136	0.439	0.171
			Unrestricted		-	-	0.873	0.37	0.445	0.159
				FARM	0.374	0.224	0.943	0.559	0.501	0.292
(C2)	100	50	Restricted	Lasso	0.834	0.801	0.77	0.188	0.666	0.210
			Unrestricted		0.857	0.347	0.805	0.174	0.758	0.199
			Restricted	DS	-	-	0.776	0.152	0.665	0.205
			Unrestricted		-	-	0.793	0.147	0.752	0.192
				FARM	1.09	1.66	0.921	0.162	0.865	0.360
	100	100	Restricted	Lasso	0.781	0.665	0.875	0.138	0.745	0.249
			Unrestricted		0.842	0.518	0.882	0.135	0.828	0.178
			Restricted	DS	-	-	0.792	0.147	0.685	0.208
			Unrestricted		-	-	0.818	0.161	0.782	0.172
				FARM	1.08	2.54	0.924	0.156	0.873	0.451
	200	50	Restricted	Lasso	0.719	0.727	0.586	0.115	0.498	0.167
			Unrestricted		0.847	0.233	0.641	0.168	0.678	0.177
			Restricted	DS	-	-	0.651	0.118	0.543	0.175
			Unrestricted		-	-	0.679	0.147	0.706	0.160
				FARM	0.681	0.676	0.787	0.183	0.624	0.202
	200	100	Restricted	Lasso	0.593	0.553	0.594	0.13	0.514	0.157
			Unrestricted		0.75	0.308	0.638	0.144	0.624	0.161
			Restricted	DS	-	-	0.642	0.12	0.549	0.170
			Unrestricted		-	-	0.671	0.123	0.64	0.150
				FARM	0.482	0.35	0.829	0.177	0.647	0.223
	500	100	Restricted	Lasso	0.459	0.382	0.449	0.089	0.384	0.115
			Unrestricted		0.808	0.336	0.526	0.16	0.594	0.148
			Restricted	DS	-	-	0.5	0.104	0.414	0.128
			Unrestricted		-	-	0.555	0.149	0.612	0.149
				FARM	0.435	0.397	0.725	0.197	0.563	0.182
	500	200	Restricted	Lasso	0.432	0.385	0.434	0.08	0.352	0.101
			Unrestricted		0.717	0.359	0.484	0.124	0.499	0.132
			Restricted	DS	-	-	0.515	0.082	0.397	0.107
			Unrestricted		-	-	0.549	0.112	0.536	0.127
				FARM	0.385	0.329	0.702	0.197	0.516	0.126

Table D.3.8: Forecasting errors of FNETS and FARM measured by (6.26) averaged over 100 realisations (also reported are the standard errors) under the model (E1) for the generation of $\boldsymbol{\xi}_t$ and (C1)–(C2) for $\boldsymbol{\chi}_t$ with varying n and p.

			Method		$\boldsymbol{\chi}_{n+1}$		$\boldsymbol{\xi}_{n+1}$		$\mathbf{X}_{n+1}$	
	n	p	$\boldsymbol{\chi}_t$	$\boldsymbol{\xi}_t$	Mean	SD	Mean	SD	Mean	SD
(C1)	100	50	Restricted	Lasso	0.677	0.316	0.957	0.063	0.79	0.144
			Unrestricted		0.676	0.249	0.972	0.081	0.787	0.129
			Restricted	DS	-	-	0.962	0.062	0.794	0.145
			Unrestricted		-	-	0.976	0.079	0.789	0.125
				FARM	1.65	1.96	1.02	0.131	1.23	0.798
	100	100	Restricted	Lasso	0.631	0.296	0.97	0.042	0.781	0.146
			Unrestricted		0.662	0.239	0.992	0.078	0.784	0.121
			Restricted	DS	-	-	0.965	0.046	0.778	0.144
			Unrestricted		-	-	0.985	0.078	0.78	0.120
				FARM	2.9	5.07	1.01	0.074	1.8	2.610
	200	50	Restricted	Lasso	0.721	0.376	0.922	0.074	0.794	0.195
			Unrestricted		0.752	0.327	0.934	0.106	0.799	0.163
			Restricted	DS	-	-	0.925	0.074	0.795	0.194
			Unrestricted		-	-	0.938	0.11	0.802	0.165
				FARM	0.756	0.442	0.986	0.118	0.843	0.230
	200	100	Restricted	Lasso	0.582	0.285	0.921	0.049	0.755	0.146
			Unrestricted		0.668	0.333	0.942	0.074	0.766	0.130
			Restricted	DS	-	-	0.922	0.042	0.755	0.148
			Unrestricted		-	-	0.942	0.069	0.767	0.131
				FARM	0.722	0.605	0.979	0.087	0.795	0.197
	500	100	Restricted	Lasso	0.524	0.327	0.882	0.055	0.7	0.160
			Unrestricted		0.612	0.261	0.9	0.064	0.718	0.140
			Restricted	DS	-	-	0.882	0.055	0.701	0.161
			Unrestricted		-	-	0.901	0.061	0.718	0.140
				FARM	0.537	0.369	0.947	0.078	0.75	0.205
	500	200	Restricted	Lasso	0.509	0.324	0.893	0.035	0.709	0.151
			Unrestricted		0.569	0.246	0.919	0.052	0.714	0.130
			Restricted	DS	-	-	0.897	0.035	0.711	0.150
			Unrestricted		-	-	0.923	0.057	0.717	0.129
				FARM	0.517	0.347	0.953	0.07	0.745	0.170
(C2)	100	50	Restricted	Lasso	0.728	0.53	0.928	0.074	0.789	0.219
			Unrestricted		0.768	0.278	0.934	0.074	0.839	0.180
			Restricted	DS	-	-	0.926	0.062	0.787	0.218
			Unrestricted		-	-	0.927	0.063	0.838	0.180
				FARM	1.21	2.4	0.969	0.068	0.946	0.457
	100	100	Restricted	Lasso	0.644	0.367	0.964	0.038	0.794	0.213
			Unrestricted		0.735	0.257	0.967	0.036	0.857	0.167
			Restricted	DS	-	-	0.943	0.044	0.781	0.211
			Unrestricted		-	-	0.951	0.048	0.846	0.168
				FARM	1.1	2.27	0.981	0.043	0.918	0.511
	200	50	Restricted	Lasso	0.575	0.43	0.887	0.084	0.708	0.214
			Unrestricted		0.749	0.26	0.917	0.108	0.817	0.161
			Restricted	DS	-	-	0.896	0.068	0.713	0.213
			Unrestricted		-	-	0.917	0.094	0.823	0.159
				FARM	0.563	0.41	0.948	0.082	0.746	0.222
	200	100	Restricted	Lasso	0.513	0.356	0.9	0.052	0.701	0.195
			Unrestricted		0.626	0.266	0.913	0.059	0.769	0.171
			Restricted	DS	-	-	0.904	0.044	0.705	0.197
			Unrestricted		-	-	0.912	0.05	0.771	0.172
				FARM	0.501	0.337	0.959	0.063	0.732	0.203
	500	100	Restricted	Lasso	0.524	0.325	0.863	0.045	0.705	0.175
			Unrestricted		0.688	0.285	0.878	0.052	0.779	0.152
			Restricted	DS	-	-	0.874	0.043	0.711	0.176
			Unrestricted		-	-	0.887	0.048	0.784	0.151
				FARM	0.524	0.352	0.921	0.058	0.741	0.183
	500	200	Restricted	Lasso	0.516	0.354	0.877	0.039	0.698	0.182
			Unrestricted		0.627	0.315	0.89	0.046	0.75	0.148
			Restricted	DS	-	-	0.887	0.033	0.703	0.182
			Unrestricted		-	-	0.896	0.038	0.757	0.149
				FARM	0.519	0.369	0.931	0.047	0.733	0.183

Table D.3.9: Forecasting errors of FNETS and FARM measured by (6.27) averaged over 100 realisations (also reported are the standard errors) under the model (E1) for the generation of $\boldsymbol{\xi}_t$ and (C1)–(C2) for $\boldsymbol{\chi}_t$ with varying n and p.

			Method		$\boldsymbol{\chi}_{n+1}$		$\boldsymbol{\xi}_{n+1}$		$\mathbf{x}_{n+1}$	
	n	p	$\boldsymbol{\chi}_t$	$\boldsymbol{\xi}_t$	Mean	SD	Mean	SD	Mean	SD
(C1)	100	50	Restricted	Lasso	0.738	0.257	0.978	0.08	0.838	0.181
			Unrestricted		0.77	0.219	1.01	0.195	0.843	0.168
			Restricted	DS	-	-	0.982	0.092	0.835	0.179
			Unrestricted		-	-	1.01	0.204	0.84	0.170
				FARM	0.983	0.586	1.02	0.147	1.03	0.441
	100	100	Restricted	Lasso	0.685	0.243	0.99	0.062	0.809	0.168
			Unrestricted		0.75	0.206	1.02	0.152	0.812	0.154
			Restricted	DS	-	-	0.984	0.069	0.811	0.165
			Unrestricted		-	-	1	0.149	0.809	0.150
				FARM	1.11	0.818	1.03	0.119	1.1	0.586
	200	50	Restricted	Lasso	0.767	0.284	0.96	0.087	0.851	0.202
			Unrestricted		0.835	0.303	0.978	0.141	0.846	0.183
			Restricted	DS	-	-	0.961	0.093	0.855	0.203
			Unrestricted		-	-	0.976	0.156	0.847	0.183
				FARM	0.746	0.264	1	0.166	0.886	0.224
	200	100	Restricted	Lasso	0.678	0.265	0.963	0.1	0.799	0.171
			Unrestricted		0.793	0.329	0.988	0.16	0.803	0.170
			Restricted	DS	-	-	0.964	0.094	0.799	0.177
			Unrestricted		-	-	0.993	0.157	0.808	0.174
				FARM	0.687	0.321	1.01	0.162	0.82	0.234
	500	100	Restricted	Lasso	0.583	0.249	0.936	0.09	0.751	0.193
			Unrestricted		0.735	0.296	0.953	0.107	0.758	0.184
			Restricted	DS	-	-	0.936	0.086	0.747	0.190
			Unrestricted		-	-	0.951	0.101	0.755	0.179
				FARM	0.535	0.251	1.02	0.213	0.782	0.220
	500	200	Restricted	Lasso	0.55	0.259	0.925	0.09	0.746	0.168
			Unrestricted		0.648	0.208	0.946	0.11	0.752	0.166
			Restricted	DS	-	-	0.93	0.093	0.746	0.169
			Unrestricted		-	-	0.951	0.117	0.751	0.168
				FARM	0.516	0.268	0.992	0.163	0.781	0.217
(C2)	100	50	Restricted	Lasso	0.839	0.261	0.95	0.097	0.903	0.142
			Unrestricted		0.9	0.16	0.951	0.094	0.928	0.117
			Restricted	DS	-	-	0.95	0.084	0.902	0.141
			Unrestricted		-	-	0.944	0.084	0.928	0.112
				FARM	0.982	0.543	0.972	0.098	0.963	0.203
	100	100	Restricted	Lasso	0.846	0.24	0.981	0.051	0.9	0.140
			Unrestricted		0.891	0.141	0.984	0.051	0.942	0.102
			Restricted	DS	-	-	0.969	0.074	0.888	0.138
			Unrestricted		-	-	0.972	0.08	0.929	0.102
				FARM	0.951	0.573	0.99	0.069	0.947	0.222
	200	50	Restricted	Lasso	0.774	0.289	0.932	0.106	0.838	0.161
			Unrestricted		0.9	0.16	0.94	0.11	0.906	0.130
			Restricted	DS	-	-	0.939	0.089	0.843	0.156
			Unrestricted		-	-	0.944	0.089	0.915	0.118
				FARM	0.762	0.262	0.964	0.088	0.874	0.172
	200	100	Restricted	Lasso	0.735	0.23	0.943	0.086	0.866	0.150
			Unrestricted		0.818	0.174	0.956	0.1	0.896	0.126
			Restricted	DS	-	-	0.946	0.074	0.87	0.147
			Unrestricted		-	-	0.953	0.085	0.894	0.119
				FARM	0.715	0.228	0.987	0.102	0.895	0.140
	500	100	Restricted	Lasso	0.726	0.221	0.916	0.084	0.856	0.152
			Unrestricted		0.855	0.168	0.924	0.093	0.884	0.131
			Restricted	DS	-	-	0.922	0.079	0.861	0.153
			Unrestricted		-	-	0.93	0.09	0.888	0.130
				FARM	0.717	0.232	0.952	0.088	0.879	0.154
	500	200	Restricted	Lasso	0.721	0.235	0.93	0.087	0.84	0.133
			Unrestricted		0.82	0.204	0.941	0.091	0.86	0.110
			Restricted	DS	-	-	0.933	0.081	0.844	0.133
			Unrestricted		-	-	0.941	0.084	0.864	0.109
				FARM	0.711	0.243	0.964	0.094	0.87	0.129

Table D.3.10: Forecasting errors of FNETS and FARM measured by (6.24) averaged over 100 realisations (also reported are the standard errors) under the models (E2)–(E4) for the generation of $\boldsymbol{\xi}_t$ and (C1) for $\boldsymbol{\chi}_t$ with varying n and p. We also report the errors of restricted and unrestricted in-sample estimators of $\boldsymbol{\chi}_t$, $1 \le t \le n$.

			Method		In-sample		$\boldsymbol{\chi}_{n+1\|n}$		$\boldsymbol{\xi}_{n+1\|n}$		$\mathbf{x}_{n+1\|n}$	
	n	p	$\boldsymbol{\chi}_t$	$\boldsymbol{\xi}_t$	Mean	SD	Mean	SD	Mean	SD	Mean	SD
(E2)	100	50	Restricted	Lasso	0.344	0.103	0.594	0.52	0.775	0.3	0.509	0.26
			Unrestricted		0.342	0.099	0.582	0.507	0.912	0.75	0.493	0.231
			Restricted	DS	-	-	-	-	0.764	0.231	0.518	0.266
			Unrestricted		-	-	-	-	0.904	0.682	0.502	0.236
	100	100	Restricted	Lasso	0.252	0.1	0.426	0.256	0.821	0.175	0.442	0.189
			Unrestricted		0.302	0.118	0.507	0.245	0.994	0.434	0.452	0.173
			Restricted	DS	-	-	-	-	0.832	0.194	0.448	0.202
			Unrestricted		-	-	-	-	1.02	0.433	0.454	0.178
	200	50	Restricted	Lasso	0.271	0.079	0.504	0.524	0.56	0.258	0.377	0.233
			Unrestricted		0.381	0.119	0.604	0.402	0.748	0.452	0.42	0.21
			Restricted	DS	-	-	-	-	0.55	0.227	0.381	0.225
			Unrestricted		-	-	-	-	0.73	0.424	0.422	0.206
	200	100	Restricted	Lasso	0.205	0.097	0.333	0.226	0.572	0.149	0.309	0.142
			Unrestricted		0.248	0.113	0.396	0.248	0.709	0.31	0.303	0.124
			Restricted	DS	-	-	-	-	0.581	0.154	0.309	0.139
			Unrestricted		-	-	-	-	0.709	0.317	0.301	0.127
	500	100	Restricted	Lasso	0.153	0.064	0.211	0.147	0.324	0.092	0.198	0.112
			Unrestricted		0.262	0.106	0.369	0.254	0.603	0.492	0.238	0.118
			Restricted	DS	-	-	-	-	0.323	0.094	0.196	0.11
			Unrestricted		-	-	-	-	0.631	0.573	0.237	0.122
	500	200	Restricted	Lasso	0.102	0.027	0.209	0.16	0.339	0.085	0.198	0.0887
			Unrestricted		0.188	0.09	0.313	0.203	0.555	0.255	0.199	0.0853
			Restricted	DS	-	-	-	-	0.348	0.099	0.2	0.0877
			Unrestricted		-	-	-	-	0.604	0.281	0.198	0.0852
(E4)	100	50	Restricted	Lasso	0.368	0.102	0.602	0.683	0.717	0.185	0.505	0.264
			Unrestricted		0.378	0.096	0.638	0.389	0.836	0.349	0.524	0.231
			Restricted	DS	-	-	-	-	0.735	0.192	0.518	0.281
			Unrestricted		-	-	-	-	0.857	0.342	0.536	0.244
	100	100	Restricted	Lasso	0.295	0.083	0.518	0.373	0.802	0.174	0.529	0.201
			Unrestricted		0.321	0.088	0.554	0.294	0.922	0.396	0.529	0.194
			Restricted	DS	-	-	-	-	0.831	0.183	0.538	0.203
			Unrestricted		-	-	-	-	0.955	0.474	0.542	0.208
	200	50	Restricted	Lasso	0.295	0.073	0.475	0.714	0.497	0.142	0.377	0.19
			Unrestricted		0.379	0.085	0.588	0.459	0.577	0.292	0.418	0.188
			Restricted	DS	-	-	-	-	0.497	0.144	0.382	0.199
			Unrestricted		-	-	-	-	0.587	0.325	0.415	0.19
	200	100	Restricted	Lasso	0.213	0.057	0.374	0.465	0.534	0.144	0.327	0.152
			Unrestricted		0.28	0.107	0.478	0.246	0.692	0.329	0.36	0.138
			Restricted	DS	-	-	-	-	0.516	0.146	0.322	0.154
			Unrestricted		-	-	-	-	0.675	0.314	0.352	0.137
	500	100	Restricted	Lasso	0.156	0.031	0.279	0.224	0.297	0.082	0.228	0.134
			Unrestricted		0.277	0.104	0.444	0.297	0.469	0.51	0.273	0.137
			Restricted	DS	-	-	-	-	0.294	0.083	0.227	0.135
			Unrestricted		-	-	-	-	0.47	0.508	0.271	0.138
	500	200	Restricted	Lasso	0.104	0.021	0.209	0.166	0.288	0.073	0.193	0.0918
			Unrestricted		0.223	0.111	0.352	0.231	0.494	0.386	0.224	0.11
			Restricted	DS	-	-	-	-	0.318	0.082	0.202	0.0986
			Unrestricted		-	-	-	-	0.533	0.352	0.235	0.115

Table D.3.11: Forecasting errors of FNETS and FARM measured by (6.25) averaged over 100 realisations (also reported are the standard errors) under the models (E2)–(E4) for the generation of $\boldsymbol{\xi}_t$ and (C1) for $\boldsymbol{\chi}_t$ with varying n and p.

			Method		$\boldsymbol{\chi}_{n+1\|n}$		$\boldsymbol{\xi}_{n+1\|n}$		$\mathbf{X}_{n+1\|n}$	
	n	p	$\boldsymbol{\chi}_t$	$\boldsymbol{\xi}_t$	Mean	SD	Mean	SD	Mean	SD
(E2)	100	50	Restricted	Lasso	0.725	0.322	0.874	0.215	0.661	0.270
			Unrestricted		0.751	0.318	0.991	0.504	0.664	0.269
			Restricted	DS	-	-	0.863	0.196	0.663	0.261
			Unrestricted		-	-	0.978	0.499	0.668	0.261
	100	100	Restricted	Lasso	0.617	0.284	0.946	0.311	0.617	0.241
			Unrestricted		0.768	0.35	1.08	0.464	0.623	0.214
			Restricted	DS	-	-	0.944	0.322	0.619	0.248
			Unrestricted		-	-	1.1	0.429	0.624	0.218
	200	50	Restricted	Lasso	0.668	0.377	0.731	0.21	0.571	0.252
			Unrestricted		0.774	0.351	0.916	0.451	0.619	0.241
			Restricted	DS	-	-	0.724	0.196	0.574	0.252
			Unrestricted		-	-	0.901	0.429	0.615	0.237
	200	100	Restricted	Lasso	0.581	0.251	0.786	0.219	0.538	0.214
			Unrestricted		0.65	0.234	0.967	0.4	0.52	0.200
			Restricted	DS	-	-	0.789	0.21	0.531	0.206
			Unrestricted		-	-	0.959	0.417	0.511	0.195
	500	100	Restricted	Lasso	0.478	0.234	0.59	0.157	0.456	0.217
			Unrestricted		0.658	0.28	0.936	0.567	0.482	0.192
			Restricted	DS	-	-	0.586	0.153	0.455	0.219
			Unrestricted		-	-	0.964	0.598	0.486	0.205
	500	200	Restricted	Lasso	0.473	0.212	0.622	0.151	0.444	0.155
			Unrestricted		0.605	0.244	0.977	0.432	0.426	0.150
			Restricted	DS	-	-	0.632	0.179	0.448	0.158
			Unrestricted		-	-	1.04	0.453	0.421	0.158
(E4)	100	50	Restricted	Lasso	0.698	0.431	0.829	0.162	0.632	0.247
			Unrestricted		0.779	0.274	0.944	0.365	0.672	0.225
			Restricted	DS	-	-	0.846	0.161	0.649	0.264
			Unrestricted		-	-	0.954	0.357	0.683	0.242
	100	100	Restricted	Lasso	0.702	0.371	0.918	0.211	0.69	0.232
			Unrestricted		0.787	0.309	1.04	0.413	0.708	0.224
			Restricted	DS	-	-	0.938	0.22	0.699	0.235
			Unrestricted		-	-	1.06	0.455	0.714	0.236
	200	50	Restricted	Lasso	0.619	0.331	0.687	0.15	0.589	0.232
			Unrestricted		0.77	0.356	0.788	0.336	0.627	0.237
			Restricted	DS	-	-	0.692	0.145	0.587	0.239
			Unrestricted		-	-	0.785	0.363	0.618	0.233
	200	100	Restricted	Lasso	0.58	0.291	0.723	0.167	0.532	0.171
			Unrestricted		0.719	0.262	0.925	0.473	0.561	0.171
			Restricted	DS	-	-	0.723	0.186	0.536	0.170
			Unrestricted		-	-	0.923	0.475	0.558	0.171
	500	100	Restricted	Lasso	0.527	0.277	0.556	0.145	0.466	0.199
			Unrestricted		0.715	0.353	0.787	0.51	0.515	0.192
			Restricted	DS	-	-	0.543	0.134	0.473	0.201
			Unrestricted		-	-	0.782	0.498	0.514	0.197
	500	200	Restricted	Lasso	0.457	0.223	0.572	0.143	0.455	0.184
			Unrestricted		0.639	0.266	0.851	0.451	0.457	0.183
			Restricted	DS	-	-	0.599	0.149	0.457	0.187
			Unrestricted		-	-	0.907	0.488	0.468	0.193

Table D.3.12: Forecasting errors of FNETS and FARM measured by (6.26) averaged over 100 realisations (also reported are the standard errors) under the models (E2)–(E4) for the generation of $\boldsymbol{\xi}_t$ and (C1) for $\boldsymbol{\chi}_t$ with varying n and p.

			Method		$\boldsymbol{\chi}_{n+1}$		$\boldsymbol{\xi}_{n+1}$		$\mathbf{X}_{n+1}$	
	n	p	$\boldsymbol{\chi}_t$	$\boldsymbol{\xi}_t$	Mean	SD	Mean	SD	Mean	SD
(E2)	100	50	Restricted	Lasso	0.731	0.349	0.962	0.073	0.809	0.172
			Unrestricted		0.7	0.301	0.99	0.164	0.802	0.165
			Restricted	DS	-	-	0.958	0.073	0.811	0.177
			Unrestricted		-	-	0.987	0.151	0.803	0.167
	100	100	Restricted	Lasso	0.678	0.318	0.97	0.044	0.791	0.197
			Unrestricted		0.729	0.261	0.994	0.085	0.796	0.166
			Restricted	DS	-	-	0.975	0.055	0.793	0.196
			Unrestricted		-	-	1	0.085	0.795	0.164
	200	50	Restricted	Lasso	0.695	0.354	0.929	0.092	0.778	0.208
			Unrestricted		0.749	0.301	0.95	0.12	0.789	0.162
			Restricted	DS	-	-	0.929	0.087	0.78	0.208
			Unrestricted		-	-	0.949	0.118	0.79	0.159
	200	100	Restricted	Lasso	0.627	0.299	0.934	0.054	0.773	0.191
			Unrestricted		0.642	0.236	0.957	0.08	0.768	0.167
			Restricted	DS	-	-	0.934	0.057	0.771	0.190
			Unrestricted		-	-	0.957	0.084	0.765	0.167
	500	100	Restricted	Lasso	0.492	0.292	0.893	0.061	0.675	0.169
			Unrestricted		0.582	0.302	0.935	0.108	0.689	0.162
			Restricted	DS	-	-	0.892	0.064	0.675	0.168
			Unrestricted		-	-	0.935	0.106	0.689	0.166
	500	200	Restricted	Lasso	0.516	0.275	0.901	0.042	0.694	0.165
			Unrestricted		0.575	0.263	0.931	0.06	0.695	0.150
			Restricted	DS	-	-	0.901	0.045	0.694	0.165
			Unrestricted		-	-	0.937	0.067	0.694	0.153
(E4)	100	50	Restricted	Lasso	0.761	0.54	0.946	0.064	0.826	0.203
			Unrestricted		0.786	0.349	0.964	0.089	0.829	0.190
			Restricted	DS	-	-	0.947	0.061	0.828	0.204
			Unrestricted		-	-	0.967	0.084	0.832	0.190
	100	100	Restricted	Lasso	0.709	0.293	0.968	0.043	0.854	0.180
			Unrestricted		0.718	0.256	0.987	0.082	0.852	0.157
			Restricted	DS	-	-	0.971	0.045	0.855	0.178
			Unrestricted		-	-	0.988	0.086	0.851	0.157
	200	50	Restricted	Lasso	0.617	0.315	0.907	0.085	0.767	0.178
			Unrestricted		0.713	0.315	0.914	0.104	0.78	0.144
			Restricted	DS	-	-	0.909	0.081	0.77	0.175
			Unrestricted		-	-	0.919	0.1	0.781	0.142
	200	100	Restricted	Lasso	0.593	0.359	0.916	0.059	0.749	0.163
			Unrestricted		0.65	0.244	0.931	0.071	0.748	0.133
			Restricted	DS	-	-	0.915	0.061	0.749	0.163
			Unrestricted		-	-	0.932	0.074	0.748	0.132
	500	100	Restricted	Lasso	0.55	0.352	0.886	0.067	0.72	0.191
			Unrestricted		0.633	0.3	0.907	0.088	0.73	0.160
			Restricted	DS	-	-	0.886	0.073	0.718	0.190
			Unrestricted		-	-	0.91	0.096	0.73	0.161
	500	200	Restricted	Lasso	0.483	0.275	0.892	0.045	0.709	0.157
			Unrestricted		0.578	0.267	0.92	0.066	0.726	0.144
			Restricted	DS	-	-	0.897	0.041	0.712	0.157
			Unrestricted		-	-	0.928	0.062	0.729	0.144

Table D.3.13: Forecasting errors of FNETS and FARM measured by (6.27) averaged over 100 realisations (also reported are the standard errors) under the models (E2)–(E4) for the generation of $\boldsymbol{\xi}_t$ and (C1) for $\boldsymbol{\chi}_t$ with varying n and p.

			Method		χ_{n+1}		ξ_{n+1}		$\mathbf{X}_{n+1}$	
	n	p	$\boldsymbol{\chi}_t$	$\boldsymbol{\xi}_t$	Mean	SD	Mean	SD	Mean	SD
(E2)	100	50	Restricted	Lasso	0.789	0.295	0.99	0.078	0.859	0.185
			Unrestricted		0.802	0.303	1.04	0.23	0.858	0.185
			Restricted	DS	-	-	0.98	0.081	0.856	0.181
			Unrestricted		-	-	1.03	0.22	0.851	0.182
	100	100	Restricted	Lasso	0.71	0.265	0.973	0.066	0.805	0.189
			Unrestricted		0.789	0.26	0.996	0.114	0.818	0.174
			Restricted	DS	-	-	0.985	0.068	0.806	0.190
			Unrestricted		-	-	1.01	0.112	0.817	0.176
	200	50	Restricted	Lasso	0.742	0.279	0.974	0.147	0.806	0.170
			Unrestricted		0.813	0.298	1	0.176	0.827	0.182
			Restricted	DS	-	-	0.978	0.127	0.804	0.173
			Unrestricted		-	-	1	0.166	0.827	0.183
	200	100	Restricted	Lasso	0.674	0.186	0.961	0.089	0.805	0.173
			Unrestricted		0.749	0.222	0.998	0.155	0.805	0.181
			Restricted	DS	-	-	0.956	0.085	0.802	0.174
			Unrestricted		-	-	0.995	0.16	0.801	0.181
	500	100	Restricted	Lasso	0.578	0.217	0.937	0.107	0.722	0.200
			Unrestricted		0.701	0.271	0.987	0.183	0.727	0.196
			Restricted	DS	-	-	0.935	0.102	0.718	0.199
			Unrestricted		-	-	0.987	0.181	0.726	0.199
	500	200	Restricted	Lasso	0.574	0.228	0.945	0.075	0.745	0.172
			Unrestricted		0.668	0.254	0.988	0.139	0.744	0.172
			Restricted	DS	-	-	0.952	0.078	0.75	0.172
			Unrestricted		-	-	1	0.155	0.748	0.172
(E4)	100	50	Restricted	Lasso	0.787	0.373	0.974	0.07	0.875	0.191
			Unrestricted		0.845	0.276	0.994	0.104	0.885	0.183
			Restricted	DS	-	-	0.972	0.07	0.878	0.195
			Unrestricted		-	-	0.995	0.104	0.886	0.182
	100	100	Restricted	Lasso	0.781	0.287	0.991	0.059	0.908	0.174
			Unrestricted		0.824	0.25	1.01	0.107	0.911	0.159
			Restricted	DS	-	-	0.985	0.056	0.907	0.171
			Unrestricted		-	-	1	0.103	0.909	0.156
	200	50	Restricted	Lasso	0.679	0.267	0.97	0.105	0.847	0.186
			Unrestricted		0.803	0.274	0.971	0.135	0.857	0.178
			Restricted	DS	-	-	0.97	0.108	0.849	0.187
			Unrestricted		-	-	0.978	0.149	0.858	0.180
	200	100	Restricted	Lasso	0.662	0.308	0.974	0.071	0.825	0.176
			Unrestricted		0.77	0.246	0.972	0.073	0.838	0.169
			Restricted	DS	-	-	0.975	0.074	0.826	0.179
			Unrestricted		-	-	0.976	0.083	0.837	0.171
	500	100	Restricted	Lasso	0.643	0.303	0.963	0.073	0.842	0.182
			Unrestricted		0.771	0.286	0.972	0.096	0.852	0.164
			Restricted	DS	-	-	0.961	0.08	0.843	0.181
			Unrestricted		-	-	0.975	0.105	0.852	0.167
	500	200	Restricted	Lasso	0.576	0.289	0.974	0.071	0.856	0.162
			Unrestricted		0.715	0.284	0.99	0.102	0.868	0.157
			Restricted	DS	-	-	0.975	0.066	0.858	0.160
			Unrestricted		-	-	0.992	0.104	0.869	0.154

BIBLIOGRAPHY

Aastveit, K. A., Carriero, A., Clark, T. E., and Marcellino, M. (2017).
Have standard VARs remained stable since the crisis?
Journal of Applied Econometrics, 32(5):931–951.

Adamek, R., Smeekes, S., and Wilms, I. (2020).
Lasso inference for high-dimensional time series.
arXiv preprint arXiv:2007.10952.

Adams, R. P. and MacKay, D. J. (2007).
Bayesian online changepoint detection.
arXiv preprint arXiv:0710.3742.

Ahelegbey, D. F., Billio, M., and Casarin, R. (2021).
Modeling turning points in the global equity market.
Econometrics and Statistics.

Ahn, S. C. and Horenstein, A. R. (2013).
Eigenvalue ratio test for the number of factors.
Econometrica, 81(3):1203–1227.

Alessi, L., Barigozzi, M., and Capasso, M. (2010).
Improved penalization for determining the number of factors in approximate factor models.
Statistics & Probability Letters, 80(23-24):1806–1813.

Ang, A. and Piazzesi, M. (2003).
A no-arbitrage vector autoregression of term structure dynamics with macroeconomic and latent variables.
Journal of Monetary economics, 50(4):745–787.

Arbelaez, P., Maire, M., Fowlkes, C., and Malik, J. (2010).
Contour detection and hierarchical image segmentation.
IEEE Transactions on Pattern Analysis and Machine Intelligence, 33(5):898–916.

Assenmacher-Wesche, K. and Pesaran, M. H. (2008).

Forecasting the swiss economy using VECX models: An exercise in forecast combination across models and observation windows.
National Institute Economic Review, 203:91–108.

Aue, A., Hörmann, S., Horváth, L., and Reimherr, M. (2009).
Break detection in the covariance structure of multivariate time series models.
The Annals of Statistics, 37(6B):4046–4087.

Avarucci, M., Cavicchioli, M., Forni, M., and Zaffaroni, P. (2022).
The main business cycle shock(s): Frequency-band estimation of the number of dynamic factors.
CEPR Discussion Paper No. DP17281.

Bai, J. (2000).
Vector autoregressive models with structural changes in regression coefficients and in variance-covariance matrices.
Annals of Economics and Finance, 1(2):303–339.

Bai, J. (2003).
Inferential theory for factor models of large dimensions.
Econometrica, 71(1):135–171.

Bai, J., Han, X., and Shi, Y. (2020a).
Estimation and inference of change points in high-dimensional factor models.
Journal of Econometrics, 219(1):66–100.

Bai, J., Li, K., and Lu, L. (2016).
Estimation and inference of FAVAR models.
Journal of Business & Economic Statistics, 34(4):620–641.

Bai, J. and Ng, S. (2002).
Determining the number of factors in approximate factor models.
Econometrica, 70(1):191–221.

Bai, J. and Ng, S. (2006).
Confidence intervals for diffusion index forecasts and inference for factor-augmented regressions.
Econometrica, 74(4):1133–1150.

Bai, J. and Ng, S. (2009).
Boosting diffusion indices.
Journal of Applied Econometrics, 24(4):607–629.

Bai, J. and Perron, P. (1998).

Estimating and testing linear models with multiple structural changes.
Econometrica, pages 47–78.

Bai, J. and Perron, P. (2003).
Computation and analysis of multiple structural change models.
Journal of Applied Econometrics, 18(1):1–22.

Bai, P. (2021).
LSVAR: Estimation of Low Rank Plus Sparse Structured Vector Auto-Regressive (VAR) Model.
R package version 1.2.

Bai, P., Safikhani, A., and Michailidis, G. (2020b).
Multiple change points detection in low rank and sparse high dimensional vector autoregressive models.
IEEE Transactions on Signal Processing, 68:3074–3089.

Bai, P., Safikhani, A., and Michailidis, G. (2021).
A fast detection method of break points in effective connectivity networks.
IEEE Transactions on Medical Imaging.

Bai, P., Safikhani, A., and Michailidis, G. (2023).
Multiple change point detection in reduced rank high dimensional vector autoregressive models.
Journal of the American Statistical Association, pages 1–17.

Bai, Y. and Safikhani, A. (2022).
A unified framework for change point detection in high-dimensional linear models.
arXiv preprint arXiv:2207.09007.

Ballarin, G. (2021).
Ridge regularized estimation of VAR models for inference.
arXiv preprint arXiv:2105.00860.

Bańbura, M., Giannone, D., Modugno, M., and Reichlin, L. (2013).
Now-casting and the real-time data flow.
In *Handbook of Economic Forecasting*, volume 2, pages 195–237. Elsevier.

Banerjee, A., Marcellino, M., and Masten, I. (2008).
Forecasting macroeconomic variables using diffusion indexes in short samples with structural change.
In *Forecasting in the presence of structural breaks and model uncertainty*. Emerald Group Publishing Limited.

Barhoumi, K., Darné, O., and Ferrara, L. (2014).
Dynamic factor models: A review of the literature.
OECD Journal: Journal of Business Cycle Measurement and Analysis, 2013(2):73–107.

Barigozzi, M. (2022).
On estimation and inference of large approximate dynamic factor models via the principal component analysis.
arXiv preprint arXiv:2211.01921.

Barigozzi, M. and Brownlees, C. (2019).
NETS: Network estimation for time series.
Journal of Applied Econometrics, 34(3):347–364.

Barigozzi, M., Cho, H., and Fryzlewicz, P. (2018).
Simultaneous multiple change-point and factor analysis for high-dimensional time series.
Journal of Econometrics, 206(1):187–225.

Barigozzi, M., Cho, H., and Owens, D. (2023).
FNETS: Factor-adjusted network estimation and forecasting for high-dimensional time series.
Journal of Business & Economic Statistics, 0(ja):1–27.

Barigozzi, M. and Trapani, L. (2020).
Sequential testing for structural stability in approximate factor models.
Stochastic Processes and their Applications, 130(8):5149–5187.

Basu, S., Li, X., and Michailidis, G. (2019).
Low rank and structured modeling of high-dimensional vector autoregressions.
IEEE Transactions on Signal Processing, 67(5):1207–1222.

Basu, S. and Michailidis, G. (2015).
Regularized estimation in sparse high-dimensional time series models.
The Annals of Statistics, 43(4):1535–1567.

Bates, B. J., Plagborg-Møller, M., Stock, J. H., and Watson, M. W. (2013).
Consistent factor estimation in dynamic factor models with structural instability.
Journal of Econometrics, 177(2):289–304.

Bauer, P. and Hackl, P. (1980).
An extension of the MOSUM technique for quality control.
Technometrics, 22(1):1–7.

Beck, A. and Teboulle, M. (2009).
A fast iterative shrinkage-thresholding algorithm for linear inverse problems.
SIAM Journal on Imaging Sciences, 2(1):183–202.

Bell, V., Co, L. W., Stone, S., and Wallis, G. (2014).
Nowcasting UK GDP growth.
Bank of England Quarterly Bulletin, page Q1.

Bergmeir, C., Hyndman, R. J., and Koo, B. (2018).
A note on the validity of cross-validation for evaluating autoregressive time series prediction.
Computational Statistics & Data Analysis, 120:70–83.

Bernanke, B. S., Boivin, J., and Eliasz, P. (2005).
Measuring the effects of monetary policy: a factor-augmented vector autoregressive (FAVAR) approach.
The Quarterly Journal of Economics, 120(1):387–422.

Billio, M., Getmansky, M., Lo, A. W., and Pelizzon, L. (2012).
Econometric measures of connectedness and systemic risk in the finance and insurance sectors.
Journal of Financial Economics, 104(3):535–559.

Bleakley, K. and Vert, J.-P. (2011).
The group fused lasso for multiple change-point detection.
arXiv preprint arXiv:1106.4199.

Breitung, J. and Eickmeier, S. (2011).
Testing for structural breaks in dynamic factor models.
Journal of Econometrics, 163(1):71–84.

Brownlees, C. (2020).
nets: Network Estimation for Time Series.
R package version 0.9.1.

Bühlmann, P. and van de Geer, S. (2011).
Statistics for High-dimensional Data: Methods, Theory and Applications.
Springer Science & Business Media.

Cai, T., Liu, W., and Luo, X. (2011).
A constrained ℓ_1 minimization approach to sparse precision matrix estimation.
Journal of the American Statistical Association, 106(494):594–607.

Cai, T. T., Liu, W., and Zhou, H. H. (2016).
Estimating sparse precision matrix: Optimal rates of convergence and adaptive estimation.
The Annals of Statistics, 44(2):455–488.

Campbell, J. Y., Lo, A. W., MacKinlay, A. C., and Whitelaw, R. F. (1998).
The econometrics of financial markets.
Macroeconomic Dynamics, 2(4):559–562.

Candes, E. and Tao, T. (2007).
The dantzig selector: Statistical estimation when p is much larger than n.
The Annals of Statistics, 35(6):2313–2351.

Cardot, H. and Degras, D. (2018).
Online principal component analysis in high dimension: Which algorithm to choose?
International Statistical Review, 86(1):29–50.

Chelani, A. (2016).
Long-memory property in air pollutant concentrations.
Atmospheric Research, 171:1–4.

Chen, J. and Chen, Z. (2008).
Extended bayesian information criteria for model selection with large model spaces.
Biometrika, 95(3):759–771.

Chen, J., Li, D., Li, Y., and Linton, O. (2023).
Estimating time-varying networks for high-dimensional time series.
arXiv preprint arXiv:2302.02476.

Chen, L., Dolado, J. J., and Gonzalo, J. (2014).
Detecting big structural breaks in large factor models.
Journal of Econometrics, 180(1):30–48.

Chen, L., Wang, W., and Wu, W. B. (2021).
Inference of breakpoints in high-dimensional time series.
Journal of the American Statistical Association, pages 1–33.

Cheng, X., Liao, Z., and Schorfheide, F. (2016).
Shrinkage estimation of high-dimensional factor models with structural instabilities.
The Review of Economic Studies, 83(4):1511–1543.

Cho, H. and Fryzlewicz, P. (2015).
Multiple-change-point detection for high dimensional time series via sparsified binary segmentation.
Journal of the Royal Statistical Society: Series B (Statistical Methodology), 77(2):475–507.

Cho, H. and Fryzlewicz, P. (2020).
Multiple change point detection under serial dependence: Wild energy maximisation and gappy schwarz criterion.
arXiv preprint arXiv:2011.13884.

Cho, H. and Kirch, C. (2021a).

Data segmentation algorithms: Univariate mean change and beyond.
Econometrics and Statistics.

Cho, H. and Kirch, C. (2021b).
Two-stage data segmentation permitting multiscale change points, heavy tails and dependence.
Annals of the Institute of Statistical Mathematics, pages 1–32.

Cho, H., Maeng, H., Eckley, I., and Fearnhead, P. (2022).
High-dimensional time series segmentation via factor-adjusted vector autoregressive modelling.
arXiv preprint arXiv:2204.02724.

Cho, H. and Owens, D. (2022).
High-dimensional data segmentation in regression settings permitting heavy tails and temporal dependence.
arXiv preprint arXiv:2209.08892.

Corradi, V. and Swanson, N. R. (2014).
Testing for structural stability of factor augmented forecasting models.
Journal of Econometrics, 182(1):100–118.

Csardi, G., Nepusz, T., et al. (2006).
The igraph software package for complex network research.
InterJournal, Complex Systems, 1695(5):1–9.

Dahlhaus, R. (2000).
Graphical interaction models for multivariate time series.
Metrika, 51(2):157–172.

Datta, A., Zou, H., and Banerjee, S. (2019).
Bayesian high-dimensional regression for change point analysis.
Statistics and Its Interface, 12(2):253.

Davis, R. A., Huang, D., and Yao, Y.-C. (1995).
Testing for a change in the parameter values and order of an autoregressive model.
The Annals of Statistics, pages 282–304.

Davis, R. A., Lee, T. C. M., and Rodriguez-Yam, G. A. (2006).
Structural break estimation for nonstationary time series models.
Journal of the American Statistical Association, 101(473):223–239.

De Mol, C., Giannone, D., and Reichlin, L. (2008).
Forecasting using a large number of predictors: Is bayesian shrinkage a valid alternative to principal components?
Journal of Econometrics, 146(2):318–328.

Diebold, F. X. and Yılmaz, K. (2014).
On the network topology of variance decompositions: Measuring the connectedness of financial firms.
Journal of Econometrics, 182(1):119–134.

Dobson, A. J. and Barnett, A. G. (2018).
An Introduction to Generalized Linear Models.
CRC press.

Doz, C., Giannone, D., and Reichlin, L. (2011).
A two-step estimator for large approximate dynamic factor models based on kalman filtering.
Journal of Econometrics, 164(1):188–205.

Doz, C., Giannone, D., and Reichlin, L. (2012).
A quasi–maximum likelihood approach for large, approximate dynamic factor models.
Review of Economics and Statistics, 94(4):1014–1024.

Duan, J., Bai, J., and Han, X. (2023).
Quasi-maximum likelihood estimation of break point in high-dimensional factor models.
Journal of Econometrics, 233(1):209–236.

Dvořák, M. (2017).
Darling-Erdös-type test for change detection in parameters and variance for stationary VAR models.
Communications in Statistics-Theory and Methods, 46(1):465–484.

Dvořák, M. and Prášková, Z. (2013).
On testing changes in autoregressive parameters of a VAR model.
Communications in Statistics-Theory and Methods, 42(7):1208–1226.

Eddelbuettel, D., François, R., Allaire, J., Ushey, K., Kou, Q., Russel, N., Chambers, J., and Bates, D. (2011).
Rcpp: Seamless R and C++ integration.

Eichinger, B. and Kirch, C. (2018).
A MOSUM procedure for the estimation of multiple random change points.
Bernoulli, 24(1):526–564.

Eichler, M. (2007).
Granger causality and path diagrams for multivariate time series.
Journal of Econometrics, 137(2):334–353.

Enikeeva, F., Klopp, O., and Rousselot, M. (2023).

Change point detection in low-rank VAR processes.
arXiv preprint arXiv:2305.00311.

Epskamp, S., Waldorp, L. J., Mõttus, R., and Borsboom, D. (2018).
The gaussian graphical model in cross-sectional and time-series data.
Multivariate Behavioral Research, 53(4):453–480.

Fan, J., Ke, Y., and Wang, K. (2020).
Factor-adjusted regularized model selection.
Journal of Econometrics, 216(1):71–85.

Fan, J., Liao, Y., and Mincheva, M. (2011).
High dimensional covariance matrix estimation in approximate factor models.
Annals of Statistics, 39(6):3320.

Fan, J., Liao, Y., and Mincheva, M. (2013).
Large covariance estimation by thresholding principal orthogonal complements.
Journal of the Royal Statistical Society: Series B (Statistical Methodology), 75(4).

Fan, J., Masini, R., and Medeiros, M. C. (2021).
Bridging factor and sparse models.
arXiv preprint arXiv:2102.11341.

Fassò, A. (2013).
Statistical assessment of air quality interventions.
Stochastic Environmental Research and Risk Assessment, 27(7):1651–1660.

Fearnhead, P. and Liu, Z. (2007).
On-line inference for multiple changepoint problems.
Journal of the Royal Statistical Society: Series B (Statistical Methodology), 69(4):589–605.

Forni, M., Giannone, D., Lippi, M., and Reichlin, L. (2009).
Opening the black box: Structural factor models with large cross sections.
Econometric Theory, 25(5):1319–1347.

Forni, M., Hallin, M., Lippi, M., and Reichlin, L. (2000).
The generalized dynamic-factor model: Identification and estimation.
Review of Economics and statistics, 82(4):540–554.

Forni, M., Hallin, M., Lippi, M., and Reichlin, L. (2004).
The generalized dynamic factor model consistency and rates.
Journal of Econometrics, 119(2):231–255.

Forni, M., Hallin, M., Lippi, M., and Reichlin, L. (2005).
The generalized dynamic factor model: one-sided estimation and forecasting.
Journal of the American Statistical Association, 100(471):830–840.

Forni, M., Hallin, M., Lippi, M., and Zaffaroni, P. (2015).
Dynamic factor models with infinite-dimensional factor spaces: One-sided representations.
Journal of Econometrics, 185(2):359–371.

Forni, M., Hallin, M., Lippi, M., and Zaffaroni, P. (2017).
Dynamic factor models with infinite-dimensional factor space: Asymptotic analysis.
Journal of Econometrics, 199(1):74–92.

Frick, K., Munk, A., and Sieling, H. (2014).
Multiscale change point inference.
Journal of the Royal Statistical Society: Series B (Statistical Methodology), 76(3):495–580.

Friedman, J., Hastie, T., and Tibshirani, R. (2010).
Regularization paths for generalized linear models via coordinate descent.
Journal of Statistical Software, 33(1):1–22.

Fryzlewicz, P. (2014).
Wild binary segmentation for multiple change-point detection.
The Annals of Statistics, 42(6):2243–2281.

Gao, F. and Wang, T. (2022).
Sparse change detection in high-dimensional linear regression.
arXiv preprint arXiv:2208.06326.

Giannone, D., Reichlin, L., and Small, D. (2008).
Nowcasting: The real-time informational content of macroeconomic data.
Journal of Monetary Economics, 55(4):665–676.

Giovannelli, A., Massacci, D., and Soccorsi, S. (2021).
Forecasting stock returns with large dimensional factor models.
Journal of Empirical Finance, 63:252–269.

Giraitis, L., Kapetanios, G., Mansur, M., and Price, S. (2015).
Forecasting under structural change.
In *Empirical Economic and Financial Research*, pages 401–419. Springer.

Gombay, E. (2008).
Change detection in autoregressive time series.
Journal of Multivariate Analysis, 99(3):451–464.

Gombay, E. and Serban, D. (2009).
Monitoring parameter change in AR(p) time series models.
Journal of Multivariate Analysis, 100(4):715–725.

Hallin, M. (2022).
Manfred deistler and the general-dynamic-factor-model approach to the statistical analysis of high-dimensional time series.
Econometrics, 10(4):37.

Hallin, M., Lippi, M., Barigozzi, M., Mario, F., and Zaffaroni, P. (2019).
Time Series in High Dimensions: The General Dynamic Factor Model.

Hallin, M. and Liška, R. (2007).
Determining the number of factors in the general dynamic factor model.
Journal of the American Statistical Association, 102(478):603–617.

Han, F., Lu, H., and Liu, H. (2015).
A direct estimation of high dimensional stationary vector autoregressions.
Journal of Machine Learning Research.

Han, L., Cribben, I., and Trueck, S. (2022).
Extremal dependence in Australian electricity markets.
arXiv preprint arXiv:2202.09970.

Han, X. and Inoue, A. (2015).
Tests for parameter instability in dynamic factor models.
Econometric Theory, pages 1117–1152.

Han, Y. and Tsay, R. S. (2020).
High-dimensional linear regression for dependent data with applications to nowcasting.
Statistica Sinica, 30(4):1797–1827.

Hännikäinen, J. (2017).
Selection of an estimation window in the presence of data revisions and recent structural breaks.
Journal of Econometric Methods, 6(1).

Harchaoui, Z. and Lévy-Leduc, C. (2010).
Multiple change-point estimation with a total variation penalty.
Journal of the American Statistical Association, 105(492):1480–1493.

Haslbeck, J. M. and Waldorp, L. J. (2020).
mgm: Estimating time-varying mixed graphical models in high-dimensional data.
Journal of Statistical Software, 93:1–46.

Hlávka, Z., Hušková, M., and Meintanis, S. G. (2017).
Change point detection with multivariate observations based on characteristic functions.
In *From Statistics to Mathematical Finance*, pages 273–290. Springer.

Hoerl, A. E. and Kennard, R. W. (1970).
Ridge regression: Biased estimation for nonorthogonal problems.
Technometrics, 12(1):55–67.

Hörmann, S. and Nisol, G. (2021).
Prediction of singular VARs and an application to generalized dynamic factor models.
Journal of Time Series Analysis, 42(3):295–313.

Huang, J. and Zhang, T. (2010).
The benefit of group sparsity.
The Annals of Statistics, pages 1978–2004.

Hušková, M. (1990).
Asymptotics for robust MOSUM.
Commentationes Mathematicae Universitatis Carolinae, 31(2):345–356.

Jenkins, N., Parfitt, H., Nicholls, M., Beckett, P., Wyche, K., Smallbone, K., Gregg, D., and Smith, M. (2020).
Estimation of changes in air pollution emissions, concentrations and exposure during the COVID-19 outbreak in the UK: Report for the air quality expert group, on behalf of defra: Analysis of air quality changes experienced in sussex and surrey since the COVID-19 outbreak.

Kaul, A., Jandhyala, V. K., and Fotopoulos, S. B. (2019a).
Detection and estimation of parameters in high dimensional multiple change point regression models via l_1/l_0 regularization and discrete optimization.
arXiv preprint arXiv:1906.04396.

Kaul, A., Jandhyala, V. K., and Fotopoulos, S. B. (2019b).
An efficient two step algorithm for high dimensional change point regression models without grid search.
Journal of Machine Learning Research, 20(111):1–40.

Killick, R., Fearnhead, P., and Eckley, I. A. (2012).
Optimal detection of changepoints with a linear computational cost.
Journal of the American Statistical Association, 107(500):1590–1598.

Kim, B., Song, J., and Baek, C. (2021).
Robust test for structural instability in dynamic factor models.
The Annals of the Institute of Statistical Mathematics, pages 1–33.

Kim, C.-J. (1994).
Dynamic linear models with markov-switching.
Journal of econometrics, 60(1-2):1–22.

Kirch, C., Muhsal, B., and Ombao, H. (2015).
Detection of changes in multivariate time series with application to EEG data.
Journal of the American Statistical Association, 110:1197–1216.

Kirch, C. and Reckrühm, K. (2022).
Data segmentation for time series based on a general moving sum approach.
arXiv preprint arXiv:2207.07396.

Knight, M., Leeming, K., Nason, G., and Nunes, M. (2020).
Generalized network autoregressive processes and the GNAR package.
Journal of Statistical Software, 96:1–36.

Kock, A. B. and Callot, L. (2015).
Oracle inequalities for high dimensional vector autoregressions.
Journal of Econometrics, 186(2):325–344.

Kock, A. B., Medeiros, M., and Vasconcelos, G. (2020).
Penalized time series regression.
Macroeconomic Forecasting in the Era of Big Data: Theory and Practice, pages 193–228.

Koo, B., Anderson, H. M., Seo, M. H., and Yao, W. (2020).
High-dimensional predictive regression in the presence of cointegration.
Journal of Econometrics, 219(2):456–477.

Koo, B. and Seo, M. H. (2015).
Structural-break models under mis-specification: Implications for forecasting.
Journal of Econometrics, 188(1):166–181.

Koo, B., Wong, B., and Zhong, Z.-Y. (2023).
Disentangling structural breaks in high dimensional factor models.
arXiv preprint arXiv:2303.00178.

Koop, G. M. (2013).
Forecasting with medium and large bayesian VARs.
Journal of Applied Econometrics, 28(2):177–203.

Korkas, K. K. and Fryzlewicz, P. (2017).
Multiple change-point detection for non-stationary time series using wild binary segmentation.
Statistica Sinica, pages 287–311.

Krampe, J. and Margaritella, L. (2021).
Dynamic factor models with sparse VAR idiosyncratic components.
arXiv preprint arXiv:2112.07149.

Krantz, S. and Bagdziunas, R. (2023).
dfms: Dynamic Factor Models.
R package version 0.2.1.

Krempl, G., Hofer, V., Webb, G., and Hüllermeier, E. (2021).
Beyond adaptation: Understanding distributional changes (dagstuhl seminar 20372).
In *Dagstuhl Reports*, volume 10. Schloss Dagstuhl-Leibniz-Zentrum für Informatik.

Lavielle, M. and Moulines, E. (2000).
Least-squares estimation of an unknown number of shifts in a time series.
Journal of time series analysis, 21(1):33–59.

Lee, S., Seo, M. H., and Shin, Y. (2016).
The lasso for high dimensional regression with a possible change point.
Journal of the Royal Statistical Society: Series B (Statistical Methodology), 78(1):193.

Leonardi, F. and Bühlmann, P. (2016).
Computationally efficient change point detection for high-dimensional regression.
arXiv preprint arXiv:1601.03704.

Li, Y., Chan, N. H., Yau, C. Y., and Zhang, R. (2020).
Group orthogonal greedy algorithm for change-point estimation of multivariate time series.
Journal of Statistical Planning and Inference.

Lippi, M., Deistler, M., and Anderson, B. (2023).
High-dimensional dynamic factor models: A selective survey and lines of future research.
Econometrics and Statistics, 26:3–16.

Liu, B., Qi, Z., Zhang, X., and Liu, Y. (2022a).
Change point detection for high-dimensional linear models: A general tail-adaptive approach.
arXiv preprint arXiv:2207.11532.

Liu, B., Zhang, X., and Liu, Y. (2021).
Simultaneous change point inference and structure recovery for high dimensional gaussian graphical models.
Journal of Machine Learning Research, 22(274):1–62.

Liu, B., Zhang, X., and Liu, Y. (2022b).
High dimensional change point inference: Recent developments and extensions.
Journal of Multivariate Analysis, 188:104833.

Liu, L. and Zhang, D. (2021a).
Robust estimation of high-dimensional vector autoregressive models.
arXiv preprint arXiv:2109.10354.

Liu, X. and Zhang, T. (2021b).
Estimating change-point latent factor models for high-dimensional time series.
Journal of Statistical Planning and Inference.

Liu, Y. and Wu, J. C. (2021).
Reconstructing the yield curve.
Journal of Financial Economics, 142(3):1395–1425.

Loh, P.-L. and Wainwright, M. J. (2012).
High-dimensional regression with noisy and missing data: Provable guarantees with nonconvexity.
The Annals of Statistics, 40(3):1637–1664.

Lu, J., Liu, A., Dong, F., Gu, F., Gama, J., and Zhang, G. (2018).
Learning under concept drift: A review.
IEEE Transactions on Knowledge and Data Engineering, 31(12):2346–2363.

Lu, Z., Banerjee, M., and Michailidis, G. (2017).
Intelligent sampling for multiple change-points in exceedingly long time series with rate guarantees.
arXiv preprint arXiv:1710.07420.

Lütkepohl, H. (2005).
New Introduction to Multiple Time Series Analysis.
Springer Science & Business Media.

Ma, S. and Su, L. (2018).
Estimation of large dimensional factor models with an unknown number of breaks.
Journal of Econometrics, 207(1):1–29.

Maciejowska, K. and Weron, R. (2013).
Forecasting of daily electricity spot prices by incorporating intra-day relationships: Evidence from the UK power market.
In *2013 10th International Conference on the European Energy Market (EEM)*, pages 1–5. IEEE.

Maeng, H., Eckley, I., and Fearnhead, P. (2021).
Collective anomaly detection in high-dimensional VAR models.
arXiv preprint arXiv:2105.07538.

Massacci, D. (2019).
Unstable diffusion indexes: With an application to bond risk premia.
Oxford Bulletin of Economics and Statistics, 81(6):1376–1400.

Massacci, D. and Kapetanios, G. (2023).
Forecasting in factor augmented regressions under structural change.
International Journal of Forecasting.

McCracken, M. W. and Ng, S. (2016).
Fred-md: A monthly database for macroeconomic research.
Journal of Business & Economic Statistics, 34(4):574–589.

McGonigle, E. T. and Cho, H. (2023).
Nonparametric data segmentation in multivariate time series via joint characteristic functions.
arXiv preprint arXiv:2305.07581.

Medeiros, M. C. and Mendes, E. F. (2016).
l1-regularization of high-dimensional time-series models with non-Gaussian and heteroskedastic errors.

Meier, A., Kirch, C., and Cho, H. (2021).
mosum: A package for moving sums in change-point analysis.
Journal of Statistical Software, 97(1):1–42.

Meinshausen, N. and Bühlmann, P. (2006).
High-dimensional graphs and variable selection with the lasso.
The Annals of Statistics, 34(3):1436–1462.

Messer, M., Albert, S., and Schneider, G. (2018).
The multiple filter test for change point detection in time series.
Metrika, 81(6):589–607.

Messer, M., Kirchner, M., Schiemann, J., Roeper, J., Neininger, R., and Schneider, G. (2014).
A multiple filter test for the detection of rate changes in renewal processes with varying variance.
The Annals of Applied Statistics, 8(4):2027–2067.

Mosley, L., Chan, T.-S., and Gibberd, A. (2023).
sparseDFM: An R Package to Estimate Dynamic Factor Models with Sparse Loadings.
arXiv preprint arXiv:2303.14125.

Nasiadka, J., Nitka, W., and Weron, R. (2022).
Calibration window selection based on change-point detection for forecasting electricity prices.
arXiv preprint arXiv:2204.00872.

Negahban, S. N., Ravikumar, P., Wainwright, M. J., and Yu, B. (2012).
A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers.
Statistical Science, 27(4):538–557.

Nicholson, W., Matteson, D., and Bien, J. (2017).
Bigvar: Tools for modeling sparse high-dimensional multivariate time series.
arXiv preprint arXiv:1702.07094.

Nicholson, W. B., Wilms, I., Bien, J., and Matteson, D. S. (2020).
High dimensional forecasting via interpretable vector autoregression.
Journal of Machine Learning Research, 21(166):1–52.

Niu, Y. S., Hao, N., and Zhang, H. (2016).
Multiple change-point detection: A selective overview.
Statistical Science, pages 611–623.

Norwood, B. and Killick, R. (2018).
Long memory and changepoint models: a spectral classification procedure.
Statistics and Computing, 28(2):291–302.

Olshen, A. B., Venkatraman, E., Lucito, R., and Wigler, M. (2004).
Circular binary segmentation for the analysis of array-based dna copy number data.
Biostatistics, 5(4):557–572.

Opgen-Rhein, R. and Strimmer, K. (2007).
Learning causal networks from systems biology time course data: an effective model selection procedure for the vector autoregressive process.
BMC bioinformatics, 8(2):1–8.

Owens, D., Cho, H., and Barigozzi, M. (2023).
fnets: An R package for network estimation and forecasting via factor-adjusted VAR modelling.
The R Journal, 15:214–239.
https://doi.org/10.32614/RJ-2023-070.

Padilla, O. H. M., Yu, Y., Wang, D., and Rinaldo, A. (2019).
Optimal nonparametric change point detection and localization.
arXiv preprint arXiv:1905.10019.

Page, E. S. (1954).
Continuous inspection schemes.
Biometrika, 41(1/2):100–115.

Palm, B. G., Alves, D. I., Vu, V. T., Pettersson, M. I., Bayer, F. M., Cintra, R. J., Machado, R., Dammert, P., and Hellsten, H. (2018).
Autoregressive model for multi-pass sar change detection based on image stacks.
In *Image and Signal Processing for Remote Sensing XXIV*, volume 10789, page 1078916. International Society for Optics and Photonics.

Peng, J., Wang, P., Zhou, N., and Zhu, J. (2009).
Partial correlation estimation by joint sparse regression models.
Journal of the American Statistical Association, 104(486):735–746.

Pesaran, M. H. and Pick, A. (2011).
Forecast combination across estimation windows.
Journal of Business & Economic Statistics, 29(2):307–318.

Pesaran, M. H., Pick, A., and Pranovich, M. (2013).
Optimal forecasts in the presence of structural breaks.
Journal of Econometrics, 177(2):134–152.

Pesaran, M. H., Schuermann, T., and Smith, L. V. (2009).
Forecasting economic and financial variables with global VARs.
International journal of forecasting, 25(4):642–675.

Pesaran, M. H. and Timmermann, A. (2007).
Selection of estimation window in the presence of breaks.
Journal of Econometrics, 137(1):134–161.

Preuss, P., Puchstein, R., and Dette, H. (2015a).
Detection of multiple structural breaks in multivariate time series.
Journal of the American Statistical Association, 110:654–668.

Preuss, P., Puchstein, R., and Dette, H. (2015b).
Detection of multiple structural breaks in multivariate time series.
Journal of the American Statistical Association, 110(510):654–668.

Qian, C., Wang, G., and Zou, C. (2023).
Reliever: Relieving the burden of costly model fits for changepoint detection.
arXiv preprint arXiv:2307.01150.

Qu, Z. and Perron, P. (2007).
Estimating and testing structural changes in multivariate regressions.
Econometrica, 75(2):459–502.

R Core Team (2020).

R: A Language and Environment for Statistical Computing.
R Foundation for Statistical Computing, Vienna, Austria.

Rapach, D. E., Strauss, J. K., and Zhou, G. (2010).
Out-of-sample equity premium prediction: Combination forecasts and links to the real economy.
The Review of Financial Studies, 23(2):821–862.

Reckrühm, K. (2019).
Estimating multiple structural breaks in time series-a generalized MOSUM approach based on estimating functions.
PhD thesis.

Rigaill, G. (2010).
Pruned dynamic programming for optimal multiple change-point detection.
arXiv preprint arXiv:1004.0887, 17.

Rinaldo, A., Wang, D., Wen, Q., Willett, R., and Yu, Y. (2021).
Localizing changes in high-dimensional regression models.
In *International Conference on Artificial Intelligence and Statistics*, pages 2089–2097. PMLR.

Ross, S. A. (1976).
The arbitrage theory of capital asset pricing.
Journal of Economic Theory, 13(3):341–360.

Rossi, B. (2021).
Forecasting in the presence of instabilities: How we know whether models predict well and how to improve them.
Journal of Economic Literature, 59(4):1135–90.

Safikhani, A., Bai, Y., and Michailidis, G. (2022).
Fast and scalable algorithm for detection of structural breaks in big VAR models.
Journal of Computational and Graphical Statistics, 31(1):176–189.

Safikhani, A. and Shojaie, A. (2022).
Joint structural break detection and parameter estimation in high-dimensional non-stationary VAR models.
Journal of the American Statistical Association, 117(537):251–264.

Sanderson, C. and Curtin, R. (2016).
Armadillo: a template-based C++ library for linear algebra.

Schwarz, G. (1978).
Estimating the dimension of a model.
The Annals of Statistics, pages 461–464.

Scott, A. J. and Knott, M. (1974).
A cluster analysis method for grouping means in the analysis of variance.
Biometrics, pages 507–512.

Shojaie, A. and Michailidis, G. (2010).
Discovering graphical granger causality using the truncating lasso penalty.
Bioinformatics, 26(18):i517–i523.

Stock, J. H. and Watson, M. (2009).
Forecasting in dynamic factor models subject to structural instability.
The Methodology and Practice of Econometrics. A Festschrift in Honour of David F. Hendry, 173:205.

Stock, J. H. and Watson, M. W. (1996).
Evidence on structural instability in macroeconomic time series relations.
Journal of Business & Economic Statistics, 14(1):11–30.

Stock, J. H. and Watson, M. W. (1999).
Forecasting inflation.
Journal of Monetary Economics, 44(2):293–335.

Stock, J. H. and Watson, M. W. (2001).
Vector autoregressions.
Journal of Economic perspectives, 15(4):101–115.

Stock, J. H. and Watson, M. W. (2002a).
Forecasting using principal components from a large number of predictors.
Journal of the American Statistical Association, 97(460):1167–1179.

Stock, J. H. and Watson, M. W. (2002b).
Has the business cycle changed and why?
NBER Macroeconomics Annual, 17:159–218.

Stock, J. H. and Watson, M. W. (2002c).
Macroeconomic forecasting using diffusion indexes.
Journal of Business & Economic Statistics, 20(2):147–162.

Stock, J. H. and Watson, M. W. (2011).
Dynamic Factor Models.
In *The Oxford Handbook of Economic Forecasting*. Oxford University Press.

Stock, J. H. and Watson, M. W. (2012).
Generalized shrinkage methods for forecasting using many predictors.
Journal of Business & Economic Statistics, 30(4):481–493.

Su, L. and Wang, X. (2017).
On time-varying factor models: Estimation and testing.
Journal of Econometrics, 198(1):84–101.

Tibshirani, R. (1996).
Regression shrinkage and selection via the lasso.
Journal of the Royal Statistical Society: Series B (Statistical Methodology), 58(1):267–288.

Tibshirani, R. (2011).
Regression shrinkage and selection via the lasso: a retrospective.
Journal of the Royal Statistical Society: Series B (Statistical Methodology), 73(3):273–282.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005).
Sparsity and smoothness via the fused lasso.
Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67(1):91–108.

Uniejewski, B., Weron, R., and Ziel, F. (2017).
Variance stabilizing transformations for electricity spot price forecasting.
IEEE Transactions on Power Systems, 33(2):2219–2229.

van de Geer, S. A. and Bühlmann, P. (2009).
On the conditions used to prove oracle results for the lasso.
Electronic Journal of Statistics, 3:1360–1392.

van den Burg, G. J. and Williams, C. K. (2020).
An evaluation of change point detection algorithms.
arXiv preprint arXiv:2003.06222.

Vazzoler, S. (2021).
sparsevar: Sparse VAR/VECM Models Estimation.
R package version 0.1.0.

Venkatraman, E. S. (1993).
Consistency results in multiple change-point problems.
PhD thesis.

Verbesselt, J., Hyndman, R., Zeileis, A., and Culvenor, D. (2010).
Phenological change detection while accounting for abrupt and gradual trends in satellite image time series.
Remote Sensing of Environment, 114(12):2970–2980.

Vostrikova, L. Y. (1981).
Detecting "disorder" in multidimensional random processes.
In *Doklady akademii nauk*, volume 259, pages 270–274. Russian Academy of Sciences.

Voynikova, D., Gocheva-Ilieva, S., Ivanov, A., and Iliev, I. (2015).
Studying the effect of meteorological factors on the SO2 and PM10 pollution levels with refined versions of the SARIMA model.
In *AIP Conference Proceedings*, volume 1684, page 100005. AIP Publishing LLC.

Wang, D. and Tsay, R. S. (2021).
Rate-optimal robust estimation of high-dimensional vector autoregressive models.
arXiv preprint arXiv:2107.11002.

Wang, D., Yu, Y., Rinaldo, A., and Willett, R. (2019).
Localizing changes in high-dimensional vector autoregressive processes.
arXiv preprint arXiv:1909.06359.

Wang, D. and Zhao, Z. (2022).
Optimal change-point testing for high-dimensional linear models with temporal dependence.
arXiv preprint arXiv:2205.03880.

Wang, D., Zhao, Z., Lin, K. Z., and Willett, R. (2021a).
Statistically and computationally efficient change point localization in regression settings.
Journal of Machine Learning Research, 22(248):1–46.

Wang, F., Padilla, O. H. M., Yu, Y., and Rinaldo, A. (2021b).
Denoising and change point localisation in piecewise-constant high-dimensional regression coefficients.
arXiv preprint arXiv:2110.14298.

Wang, S., Cui, G., and Li, K. (2015).
Factor-augmented regression models with structural change.
Economics Letters, 130:124–127.

Welch, I. and Goyal, A. (2008).
A comprehensive look at the empirical performance of equity premium prediction.
The Review of Financial Studies, 21(4):1455–1508.

Wilms, I., Basu, S., Bien, J., and Matteson, D. (2021).
bigtime: Sparse estimation of large time series models, R package version 0.2.1.

Wong, K. C., Li, Z., and Tewari, A. (2020).
Lasso guarantees for β-mixing heavy-tailed time series.
The Annals of Statistics, 48(2):1124–1142.

Woodbury, M. A. (1950).
Inverting modified matrices.
Statistical Research Group.

Wu, W.-B. and Wu, Y. N. (2016).
Performance bounds for parameter estimates of high-dimensional linear models with correlated errors.
Electronic Journal of Statistics, 10(1):352–379.

Xie, L., Zou, S., Xie, Y., and Veeravalli, V. V. (2021).
Sequential (quickest) change detection: Classical results and new directions.
IEEE Journal on Selected Areas in Information Theory, 2(2):494–514.

Xu, H., Wang, D., Zhao, Z., and Yu, Y. (2022).
Change point inference in high-dimensional regression models under temporal dependence.
arXiv preprint arXiv:2207.12453.

Yao, Y.-C. (1988).
Estimating the number of change-points via schwarz'criterion.
Statistics & Probability Letters, 6(3):181–189.

Yau, C. Y. and Zhao, Z. (2016).
Inference for multiple change points in time series via likelihood ratio scan statistics.
Journal of the Royal Statistical Society: Series B, 78:895–916.

Yin, H., Safikhani, A., and Michailidis, G. (2021).
A general modeling framework for network autoregressive processes.
arXiv preprint arXiv:2110.09596.

Yu, Y. (2020).
A review on minimax rates in change point detection and localisation.
arXiv preprint arXiv:2011.01857.

Yu, Y., Wang, T., and Samworth, R. J. (2015).
A useful variant of the Davis–Kahan theorem for statisticians.
Biometrika, 102(2):315–323.

Yuan, M. and Lin, Y. (2006).
Model selection and estimation in regression with grouped variables.
Journal of the Royal Statistical Society: Series B (Statistical Methodology), 68(1):49–67.

Zhang, B., Geng, J., and Lai, L. (2015).
Change-point estimation in high dimensional linear regression models via sparse group lasso.
In *2015 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 815–821. IEEE.

Zhang, D. and Wu, W. B. (2017).

Gaussian approximation for high dimensional time series.
The Annals of Statistics, 45(5):1895–1919.

Zhang, D. and Wu, W. B. (2021).
Convergence of covariance and spectral density estimates for high-dimensional locally stationary processes.
The Annals of Statistics, 49(1):233–254.

Zhang, H. (2023).
Inference on structural changes in high dimensional linear regression models.
PhD thesis, Washington State University.

Zhao, Z., Jiang, F., and Shao, X. (2022).
Segmenting time series via self-normalisation.
Journal of the Royal Statistical Society Series B: Statistical Methodology, 84(5):1699–1725.

Zheng, Y. (2022).
An interpretable and efficient infinite-order vector autoregressive model for high-dimensional time series.
arXiv preprint arXiv:2209.01172.

Zou, C., Wang, G., and Li, R. (2020).
Consistent selection of the number of change-points via sample-splitting.
The Annals of Statistics, 48(1):413.

Zou, H. (2006).
The adaptive lasso and its oracle properties.
Journal of the American Statistical Association, 101(476):1418–1429.

Zou, H. and Hastie, T. (2005).
Regularization and variable selection via the elastic net.
Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67(2):301–320.