MOSUM Methods for Multiple Change-Point Analysis in Causal Networks

Dominic Owens

TB2 2019-2020

Abstract

We propose a moving sum method to detect and locate multiple change-points in multiple time series, approximated by a vector autoregressive model, based on results from [Rec19]. Non-parametric test statistics are derived using the VAR estimating function. Under the null hypothesis, an asymptotic distribution is given. Under the alternative hypothesis of multiple changes, a test with asymptotic power one is proposed; this leads naturally to a procedure for locating changes, which consistently estimates the number and location of changes. A method for estimating causal networks from stationary segments is adapted from [SM10]. Using simulated data, we demonstrate the performance of the change-point methods, and with real data on systemic financial risk we showcase a potential application of the methodology. This method is implemented in **R**, and will be available in the package 'VAR_MOSUM'.

Contents

Chapter 1

Introduction and Motivation

1.1 Introduction

Change-point analysis is the study of the detection and location of **change-points** (variously referred to as **structural breaks**) in the stochastic process generating an observed time series. Acknowledging structural breaks can allow more accurate estimation of a model; with the aim of forecasting in mind, this increases the predictive power of the model from the most recent stationary segment; for structural analysis, the model will do a better job of reflecting the true nature of the process, and we might give meaningful interpretation to structural breaks with domain knowledge. Allowing for the presence of structural breaks is perhaps the simplest divergence from the standard assumption made in classical time series analysis of global stationarity: we instead assume the underlying process consists of piecewise-stationary segments between the change-points.

Often we are interested in the first- or second-order structure of a stochastic process. Changes in these can be characterised by changes in a population parameter. Changes in the population mean, covariance, or autocovariance can be detected and located via the mean and covariance parameters of a model.

Vector Autoregression (VAR) models are a popular and general means of modelling multiple time series. They provide a reasonable approximation to nonlinear autoregressive processes, as well as to non-autoregressive specifications of processes. These can be estimated tractably and efficiently, admit clear interpretation, and give simple forecasting procedures [Lüt05]. To the end of identifying changes in structure, these balance simplicity with sufficient parameterisation such that inference can be meaningfully conducted.

Independently, **network analysis** is the study of collections of objects and their interrelationships. One important topic is that of **network inference**, in which we have observations

of **vertex attributes** associated with each object but we are unable to observe the structure of the network itself. We want to use the observations to select a graph structure which best explains the observations. In particular, **association networks** have edges defined by a measure of association between random variables, which serve as nodes. Possible association measures include covariance, correlation, and Granger-causal dependence [Gra88].

Dynamic networks have time-varying relational structures. As with any stochastic process observed over time, these can be modelled with breaks in the generating structure, giving rise to **network change-point analysis**. Here, too, the focus is to detect and locate changes to improve the ability of the model to describe the underlying process.

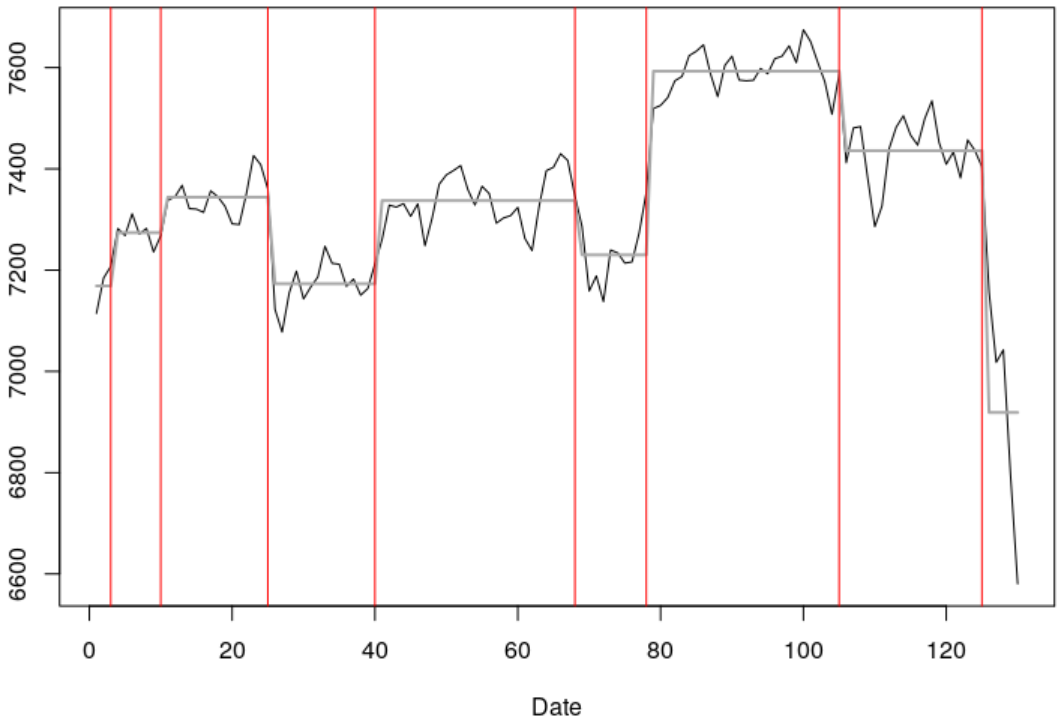


Figure 1.1: FTSE100 until 28/02/2020 - Multiple Changes in Mean

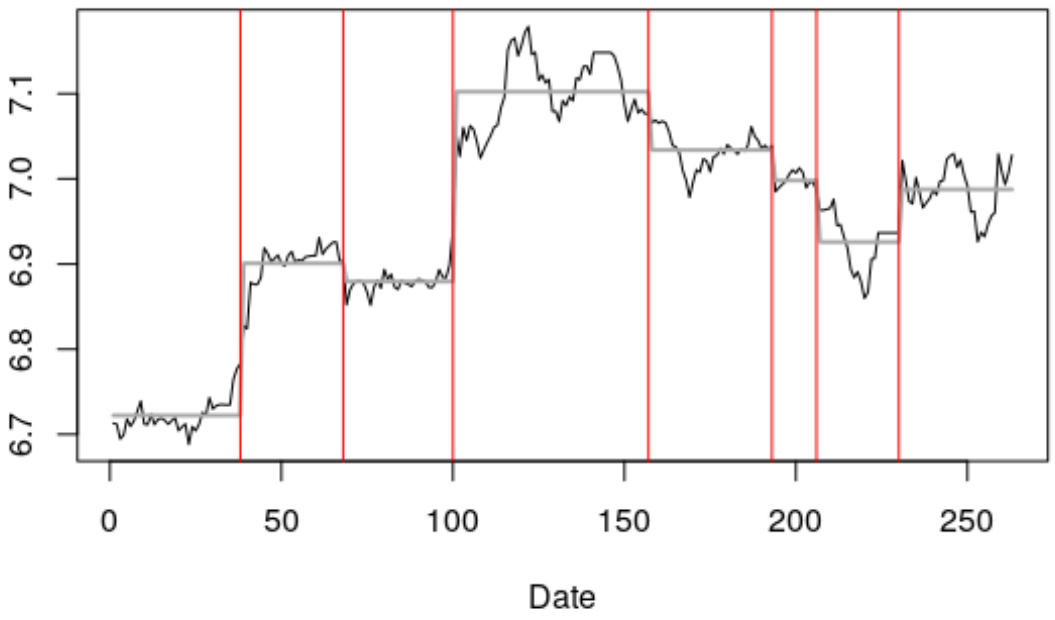


Figure 1.2: USD - CNY, 2019/03/18 - 2020/03/18

1.2 Literature Review

Change-Point Analysis for Time Series Data

The analysis of change-points in univariate time series data is a well-studied problem [AC17, AH13]. The classical statistical setting for problems with at most one change (AMOC) posits a null hypothesis of no change, against an alternative hypothesis of a single change existing [Pic85]. The simplest problem to consider is the change in mean of a parametric model, which can be easily extended to changes in other quantities (such as variance or conditional mean).

In many observed series, multiple change points may be present. These again can be approached in a hypothesis testing framework [Ven93], leading to algorithmic solutions for detection and location [F+14, KFE12, EK+18] (including for variance [IT94]). For an insightful review discussing a variety of popular algorithms and their properties, see [NHZ16].

Alternatively, this can be phrased as a model choice problem, minimising an objective function based on information criteria, constructing a model which describes the process sufficiently well and controls the number of change-points [Yao88, TOV19, PC06].

Machine-learning approaches to detecting and locating one change have been proposed via density ratio estimation [KS09, LYCS13]. These are flexible methods and can identify changes in higher-order properties of the underlying distribution, but do not offer statistical guarantees.

In contrast to the offline setting, in which all historical data is readily available, we are sometimes interested in the online setting. Here, observations arrive sequentially in a stream, and only a specified window of data is stored. Bayesian updating methods are a natural approach to this [FL07], which have been extended to include change points in Gaussian Process models [STR10]. Frequentist testing procedures have been proposed using sequential testing [Web17], as well as machine learning approaches [KS12].

Extending methods to analyse change-points in multiple (possibly high-dimensional) series is a non-trivial problem, given we should want to use multivariate dependence information to inform our inference. Univariate methods can be adapted to account for dependence structures [CF15, WS18], or distinct approaches directly utilising the multivariate structure can be employed [SS17].

Moving Sum Procedures

A hypothesis-testing procedure using a general class of score-type statistics has been proposed for the AMOC and epidemic alternatives in multivariate series [KMO15]. This relies on the estimation of the long-run covariance matrix, restricting the extension to high-dimensional scenarios and computational tractability.

For multiple changes, moving sum (MOSUM) procedures [EK$^+$18] extend this, accounting for the possibility of contagion from nearby changes by using a restricted window of observations.

A framework for multiple changes in multiple series has also been proposed, based on generalised MOSUM statistics and a new concept, estimating functions [Rec19], which allow us to detect changes in quantities other than the mean. Consistency results for existence, number and location are given for general conditions under relaxed distributional assumptions. This method again relies on estimating and inverting the long-run covariance matrix.

Vector Autoregressive Models

VAR models are perhaps the most popular means for analysing multivariate series, not least given their ability to approximate the second-order structure of a rich class of more complicated processes. Indeed, they are especially popular among macroeconomists for description, forecasting, structural analysis and policy analysis [SW01].

Time-varying parameters have been proposed as a means of capturing non-stationarity

[LM15], though these suffer from over-parameterisation leading to overfitting the data.

Within the global optimisation paradigm, a method for estimating multiple change-point locations and AR parameters under sparsity has been proposed [SS17], and under varying sparsity in [WYRW19]. A sequential scheme based on quasi-likelihoods was proposed by [Prá18], and methods for processes with infinite variance were considered in [ADL18].

We can identify no method in the literature for multiple change-point analysis in a hypothesis testing framework.

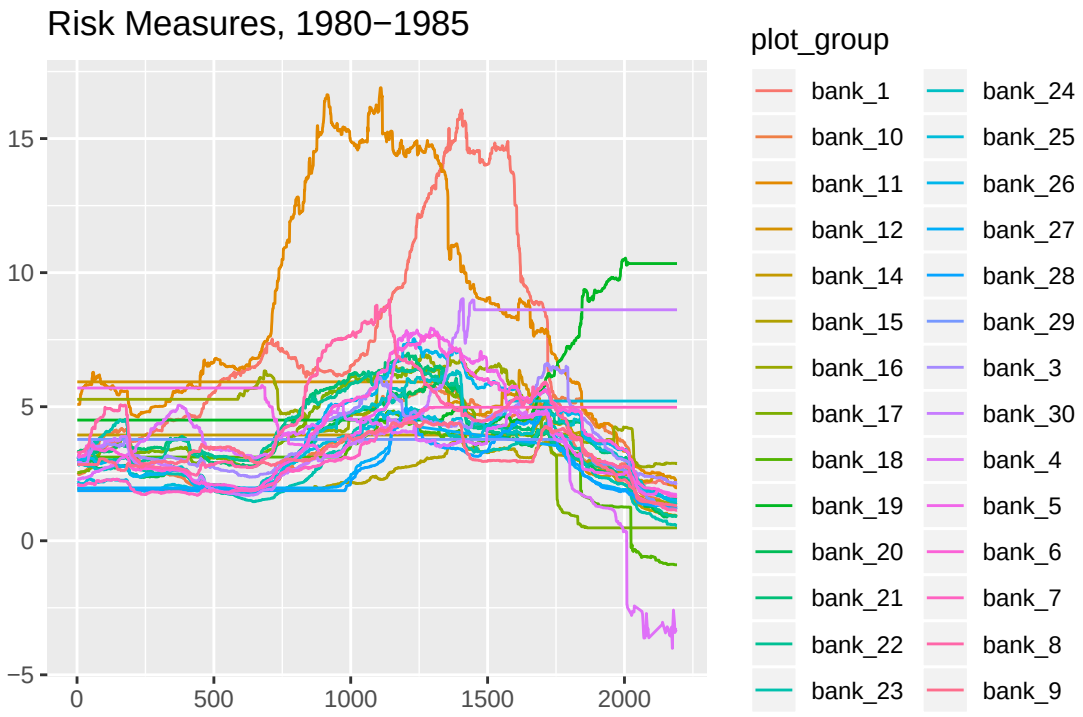


Figure 1.3: An Example of Multivariate Time Series

Network Change-Point Analysis and Causal Networks

A diverse range of methods for change point analysis of networks exist in the literature, differing strongly in the assumptions placed on the network. Three such examples are [HMPYP19], which considers random dot product graphs, [RAM17] considering Markov random-field models, and [WYR18] considering sparse bernoulli networks.

Currently, no method has been proposed for change-point analysis of causal networks, al-

though any VAR change-point detection method could be used to detect changes in the time series from which the network is inferred.

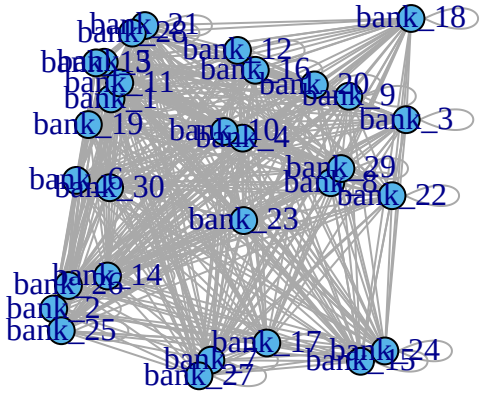


Figure 1.4: An Example Network Defined by Covariance

Applications

Change-point analysis is of eminent interest to practitioners in many fields, particularly in the sciences. Neuroscientists have made use of structural breaks for analysing electroencephalogram (EEG) data [KMO15] to understand brain function, while climate scientists allow for structural breaks in meteorological studies for the purpose of identifying shifts in long-term weather patterns [AMB+17]. Epidemiologists have considered using change-points as indicators of disease outbreaks [TFP+16].

Multiple scenarios appropriate for the application of change-point methods arise in finance and economics [SW96], for example in modelling currency exchange prices [AG02], real-time financial surveillance [PP15], and managing tail risk in dependent series [FGP18].

In the field of computer science, an online method for anomaly detection in computer networks was proposed in [LLR+09, TRBK06], as was a natural language processing method

for verifying authorship [VA16].

Vector autoregressions are the go-to model in many areas of economics. Particular applications include fiscal [Per05] and monetary [KLGS09] policy analysis.

Causal networks arose from econometrics [BGLP12], but are also used to define gene regulatory networks [ARK18], in neuroscience to understand the behaviour of neural systems [Set08], and in energy modelling to understand trade flows [PDK+20].

1.3 Motivation

The overall aim of this work is to propose a hypothesis testing approach to multiple change-point analysis in both vector autoregression models, and causal networks derived from multivariate time series. By approximating a process with a vector autoregressive model, and using moving sum statistics, we aim to detect and locate multiple changes in the second order structure of a network inferred from multiple time series, where association is defined by Granger causality.

Our motivation is given by application to risk management. Monitoring a panel of risk measures for financial institutions allows us to understand the dependencies between them, and provides useful information for forecasting the returns an instrument might incur. Historical evidence suggests many risk relationships are both non-linear and non-stationary. This suggests that analysts should allow their risk models to change over time to accurately reflect macro-level structure. In particular, by allowing for multiple changes to occur in the dependence structure, we can identify the impact of exogenous events on risk.

1.4 Report Outline

- Chapter 1 introduces the report, offers a review of literature relevant to change-point analysis for network time series, and motivates the investigation.
- Chapter 2 showcases three simple, related change-point analysis problems for univariate series, offering preliminary results on asymptotic distributions, location procedures and convergence rates.
- Chapter 3 details our new method. The chapter starts with a map of relevant results from [Rec19] and [KMO15], then assumptions for the test (and in particular, on the generating process) are outlined. The MOSUM Score-type and Wald-type statistics for the VAR(p) model are proposed, along with results on consistency and a procedure

for locating changes. Assumptions are verified to hold for the model, and choices for estimators are discussed.

- Chapter 4 outlines two procedures for causal network inference from stationary time series.
- Chapter 5 gives a simulation study, showing the performance of the method, and discusses computational aspects of the procedure. An empirical study on financial risk data shows how we can use the methods in practice.
- Chapter 6 concludes the report, collecting findings and discussing potential further work.

Chapter 2

Change-point Analysis for Univariate Series

Our aim is to detect and locate multiple change-points in autoregression for multivariate series. We begin with the simplest problem, a single change in mean for a univariate series, and extend this to multiple changes in a single series. We introduce two useful test statistics, and consider their properties.

In this chapter, the stochastic process $\{X_t : t \geq 1\}$ initiated at time $t = 1$ takes values in $\mathbb{R}$.

2.1 At-Most-One Change

We start by considering the **At Most One Change-in-mean model (1CM)**:

$$X_t = \begin{cases} \mu + \varepsilon_t, & 1 \leq t \leq \tilde{k} \\ \mu + d + \varepsilon_t, & \tilde{k} < t \leq n \end{cases} \tag{2.1}$$

The change point $1 \leq \tilde{k} \leq n$, and the step size $d \neq 0$ are unknown. The errors $\varepsilon_t \sim (0, \sigma^2) t \geq 1$ are I.I.D. unobservable white noise. When $\tilde{k} = n$, no change occurs; when $\tilde{k} < n$, we have a change in regime starting at $\tilde{k} + 1$.

We phrase this as a hypothesis test. To detect one change point is to test the following hypotheses:

$\boldsymbol{H_0} : \tilde{k} = n$ - there is no change, against

$\boldsymbol{H_1} : \tilde{k} < n$ - the mean changes starting at $\tilde{k} + 1$.

The likelihood-ratio test would be the uniformly most powerful test here. However, we cannot

obtain it without making more restrictive assumptions on the distribution of the errors, which would restrict the versatility of the method. To approach this in a nonparametric manner, we derive the *pseudo-likelihood-ratio statistic*, by first obtaining the statistic under normal errors and then proving results hold for more general distributions.

Recall that the likelihood ratio for two consecutive samples, delineated at time k, is given by

$$\begin{aligned} & L_k(X_1, \ldots, X_n) \\ & = \log \frac{\sup_{a,b} \prod_{t=1}^{k} \phi(X_t - a) \prod_{t=k+1}^{n} \phi(X_t - b)}{\sup_a \prod_{t=1}^{n} \phi(X_t - a)} \\ & = \log \frac{\prod_{i=t}^{k} \phi\left(X_t - \bar{X}_k\right) \prod_{t=k+1}^{n} \phi\left(X_t - \bar{X}_k^0\right)}{\prod_{t=1}^{n} \phi\left(X_t - \bar{X}_n\right)} \end{aligned} \tag{2.2}$$

where ϕ is the standard normal density, and we have sub-sample means $\bar{X}_k = \frac{1}{k}\sum_{t=1}^{k} X_t$, $\bar{X}_k^0 = \frac{1}{n-k}\sum_{t=k+1}^{n} X_t$.

Then, denoting partial sums $S_l = \sum_{t=1}^{l} X_t$, we can express the ratio as follows:

$$\begin{aligned} & L_k(X_1, \ldots, X_n) \\ & = \frac{1}{2}\left(\sum_{t=1}^{n}\left(X_t - \bar{X}_n\right)^2 - \sum_{t=1}^{k}\left(X_t - \bar{X}_k\right)^2 - \sum_{t=k+1}^{n}\left(X_t - \bar{X}_k^0\right)^2\right) \\ & = \frac{1}{2}\left(k\bar{X}_k^2 + (n-k)\left(\bar{X}_k^0\right)^2 - n\bar{X}_n^2\right) \\ & = \frac{1}{2}\left(-\frac{1}{n}S_n^2 + \frac{1}{k}S_k^2 + \frac{1}{n-k}\left(S_n^2 - 2S_nS_k + S_k^2\right)\right) \\ & = \frac{1}{2}\left(\frac{n}{k(n-k)}S_k^2 + \frac{k}{n(n-k)}S_n^2 - 2\frac{1}{n-k}S_nS_k\right) \\ & = \frac{1}{2}\frac{n}{k(n-k)}\left(S_k - \frac{k}{n}S_n\right)^2 = \frac{1}{2}\frac{n}{k(n-k)}\left(\sum_{t=1}^{k}\left(X_t - \bar{X}_n\right)\right)^2 \end{aligned} \tag{2.3}$$

Maximising the ratio with respect to the unknown change-point, we obtain the **Weighted CUSUM Statistic**:

$$
\begin{aligned}
T_n^w = T_n^w(X_1, \ldots, X_n) &:= \max_{1 \le k < n} \sqrt{\frac{n}{k(n-k)}} \left| \sum_{i=1}^{k} \left(X_t - \bar{X}_n \right) \right| \\
&= \max_{1 \le k < n} \sqrt{\frac{k(n-k)}{n}} \left| \frac{1}{k} \sum_{t=1}^{k} X_t - \frac{1}{n-k} \sum_{t=k+1}^{n} X_t \right|
\end{aligned} \tag{2.4}
$$

To give some intuition of how this works, consider the following diagram showing a piecewise-Gaussian series, with a change in mean at $k = 100$ and constant variance.

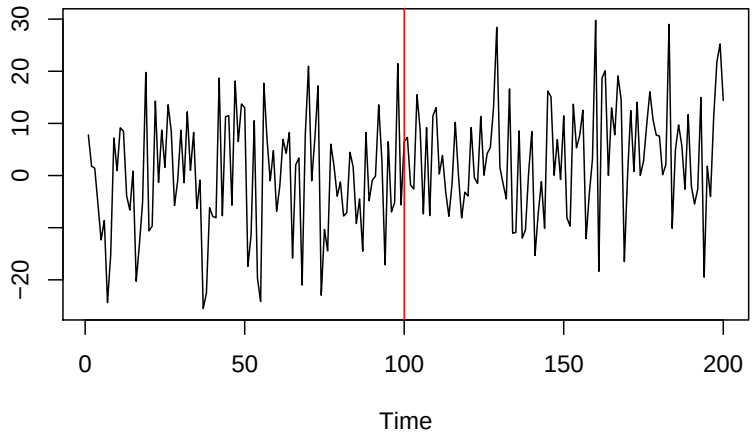


Figure 2.1: Single Change in Mean

Evaluating the statistic at each time point gives the following plot, with the maximum value marked in red. As would be expected, the statistic appears to be white noise far from the change point, peaking around the true value.

To conduct inference, we need to derive the distribution of this statistic. This depends on the distribution of the errors; regardless of whether these are specified, the exact distribution of the statistic is unknown, so we must instead find the limiting distribution and rely on asymptotic inference.

We specify some assumptions on our series, namely on the finitude of moments and strong mixing, as necessary for obtaining asymptotic results.

Assumption 2.1 (X). *Let* $\{\mathbf{X}_t\}$ *be a* $\mathbb{R}^d$ *-valued strictly stationary sequence of random*

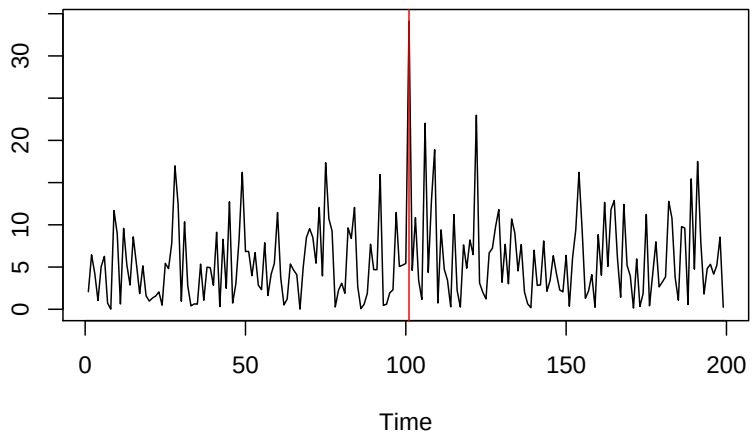


Figure 2.2: CUSUM Statistic Distribution for Single Change in Mean

vectors with

$$\mathbb{E}\left[\mathbf{X}_1\right] = 0$$

$$\mathbb{E}\left\|\mathbf{X}_1\right\|^{4+\nu} < \infty$$

for some $\nu > 0$ *satisfying a strong mixing condition with mixing rate* $\alpha(n) = O\left(n^{-\beta}\right)$ *for some* $\beta > \max(3, (4+\nu)/\nu)$

Under a parametric distribution for the errors, we could establish a form for the distribution of the test statistic [CH97, Chapter 1]. With our weaker assumptions, the initial problem to overcome is the lack of almost-sure convergence of the statistic:

Lemma 2.1. *Under assumption 2.1 it holds under the null hypothesis*

$$T_n^w(X_1, \ldots, X_n) \to \infty \qquad a.s.$$

We do, however, have convergence in distribution for extreme values.

Theorem 2.2 (Extreme Value Distribution)**.** *Under assumption 2.1*

$$\alpha(\log n)\frac{T_n^w(X_1, \ldots, X_n)}{\hat{\sigma}_n} - \beta(\log n) \xrightarrow{\mathcal{D}} G_2$$

where G_2 is a Gumbel extreme value distribution, with law $P(G_2 \leq x) = \exp(-2\exp(-x))$ and $\alpha(x) = \sqrt{2\log x}, \quad \beta(x) = 2\log x + \frac{1}{2}\log\log x - \frac{1}{2}\log\pi$

For the proof, see [C.08].

It is also important to think about the behaviour of the statistic under an alternative hypothesis. We suppose a change does occur:

Assumption 2.2 (Location of Change)**.** *Let $\tilde{k} = \lfloor \theta n \rfloor$ where $0 < \theta < 1$*

Intuitively, we would want the test to always give true positives at the limit. We establish the conditions for the test to have asymptotic power 1, that is under H_1

$$P(T_n > c_\alpha) \to 1$$

For the size of the change d_n, we can allow shrinking with respect to n: $d_n \to 0$ or a fixed change: $d_n = d$.

Theorem 2.3 (Asymptotic Power 1)**.** *Under assumptions (2.3) and (2.2), if*

$$\sqrt{\frac{n}{\log\log n}}|d_n| \xrightarrow{P} \infty$$

then

$$(\log\log n)^{-1/2} T_n^w \xrightarrow{P} \infty$$

For a proof, see [C.08].

The results seen here are demonstrative of the results we would want to find in most change-point analysis problems, and we will see similar results for different contexts frequently throughout this report.

2.2 Multiple Changes

We have considered the simplest case of possibly one change in mean. What if there are multiple changes present in the observed series?

2.2.1 Multiple Changes in Mean

Extending the AMOC-in-mean model (2.1), we consider the **Multiple Change in Mean (MCM)** model:

$$X_t = \begin{cases} \mu_1 + \varepsilon_t, & 1 \leq t \leq \tilde{k_1} \\ \mu_2 + \varepsilon_t, & \tilde{k_1} < t \leq \tilde{k_2} \\ \dots \\ \mu_q + \varepsilon_t, & \tilde{k}_{q-1} \leq t \leq \tilde{k}_q \end{cases} \tag{2.5}$$

Where $q \in \mathbb{N}$ is the number of change points $0 = \tilde{k}_0 < \tilde{k}_1 = \lfloor \theta_1 n \rfloor \leq \cdots \leq \tilde{k}_q = \lfloor \theta_q n \rfloor \leq \tilde{k}_{q+1} = n$, with proportions $0 < \theta_1 \leq \cdots \leq \theta_q \leq 1$ and unknown means $\{\mu_i : 1 \leq i \leq q\}$.

In the standard specification, q does not grow in relation to n. The proportions $\theta_i, i = 1, \dots, q$ remain fixed but the change point locations $\tilde{k}_i$ increase with n.

The simplest specification is to treat the unobservable errors $\{\varepsilon_t : t \geq 1\}$ as I.I.D. This is a very strong assumption, and is likely violated for any application we might be interested in. It is convention in the literature to provide results with relaxed conditions. Consider the following from [Rec19], in which we have time-series errors with strongly-mixing dependence:

Assumption 2.3 (E1)**.** *Let the errors* $\{\varepsilon_t : 1 \leq t \leq n\}$ *be a strictly stationary sequence with*

$$E\varepsilon_1 = 0, \quad 0 < \sigma^2 = \mathbb{E}\varepsilon_1^2 < \infty, \quad \mathbb{E}\left|\varepsilon_1\right|^{2+\nu} < \infty$$

for some $\nu > 0$

$$\sum_{h \geq 0} |\gamma(h)| < \infty,$$

where $\gamma(h) = \operatorname{cov}(\varepsilon_1, \varepsilon_{1+h})$ *and long-run variance*

$$\tau^2 := \sigma^2 + 2\sum_{h>0} \gamma(h) > 0$$

We are testing the hypotheses

$H_0 : \mathbb{E}X_1 = \mathbb{E}X_2 = \cdots = \mathbb{E}X_n$

H_1 : There is at least one change in expectation, i.e.

$\exists q \geq 1$ s.t. $\mathbb{E}X_j \neq \mathbb{E}X_{j+1}, j = 1, \dots, q-1$

The CUSUM procedure is still consistent under our new alternative hypothesis [Vos81], converging in probability at rate $\sqrt{\log \log n}$. The immensely popular binary segmentation procedure for multiple mean changes [SS75] makes use of this fact to extend the CUSUM method. However, controlling the significance level under multiple testing is difficult, as is selecting an appropriate variance estimator. This has proved a somewhat harder problem to overcome, motivating multiple approaches [F+14] including the MOSUM procedure, which we focus on.

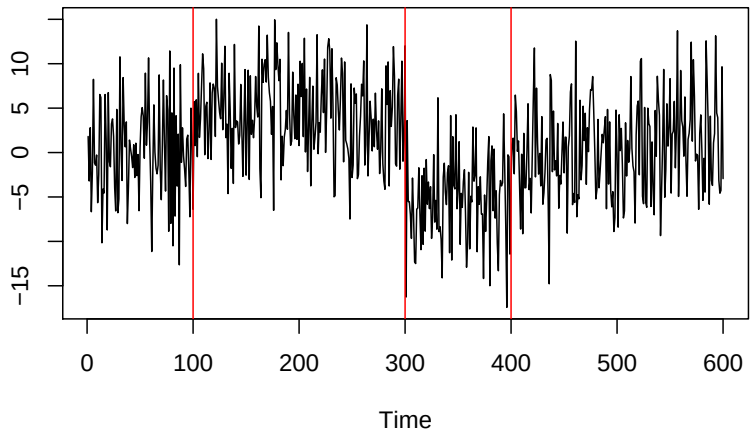
MOSUM for Detection

We consider the **MOSUM** statistic $T_n(G)$, which is specially adapted to this problem.

$$\tilde{T}_k(G) := \frac{1}{\tau\sqrt{2G}} \left| \sum_{t=k+1}^{k+G} X_t - \sum_{t=k-G}^{k} X_t \right| \quad T_n(G) := \max_{G<k<n-G} \tilde{T}_k(G) \tag{2.6}$$

Note the similarity to the CUSUM statistic (2.4). By restricting our summation at time k to the interval $t = k - G, \ldots, k + G$ we prevent, when detecting one change, contamination from other changes later or earlier in the data. This approximates the Wald statistic for the size-$2G$ subsample centred at k.

To give some intuition of how this works, consider the following diagram of a piecewise-Gaussian series with mean changes at $k = 100, 300, 400$ and constant variance.

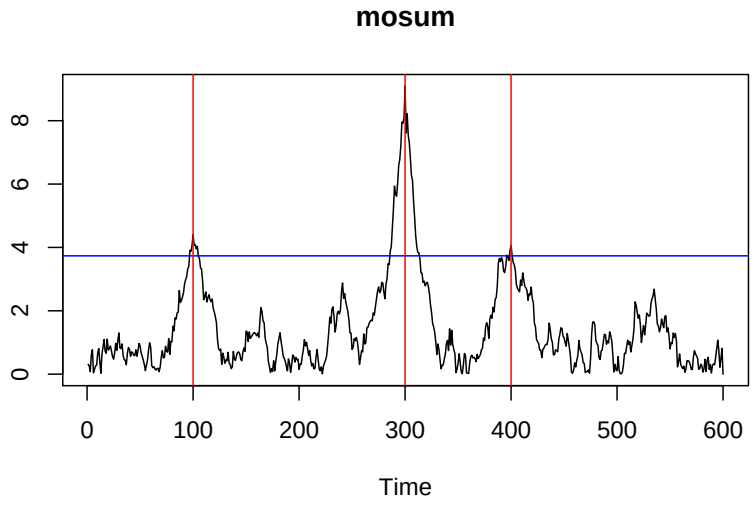


The MOSUM statistic scans over the moving window (Here, $k \pm 40$), meaning the statistic is only influenced by the unique change contained within each window. A CUSUM on this problem would be influenced by all three changes at all times.

Given we calculate statistics on subsamples, as opposed to the entire sample, we need to control the subsample size. We allow the bandwidth $G = G(n)$ to decrease with respect to n, but not too quickly:

$$\frac{n}{G} \to \infty \quad \text{and} \quad \frac{n^{\frac{2}{2+\nu}} \log n}{G} \to 0 \tag{2.7}$$

With this condition in place, we can derive familiar-looking asymptotic results:



Theorem 2.4 (MOSUM Null Asymptotics)**.** *Let H_0 hold. Let the errors fulfil (2.3) and the bandwidth G behave as in (2.7). Then*

$$a(n/G)T_n(G) - b(n/G) \xrightarrow{\mathcal{D}} G_2$$

where G_2 is a Gumbel extreme value distribution, with law $P(G_2 \leq x) = \exp(-2\exp(-x))$ and $a(x) = \sqrt{2\log x}, \quad b(x) = 2\log x + \frac{1}{2}\log\log x - \frac{1}{2}\log\pi$

This also holds when τ is replaced by an estimator $\hat{\tau}_{k,n}$ fulfilling, under H_0, the convergence

$$\max_{G \leq k \leq n-G} \left|\hat{\tau}_{k,n}^2 - \tau^2\right| = o_P\left((\log(n/G))^{-1}\right)$$

Note these results are similar to 2.2, with the distribution depending instead on n/G.

We hence have an asymptotic level-α test, rejecting the null when $T_n(G) > D_n(G;\alpha)$ where $D_n(G;\alpha) = \frac{b(n/G)+c_\alpha}{a(n/G)}, \quad c_\alpha = -\log\log\frac{1}{\sqrt{1-\alpha}}$

It is equally important that we understand how the MOSUM statistic behaves under the alternative hypothesis. To do so, it is crucial that we identify which changes are identifiable, which is determined by the relationship between the minimum distance between changes, and the size of the jumps in location.

We make the following assumption relating these two sizes:

Assumption 2.4 (Size)**.**

$$P\left(\min_{1 \leq j \leq q+1} |k_j - k_{j-1}| > CG\right) \to 1$$

for some constant $C \geq 2$

We find that the procedure has asymptotic power one under these assumptions.

Theorem 2.5 (Asymptotic Power One)**.** *[EK+18] Let the assumptions detailed in this chapter hold. Under H_1, we obtain for any $z \in R$*

$$\lim_{n\to\infty} P\left(a(n/G)T_n(G) - b(n/G) \geq z\right) = 1$$

MOSUM for Location

It is reasonable to want our procedure to consistently estimate the number and location of the changes.

The MOSUM method permits a procedure for estimating the number and locations of changes, as follows:

We find all pairs of indices $v_j, w_j, j = 1, \ldots, \widehat{q}$, such that

1) $(w_j - v_j) \geq CG$ (where $C \geq 2$ arbitrary but fixed) - sufficiently far apart

2) $\tilde{T}_k \geq c\left(\alpha_n\right)$ for $v_j < k < w_j$ - statistic is critical in between

3) $\tilde{T}_k < c\left(\alpha_n\right)$ for $k = v_j, w_j$ - statistic is not critical at ends

for the critical value,

$$c\left(\alpha_n\right) = \frac{\sqrt{2G}}{\sqrt{2\log\left(\frac{n}{G}\right)}}\left(2\log\left(\frac{n}{G}\right) + \frac{1}{2}\log\log\left(\frac{n}{G}\right) - \frac{1}{2}\log\left(\frac{2\pi}{9}\right) - \log\left(\frac{1}{1-\alpha_n}\right)\right)$$

at level α_n.

Define estimators

$$\begin{aligned} \hat{q} &:= \text{number of pairs } v_j, w_j \\ \hat{k}_j &:= \arg\max_{v_j \leq k \leq w_j} \frac{|T_{k,n}(G)|}{\hat{\tau}_{k,n}} \end{aligned} \tag{2.8}$$

Theorem 2.6 (Consistency in number and location)**.** *Under conditions, as detailed in the preceding chapter, (a)* $P\left(\hat{q}_n = q_n\right) \to 1$

(b) $P\left(\max_{1\leq j\leq q_n}\left|\hat{k}_j 1_{(j\leq \hat{q}_n|} - k_j\right| \geq G\right) \to 0$

2.2.2 Multiple Changes in Autoregression

We have seen how the method looks for multiple changes in the mean. For our task of inferring changes in the structure of an autoregressive network, we need to understand the case of changes in autoregression.

We should begin with the globally stationary, univariate linear **Autoregressive Order** p **(AR(p))** model:

$$X_t = \nu + \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + \cdots + \alpha_p X_{t-p} + \varepsilon_t = \boldsymbol{\alpha}^T \mathbb{X}_t + \varepsilon_t$$

with parameters $\boldsymbol{\alpha} = (\nu, \alpha_1, \ldots, \alpha_p)^T$, regressors $\mathbb{X}_t = (1, X_{t-1}, X_{t-2} \ldots, X_{t-p})^T$, and stationary, mean 0 errors $\{\varepsilon_t : t \geq 1\}$ (these are usually specified as white noise).

This represents the values of the process X_t at time t as a linear combination of the values at p previous time steps and an intercept ν,augmented by errors.

The parameter $\boldsymbol{\alpha}$ determines whether the process is stationarity. Assuming the process is indeed stationary is equivalent to assuming that the roots of the polynomial

$$1 - \alpha_1 t - \alpha_2 t^2 - \cdots - \alpha_p t^p$$

lie outside the unit circle. We can find $\boldsymbol{\alpha}$ through least squares when operating under the squared loss function, with the closed-form solution

$$\boldsymbol{\alpha} = C^{-1}\boldsymbol{\gamma}$$

letting C be the autocovariance of the process $\{X_t : t \geq 1\}$, so that $C = \mathbb{E}[\mathbb{X}_t \mathbb{X}_t^T]$, and letting $\boldsymbol{\gamma} = \mathbb{E}[\mathbb{X}_t X_t]$ be the vector of the first p-many autocovariances.

If we allow the parameters to vary over time, we obtain the **Multiple Changes in Autoregression (MCAR)** model

$$X_t = \begin{cases} \boldsymbol{\alpha}_1^T \mathbb{X}_t^{(1)} + \varepsilon_t, & 1 \leq t \leq \tilde{k}_1 \\ \boldsymbol{\alpha}_2^T \mathbb{X}_t^{(2)} + \varepsilon_t, & \tilde{k}_1 < t \leq \tilde{k}_2 \\ \cdots \\ \boldsymbol{\alpha}_{q+1}^T \mathbb{X}_t^{(q+1)} + \varepsilon_t, & \tilde{k}_q < t \leq n \end{cases}$$

with unknown change-points $\tilde{k}_i, i = 1, \ldots, q$, and parameter vectors $\boldsymbol{\alpha}_1 \neq \boldsymbol{\alpha}_2 \neq \cdots \neq \boldsymbol{\alpha}_{q+1}$. The errors $\{\varepsilon_t : t \geq 1\}$ follow assumption (2.3).

Results on asymptotic distributions, power under the alternative, and estimator consistency for number and location, can be derived for this simple specification. We are capable of extending this model to multivariate series, and so give these results in the more general case, which is the focus of the next chapter.

Chapter 3

Multiple Change-Point Analysis for Multivariate Series

We now understand the simpler, related settings of both a single change and multiple changes in the mean of univariate series. We also looked at univariate autoregressive series, and we will propose methods for the multivariate generalisation of this model within this chapter, to achieve our goal of detecting and locating multiple changes in a dynamic network.

To do so, we introduce the VAR model for multiple time series. This model is chosen for two reasons. Firstly, it can approximate any stationary, multivariate process under reasonable conditions (i.e. the process is purely deterministic, in the sense of the Wold decomposition [Lüt05, p.25]). Also, the VAR(p) model is parsimonious relative to other possible models (including more specific linear models like the VARMA, or non-linear volatility models including the MGARCH). This allows us to capture second-order structure and identify changes efficiently, without introducing error in estimation and computation.

Consider the single-regime, order p **Vector Autoregression (VAR(p))** model

$$\boldsymbol{Y}_t = \boldsymbol{\nu} + \boldsymbol{A}_1\boldsymbol{Y}_{t-1} + \boldsymbol{A}_2\boldsymbol{Y}_{t-2} + \cdots + \boldsymbol{A}_p\boldsymbol{Y}_{t-p} + \boldsymbol{\varepsilon}_t \tag{3.1}$$

with intercept $\boldsymbol{\nu}$, parameterised by transfer matrices $\boldsymbol{A}_i \in \mathbb{R}^{d\times d}, i = 1, \ldots, q$ with white noise error vectors $\{\boldsymbol{\varepsilon}_t \in \mathbb{R}^d : t \geq 1\}$. This has expectation $E(\boldsymbol{Y}_1) = \boldsymbol{\mu} = (\boldsymbol{I} - \boldsymbol{A}_1 - \cdots - \boldsymbol{A}_p)^{-1}\boldsymbol{\nu}$.

For notational purposes, we use a vector representation as in [KMO15].

$$\boldsymbol{Y}_t = \boldsymbol{\Phi}\mathbb{Y}_t + \boldsymbol{\varepsilon}_t \tag{3.2}$$

Define the autoregression parameter vector for d dimensions and p lagged time steps

$$\boldsymbol{\Phi} = \begin{pmatrix} \boldsymbol{\Phi}^T(1) \\ \dots \\ \boldsymbol{\Phi}^T(d) \end{pmatrix} \in \mathbb{R}^{d \times dp} \tag{3.3}$$

The correspondence between the intercept parameters $\boldsymbol{\nu}$, entries of the transfer matrices $\boldsymbol{A}_i, i = 1, \dots,$ and the entries of the vector $\boldsymbol{\Phi}(i)$ is

$$\begin{aligned} \boldsymbol{\Phi}^T(i) = \Big(\nu_i, \quad & \boldsymbol{A}_1[i,1], \boldsymbol{A}_2[i,1], \dots, \boldsymbol{A}_p[i,1], \\ & \boldsymbol{A}_1[i,2], \boldsymbol{A}_2[i,2], \dots, \boldsymbol{A}_p[i,2], \\ & \dots \\ & \boldsymbol{A}_1[i,d], \boldsymbol{A}_2[i,d], \dots, \boldsymbol{A}_p[i,d]\Big) \in \mathbb{R}^{1 \times (dp+1)} \end{aligned} \tag{3.4}$$

In particular, for the VAR(1) model this simplifies to

$$\boldsymbol{\Phi}^T(i) = (\nu_i, \quad \boldsymbol{A}_1[i,.]) \in \mathbb{R}^{1 \times (d+1)} \tag{3.5}$$

The stacked observation vector for d dimensions and p lagged time steps is

$$\mathbb{Y}_{t-1} = \begin{bmatrix} 1 \\ \mathbb{Y}_{t-1}(1) \\ \dots \\ \mathbb{Y}_{t-1}(d) \end{bmatrix} \in \mathbb{R}^{(dp+1) \times 1} \tag{3.6}$$

where for an individual channel $i = 1, \dots, d$ we have the lagged observation vector $\mathbb{Y}_{t-1}(i) = [Y_{t-1}(i), \dots, Y_{t-p}(i)]^\top \in \mathbb{R}^{(p \times 1}$

We assume the process is **stable** in the sense that

$$det\left(I_{dp+1} - \boldsymbol{\Phi}\boldsymbol{z}\right) \neq 0 \quad \text{for } |\boldsymbol{z}| \leq 1 \tag{3.7}$$

and thus the process is stationary [Lüt05, prop. 2.1].

Allowing the parameters to vary over time, we obtain the **Multiple Changes in Vector Autoregression (MCVAR)** model

$$\boldsymbol{Y}_t = \begin{cases} \boldsymbol{\Phi}_1 \mathbb{Y}_{t-1} + \boldsymbol{\varepsilon}_t, & 1 \leq t \leq \tilde{k}_1 \\ \boldsymbol{\Phi}_2 \mathbb{Y}_{t-1} + \boldsymbol{\varepsilon}_t, & \tilde{k}_1 < t \leq \tilde{k}_2 \\ \dots \\ \boldsymbol{\Phi}_q \mathbb{Y}_{t-1} + \boldsymbol{\varepsilon}_t, & \tilde{k}_{q-1} < t \leq \tilde{k}_q \end{cases} \tag{3.8}$$

with unknown change-points $\tilde{k}_i = \lfloor \lambda_i n \rfloor, i = 1, \ldots, q$, parameterised by autoregression matrices $\mathbf{\Phi}_1 \neq \mathbf{\Phi}_2 \neq \cdots \neq \mathbf{\Phi}_{q+1}$, all of dimension $d \times (dp+1)$. The error vectors $\{\boldsymbol{\varepsilon}_t \in \mathbb{R}^d : t \geq 1\}$ follow assumption (2.3).

Allowing $\boldsymbol{Y}_i^{(j)} = \mathbf{\Phi}_j \mathbb{Y}_{t-1} + \boldsymbol{\varepsilon}_t$ to be the process under regime j, we assume this series is stationary.

Dimension Reduction

Throughout this chapter we are concerned with the dimensionality of the model, as we will see this can lead to estimation issues and difficulty controlling the size of a test. We may reduce the dimension by incorporating prior knowledge of the process into the model design.

For the testing procedure, we follow the setup of [KMO15]. We consider the parameter values which *could* influence channel i:

$$\mathbf{\Phi}^T(i) = [1, \phi(i,1), \ldots, \phi(i,p), \ldots, \phi(i,dp)]^T \in \mathbb{R}^{dp+1}$$

Using the **Indicator set** of coefficients which *don't* influence channel i:

$$\mathcal{I}(i) = \{r : \phi(i,r) = 0\}$$

And the **Projection operator** determined by this set, for a vector $\mathcal{Y} \in \mathbb{R}^{dp+1}$:

$$\mathcal{P}_{\mathcal{I}}(\mathcal{Y}) = [Y_r, r \notin \mathcal{I}]^\top$$

We have the reduced vectors, with dimension $1 + dp - |\mathcal{I}(i)|$

$$\boldsymbol{a}(i) = \mathcal{P}_{\mathcal{I}(i)}(\mathbf{\Phi}(i)) \text{ and } \mathbb{X}_{t-1}(i) = \mathcal{P}_{I(i)}\left(\mathbb{Y}_{t-1}\right)$$

Giving the reduced model for channel i:

$$Y_t(i) = \boldsymbol{a}(i)^T \mathbb{X}_{t-1}(i) + e_t(i) \tag{3.9}$$

Note this is class of models contains the univariate AR(p), but in general still permits cross-dependence. The model is permitted to be misspecified in exchange for our simplifying approximation, so we denote the dependent time series of errors in channel i as $\{e_t(i)\}$. These errors may contain cross-dependence structure not contained in $\boldsymbol{a}(i)$, in particular when our chosen lag $\hat{p}$ underestimates the true lag p, although if dimensionality is not an issue for computational or estimation reasons, we can safely use $\hat{p}$ as determined by some information criterion, which will account for this form of misspecification.

Without prior knowledge, we have no reason to believe that an arbitrary channel i has a different dependence structure from another channel $i' \neq i$, or indeed that channel i has any independence from a previous observation in any channel, so we would treat the indicator set as empty (meaning $\mathbb{X}_t(i) \equiv \mathbb{X}_t(i') \equiv \mathbb{Y}_t$).

When might we have prior knowledge on dependence? If we think that the series has constant expectation $E(\boldsymbol{Y}_t) = \mathbf{0}$, for example with financial asset log-returns series, we can remove the intercept term. In Econometrics, structural VAR models (see [Lüt05, Chapter 9]) restrict possible dependence structures to reflect macroeconomic theory. For example, [Ber86] restricts dependence to fit a monetary model.

Note that in deriving asymptotics, we use the adapted form of the model (3.9):

$$Y_t(i) = \boldsymbol{a}_j(i)^T \mathbb{X}_{t-1}^{(j)}(i) + \varepsilon_t(i) \tag{3.10}$$

the difference here being that $\varepsilon_t(i) \sim (0, \sigma^2(i))$ are i.i.d., and uncorrelated across channels (See assumptions (R4), (R4*)).

Imposing the same reduction as in the single-regime case, we have the reduced multiple structural change model for each channel:

$$Y_t(i) = \begin{cases} \boldsymbol{a}_1(i)^T \mathbb{X}_{t-1}^{(1)}(i) + e_t(i), & 1 \leq t \leq \tilde{k}_1 \\ \boldsymbol{a}_2(i)^T \mathbb{X}_{t-1}^{(2)}(i) + e_t(i), & \tilde{k}_1 < t \leq \tilde{k}_2 \\ \dots \\ \boldsymbol{a}_q(i)^T \mathbb{X}_{t-1}^{(q)}(i) + e_t(i), & \tilde{k}_{q-1} < t \leq \tilde{k}_q \end{cases}$$

with unknown change-points

$$\tilde{k}_i = \lfloor \lambda_i n \rfloor, i = 1, \dots, q$$

Under regime j, dependence is determined by the parameter vector

$$\boldsymbol{a}_j(i) = \mathcal{P}_{\mathcal{I}(i)}(\boldsymbol{\Phi}_j(i)) = \boldsymbol{C}_j^{-1}(i)\boldsymbol{c}_j(i)$$

Where

$$\boldsymbol{C}_j(i) = \mathbb{E}\left(\mathbb{X}_{t-1}^{(j)}(i)\left(\mathbb{X}_{t-1}^{(j)}(i)\right)^\top\right), \quad \boldsymbol{c}_j(i) = \mathbb{E}\left(\mathbb{X}_{t-1}^{(j)}(i) Y_t^{(j)}(i)\right)$$

are the (uncentred) covariance matrix and autocovariance vector, respectively. This choice of $\boldsymbol{a}_j(i)$ is optimal in the sense that it minimises the expected squared error; in the case that the model is well-specified, this parameter determines the data generating process.

Our task is to find a test statistic with which we can phrase our inference problem. Under parametric assumptions on the error distribution, we could write down a likelihood function and possibly derive a statistic from this through maximum likelihood. However, in our

non-parametric setting, this is not possible. We can generalise maximum likelihood estimators while preserving some of their desirable properties with **Estimating Functions** (also referred to as **Z-estimators** [VdV00, p.41]).

These functions are solutions to **Estimating Equations**, which are systems of the sum of all partial derivatives set equal to zero. Sample properties (in particular, mean and variance) are often consistent for population properties, and are asymptotically normal, meaning we can conduct asymptotic inference with these functions in a straightforward manner.

In particular, we can cast the simple parameter-change problems discussed in Chapter 2 into a more general problem of a change in the expectation of the estimating function (see [Rec19, Section 1.1]), and we can do the same for a VAR(p) model.

In the single channel i under regime j, where $\boldsymbol{a}_j(i)$ is the best approximating vector, and $\tilde{\boldsymbol{a}}_j(i)$ is the plug-in estimator, the (negative) **VAR estimating function** is

$$\begin{aligned} -\boldsymbol{H}_i &= -\boldsymbol{H}_i\left(Y_t(i), \mathbb{X}_{t-1}^{(j)}(i), \tilde{\boldsymbol{a}}_j(i)\right) \\ &= \left(Y_t(i) - \tilde{\boldsymbol{a}}_j(i)^T \mathbb{X}_{t-1}^{(j)}(i)\right) \mathbb{X}_{t-1}^{(j)}(i) \\ &= Y_t(i)\mathbb{X}_{t-1}^{(j)}(i) - \tilde{\boldsymbol{a}}_j(i)^T \mathbb{X}_{t-1}^{(j)}(i)(\mathbb{X}_{t-1}^{(j)}(i))^T \\ &= (\boldsymbol{a}_j(i) - \tilde{\boldsymbol{a}}_j(i))^T \mathbb{X}_{t-1}^{(j)}(i)(\mathbb{X}_{t-1}^{(j)}(i))^T + e_t(i)\mathbb{X}_{t-1}^{(j)}(i) \in \mathbb{R}^{dp-|\mathcal{I}(i)|} \end{aligned} \tag{3.11}$$

Note the representation on the second line, which tells us the estimating function is equal to the regressor vector multiplied by the residual for the channel.

We need to evaluate the function for all d dimensions of the variable $\boldsymbol{Y}_t$, so we stack values across individual channels into a single vector for both the autoregression parameter and the estimating function:

$$\tilde{\boldsymbol{a}}_j = \begin{pmatrix} \tilde{\boldsymbol{a}}_j(1) \\ \tilde{\boldsymbol{a}}_j(2) \\ \dots \\ \tilde{\boldsymbol{a}}_j(d) \end{pmatrix} \quad \mathcal{X}_{t-1}^{(j)} = \begin{pmatrix} \mathbb{X}_{t-1}^{(j)}(1) \\ \mathbb{X}_{t-1}^{(j)}(2) \\ \dots \\ \mathbb{X}_{t-1}^{(j)}(d) \end{pmatrix} \quad \boldsymbol{H}\left(\boldsymbol{Y}_t, \mathcal{X}_{t-1}^{(j)}, \tilde{\boldsymbol{a}}_j\right) = \begin{pmatrix} \boldsymbol{H}_1 \\ \boldsymbol{H}_2 \\ \dots \\ \boldsymbol{H}_d \end{pmatrix} \tag{3.12}$$

These vectors have dimension $d(dp+1) - I$, where $I = \sum_{i=1}^{d} |\mathcal{I}(i)|$ is the total number of parameters omitted by the reduction projection, across all channels. Let these estimating function evaluations be defined analogously for the single-regime model, written without scripts j.

We use estimating functions to define two types of test statistic.

For the **Score-Type Statistic** with bandwidth $2G$ and global parameter $\tilde{\boldsymbol{a}}$, we have

$$T_n(G, \tilde{\boldsymbol{a}}) = \max_{G \le k \le n-G} T_{k,n}(G, \tilde{\boldsymbol{a}}), \qquad T_{k,n}(G, \tilde{\boldsymbol{a}}) = \frac{1}{\sqrt{2G}} \left\| \boldsymbol{\Sigma}_k^{-1/2} \boldsymbol{A}_{\tilde{\boldsymbol{a}},k} \right\| \tag{3.13}$$

Here the **Difference Vector** at time k, evaluated with parameter $\tilde{\boldsymbol{a}}$ is

$$\boldsymbol{A}_{\tilde{\boldsymbol{a}},k} = \sum_{t=k+1}^{k+G} \boldsymbol{H}\left(\boldsymbol{Y}_t, \mathcal{X}_{t-1}, \tilde{\boldsymbol{a}}\right) - \sum_{t=k-G+1}^{k} \boldsymbol{H}\left(\boldsymbol{Y}_t, \mathcal{X}_{t-1}, \tilde{\boldsymbol{a}}\right) \tag{3.14}$$

and $\boldsymbol{\Sigma}_k$ is the covariance of $\mathcal{X}_{k-1}$.

Given that a vector autoregression model amounts to fitting d-many separate regressions, one for each channel, we can express the statistic in the following way:

$$T_{k,n}(G, \tilde{\boldsymbol{a}}) = \frac{1}{\sqrt{2G}} \sum_{i=1}^{d} \left\| \boldsymbol{\Sigma}_k^{-1/2}(i) \boldsymbol{A}_{\tilde{\boldsymbol{a}},k}(i) \right\| \tag{3.15}$$

where $\boldsymbol{\Sigma}_k(i)$ is the covariance of $\mathbb{X}_{k-1}(i)$, and also the i-th diagonal block matrix of $\boldsymbol{\Sigma}_k$; this depends on our assumption of zero error correlation between channels (see 3.25).

Also,

$$\boldsymbol{A}_{\tilde{\boldsymbol{a}}(i),k}(i) = \sum_{t=k+1}^{k+G} \boldsymbol{H}_i\left(Y_t(i), \mathbb{X}_{t-1}(i), \tilde{\boldsymbol{a}}(i)\right) - \sum_{t=k-G+1}^{k} \boldsymbol{H}_i\left(Y_t(i), \mathbb{X}_{t-1}(i), \tilde{\boldsymbol{a}}(i)\right) \tag{3.16}$$

is the difference vector for channel i.

The **Wald-Type Statistic** with bandwidth $2G$ is

$$\begin{aligned} W_n(G) &= \max_{G \le k \le n-G} W_{k,n}(G) \\ W_{k,n}(G) &= \sqrt{\frac{G}{2}} \sqrt{\left(\tilde{\boldsymbol{a}}_{k+1,k+G} - \tilde{\boldsymbol{a}}_{k-G+1,k}\right)^T \boldsymbol{\Gamma}_k^{-1} \left(\tilde{\boldsymbol{a}}_{k+1,k+G} - \tilde{\boldsymbol{a}}_{k-G+1,k}\right)} \\ &= \sqrt{\frac{G}{2}} \left\| \boldsymbol{\Gamma}_k^{-1/2} \left(\tilde{\boldsymbol{a}}_{k+1,k+G} - \tilde{\boldsymbol{a}}_{k-G+1,k}\right) \right\| \end{aligned} \tag{3.17}$$

The parameter vector $\tilde{\boldsymbol{a}}_{l,u}$ is the unique solution of

$$\sum_{t=l}^{u} \boldsymbol{H}\left(\boldsymbol{Y}_t, \mathcal{X}_{t-1}, \tilde{\boldsymbol{a}}_{l,u}\right) = \boldsymbol{0}$$

which is a stacked vector

$$\tilde{\boldsymbol{a}}_{l,u} = \begin{pmatrix} \tilde{\boldsymbol{a}}_{l,u}(1) \\ \tilde{\boldsymbol{a}}_{l,u}(2) \\ \dots \\ \tilde{\boldsymbol{a}}_{l,u}(d) \end{pmatrix} \tag{3.18}$$

of the solutions to the individual-channel least squares problems:

$$\tilde{\boldsymbol{a}}_{l,u}(i) = \left(\sum_{t=l}^{u} Y_t(i)\mathbb{X}_{t-1}^T(i)\right)\left(\sum_{t=l}^{u} \mathbb{X}_{t-1}(i)\mathbb{X}_{t-1}^T(i)\right)^{-1} \quad i = 1, \ldots, d \tag{3.19}$$

The matrix $\boldsymbol{\Gamma}_k$ is the asymptotic covariance matrix of $\sqrt{G}\tilde{\boldsymbol{a}}_{k-G+1,k}$:

$$\boldsymbol{\Gamma}_k = \lim_{n\to\infty, G\to 0} GE[(\tilde{\boldsymbol{a}}_{k-G+1,k} - E[\tilde{\boldsymbol{a}}_{k-G+1,k}])(\tilde{\boldsymbol{a}}_{k-G+1,k} - E[\tilde{\boldsymbol{a}}_{k-G+1,k}])^T] \tag{3.20}$$

This is also specified in [Rec19, Section 3.1].

As in the case of 3.15, we can express the Wald statistic as a sum across channels:

$$W_{k,n}(G) = \frac{1}{\sqrt{2G}} \sum_{i=1}^{d} \left\| \boldsymbol{\Gamma}_k^{-1/2}(i) \left(\tilde{\boldsymbol{a}}_{k+1,k+G}(i) - \tilde{\boldsymbol{a}}_{k-G+1,k}(i)\right) \right\| \tag{3.21}$$

where $\boldsymbol{\Gamma}_k^{-1/2}(i)$ is the asymptotic covariance matrix of $\sqrt{G}\tilde{\boldsymbol{a}}_{k-G+1,k}(i)$.

3.1 Results in the Literature

We now have the multiple change in autoregression problem fully established. As mentioned in the literature review 1.2, theoretical results for a more general problem based on estimating functions were proposed in [Rec19]. The single change case for a VAR model was considered in [Muh13], and the corresponding paper [KMO15]. Our work depends strongly on the results from the given texts, so their contributions are mapped out in the following.

After establishing the problem, [Rec19] proceeds as follows.

Chapter 2 concerns MOSUM score-type statistics.

- The score-type statistic is proposed (Def 2.0.1)
- Under the null, assumptions A.1.1 on the bandwidth convergence, A.1.2 on stationarity, misspecification, and estimating function covariance, A.1.3 on strong invariance, A.1.4 on the parameter estimator sequence, and A.1.5 on the long-run covariance estimator are made.
- The Gumbel(2) null limit distribution for the transformed statistic, evaluated with parameter and covariance estimators, is derived (Thm 2.1.1)

- Under the alternative, assumptions A.2.1 on the number of changes, and A.2.2-A.2.5 mirroring A.1.2-A.1.5 are made.
- Assumption A.2.6, that the expectation of the estimating function series changes at change points, permits an asymptotic level-α test. This assumption plays a similar role to an assumption on the jump sizes and the minimum distance between changes, in that it permits identification of a change.
- Asymptotic power 1 for the test, evaluated with parameter and covariance estimators, is derived (Thm 2.1.5)
- A procedure for estimating the number and locations of changes is given (Section 2.1.3.1)
- Assumptions A.2.7 specifying the identifiable set of changes, A.2.8 on the sequence of significance levels, and A.2.9 on the parameter estimators are made
- Consistency in estimating the number of changes is derived, with estimators for parameters and long-run covariance (Thm 2.1.8), which implies consistency in location (Cor. 2.1.11)
- Stronger assumptions A.2.10 with Hajek-Renyi inequalities and A.2.11 on the parameter estimator sequence are made, giving locational convergence in probability with estimators for the parameters and the covariance (Thm 2.2.5)
- Crucially, in section 2.3 assumptions are verified for a stationary, strongly-mixing sequence. Assumptions B.1.1-B.1.6 are made on the moments of the parameter estimator sequence under the null. A.1.3, A.1.4 are shown to hold.
- Asymptotic normality of Z-estimators (which includes estimating functions) is shown under assumptions B.1.1-B.1.6 and the null (Thm 2.3.5)
- Likewise under the alternative, assumptions B.2.1-B.2.6 are made on the moments of the parameter estimator sequence. Assumptions A.2.3, A.2.4, A.2.9 are shown to hold (Lemma 2.3.7).
- $\sqrt{n}$-consistency of the parameter estimator under the moment assumptions and the alternative is derived (Thm 2.3.9)
- Two related problems with the procedure, namely the choice of bandwidth in relation to the minimum gap between changes, and detectability of changes, are discussed (Section 2.4).

Chapter 3 concerns MOSUM Wald-Type statistics.

- The MOSUM Wald-type statistic is introduced
- Additional assumptions B.1.7 and B.1.8 are placed on the first and second derivative matrices
- Under these assumptions, asymptotic normality of a local parameter estimator is derived for the strongly mixing process (Thm 3.1.3)
- The asymptotic covariance matrix is specified (Equation 3.1)
- It is derived that a transformed statistic, under the null and few assumptions, is asymptotically distributed as Gumbel(2), using estimators for the long-run covariance and the first derivative matrix. (Thm 3.1.8)
- For the alternative, additional assumptions B.2.7-B.2.11 are placed on the derivative matrices.
- It is shown that an asymptotic level-α test has asymptotic power 1, under some assumptions and using an estimator for the asymptotic covariance (Thm 3.1.12)
- A procedure for estimating the locations and number of changes is proposed (Section 3.1.2.2). It is shown to be consistent in number (Thm 3.1.15) and weakly consistent in location (Cor. 3.1.16) with an estimator for the long-run covariance.
- The particular setting of the linear regression model with random design is considered (Section 3.2), and results are shown under further assumptions (R1)-(R6), (R1*)-(R7*). An explicit form for the asymptotic covariance is given (Equation 3.38)

Chapter 4 gives simulation studies for linear regression and Poisson autoregression models.

- For linear regression: Assumptions (R1)-(R6), (R1*)-(R7*) are verified for the Wald-type statistic
- Assumptions for Thms. 2.1.1 and 2.1.8 are verified for the score-type statistic. For the null, these are A.1.3 of strong invariance of the estimating function and A.1.4 of convergence in the difference matrix. For the alternative, these are A.2.3 of strong invariance of the estimating function, with A.2.4 and A.2.9 of convergence in the difference matrix.
- A simplified form for estimating the covariance of the estimating function is given, along with forms for both statistics, and some estimators for the variance are discussed (Section 4.1.3)

- For Poisson autoregression: Assumptions for a strongly mixing series are verified, namely the assumptions B.1.1-B.2.11. A MOSUM estimator for the long-run covariance is given, with the asymptotic covariance matrix of the parameters, giving a MOSUM estimator for the long-run autocovariance (Section 4.2.3)

We are interested particularly in the VAR setting. We consider the contributions of [KMO15] to the single-change problem for VAR models.

- Section 1: The VAR model is constructed, isolating the regression problem for each individual channel
- The reduced model is proposed to account for the dimensionality of the long-run covariance matrix
- The model is allowed to be misspecified, allowing use as a simple approximation to more complex processes
- Section 2: Assumptions are placed on the model. 2.1 supposes the process is strongly mixing, corresponding to A.1.3. 2.2 supposes the process is ergodic, corresponding to (R1), and that covariance and autocovariance estimators are consistent, corresponding to (R5) and A.1.2 (while not explicitly referencing estimating functions).
- Two general forms for statistics are given, maximum- and sum-types, for the single change problem (Section 2.1.4). These depend on a weight matrix and weight function, choices for which are discussed (Section 2.1.6). It is proven that these converge to functions of Brownian bridges under the null (Thm 2.1). Under the alternative and assumption (2.13) on identifiability of changes, the test has asymptotic power 1 (Thm 2.3). With an estimator for the weight matrix, the procedure is consistent in location (Thm 2.4).

3.2 Assumptions

Before we propose the Score- and Wald-type statistics for the VAR model, we outline necessary assumptions modified from [Rec19], explaining what they mean and how they are used. Assumption which can and will be verified are labelled in cyan, those which cannot are labelled in red.

3.2.1 Assumptions for the Score-type Statistic

The following assumptions are used in deriving the asymptotic null distribution of the test statistic [Rec19, Theorem 2.1.1].

Under the Null Hypothesis

We begin by assuming that the bandwidth grows with n, but not too quickly.

Assumption 3.1. *[Rec19, Assumption A.1.1.] Let the bandwidth G depend on n, i.e. $G = G(n)$. Furthermore, for $\nu > 0$ assume that*

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

We assume the series is well-behaved, has some optimal parameter, and the estimating function is stationary with an existing long-run covariance matrix.

Assumption 3.2. *[Rec19, Assumption A.1.2.] Let $\{\boldsymbol{Y}_t : t \geq 1\}$ be a stationary series with expectation $E(\boldsymbol{Y}_1) = \boldsymbol{\mu}$ following a distribution determined by $\tilde{\boldsymbol{a}}_0$ in a correctly specified model. Under misspecification let $\tilde{\boldsymbol{a}}_0$ be the best approximating parameter for $\{\boldsymbol{Y}_t : t \geq 1\}$ in the sense of $E\left(\boldsymbol{H}\left(\boldsymbol{Y}_t, \mathcal{X}_{t-1}, \tilde{\boldsymbol{a}}_0\right)\right) = \boldsymbol{0}$*

Furthermore, we assume that the stationary sequence $\{\boldsymbol{H}\left(\boldsymbol{Y}_t, \mathcal{X}_{t-1}, \tilde{\boldsymbol{a}}\right) : t \geq 1\}$ has a positive definite long-run covariance matrix $\boldsymbol{\Sigma}(\tilde{\boldsymbol{a}}) = \boldsymbol{\Sigma}$

We assume the sums of the estimating function series fulfil a Gaussian approximation.

Assumption 3.3. *[Rec19, Assumption A.1.3.] Let $\boldsymbol{S}(k, \tilde{\boldsymbol{a}}) = \sum_{t=1}^{k} \boldsymbol{H}\left(\boldsymbol{Y}_t, \mathcal{X}_{t-1}, \tilde{\boldsymbol{a}}\right)$ fulfil a strong invariance principle. So possibly after changing the probability space there exists a d-dimensional standard Wiener process $\{\boldsymbol{W}(k) : k \geq 0\}$ with identity matrix $\boldsymbol{I}_d$ as covariance matrix and $\nu > 0$ such that*

$$\left\| \boldsymbol{\Sigma}^{-1/2} \left(\boldsymbol{S}(k, \tilde{\boldsymbol{a}}) - E(\boldsymbol{S}(k, \tilde{\boldsymbol{a}}))\right) - \boldsymbol{W}(k) \right\| = O\left(k^{1/(2+\nu)}\right) \text{ a.s.}$$

as k goes to infinity.

We assume that the difference matrix with an estimating sequence for the parameter converges in probability to the true difference matrix.

Assumption 3.4. *[Rec19, Assumption A.1.4.] Let*

$$\begin{aligned}
&\max_{G\le k\le n-G} \frac{1}{\sqrt{2G}}\|\boldsymbol{A}_{\widehat{\boldsymbol{a}}_n,k} - \boldsymbol{A}_{\tilde{\boldsymbol{a}},k}\| \\
&= \max_{G\le k\le n-G} \frac{1}{\sqrt{2G}}\| \sum_{t=k+1}^{k+G} \Big(\boldsymbol{H}(\boldsymbol{Y}_t, \mathcal{X}_{t-1}, \widehat{\boldsymbol{a}}_n) - \boldsymbol{H}(\boldsymbol{Y}_t, \mathcal{X}_{t-1}, \tilde{\boldsymbol{a}})\Big) - \\
&\qquad\qquad \sum_{i=k-G+1}^{k} \Big(\boldsymbol{H}(\boldsymbol{Y}_t, \mathcal{X}_{t-1}, \widehat{\boldsymbol{a}}_n) - \boldsymbol{H}(\boldsymbol{Y}_t, \mathcal{X}_{t-1}, \tilde{\boldsymbol{a}})\Big)\| \\
&\qquad\qquad = o_P\left((\log(n/G))^{-1/2}\right)
\end{aligned}$$

hold for some $\tilde{\boldsymbol{a}}$*, where* $\{\widehat{\boldsymbol{a}}_n\}_{n\in\mathbb{N}}$ *is an estimator sequence*

We assume the estimator of the long-run covariance converges to the true long-run covariance.

Assumption 3.5. *[Rec19, Assumption A.1.5.] The estimator* $\widehat{\boldsymbol{\Sigma}}_{k,n}$ *of the long-run covariance matrix* $\boldsymbol{\Sigma}$ *can depend on* k *and satisfies*

$$\max_{G\le k\le n-G} \left\|\widehat{\boldsymbol{\Sigma}}_{k,n}^{-1/2} - \boldsymbol{\Sigma}^{-1/2}\right\|_F = o_P\left((\log(n/G))^{-1}\right)$$

under the null hypothesis.

Under the Alternative Hypothesis

The following assumptions are used in showing that the level-α testing procedure has asymptotic power 1 [Rec19, Theorem 2.1.5], and hence in showing the estimation procedure for the number and locations of any changes is consistent [Rec19, Theorem 2.1.8, Cor. 2.1.11].

We assume possibly multiple changes exist, and scale them into the unit interval.

Assumption 3.6. *[Rec19, Assumption A.2.1] Let* q *be the number of change points, occuring in the time period, which is unknown but fixed. Furthermore, let* $k_{1,n} < \ldots < k_{q,n}$ *be the change points depending on the sample size* n *in the following way:* $k_{j,n} = \lfloor \lambda_j n \rfloor$ *with* λ_j *as rescaled change point being a constant but unknown value in* $(0,1)$*, for* $j = 1, \ldots, q$

We model the series as piecewise stationary between the changes, and assume it is determined by some optimal parameter.

Assumption 3.7. *[Rec19, Assumption A.2.2] Let* $\{\boldsymbol{Y}_t : t \geq 1\}$ *be a piecewise stationary series such that*

$$\boldsymbol{Y}_t = \begin{cases} \boldsymbol{Y}_t^{(1)} & \text{if } 1 \leq t \leq k_{1,n} \\ \boldsymbol{Y}_t^{(2)} & \text{if } k_{1,n} < t \leq k_{2,n} \\ \vdots & \\ \boldsymbol{Y}_t^{(q+1)} & \text{if } k_{1,n} < t \leq n \end{cases}$$

where $\left\{\boldsymbol{Y}_t^{(j)} : t \geq 1\right\}$ *is a stationary series with expectation* $E(\boldsymbol{Y}_t^{(j)}) = \boldsymbol{\mu^{(j)}}$*. The regressor series* $\{\mathbb{X}_t : t \geq 1\}$ *is stationary, following a distribution determined by* $\tilde{\boldsymbol{a}}_j$ *in a correctly specified model. Define the stacked vector of these as* $\mathcal{X}^{(j)}$*. Under misspecification let* $\tilde{\boldsymbol{a}}_j$ *be the best approximating parameter for* $\left\{\mathbb{X}_t^{(j)} : t \geq 1\right\}$ *in the sense of* $E\left(\boldsymbol{H}\left(\boldsymbol{Y}_t, \mathcal{X}_{t-1}^{(j)}, \tilde{\boldsymbol{a}}_j\right)\right) = \boldsymbol{0}$.

Furthermore, we assume that the stationary sequence $\left\{\boldsymbol{H}\left(\boldsymbol{Y}_t, \mathcal{X}_{t-1}^{(j)}, \tilde{\boldsymbol{a}}_j\right) : t \geq 1\right\}$ *has a positive definite long-run covariance matrix* $\boldsymbol{\Sigma}_{(j)}(\tilde{\boldsymbol{a}}) = \boldsymbol{\Sigma}_{(j)}$

This strong invariance principle is needed in proofs.

Assumption 3.8. *[Rec19, Assumption A.2.3] Let* $\boldsymbol{S}(j,k,\tilde{\boldsymbol{a}}) = \sum_{t=1}^{k} \boldsymbol{H}\left(\boldsymbol{Y}_t, \mathcal{X}_{t-1}^{(j)}, \tilde{\boldsymbol{a}}\right)$ *fulfil a strong invariance principle. So possibly after changing the probability space there exists a* d*-dimensional standard Wiener process* $\{\boldsymbol{W}(k) : k \geq 0\}$ *with identity matrix* $\boldsymbol{I}_d$ *as covariance matrix and* $\nu > 0$ *such that*

$$\left\| \boldsymbol{\Sigma}_{(j)}^{-1/2} \left(\boldsymbol{S}(j,k,\tilde{\boldsymbol{a}}) - E(\boldsymbol{S}(j,k,\tilde{\boldsymbol{a}}))\right) - \boldsymbol{W}(k) \right\| = O\left(k^{1/(2+\nu)}\right) \quad a.s.$$

as k *goes to infinity.*

We wish to use an estimator sequence for the parameters, so require the sequence to be bounded in the size of the difference vector.

Assumption 3.9. *[Rec19, Assumption A.2.4] Let* $\{\widehat{\boldsymbol{a}}_n\}_{n \in N}$ *be a sequence of estimators fulfilling, for some* $\tilde{\boldsymbol{a}}$

$$\max_{G \leq k \leq n-G} \frac{1}{\sqrt{2G}} \left\| \boldsymbol{A}_{\widehat{\boldsymbol{a}}_n, k} - \boldsymbol{A}_{\tilde{\boldsymbol{a}}, k} \right\| = O_P(\sqrt{\log(n/G)})$$

The following set of points which are outside of the G-window around every changepoint is used in proofs:

$$A_{n,G} := \{k \in \{G, \ldots, n-G\} : |k - \lfloor \lambda_j n \rfloor| \geq G \forall j \in \{1, \ldots, q\}\} \tag{3.22}$$

The complementary set contains all time points close enough to a change-point to influence a MOSUM statistic:

$$B_{n,G} := \{k \in \{G, \dots, n-G\} : \exists j \in \{1, \dots, q\} : |k - k_{j,n}| \leq G\} \tag{3.23}$$

We also want to use an estimator for the long-run covariance matrix, so make the following assumptions on its' size

Assumption 3.10. *[Rec19, Assumption A.2.5] The estimator $\widehat{\mathbf{\Sigma}}_{k,n}$ of the long run covariance matrix $\mathbf{\Sigma}_k$ is positive definite and satisfies*

(a) $\max_{G \leq k \leq n-G} \left\| \widehat{\mathbf{\Sigma}}_{k,n}^{-1/2} \right\|_F = O_P(1)$

(b) $\max_{k \in A_{n,G}} \left\| \widehat{\mathbf{\Sigma}}_{k,n}^{-1/2} - \mathbf{\Sigma}_k^{-1/2} \right\|_F = o_P\left(\log(n/G)^{-1}\right)$ *with $A_{n,G}$ as in (3.22)*

(c) $\max_{k \in B_{n,G}} \left\| \widehat{\mathbf{\Sigma}}_{k,n}^{-1/2} - \mathbf{\Sigma}_{A,k}^{-1/2} \right\|_F = o_P(1)$, *with $B_{n,G}$ as in (3.23) and $\{\mathbf{\Sigma}_{A,k}\}$ is a sequence of positive definite matrices fulfilling* $\sup_n \sup_{k \in B_{n,G}} \|\mathbf{\Sigma}_{A,\boldsymbol{k}}\|_F < \infty$

For consistency of the test, we need the following assumption on the identifiability of changes

Assumption 3.11. *[Rec19, Assumption A.2.6] For at least one $j \in \{1, \dots, q\}$ it holds that*

$$E\left(\boldsymbol{H}\left(\boldsymbol{Y}_t, \mathcal{X}_{t-1}^{(j)}, \tilde{\boldsymbol{a}}\right)\right) \neq E\left(\boldsymbol{H}\left(\boldsymbol{Y}_t, \mathcal{X}_{t-1}^{(j+1)}, \tilde{\boldsymbol{a}}\right)\right)$$

We collect these changes into a single set

Assumption 3.12. *[Rec19, Assumption A.2.7] Let $\widetilde{Q} = \widetilde{Q}(\tilde{\boldsymbol{a}})$ be the set of indices of all rescaled change points causing a change in the expected value of the transformed series (detectable changes), i.e.*

$$E\left(\boldsymbol{H}\left(\boldsymbol{Y}_t, \mathcal{X}_{t-1}^{(j)}, \tilde{\boldsymbol{a}}\right)\right) \neq E\left(\boldsymbol{H}\left(\boldsymbol{Y}_t, \mathcal{X}_{t-1}^{(j+1)}, \tilde{\boldsymbol{a}}\right)\right)$$

for all $j \in \widetilde{Q}$ and

$$E\left(\boldsymbol{H}\left(\boldsymbol{Y}_t, \mathcal{X}_{t-1}^{(j)}, \tilde{\boldsymbol{a}}\right)\right) = E\left(\boldsymbol{H}\left(\boldsymbol{Y}_t, \mathcal{X}_{t-1}^{(j+1)}, \tilde{\boldsymbol{a}}\right)\right)$$

for all $j \in \{1, \dots, q\} \backslash \widetilde{Q}$. Furthermore, let $\widetilde{q} = \tilde{q}(\boldsymbol{a})$ be the number of elements of $\widetilde{Q}$ which is the number of detectable changes.

We distinguish between the total number of actual changes, and the total number of *detectable* changes, with the following sets

$$\begin{aligned}\widetilde{A}_{n,G} &:= \left\{k \in \{G, \dots, n-G\} : |k - k_{j,n}| \geq G \forall j \in \widetilde{Q}\right\} \\ \bar{B}_{n,G} &:= \left\{k \in \{G, \dots, n-G\} : \exists j \in \widetilde{Q} : |k - k_{j,n}| < (1-\varepsilon)G\right\}\end{aligned} \tag{3.24}$$

In proofs, it is required that the significance level is not fixed, but converges to 0.

Assumption 3.13. *[Rec19, Assumption A.2.8] Let the sequence of significance levels α_n fulfill*

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

We again wish to use an estimator sequence for the parameters, so we assume these determine difference vectors that converge.

Assumption 3.14. *[Rec19, Assumption A.2.9] Let $\{\hat{\boldsymbol{a}}_n\}_{n\in\mathbb{N}}$ be a sequence of estimators fulfilling*

(I) $\max_{k\in A_{n,G}} \frac{1}{\sqrt{2G}} \|\boldsymbol{A}_{\hat{\boldsymbol{a}}_n,k} - \boldsymbol{A}_{\tilde{\boldsymbol{a}},k}\| = o_P\left((\log(n/G))^{-1/2}\right)$ *with $A_{n,G}$ as in 3.22*

(II) $\max_{k\in \tilde{A}_{n,G}} \frac{1}{\sqrt{2G}} \|\boldsymbol{A}_{\hat{\boldsymbol{a}}_n,k} - \boldsymbol{A}_{\tilde{\boldsymbol{a}},k}\| = o_P(\sqrt{\log(n/G)})$ *with $\widetilde{A}_{n,G}$ as in 3.24*

The final three assumptions [Rec19, Assumptions A.2.10, A.2.11, A.2.12] of the corresponding subsection are supplementary, and used to determine convergence rates; we do not detail these here for brevity.

3.2.2 Assumptions for the Score-type Statistic: Strongly-Mixing VAR Processes

For the particular case of the strongly-mixing sequence, which can include VAR processes, we make specific assumptions. These are verified in [Rec19, Section 2.3].

Under the Null

As a higher-level condition, we assume the series $\{\mathbb{X}_t : t \geq 1\}$ is stationary and strongly mixing with a mixing rate $\alpha(n)$ satisfying $\alpha(n) = O\left(n^{-\beta}\right)$ for some $\beta > 1 + 2/\nu$. The estimator sequence $\hat{\boldsymbol{a}}_n$ is $\sqrt{n}$-consistent. Let $\Theta \subset \mathbb{R}^{d(dp+1)-I}$ be the regression parameter space.

We make assumptions on the moments of the series, the gradient, and the Hessian.

The expectation is finite:

Assumption 3.15. *[Rec19, Assumption B.1.1] Let* $E\left(\|\boldsymbol{H}\left(\boldsymbol{Y}_t, \mathcal{X}_{t-1}, \tilde{\boldsymbol{a}}\right)\|\right) < \infty$ *for all* $\tilde{\boldsymbol{a}} \in \Theta$

The second moment is finite:

Assumption 3.16. *[Rec19, Assumption B.1.2] Let* $E\left(\|\boldsymbol{H}\left(\boldsymbol{Y}_t, \mathcal{X}_{t-1}, \tilde{\boldsymbol{a}}\right)\|^2\right) < \infty$

The expected supremum of the gradient (matrix) is finite.

Assumption 3.17. *[Rec19, Assumption B.1.3] Let* $E\left(\sup_{\boldsymbol{a}\in\Theta}\|\nabla\boldsymbol{H}\left(\boldsymbol{Y}_t, \mathcal{X}_{t-1}, \tilde{\boldsymbol{a}}\right)\|_F\right) < \infty$

The expected supremum of the Hessian (tensor) is finite.

Assumption 3.18. *[Rec19, Assumption B.1.4] Let* $E\left(\sup_{\boldsymbol{a}\in\Theta}\|\nabla^2\boldsymbol{H}\left(\boldsymbol{Y}_t, \mathcal{X}_{t-1}, \tilde{\boldsymbol{a}}\right)\|_F\right) < \infty$

The $2+\nu$-th uncentred moment is finite.

Assumption 3.19. *[Rec19, Assumption B.1.5] There exists* $\nu > 0$ *such that*

$$E\left(\|\boldsymbol{H}\left(\boldsymbol{Y}_t, \mathcal{X}_{t-1}, \tilde{\boldsymbol{a}}\right)\|^{2+\nu}\right) < \infty$$

The gradient (matrix) has a finite $2+\nu$-th uncentred moment.

Assumption 3.20. *[Rec19, Assumption B.1.6] There exists* $\nu > 0$ *such that*

$$E\left(\|\nabla\boldsymbol{H}\left(\boldsymbol{Y}_t, \mathcal{X}_{t-1}, \tilde{\boldsymbol{a}}\right)\|_F^{2+\nu}\right) < \infty$$

holds for all $\tilde{\boldsymbol{a}} \in \Theta$

Under the Alternative

Under the alternative, we treat $\{\mathbb{X}_t : t \geq 1\}$ as piecewise stationary within each regime, as per 3.7. We assume each segment $\left\{\mathbb{X}_t^{(j)} : t \geq 1\right\}$ is strongly mixing.

Further, we make assumptions [Rec19, Assumptions B.2.1-B.2.6] on the moments of the series and its' derivatives mirroring those in 3.2.2 for each segment.

3.2.3 Assumptions for the Wald-type Statistic

We have separate assumptions for deriving results for the Wald-type statistic. In particular, these are further assumptions on the moments of the process and its' derivatives. Many of the assumptions in 3.2.2 are also placed on the process here.

Under the Null Hypothesis

The Hessian (tensor) is bounded in the $2+\nu$-th moment

Assumption 3.21. *[Rec19, Assumption B.1.7] Let*

$$E\left(\sup_{\boldsymbol{a}\in\Theta}\left\|\nabla^2\boldsymbol{H}\left(\boldsymbol{Y}_t,\mathcal{X}_{t-1},\boldsymbol{a}\right)\right\|_F^{2+\nu}\right)<\infty$$

hold for some $\nu>0$.

The inverse of the gradient matrix is regular and all entries are finite

Assumption 3.22. *[Rec19, Assumption B.1.8] Let $\boldsymbol{V}(\boldsymbol{a})=E\left(\nabla\boldsymbol{H}\left(\boldsymbol{Y}_t,\mathcal{X}_{t-1},\boldsymbol{a}\right)\right)^T$ be a regular matrix for all $\boldsymbol{a}$ and let*

$$\sup_{\boldsymbol{a}\in\Theta}\left\|\boldsymbol{V}(\boldsymbol{a})^{-1}\right\|_F<\infty$$

We have an additional assumption on the estimator $\widehat{\boldsymbol{\Gamma}}_{k,n}$ (see assumption 3.25).

Under the Alternative Hypothesis

Under the alternative we repeat the assumptions 3.2.3 for the piecewise stationary process, as in [Rec19, Assumptions B.2.7, B.2.8]. [Rec19, Assumption B.2.9] defines $\boldsymbol{V}_j$ analogously to 3.1, which is the expectation of the first derivative under regime j.

We place the following assumption on convex mixtures of first derivative matrices for neighbouring regimes. This ensures any Wald statistic is regular and finite in size.

Assumption 3.23. *[Rec19, Assumption B.2.10] Let $\delta\boldsymbol{V}_j(\boldsymbol{a})+(1-\delta)\boldsymbol{V}_{j+1}(\boldsymbol{a})$ be a regular matrix for all $\boldsymbol{a}\in\Theta$ and all $\delta\in[0,1]$ and let*

$$\sup_{\delta\in[0,1]}\sup_{\boldsymbol{a}\in\Theta}\left\|\left(\delta\boldsymbol{V}_j(\boldsymbol{a})+(1-\delta)\boldsymbol{V}_{j+1}(\boldsymbol{a})\right)^{-1}\right\|_F<\infty,\quad j=1,\ldots,q$$

We assume the $2+\nu$-th moment of the gradient is bounded across regimes.

Assumption 3.24. *[Rec19, Assumption B.2.11] There exists $\nu > 0$ such that*

$$E\left(\sup_{\boldsymbol{a}\in\Theta}\|\nabla \boldsymbol{H}\left(\boldsymbol{Y}_t, \mathcal{X}_{t-1}^{(j)}, \tilde{\boldsymbol{a}}\right)\|^{2+\nu}\right) < \infty$$

for $j = 1, \ldots, q+1$

3.2.4 Assumptions for the Wald-type Statistic: The Autoregressive Model

For the Wald-type statistic, we have assumptions on the observed and estimating function processes, modified from [Rec19, p.105]. We verify that these hold for the VAR(p) model further into the chapter.

VAR Under the Null Hypothesis

(R1) The sequence $\{\boldsymbol{Y}_t\}_{t\geq 1}$ is stationary and ergodic with $E\left(\|\boldsymbol{Y}_1\|\right) < \infty, E\left(\boldsymbol{Y}_1\right) = \boldsymbol{\mu} \in \mathbb{R}^d$

(R2) Let $\mathcal{F}_t = \sigma\left(\boldsymbol{Y}_j, \varepsilon_{j-1}(i), j \leq t, i = 1, \ldots, d\right)$. We assume that $\boldsymbol{\varepsilon}_t$ and $\mathcal{F}_t$ are independent.

(R3) $\boldsymbol{C} := E\left(\mathcal{X}_1\mathcal{X}_1^T\right) \in \mathbb{R}^{d(dp+1)-I\times d(dp+1)-I}$ (thus $\boldsymbol{C}(i) := E\left(\mathbb{X}_1(i)\mathbb{X}_1(i)^T\right)$) is a positive definite matrix.

(R4) The sequence $\{\boldsymbol{\varepsilon}_t\}_{t\geq 1}$ is i.i.d. with $E\left(\boldsymbol{\varepsilon}_1\right) = \boldsymbol{0}$ and

$E\left(\boldsymbol{\varepsilon}_1\boldsymbol{\varepsilon}_1^T\right) := diag(\sigma^2(i), i = 1, \ldots, d) = \mathcal{S}$, where $0 < \sigma^2(i) < \infty$ for each i

(R5) Let the components of $\left\{\mathcal{X}_t\mathcal{X}_t^T - \boldsymbol{C}\right\}_{t\geq 1}$ (thus components of block-diagonal matrices $\left\{\mathbb{X}_t(i)\mathbb{X}_t^T(i) - \boldsymbol{C}(i)\right\}_{t\geq 1}$) satisfy a strong invariance principle similar to that in assumption 3.3.

(R6) For each $i = 1, \ldots, d$ let $\{\mathbb{X}_{t-1}(i)\varepsilon_t(i)\}_{t\geq 1}$ be a series with positive definite long-run covariance matrix $\boldsymbol{\Sigma}(i)$ satisfying a strong invariance principle similar to that in assumption 3.3.

Note that if we stack $\{\mathbb{X}_{t-1}(i)\varepsilon_t(i), i = 1, \ldots, d\}$ into a vector according to the Block-Kronecker product $\boldsymbol{\varepsilon}_t * \mathcal{X}_t$, (R6) tells us the covariance of this vector is the blockwise-diagonal matrix

$$\begin{aligned}\boldsymbol{\Sigma} &= \mathcal{S} * \boldsymbol{C} \\ &= diag(\sigma^2(i)E(\mathbb{X}_1(i)\mathbb{X}_1(i)^T), i = 1, \ldots, d) \\ &= diag(\boldsymbol{\Sigma}(i), i = 1, \ldots, d)\end{aligned} \tag{3.25}$$

VAR Under the Alternative Hypothesis

Under the alternative hypothesis, we modify these as in [Rec19, p.111] for

(R1*) The sequence $\left\{\boldsymbol{Y}_t^{(j)}\right\}_{t\geq 1}$ is stationary and ergodic with $E\left(\left\|\boldsymbol{Y}_1^{(j)}\right\|\right) < \infty, E\left(\boldsymbol{Y}_1^{(j)}\right) = \boldsymbol{\mu}^{(j)} \in \mathbb{R}^d$ for $j = 1, \ldots, q+1$

(R2*) Let $\mathcal{F}_t = \sigma\left(\boldsymbol{Y}_j, \varepsilon_{j-1}(i), j \leq t, i = 1, \ldots, d\right)$ We assume that $\boldsymbol{\varepsilon}_t$ and $\mathcal{F}_t$ are independent.

(R3*) $\boldsymbol{C}_{(j)} := E\left(\mathcal{X}_t^{(j)}\mathcal{X}_t^{(j)T}\right) \in \mathbb{R}^{d(dp+1)-I \times d(dp+1)-I}$ (thus $\boldsymbol{C}^{(j)}(i) := E\left(\mathbb{X}_1^{(j)}(i)\mathbb{X}_1(i)^{(j)T}\right)$) is a positive definite matrix, for $j = 1, \ldots, q+1$

(R4*) The sequence $\left\{\boldsymbol{\varepsilon}_t^{(j)}\right\}_{t\geq 1}$ is i.i.d. with $E\left(\boldsymbol{\varepsilon}_t^{(j)}\right) = \mathbf{0}$

and $E\left(\boldsymbol{\varepsilon}_t^{(j)}\boldsymbol{\varepsilon}_t^{(j)T}\right) := diag(\sigma_j^2(i), i = 1, \ldots, d) = \mathcal{S}_{(j)}$, where $0 < \sigma^2(i) < \infty$ for each i.

(R5*) Let the components of $\left\{\mathcal{X}_t^{(j)}\mathcal{X}_t^{(j)T} - \boldsymbol{C}_{(j)}\right\}_{t\geq 1}$ (thus components of block-diagonal matrices $\left\{\mathbb{X}_t^{(j)}(i)\mathbb{X}_t^{(j)T}(i) - \boldsymbol{C}^{(j)}(i)\right\}_{t\geq 1}$) satisfy a strong invariance principle similar to that in assumption 3.3 for $j = 1, \ldots, q+1$

(R6*) For each channel $i = 1, \ldots, d$ and regime $j = 1, \ldots, q+1$, let $\left\{\mathbb{X}_{t-1}^{(j)}(i)\varepsilon_t^T(i)\right\}_{i\geq 1}$ be a series with positive definite long-run covariance matrix $\boldsymbol{\Sigma}_{(j)}(i)$ satisfying a strong invariance principle similar to that in assumption 3.3.

(R7*) Let the matrix $\delta\boldsymbol{C}_{(j)} + (1-\delta)\boldsymbol{C}_{(j+1)}$ be positive definite for all $\delta \in [0,1]$ and assume that $\sup_{\delta\in[0,1]} \left\|\left(\delta\boldsymbol{C}_{(j)} + (1-\delta)\boldsymbol{C}_{(j+1)}\right)^{-1}\right\|_F < \infty$, for all $j = 1, \ldots, q$. Here, $\|.\|_F$ is the Frobenius norm.

We have a similar result to (3.25) for the within-regime covariance $\boldsymbol{\Sigma}_{(j)}$.

$$\begin{aligned}\boldsymbol{\Sigma}_{(j)} &= \mathcal{S}_{(j)} * \boldsymbol{C}_{(j)} \\ &= diag(\sigma_{(j)}^2(i)E(\mathbb{X}_{t-1}(i)\mathbb{X}_{t-1}(i)^T), i = 1, \ldots, d) \\ &= diag(\boldsymbol{\Sigma}_{(j)}(i), i = 1, \ldots, d)\end{aligned} \tag{3.26}$$

3.3 The MOSUM Score-Type Statistic

In this section we give key results for outlining a testing and estimation procedure, based on the Score-type statistic.

Recall the definition given in 3.13:

$$T_n(G, \tilde{\boldsymbol{a}}) = \max_{G \le k \le n-G} T_{k,n}(G, \tilde{\boldsymbol{a}}), \qquad T_{k,n}(G, \tilde{\boldsymbol{a}}) = \frac{1}{\sqrt{2G}} \left\| \boldsymbol{\Sigma}_k^{-1/2} \boldsymbol{A}_{\tilde{\boldsymbol{a}},k} \right\| \tag{3.27}$$

We are interested in testing the following hypotheses:

$H_0 : q = 0$, i.e. no changes occur

against

$H_1 : q \ge 1$, i.e. at least one change occurs

In practice, we use estimators $\hat{\boldsymbol{a}}_{1,n}$ (the global least-squares estimator 3.39) for the parameters and $\widehat{\boldsymbol{\Sigma}}_{n,k}$ (see section 3.4.3) for the long-run covariance. Evaluated with estimators, the test statistic is

$$\widehat{T}_{k,n}(G, \hat{\boldsymbol{a}}_{1,n}) = \frac{1}{\sqrt{2G}} \left\| \widehat{\boldsymbol{\Sigma}}_k^{-1/2} \boldsymbol{A}_{\hat{\boldsymbol{a}}_{1,n}} \right\| \tag{3.28}$$

According to [Rec19, Thm 2.1.1], under assumptions 3.1 on the bandwidth G, 3.2 3.3 and 3.4 on $\hat{\boldsymbol{a}}_n$, and 3.5 on $\widehat{\boldsymbol{\Sigma}}_{k,n}$ (we allow an estimator to be used in place of the true $\boldsymbol{\Sigma}$) we have under H_0 the **Limit Distribution**

$$a(n/G) T_n(G, \hat{\boldsymbol{a}}_n) - b(n/G) \xrightarrow{\mathcal{D}} G_2 \tag{3.29}$$

where G_2 is distributed according to a Gumbel(2) distribution, i.e. $P(G_2 \le x) = \exp(-2\exp(-x))$ and

$$\begin{aligned} a(x) &= \sqrt{2\log(x)} \\ b(x) &= 2\log(x) + \frac{d}{2}\log(\log(x)) - \log\left(\frac{2}{3}\Gamma\left(\frac{d(dp+1)}{2}\right)\right) \end{aligned} \tag{3.30}$$

where $d(dp+1)$ is the dimension of the parameter space (using the model specified without dimension reduction), as determined by the dimension d of $\boldsymbol{Y}_t$ and the order p of the model, and Γ denotes the Gamma function.

Assumptions A.1.3, A.1.4 are shown to hold for a strongly-mixing sequence such as ours in [Rec19, Section 2.3].

We can immediately see that under H_0

$$P\left(a(n/G) T_n(G, \hat{\boldsymbol{a}}_n) - b(n/G) > c_\alpha\right) \to \alpha, \tag{3.31}$$

with the **Critical Value**

$$c_\alpha := -\log\log\frac{1}{\sqrt{1-\alpha}}$$

which is the $(1-\alpha)$ quantile of the Gumbel(2) distribution.

This leads to a **Testing Procedure** with asymptotic level-α

$$\begin{aligned} &\text{Reject } H_0 \text{ if } T_n(G, \hat{\boldsymbol{a}}_n) > D_n(G, \alpha) \\ &\text{with } D_n(G, \alpha) = \frac{b(n/G) + c_\alpha}{a(n/G)} \end{aligned} \tag{3.32}$$

As per [Rec19, Thm 2.1.5] this test has asymptotic power 1. Under the alternative, as well as assumptions 3.1 on the bandwidth, 3.6 on the scaled changes, 3.7 on the piecewise stationary sequence, 3.8 on strong invariance and 3.11 on identifiability of changes, we get the following result. We use estimators $\hat{\boldsymbol{a}}_n$ meeting assumption 3.9 and $\widehat{\boldsymbol{\Sigma}}_{n,k}$ meeting assumption 3.10.

$$\forall z \in \mathbb{R}, \lim_{n\to\infty} P\left(a(n/G)T_n(G, \hat{\boldsymbol{a}}_n) - b(n/G) \geq z\right) = 1 \tag{3.33}$$

Given multiple changes have occurred, we would naturally want to estimate the number that have occurred and to find the location of the changes. To do so, we confer the **MOSUM Procedure** of [EK+18]:

Consider all pairs of time points $(v_{j,n}, w_{j,n})$ with

$$\begin{aligned} &T_{k,n}(G, \hat{\boldsymbol{a}}_n) \geq D_n\left(\alpha_n, G\right) \text{ for } v_{j,n} \leq k \leq w_{j,n} \\ &T_{k,n}(G, \hat{\boldsymbol{a}}_n) < D_n\left(\alpha_n, G\right) \text{ for } k = v_{j,n} - 1, w_{j,n} + 1 \end{aligned} \tag{3.34}$$

where $w_{j,n} - v_{j,n} \geq \epsilon G$ with $0 < \epsilon < 1/2$ arbitrary but fixed. This condition prevents spurious over-estimation of the number of changes, given that close to true changes, it is likely that noise will cause additional spikes in the test statistic.

We take the number of these pairs as an estimator for the number of changes:

$$\widehat{q}_n = \widehat{q}_n(\hat{\boldsymbol{a}}_n) \widehat{=} \text{ number of pairs } (v_{j,n}, w_{j,n}) \tag{3.35}$$

Furthermore, we determine the local maxima between $v_{j,n}$ and $w_{j,n}, j = 1, \ldots, \widehat{q}_n$, and use them as estimators for the locations of the change points:

$$\widehat{k}_{j,n} = \widehat{k}_{j,n}(\hat{\boldsymbol{a}}_n) := \underset{v, i<k<w_i,}{\arg\max} T_{k,n}(G, \hat{\boldsymbol{a}}_n) \tag{3.36}$$

Accordingly with [Rec19, Thm 2.1.8], this procedure consistently estimates the number of changes. Let the Assumptions 3.1, 3.6, 3.7, 3.8 and 3.12 hold for some $\tilde{\boldsymbol{a}}$. Furthermore, assume that the sequence $\{\alpha_n\}_{n\in N}$ fulfills Assumption 3.13. Let the estimator sequence $\hat{\boldsymbol{a}}_n$ meet assumptions 3.9 and 3.14, and the sequence $\widehat{\boldsymbol{\Sigma}}_{n,k}$ meet assumption 3.10.

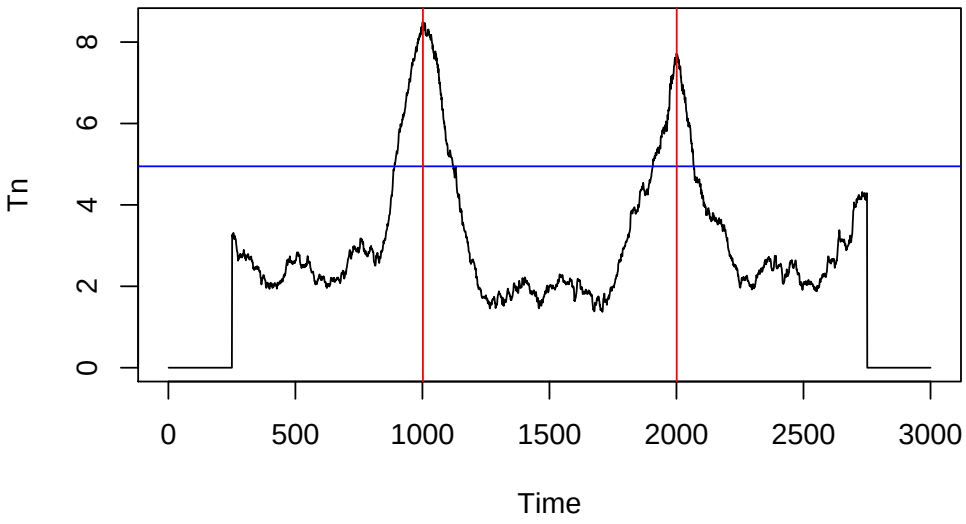


Figure 3.1: Two Changes in Autoregression

Then, for any $\tilde{\boldsymbol{a}}$ with corresponding $\tilde{q} = \tilde{q}(\tilde{\boldsymbol{a}})$

$$P\left(\widehat{q}_n(\hat{\boldsymbol{a}}_n) = \tilde{q}\right) \to 1 \quad \text{as } n \to \infty \tag{3.37}$$

Assumption 3.14 is shown to hold for a strongly-mixing sequence such as ours in [Rec19, Section 2.3].

The procedure is also weakly consistent for the locations of changes. Under the same assumptions as the prequel, with $\tilde{Q}$ the set of detectable change points as in 3.12, we have

$$P\left(\max_{j\in\tilde{Q}} \min_{1\le l\le \hat{q}_n} \left|\widehat{k}_{l,n} - k_{j,n}\right| < G\right) \to 1 \tag{3.38}$$

That is, the probability of every detectable change having at least one estimator within distance G tends to 1.

For improved results on convergence rates under stronger assumptions, confer [Rec19, Section 2.2].

3.3.1 Verifying Assumptions

For the the score-type procedure to be valid, we need to verify the assumptions of the key results (3.29) and (3.37).

Assumptions 3.4 and 3.3 underpin (3.29). We follow the arguments of [Rec19, p.122]. Under the null hypothesis, we use the global least squares estimator $\hat{\boldsymbol{a}}_{1,n}$ calculated over the whole sample, and $\tilde{\boldsymbol{a}} = \boldsymbol{a}_0$ is the true parameter.

We have that

$$-\boldsymbol{H}_i(Y_t(i), \mathbb{X}_{t-1}(i), \boldsymbol{a}_0(i)) = \mathbb{X}_{t-1}(i)\varepsilon_t(i)$$

so (R6) implies 3.3.

$\hat{\boldsymbol{a}}_{1,n}$ is $\sqrt{n}$-consistent for $\boldsymbol{a}_0$ with stochastic regressors, as per the argument in [Lüt05, Section 3.2.2], as we have a stable process with standard white noise residuals, where the residual components are independent. Combining this with (R5) and 3.1, we have assumption 3.4.

Assumptions 3.8, 3.9 and 3.14 underpin (3.37). Assumptions 3.8, 3.9 are also shown to hold for a strongly-mixing sequence with deterministic regressors in [Rec19, Section 2.3]; we adapt this to the case with stochastic regressors by again using the argument in [Lüt05, Section 3.2.2], combined with the piecewise-stability of the sequence under the alternative, so $\hat{\boldsymbol{a}}_{1,n}$ is $\sqrt{n}$-consistent for $\sum_{j=1}^{q+1}(\lambda_j - \lambda_{j-1})\boldsymbol{a}_j$.

All assumptions 3.15-3.22, 3.2.2-3.24 on moments are demonstrated to hold for a strongly-mixing sequence on a compact parameter space. The standard parameter space $\mathbb{R}^d$ is not compact; however given that the model is stable (3.7), we have restricted to a closed and bounded parameter space $\Theta \subset \mathbb{R}^d$.

3.3.2 Estimators

In estimating the parameter vector $\tilde{\boldsymbol{a}}$ we use the global least-squares solution

$$\hat{\boldsymbol{a}}_{1,n} = \begin{pmatrix} \hat{\boldsymbol{a}}_{1,n}(1) \\ \hat{\boldsymbol{a}}_{1,n}(2) \\ \dots \\ \hat{\boldsymbol{a}}_{1,n}(d) \end{pmatrix}, \quad \hat{\boldsymbol{a}}_{1,n}(i) = \left(\sum_{t=1}^{n} Y_t(i)\mathbb{X}_{t-1}^T(i)\right)\left(\sum_{t=1}^{n} \mathbb{X}_{t-1}(i)\mathbb{X}_{t-1}^T(i)\right)^{-1} \quad i = 1, \dots, d \tag{3.39}$$

To evaluate the Score statistic, we require an estimator $\widehat{\boldsymbol{\Sigma}}_{n,k}$ for the covariance $\boldsymbol{\Sigma}$ which is consistent under both the null hypothesis 3.25 and the alternative 3.26. We propose three possible options.

Diagonal-C:

$$\widehat{\boldsymbol{\Sigma}}_{n,k} = diag\Big[\hat{\sigma}^2_{n,k}(i)\widehat{\boldsymbol{C}}_{k-G+1,k+G}(i), i = 1, \dots, d\Big] \tag{3.40}$$

This is a MOSUM estimator with block-diagonal entries calculated as the product of the channel-specific variance estimator and the empirical covariance of $\mathbb{X}_{t-1}(i)$. Under the diagonal assumption (R4)/(R4*), this is consistent under both hypotheses.

Denote the estimator of the (uncentred) covariance of the i-th block matrix as

$$\widehat{\boldsymbol{C}}_{l,u}(i) = \frac{1}{u-l}\sum_{t=l}^{u} \mathbb{X}_{t-1}(i)\mathbb{X}^T_{t-1}(i) \tag{3.41}$$

Note that by the assumption 3.1 on G and (R5)/(R5*), each estimator is consistent for $\boldsymbol{C}(i)$ under H_0 and $\boldsymbol{C}_{(j)}(i)$ under H_1. We have to localise the estimator, since our regressors are dependent in the strongly-mixing sense. This is in contrast to [Rec19, p.121], which uses I.I.D. regressors.

We use the following estimator for the channel-specific variance $\sigma^2(i)$

GLOBAL:

$$\hat{\sigma}^2_{n,k}(i) = \frac{1}{n-p-2}\sum_{t=p+1}^{n}\big(Y_t(i) - \hat{\boldsymbol{a}}^T_{1,n}(i)\mathbb{X}_{t-1}(i)\big)^2 \tag{3.42}$$

Using a global estimator $\hat{\boldsymbol{a}}^T_{1,n}(i)$ for the parameters, we compute residuals over the whole data.

Using well-known results, the estimator $\hat{\sigma}^2_{n,k}(i)$ from 3.42 consistently estimates $\sigma^2_{n,k}(i)$ under H_0. Under H_1, this will overestimate the error variance.

Full-H:

$$\begin{aligned}\widehat{\boldsymbol{\Sigma}}_{n,k} = \frac{1}{2G}\Big[\Big(&\sum_{t=k-G+1}^{k}\big(\boldsymbol{H}\left(\boldsymbol{Y}_t, \mathcal{X}_{t-1}, \hat{\boldsymbol{a}}_{1,n}\right) - \bar{\boldsymbol{H}}_{k-G+1,k}\big)\boldsymbol{H}\left(\boldsymbol{Y}_t, \mathcal{X}_{t-1}, \hat{\boldsymbol{a}}_{1,n}\right) - \bar{\boldsymbol{H}}_{k-G+1,k}\big)^T + \\ &\sum_{t=k+1}^{k+G}\big(\boldsymbol{H}\left(\boldsymbol{Y}_t, \mathcal{X}_{t-1}, \hat{\boldsymbol{a}}_{1,n}\right) - \bar{\boldsymbol{H}}_{k+1,k+G}\big)\big(\boldsymbol{H}\left(\boldsymbol{Y}_t, \mathcal{X}_{t-1}, \hat{\boldsymbol{a}}_{1,n}\right) - \bar{\boldsymbol{H}}_{k+1,k+G}\big)^T\Big)\Big]\end{aligned} \tag{3.43}$$

This is a MOSUM estimator calculated as the empirical covariance of $\boldsymbol{H}\left(\boldsymbol{Y}_t, \mathcal{X}_{t-1}, \hat{\boldsymbol{a}}_{1,n}\right)$, the estimating function evaluated with the global parameter estimate $\hat{\boldsymbol{a}}_{1,n}$.Under the assumptions (R6) and (R6*), this is consistent under both hypotheses.

This estimator poses practical issues, in that we are required to find the negative square root (as per 3.28), which is highly expensive for a dense matrix of dimension $d(dp+1) - I \times$

$d(dp+1)-I$, meaning evaluating this introduces numerical error for large values of p and d. For some solution to this, see 5.2.

Here, $\bar{\boldsymbol{H}}_{l,u}$ is the sample mean of $\boldsymbol{H}\left(\boldsymbol{Y}_t, \mathcal{X}_{t-1}, \hat{\boldsymbol{a}}_{1,n}\right)$ for $t = l, \ldots, u$.

Diagonal-H:

$$\widehat{\boldsymbol{\Sigma}}_{n,k} = \frac{1}{2G} diag\Big[\sum_{t=k-G+1}^{k} \left(\boldsymbol{H}_i\left(Y_t(i), \mathbb{X}_{t-1}(i), \hat{\boldsymbol{a}}_{1,n}(i)\right) - \bar{\boldsymbol{H}}_{k-G+1,k}(i)\right)\left(\boldsymbol{H}_i\left(Y_t(i), \mathbb{X}_{t-1}(i), \hat{\boldsymbol{a}}_{1,n}(i)\right) - \bar{\boldsymbol{H}}_{k-G+1,k}(i)\right)^T + \sum_{t=k+1}^{k+G} \left(\boldsymbol{H}\left(Y_t(i), \mathbb{X}_{t-1}(i), \hat{\boldsymbol{a}}_{1,n}(i)\right) - \bar{\boldsymbol{H}}_{k+1,k+G}(i)\right)\left(\boldsymbol{H}_i\left(Y_t(i), \mathbb{X}_{t-1}(i), \hat{\boldsymbol{a}}_{1,n}(i)\right) - \bar{\boldsymbol{H}}_{k+1,k+G}(i)\right)^T, \quad i = 1, \ldots, d\Big] \tag{3.44}$$

This adapts 3.43 into a blockwise-diagonal matrix, introducing sparsity and thus reducing the complexity of calculating the negative square root matrix, along with the possibility of numerical error. Under the diagonal assumption (R4)/(R4*), this is still consistent.

Here, $\bar{\boldsymbol{H}}_{l,u}(i)$ is the sample mean of $\boldsymbol{H}_i\left(Y_t(i), \mathbb{X}_{t-1}(i), \hat{\boldsymbol{a}}_{1,n}(i)\right)$ for $t = l, \ldots, u$.

3.4 The MOSUM Wald-type Statistic

In this section we propose an alternative test statistic based on estimating functions, with key results for the testing and estimation procedure. Estimators for each relevant quantity are given in subsection 3.4.3.

Recall the Wald statistic, as defined in 3.17:

$$\begin{aligned} W_n(G) &= \max_{G \le k \le n-G} W_{k,n}(G) \\ W_{k,n}(G) &= \sqrt{\frac{G}{2}} \sqrt{\left(\tilde{\boldsymbol{a}}_{k+1,k+G} - \tilde{\boldsymbol{a}}_{k-G+1,k}\right)^T \boldsymbol{\Gamma}_k^{-1} \left(\tilde{\boldsymbol{a}}_{k+1,k+G} - \tilde{\boldsymbol{a}}_{k-G+1,k}\right)} \\ &= \sqrt{\frac{G}{2}} \left\| \boldsymbol{\Gamma}_k^{-1/2} \left(\tilde{\boldsymbol{a}}_{k+1,k+G} - \tilde{\boldsymbol{a}}_{k-G+1,k}\right) \right\| \end{aligned} \tag{3.45}$$

We find the following results on the derivatives of the estimating function:

Lemma 3.1. *a) The expectation of the first derivative of the estimating function is*

$$\boldsymbol{V}(\tilde{\boldsymbol{a}}) = E\left(\nabla \boldsymbol{H}\left(\boldsymbol{Y}_t, \mathcal{X}_{t-1}, \tilde{\boldsymbol{a}}\right)\right)$$

$$= E(diag(\mathbb{X}_{t-1}(i)\mathbb{X}_{t-1}(i)^T, i = 1, \dots, d)) = \boldsymbol{C}(i) \in \mathbb{R}^{d(dp+1)-I \times d(dp+1)-I}$$

b) $\boldsymbol{V}(\tilde{\boldsymbol{a}})$ *is constant with respect to* $\tilde{\boldsymbol{a}}$ *and regular.*

c) The expectation of the second derivative (Hessian tensor) is zero:

$$E\left(\nabla^2 \boldsymbol{H}\left(\boldsymbol{Y}_t, \mathcal{X}_{t-1}, \tilde{\boldsymbol{a}}\right)\right) = \boldsymbol{0} \in \mathbb{R}^{d(dp+1)-I \times d(dp+1)-I \times d(dp+1)-I}$$

Proof. a) Omitted

b) $\boldsymbol{V}(\tilde{\boldsymbol{a}})$ is blockwise diagonal, consisting of uncentred covariance matrix blocks $E(\mathbb{X}_{t-1}(i)\mathbb{X}_{t-1}(i)^T)$ which are positive definite by assumption (R3), hence $\boldsymbol{V}(\tilde{\boldsymbol{a}})$ is regular.

c) This follows from (b) □

Note that in the particular case where $\mathbb{X}_{t-1}(i) = \mathbb{X}_{t-1}$ for all channels i, i.e. when we incorporate no prior information into the regression design, this matrix simplifies to

$$\boldsymbol{V}(\tilde{\boldsymbol{a}}) = diag(E(\mathbb{X}_{t-1}\mathbb{X}_{t-1}^T), i = 1, \dots, d) \in \mathbb{R}^{d(dp+1) \times d(dp+1)}$$

In the general case, denoting $\boldsymbol{V}_k = \boldsymbol{V}(\tilde{\boldsymbol{a}}_{k-G+1,k})$ as the expectation calculated on the window before time k, we find

$$\boldsymbol{\Gamma}_k = \boldsymbol{V}_k^{-1}\boldsymbol{\Sigma}_k\boldsymbol{V}_k^{-T} \tag{3.46}$$

We can then express the statistic as

$$W_{k,n}(G) = \sqrt{\frac{G}{2}}\|\boldsymbol{\Sigma}_k^{-1/2}\boldsymbol{V}_k\left(\tilde{\boldsymbol{a}}_{k+1,k+G} - \tilde{\boldsymbol{a}}_{k-G+1,k}\right)\| \tag{3.47}$$

3.4.1 Results and Procedure

We have a **Limiting Distribution** as per [Rec19, Thm 3.1.8]. Assume a strongly mixing series satisfying assumptions 3.19-**??**, with bandwidth $G(n)$ satisfying 3.1. We use estimators $\widehat{\boldsymbol{\Sigma}}_{n,k}$ and $\widehat{\boldsymbol{C}}_{n,k}$.

Then, under H_0

$$a(n/G)W_n(G) - b(n/G) \xrightarrow{D} G_2 \tag{3.48}$$

where G_2 is a Gumbel(2) random variable and with $a(x)$ and $b(x)$ as in 3.29.

This leads to a **Testing Procedure** with asymptotic level α:

Reject H_0 if

$$W_n(G) > D_n(G, \alpha) \text{ with } D_n(G, \alpha) = \frac{b(n/G) + c_\alpha}{a(n/G)} \tag{3.49}$$

with the **Critical Value**

$$c_\alpha := -\log\log\frac{1}{\sqrt{1-\alpha}}$$

which is the $(1-\alpha)$ quantile of the Gumbel(2) distribution.

Following [Rec19, Thm 3.1.12], this procedure has asymptotic power 1. Suppose we have a strongly-mixing, piecewise-stationary series meeting assumptions 3.2.2, 3.2.3, 3.2.3 and 3.2.3 and bandwidth meeting assumption 3.1, as well as 3.6 on the scaled changes.

Then, under H_1, we obtain for any $z \in \mathbb{R}$

$$\lim_{n\to\infty} P\left(a(n/G)W_n(G) - b(n/G) \geq z\right) = 1 \tag{3.50}$$

This holds with a consistent estimator $\widehat{\mathbf{\Gamma}}_{k,n}$ meeting the following assumption

Assumption 3.25. *[Rec19, Thm 3.1.12] (I)* $\max_{k\in B_{n,G}} \left\|\widehat{\Gamma}_{k,n}^{-1/2} - \Gamma_{A,k}^{-1/2}\right\|_F = o_P(1)$, *where* $\{\Gamma_{A,k}\}_{k\geq 1}$ *is a sequence of positive definite matrices fulfilling* $\sup_k \|\Gamma_{A,k}\|_F < \infty$ *and* $\sup_k \left\|\Gamma_{A,k}^{-1/2}\right\|_F < \infty$

(II) $\max_{k\in A_{n,G}} \left\|\widehat{\mathbf{\Gamma}}_{k,n}^{-1/2} - \mathbf{\Gamma}_k^{-1/2}\right\|_F = o_P\left(\log(n/G)^{-1}\right)$

We have an **Estimation Procedure** for the number and locations of changes, analogous to (3.34).

We consider all pairs of time points $(v_{j,n}, w_{j,n})$ with

$$\begin{aligned} &W_{k,n}(G) \geq D_n(\alpha_n, G) \text{ for } v_{j,n} \leq k \leq w_{j,n} \\ &W_{k,n}(G) < D_n(\alpha_n, G) \text{ for } k = v_{j,n} - 1, w_{j,n} + 1 \\ &w_{j,n} - v_{j,n} \geq \epsilon G \quad \text{with } 0 < \epsilon < 1/2 \text{ fixed.} \end{aligned} \tag{3.51}$$

The estimator for the number of changes $\widehat{q}_n$ is given by the number of pairs and we take the maximal points of these exceeding intervals $[v_{j,n}, w_{j,n}]$ as estimators for the location of the change points:

$$\widehat{k}_{j,n} := \underset{v_{j,n}\leq k\leq w_{j,n}}{\arg\max}\ W_{k,n}(G) \tag{3.52}$$

The estimator for the number of changes is consistent, by [Rec19, Thm 3.1.15]. Let assumptions 3.1 and 3.6 hold, and suppose we have a strongly mixing sequence satisfying assumptions 3.2.3, 3.2.3, 3.23 and 3.23. Allow the significance sequence to fulfill assumption 3.13.

$$P\left(\widehat{q}_n = q\right) \to 1, \quad \text{as } n \to \infty \tag{3.53}$$

The above holds when using an estimator $\widehat{\mathbf{\Gamma}}_{k,n}$ meeting assumption 3.25.

Under the same conditions, we have a weak consistency in location from [Rec19, Cor. 3.1.16]

$$P\left(\max_{1\leq j\leq q, 1\leq l\leq q_n}\left|\widehat{k}_{l,n}-k_{j,n}\right|<G\right)\to 1 \tag{3.54}$$

3.4.2 Verifying Assumptions

Moment assumptions 3.21, 3.22 were shown to hold in 3.3.1.

We verify the assumptions of the Wald statistic conferred from [Rec19, p.105], generalised to d dimensions (see 7)

- Due to the stationarity of $\{\boldsymbol{Y}_t : t\geq 1\}$ and $\left\{\boldsymbol{Y}_t^{(j)} : t\geq 1\right\}$ for all regimes j, we have the equivalences
$$(R1) \iff (R1*),\ldots,(R6) \iff (R6*)$$
- (R7*) simplifies to (R3). Given the fact that $\boldsymbol{C}_{(j)}$ is positive definite iff $\boldsymbol{z}^T\boldsymbol{C}_{(j)}\boldsymbol{z}>0$ for all conformable vectors $\boldsymbol{z}\in\mathbb{R}^{d(dp+1)-I}$, it follows that $\boldsymbol{z}^T\left(\delta\boldsymbol{C}_{(j)}+(1-\delta)\boldsymbol{C}_{(j+1)}\right)\boldsymbol{z}>0$
- (R1), (R2) and (R4) follow by definition of $\{\boldsymbol{Y}_t : t\geq 1\}$
- The sequences $\{\mathbb{X}_{t-1}(i)\varepsilon_t(i) : t\geq 1\}$ are strongly mixing with all moments existing, for all channels i. [E+89, Theorem 2] implies (R6).
- The components of the sequence $\left\{\mathcal{X}_t\mathcal{X}_t^T-\boldsymbol{C} : t\geq 1\right\}$ have expectation 0 and all moments finite, so (R5) follows from the invariance principle as in [Rec19, Section 4.1.1].
- For (R3), the matrix $\boldsymbol{C}=E(\mathcal{X}_1\mathcal{X}_1^T)$ is positive definite with probability 1

3.4.3 Estimators

In this subsection, we propose estimators which meet the required consistency conditions.

In finding an estimator for $\boldsymbol{\Gamma}$, we observe [Rec19, Remark 3.1.13.]. The assumption on the estimator sequence of the long-run covariance matrix in 3.25 is fulfilled if

$$\widehat{\boldsymbol{\Gamma}}_{k,n}=\widehat{\boldsymbol{V}}_{k,n}^{-1}\widehat{\boldsymbol{\Sigma}}_{k,n}\left(\widehat{\boldsymbol{V}}_{k,n}^{-1}\right)^T,$$

where $\{\widehat{\boldsymbol{\Sigma}}_{k,n}\}$ is a positive definite estimator sequence satisfying assumption 3.10, as shown above,

and $\{\hat{\boldsymbol{V}}_{k,n}\}$ is a regular estimator sequence satisfying:

$$\max_{k\in B_{n,a}} \left\| \hat{\boldsymbol{V}}_{k,n} - \boldsymbol{V}_{A,k} \right\|_F = o_P(1),$$

with $\{\boldsymbol{V}_{A,k}\}_{k\geq 1}$ denoting a sequence of regular matrices fulfilling $\sup_k \left\|\boldsymbol{V}_{A,k}^{-1}\right\|_F < \infty$ and $\sup_k \|\boldsymbol{V}_{A,k}\|_F < \infty$ with A as in 3.22.

Thus we estimate each diagonal block as $\widehat{\boldsymbol{V}}_{k,n}(i) = \widehat{\boldsymbol{C}}_{k-G+1,k}(i)$ 3.41.

In the case without prior information on the dependence of each channel, we can evaluate the test statistic at time k with

$$\widehat{W}_{k,n}(G) = \sqrt{\frac{G}{2}} \|\widehat{\boldsymbol{\Gamma}}_k^{-1/2}\big(\hat{\boldsymbol{a}}_{k+1,k+G} - \hat{\boldsymbol{a}}_{k-G+1,k}\big)\| = \sqrt{\frac{G}{2}} \|\widehat{\boldsymbol{\Sigma}}_k^{-1/2}\widehat{\boldsymbol{V}}_k\big(\hat{\boldsymbol{a}}_{k+1,k+G} - \hat{\boldsymbol{a}}_{k-G+1,k}\big)\| \tag{3.55}$$

with $\hat{\boldsymbol{a}}_{l,u}(i) = \tilde{\boldsymbol{a}}_{l,u}(i)$ as in (3.18).

There are many possible estimators to choose from when evaluating the test statistic; a select few are discussed here.

We can directly estimate $\widehat{\boldsymbol{\Sigma}}_k$ using **Full-H** (3.43) and **Diagonal-H** (3.44) as we did for the Score statistic. These should differ from those used before by using a local estimator $\hat{\boldsymbol{a}}_{l,u}$ for the regression parameters, giving a fully MOSUM-type estimator.

These are:

Diagonal-H:

$$\begin{aligned}
&\widehat{\boldsymbol{\Sigma}}_{n,k} = \frac{1}{2G} diag\Big[\\
&\sum_{t=k-G+1}^{k} \big(\boldsymbol{H}_i\left(Y_t(i), \mathbb{X}_{t-1}(i), \hat{\boldsymbol{a}}_{k-G+1,k}(i)\right) - \bar{\boldsymbol{H}}_{k-G+1,k}(i)\big)\big(\boldsymbol{H}_i\left(Y_t(i), \mathbb{X}_{t-1}(i), \hat{\boldsymbol{a}}_{k-G+1,k}(i)\right) - \bar{\boldsymbol{H}}_{k-G+1,k}(i)\big)^T \\
&\sum_{t=k+1}^{k+G} \big(\boldsymbol{H}\left(Y_t(i), \mathbb{X}_{t-1}(i), \hat{\boldsymbol{a}}_{k+1,k+G}(i)\right) - \bar{\boldsymbol{H}}_{k+1,k+G}(i)\big)\big(\boldsymbol{H}_i\left(Y_t(i), \mathbb{X}_{t-1}(i), \hat{\boldsymbol{a}}_{k+1,k+G}(i)\right) - \bar{\boldsymbol{H}}_{k+1,k+G}(i)\big)^T, \\
&i = 1, \ldots, d\Big]
\end{aligned} \tag{3.56}$$

Full-H:

$$\begin{aligned}
\widehat{\boldsymbol{\Sigma}}_{n,k} = \frac{1}{2G}\Big[\Big(&\sum_{t=k-G+1}^{k} \big(\boldsymbol{H}\left(\boldsymbol{Y}_t, \mathcal{X}_{t-1}, \hat{\boldsymbol{a}}_{k-G+1,k}\right) - \bar{\boldsymbol{H}}_{k-G+1,k}\big)\boldsymbol{H}\left(\boldsymbol{Y}_t, \mathcal{X}_{t-1}, \hat{\boldsymbol{a}}_{k-G+1,k}\right) - \bar{\boldsymbol{H}}_{k-G+1,k}\big)^T + \\
&\sum_{t=k+1}^{k+G} \big(\boldsymbol{H}\left(\boldsymbol{Y}_t, \mathcal{X}_{t-1}, \hat{\boldsymbol{a}}_{k+1,k+G}\right) - \bar{\boldsymbol{H}}_{k+1,k+G}\big)\big(\boldsymbol{H}\left(\boldsymbol{Y}_t, \mathcal{X}_{t-1}, \hat{\boldsymbol{a}}_{k+1,k+G}\right) - \bar{\boldsymbol{H}}_{k+1,k+G}\big)^T\Big)\Big]
\end{aligned} \tag{3.57}$$

Alternatively, we can adapt **Diagonal-C** (3.40) with a local estimator for the channel error variance.

Diagonal-C:

$$\begin{aligned}\widehat{\boldsymbol{\Sigma}}_{n,k} &= diag\Big[\hat{\sigma}^2_{n,k}(i)\widehat{\boldsymbol{C}}_{k-G+1,k+G}(i), i = 1,\dots,d\Big] \\ &= \hat{\mathcal{S}}_{n,k} * \widehat{\boldsymbol{C}}_{k-G+1,k+G}\end{aligned} \tag{3.58}$$

Where

$$\widehat{\mathcal{S}}_{n,k} := diag(\hat{\sigma}^2_{n,k}(i), i = 1,\dots,d) \in \mathbb{R}^{d\times d} \tag{3.59}$$

The error covariance estimator $\hat{\mathcal{S}}_{n,k}$ for $\mathcal{S}$ generalises that discussed in [Rec19, p.124]. Given possible estimators are based on estimated residuals, these will be contaminated under the alternative hypothesis where the parameters change, motivating the use of a MOSUM-type estimator over a restricted window. The proposed choice for each channel is

LOCAL1:

$$\hat{\sigma}^2_{n,k}(i) \quad := \frac{1}{2G}\left[\sum_{t=k+1}^{k+G}\left(Y_t(i) - \hat{\boldsymbol{a}}^T_{k+1,k+G}(i)\mathbb{X}_{t-1}(i)\right)^2 \; + \sum_{t=k-G+1}^{k}\left(Y_t(i) - \hat{\boldsymbol{a}}^T_{k-G+1,k}(i)\mathbb{X}_{t-1}(i)\right)^2\right] \tag{3.60}$$

We average the covariance from residuals on the left side of the window, based on the local parameter estimator $\hat{\boldsymbol{a}}^T_{k-G+1,k}(i)$ from the same range, with corresponding residuals from the right of the window based on a parameter estimator $\hat{\boldsymbol{a}}^T_{k+1,k+G}(i)$ from the right.

Using well-known results, the estimator $\hat{\sigma}^2_{n,k}(i)$ from 3.42 consistently estimates $\sigma^2_{n,k}(i)$ under H_0, and the estimator 3.4.3 is consistent under H_1.

We can state the following consistency results in the vein of assumption 3.10:

Lemma 3.2. *(a)* $\max_{G\le k\le n-G}\left\|\widehat{\mathcal{S}}^{-1/2}_{k,n}\right\|_F = O_P(1)$

(b) $\max_{k\in A_{n,G}}\left\|\widehat{\mathcal{S}}^{-1/2}_{k,n} - \mathcal{S}^{-1/2}_{k}\right\|_F = o_P\left(\log(n/G)^{-1}\right)$ *with* $A_{n,G}$ *as in (3.22)*

(c) $\max_{k\in B_{n,G}}\left\|\widehat{\mathcal{S}}^{-1/2}_{k,n} - \mathcal{S}^{-1/2}_{A,k}\right\|_F = o_P(1)$, *with* $B_{n,G}$ *as in (3.23) and* $\{\mathcal{S}_{A,k}\}$ *is a sequence of positive definite matrices fulfilling sup* $\sup_{\boldsymbol{n}}\|\mathcal{S}_{A,\boldsymbol{k}}\|_F < \infty$

Using results from Lemma 3.2, , and the properties of the Frobenius norm, Diag-C is consistent for $\boldsymbol{\Sigma}_k$ in the same sense, verifying assumption 3.5.

3.5 Practical Considerations

In this section, we give some thought to how we should use the procedures we have described in practice. Of particular concern are finding critical values, and computing estimators.

3.5.1 Critical Values

When performing tests in a finite sample case, we should be aware of how the procedure responds to changes in dimensionality. Consider again the transformed critical value $D_n(G, \alpha)$ from 3.32 and 3.49. Denoting the dimensionality of the parameter space as $\beta = d(dp + 1)$, we have that D_n depends on β through an evaluation of $\Gamma(\beta/2)$. The Gamma function itself grows near-logarithmically, and β grows quadratically with d, so D_n is highly responsive to dimensionality.

Indeed, as demonstrated in figure 3.2, D_n can take negative values for seemingly large values of n under moderate values for β, meaning our test is invalid.

Assuming $G << n$, β fixed and we have significance level α such that $c_\alpha / a(n/G) = \mathcal{O}(1)$ it can be derived that

$$D_n(G, \alpha) = \mathcal{O}(\sqrt{2\log(n)} + \frac{c_\alpha}{\sqrt{2\log(n)}} \tag{3.61}$$

We hence propose a practical solution of using the transformed critical value

$$\widetilde{D}_n(G, \alpha) = \max\left\{ D_n(G, \alpha), \sqrt{2\log(n)} + \frac{c_\alpha}{\sqrt{2\log(n)}} \right\} \tag{3.62}$$

this allows some notion of size control through c_α, and is asymptotically equal to $D_n(G, \alpha)$, but will ensure the threshold is always both positive and close to a "reasonable" value (that is, useful for estimation).

3.5.2 Computing Estimators

Recall the estimators Diagonal-C 3.40, Diagonal-H3.44 and Full-H 3.43 for $\mathbf{\Sigma}_k$ in the Score-type procedure. When evaluating the test statistic, we require either $\widehat{\mathbf{\Sigma}}_k^{-1}$ or $\widehat{\mathbf{\Sigma}}_k^{-1/2}$. Inverting $(n - 2G)$-many matrices will be extremely costly (see 5.2), but by noting the relationship between $\widehat{\mathbf{\Sigma}}_k$ and $\widehat{\mathbf{\Sigma}}_{k+1}^{-1/2}$, we can invoke the Woodbury formula [She] and reduce the amount of operations required.

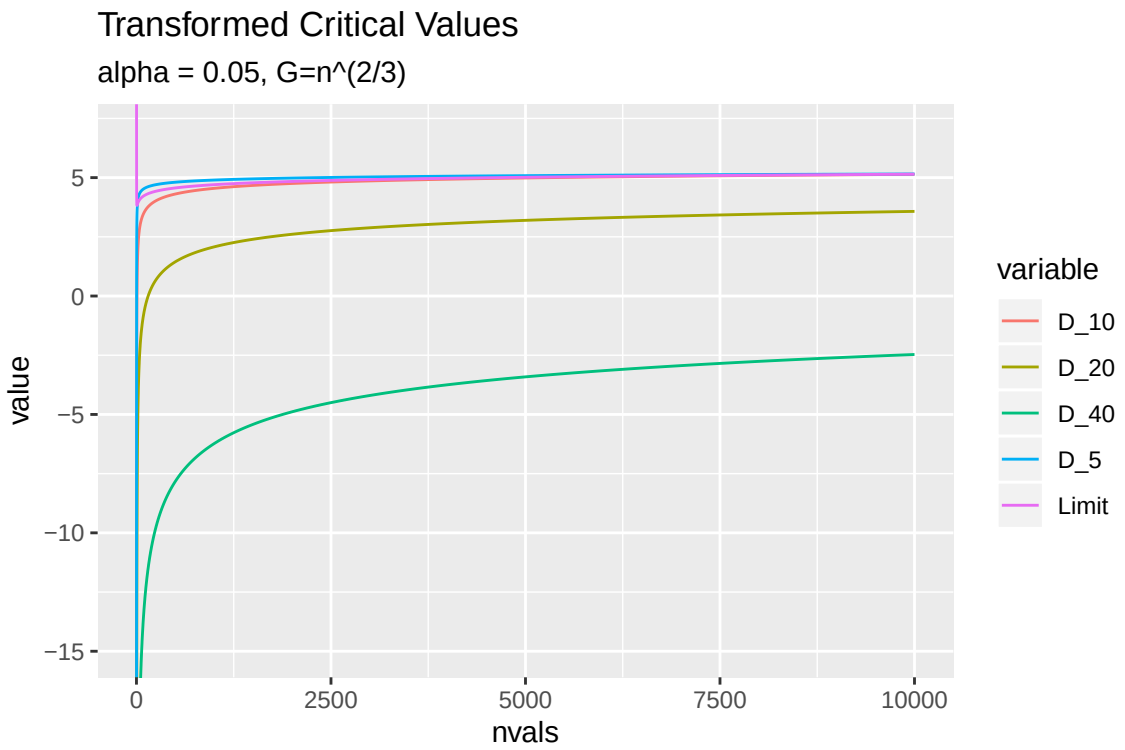


Figure 3.2: Theoretical critical values against sample size, by dimensionality; Limit = $\sqrt{2\log(n)} + \frac{c_\alpha}{\sqrt{2\log(n)}}$

For example, in Full-H 3.43, we have that

$$\begin{aligned}\widehat{\boldsymbol{\Sigma}}_{k+1} = \widehat{\boldsymbol{\Sigma}}_k &+ \boldsymbol{H}\left(Y_{k+1+G}, \mathcal{X}_{k+G}, \hat{\boldsymbol{a}}_{1,n}\right)\boldsymbol{H}\left(Y_{k+1+G}, \mathcal{X}_{k+G}, \hat{\boldsymbol{a}}_{1,n}\right)^T - \\ &\boldsymbol{H}\left(Y_{k-G}, \mathcal{X}_{k-1-G}, \hat{\boldsymbol{a}}_{1,n}\right)\boldsymbol{H}\left(Y_{k-G}, \mathcal{X}_{k-1-G}, \hat{\boldsymbol{a}}_{1,n}\right)^T - \\ &2\boldsymbol{H}\left(Y_k, \mathcal{X}_{k-1}, \hat{\boldsymbol{a}}_{1,n}\right)\boldsymbol{H}\left(Y_k, \mathcal{X}_{k-1}, \hat{\boldsymbol{a}}_{1,n}\right)^T\end{aligned} \tag{3.63}$$

Denote $h_k = \boldsymbol{H}\left(Y_k, \mathcal{X}_{k-1}, \hat{\boldsymbol{a}}_{1,n}\right), U = (h_{k+1+G}, -h_{k-G}, 2h_k), V = (h_{k+1+G}, h_{k-G}, h_k)^T$, then we can express this update in matrix form:

$$\widehat{\boldsymbol{\Sigma}}_{k+1} = \widehat{\boldsymbol{\Sigma}}_k + UV \tag{3.64}$$

So

$$\widehat{\boldsymbol{\Sigma}}_{k+1}^{-1} = \widehat{\boldsymbol{\Sigma}}_k^{-1} - \widehat{\boldsymbol{\Sigma}}_k^{-1}U(I_3 + V\widehat{\boldsymbol{\Sigma}}_k^{-1}U)^{-1}V\widehat{\boldsymbol{\Sigma}}_k^{-1} \tag{3.65}$$

and we can start the iteration from $\widehat{\boldsymbol{\Sigma}}_{G+1}^{-1}$.

Again, for complexity analysis of this procedure, see 5.2.

This can also be applied to the expectation matrix component of 3.58. Unfortunately the finding the procedures for the Wald-type estimators 3.56 and 3.57 is non-trivial, since each evaluation of $\widehat{\boldsymbol{\Sigma}}_k$ depends on new parameters $\hat{\boldsymbol{a}}_{k+1,k+G}$ and $\hat{\boldsymbol{a}}_{k-G+1,k}$. Procedural updating of $\hat{\boldsymbol{a}}_{l,u}$ is possible, so with more work we might find a recursive formula relating the covariance estimators.

Chapter 4

Inferring Network Structure from Autoregressive Time Series

Now that we have a means with which to identify and locate changes in multiple time series, we are ready to infer network structure from the stationary segments. In this chapter, we introduce the notion of Granger causality, describe how it can be used to define a network structure from a vector autoregression model, and propose methods for inferring networks.

4.1 Granger Causality

The idea of Granger causality is that a cause must precede an effect in time. Hence, if a variable X_t has some effect on another Z_t, knowledge of Z_{t-s} for some lag $s \in \mathbb{N}_0$ will improve our predictions of Z_t. Bear in mind, this relationship is not necessarily causal in the strict sense (for example, there may be another variable, possibly omitted from the analysis, which actually has a causal relationship with Z_t), but as with any regression task, we can use this relationship for prediction or understanding association.

To define this idea formally, we confer [Lüt05, Section 2.3.1]. Let Ω_t be the set of all relevant information up to time t, and let $Z_t(h|\Omega_t)$ be the MSE-optimal h-step predictor of a univariate process Z_t at time t, based on the information in Ω_t - that is, $MSE_Z(h|\Omega_t)$ is minimised.

A process X_t **Granger-causes** Z_t if

$$MSE_Z(h|\Omega_t) < MSE_Z(h|\Omega_t \backslash \{X_s : s \leq t\}) \text{ for at least one } h = 1, 2, \ldots \tag{4.1}$$

A multivariate process $\boldsymbol{X}_t$ **Granger-causes** another, $\boldsymbol{Z}_t$, if

$$\boldsymbol{MSE}_{\boldsymbol{Z}}(h|\Omega_t) \neq \boldsymbol{MSE}_{\boldsymbol{Z}}(h|\Omega_t \backslash \{X_s : s \leq t\}) \text{ for at least one } h = 1, 2, \ldots \tag{4.2}$$

that is, the difference between the two matrices is positive semidefinite and the two $\boldsymbol{MSE}$ matrices are not equal.

In particular, for a stationary, stable VAR(p) model,

$$\begin{aligned} \boldsymbol{Y}_t = \begin{bmatrix} \boldsymbol{Z}_t \\ \boldsymbol{X}_t \end{bmatrix} = \begin{bmatrix} \boldsymbol{\nu}_1 \\ \boldsymbol{\nu}_2 \end{bmatrix} + \begin{bmatrix} A_{11,1} & A_{12,1} \\ A_{21,1} & A_{22,1} \end{bmatrix} \begin{bmatrix} \boldsymbol{Z}_{t-1} \\ \boldsymbol{X}_{t-1} \end{bmatrix} + \cdots \\ + \begin{bmatrix} A_{11,p} & A_{12,p} \\ A_{21,p} & A_{22,p} \end{bmatrix} \begin{bmatrix} \boldsymbol{Z}_{t-p} \\ \boldsymbol{X}_{t-p} \end{bmatrix} + \begin{bmatrix} \boldsymbol{\varepsilon}_{1t} \\ \boldsymbol{\varepsilon}_{2t} \end{bmatrix} \end{aligned} \tag{4.3}$$

we have that

$$\boldsymbol{Z}_t\left(h \,|\{\boldsymbol{Y}_s : s \leq t\}\right) = \boldsymbol{Z}_t\left(h \,|\{\boldsymbol{Z}_s : s \leq t\}\right) \text{ for } h = 1, 2, \ldots, \quad \Leftrightarrow \quad A_{12,l} = 0 \quad \text{for } l = 1, 2, \ldots, p \tag{4.4}$$

that is, $\boldsymbol{X}_t$ is non-causal for $\boldsymbol{Z}_t$ only when the regression coefficients from $\boldsymbol{X}_{t-l}$ to $\boldsymbol{Z}_t$ are 0 for all considered lags $l = 1, \ldots, p$.

Using stacked vector notation (as in 3.2) to represent our time series $\boldsymbol{Y}_t$, recall we have the regression problem for each channel (as in 3.9):

$$Y_t(i) = \boldsymbol{a}(i)^T \mathbb{X}_{t-1}(i) + e_t(i) \tag{4.5}$$

We have that the r-th entry $Y_{t-h}(i') = \mathbb{X}_{t-1}(i)[r]$ (where $h \leq p$) of $\mathbb{X}_{t-1}(i)$ is Granger-causal for $Y_t(i)$ if and only if the r-th entry $a(i)[r]$ of $\boldsymbol{a}(i)$ is non-zero (i.e. $a(i)[r] \neq 0$), and denote this

$$Y_{t-h}(i') \to Y_t(i) \tag{4.6}$$

Note that we exclude intercept terms in the definition of our network.

By assuming diagonal error covariance, as in (R4) and (R4*), we assume instantaneous non-causality in $\boldsymbol{Y}_t$ [Lüt05, Equation 2.3.3].

4.1.1 Causal Networks

Multivariate causality enables us to define a network structure [BSM15]. A **Causal Network** is a pair $\mathcal{N} = (\mathcal{V}, \mathcal{E})$ with a **Vertex Set**

$$\mathcal{V} \subset \{Y_{t-l}(i) : i = 1, \ldots, d; l = 0, 1, \ldots, p\} \tag{4.7}$$

of random variables, and an **Edge Set**

$$\mathcal{E} \subset \{(Y_{t-l}(i), Y_{t-l'}(i')) : i, i' = 1, \ldots, d, \quad l, l' = 0, 1, \ldots, p \quad l > l'\} \tag{4.8}$$

We define directed edges $Y_{t-l}(i') \to Y_{t-l'}(i)$ where $Y_{t-l}(i')$ is Granger-causal for $Y_{t-l'}(i)$, as in 4.6.

Of course, naively estimating a network defined by least-squares estimates of coefficients will return a dense network with all possible edges with probability 1. To accurately reflect the true underlying structure, we must impose some constraints on the edges we want to include, which is the focus of the next section.

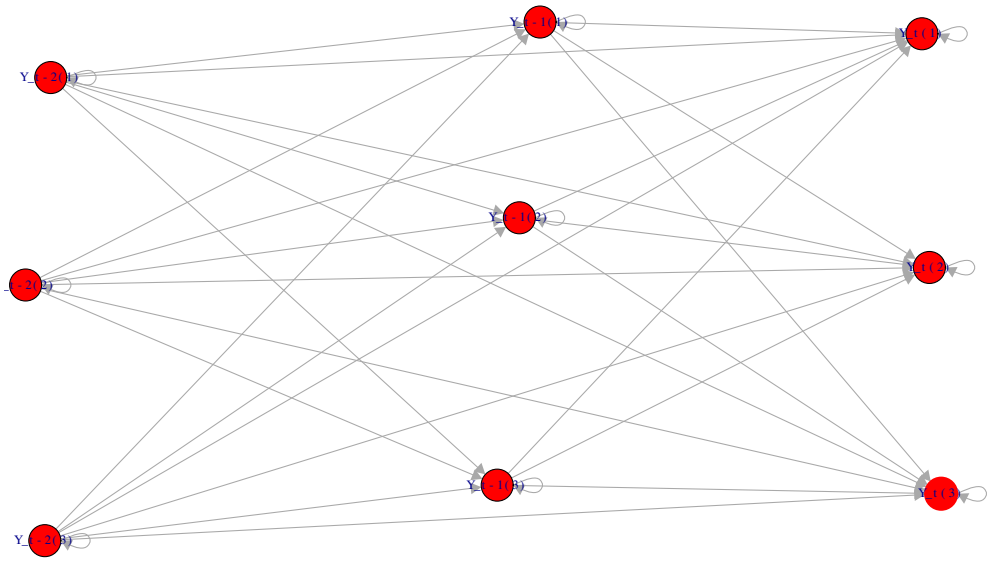


Figure 4.1: Causal network, $d = 3$, $p = 2$

4.2 Network Inference

Broadly speaking, there are two classes of method for inferring a causal network from a VAR model, as per [SKTK19]. The first adds edges to the network if they pass parametric hypothesis testing. The second class uses regularisation, either to control the density of the network or to control the magnitudes of causality permitted. These can either implicitly or explicitly control the significance level of the network inference procedure. For a review of methods for inferring causal networks from linear autoregressive models in the context of gene regulatory networks [MdB13, Section 4].

4.2.1 Network Inference Via Hypothesis Testing

First, we consider a hypothesis-testing method. In [MC07], a means of defining an atemporal network with variables as vertices is proposed. Causation between a pair of vertices is allowed to be in a single direction, that with the greatest autoregression parameter, meaning $d(d+1)/2$ tests for potential edges must take place. Using the false discovery correction of [BH95], the null hypothesis of no edge existing is rejected for the subset of p-values meeting the correction criteria.

We could easily adapt this to a temporal network. Given that causation can only happen forwards in time, and ignoring intercept terms, we have d^2p-many hypotheses to test. Clearly, this grows very quickly with dimensionality and so our size (false discovery rate) will be controlled by the correction criteria, power will diminish as a trade-off. This can be appropriate for models with small dimensions.

Consider, for each edge $(Y_{t-l}(i), Y_{t-l'}(i'))$ which could possibly be in $\mathcal{E}$, the hypotheses

$$\begin{aligned} H_0 &: \text{ The edge } (Y_{t-l}(i), Y_{t-l'}(i')) \text{ is not an element of } \mathcal{E} \\ H_1 &: \text{ The edge } (Y_{t-l}(i), Y_{t-l'}(i')) \text{ is an element of } \mathcal{E} \end{aligned} \tag{4.9}$$

For a Wald test of multi-step Granger causality, we confer [Lüt05, Section 3.6.4].

4.2.2 Network Inference Via Regularisation

Alternatively, and more appropriately when dimensionality is large, we can use regularisation methods to infer a network. We should bear in mind, given our motivating application of analysing risk measure networks, that these methods have been developed with sparse data in mind [BA99], and for our problems similar to our application often the opposite is the case [GLP17].

Methods are distinguished by the choice of regularisation penalty. In our case, we wish to determine the correct order of autoregression and to allow causal effects to vary with respect to their lag, so we employ a version of the **Truncating Lasso** from [SM10].

The following loss function is minimised over individual channels $i = 1, \dots, d$:

$$\begin{aligned} &\min_{\boldsymbol{a}(i)\in\Theta} \frac{1}{n} \sum_{t=1}^{n} \|Y_t(i) - \boldsymbol{a}^T(i)\mathbb{X}_{t-1}\|_2^2 + \lambda \sum_{l=1}^{p} \Psi_l \sum_{j=1}^{dp+1} |a(i)[j]| \\ &\Psi_1 = 1 \\ &\Psi_l = M^{\mathbb{I}\left\{\|A_{(l-1)}\|_0 < d^2\beta\right\}}, l \geq 2 \end{aligned} \tag{4.10}$$

where M is a large constant, λ is the regularisation hyperparameter, and β is the permitted false discovery (type II error) rate.

The rationale behind the truncation provided by Ψ_l is that the size and number of causal effects should diminish over time. We impose the sparsity condition $\left\|A_{(l-1)}\right\|_0 < d^2\beta$ on the zero-norm of the coefficient matrix one step more recent than $A_{(l)}$. If the number of edges in $A_{(l)}$is not significant, as controlled by β, the edges from matrices of all subsequent lags are set to zero, thus determining the maximum lag. This feature means we can set an initial choice for the lag p to be much higher than we suspect the true value to be (if computation permits us) and recover a network with a data-driven choice for the lag. As shown in [SM10], this is asymptotically consistent in variable selection for high-dimensional sparse models.

In selecting hyperparameter values, the authors recommend

$$\lambda = 2n^{-1/2} Z^*_{\alpha/2d(dp+1)} \tag{4.11}$$

where Z^*_w is the $(1-w)$-th percentile of a standard normal distribution. This controls false positives at rate α. We can also select this through cross-validation for predictive purposes.

For β, we should bear in mind the level of sparsity we want to impose on the network. For $\beta = 1$, any coefficient matrix with a single zero entry will cause all subsequent matrices to be set to zero; we are choosing to always say an edge doesn't exist, even if it does in the true model. In financial applications, we want our network to be relatively dense, so would select β to be close to zero, and conversely we are happy to claim an edge does exist when in fact it does not.

Chapter 5

Simulation Study and Data Analysis

5.1 Simulation Study

We have given theoretical large-sample results for both MOSUM statistics in Chapter 3. Deriving finite-sample results analytically is extremely difficult, so we provide empirical results with synthetic data which should resemble situations which might be seen in practice, and evaluate the performance of the procedures over multiple replicates.

The metrics determining performance here are empirical size (average observed number of false positives) and empirical power (average observed true negatives). We are also interested in knowing how well the procedure estimates the number and location of changes. The sample mean and sample standard deviation are reported for the estimated number of changes, along with the percentage of locations estimated to be within $k_{i,2000} \pm 40$ (using intervals $\pm 0.05 \times n$).

To investigate how the methods perform for different model specifications, especially those with many parameters, we consider three processes. First a VAR(p) data-generating process with order $p = 1$ and dimensions $d = 4$. Secondly, we consider a process with order $p = 2$ and dimension $d = 3$. Thirdly, for results from a smaller model, we simulate from a model with $d = 2$ and $p = 1$. For each process, we use errors $\varepsilon_t(i) \sim N(0, 0.5^2)$ and $n = 2000$ time steps, with $N = 100$ simulation replicates.

Under the null hypothesis, each process follows the first described regime for the entire sample. Under the alternative, there are $q = 3$ changes (respectively, $q + 1 = 4$ regimes), and changes occur at $k_{1,2000} = 500$,$k_{2,2000} = 1000$, and $k_{3,2000} = 1500$.

For test tuning parameters, we use significance level $\alpha = 0.05$ and location neighbourhood parameter $\epsilon = 0.25$ (see 3.34, 3.51).

5.1.1 Simulation 1: $d = 4$, $p = 1$

We simulate a process with $d = 4$, $p = 1$. The dimensionality ($4 \times 5 = 20$) of this problem is large enough to cause instability in the testing procedure, as demonstrated in 3.5 and 5.2.

For each regime, we use the following parameter matrices

$$\text{Regime 1} : \boldsymbol{A} = \begin{pmatrix} 0.7 & -0.1 & -0.1 & -0.1 \\ -0.1 & 0.7 & -0.1 & -0.1 \\ -0.1 & -0.1 & 0.7 & -0.1 \\ -0.1 & -0.1 & -0.1 & 0.7 \end{pmatrix} \text{Regime 2} : \boldsymbol{A} = \begin{pmatrix} 0.6 & 0.1 & 0.1 & 0.1 \\ 0.1 & 0.6 & 0.1 & 0.1 \\ 0.1 & 0.1 & 0.6 & 0.1 \\ 0.1 & 0.1 & 0.1 & 0.6 \end{pmatrix}$$

$$\text{Regime 3} : \boldsymbol{A} = \begin{pmatrix} 0.5 & 0.1 & 0.1 & -0.1 \\ 0.1 & 0.5 & -0.1 & 0.1 \\ 0.1 & -0.1 & 0.5 & 0.1 \\ -0.1 & 0.1 & 0.1 & 0.5 \end{pmatrix} \text{Regime 4} : \boldsymbol{A} = \begin{pmatrix} 0.7 & -0.1 & -0.1 & -0.1 \\ -0.1 & 0.7 & -0.1 & -0.1 \\ -0.1 & -0.1 & 0.7 & -0.1 \\ -0.1 & -0.1 & -0.1 & 0.7 \end{pmatrix} \tag{5.1}$$

Table 5.1: Score-type Statistic; d=4, p=1

	H_0 Empirical Size	H_1 Empirical Power	Estimated Number		Estimated Location		
			Mean	S.D.	$[500 \pm 40]$	$[1000 \pm 40]$	$[1500 \pm 40]$
DIAG-C							
G=100	1.00	1.00	8.92	1.91	0.83	0.67	0.85
G=150	0.97	1.00	5.96	1.21	0.84	0.72	0.81
G=200	0.95	1.00	4.49	1.01	0.87	0.66	0.87
DIAG-H							
G=100	0.52	1.00	5.52	1.85	0.84	0.41	0.53
G=150	0.34	1.00	4.43	1.22	0.92	0.53	0.76
G=200	0.16	1.00	3.87	0.92	0.88	0.62	0.83
FULL-H							
G=100	0.49	1.00	5.44	1.62	0.87	0.42	0.63
G=150	0.30	1.00	4.59	1.36	0.90	0.57	0.78
G=200	0.30	1.00	3.97	0.86	0.90	0.72	0.92

For the Score procedure, with results in table 5.1, we can see that the DIAG-C estimator does not control the empirical size at all (confer 5.1); this is likely due to how the error variances $\sigma^2(i)$ are underestimated by 3.42. To rectify this, we should consider instead implementing a MOSUM-type estimator, along the lines of LOCAL-2 in [Rec19, Section 4].

Both DIAG-H and FULL-H do a poor job of controlling empirical size, though we can see some decreasing effect with G, so we should consider using larger values for G. In terms

of estimation, these perform reasonably well, and accuracy improves with G, though the number is always over-estimated, suggesting the test is liberal. Reporting size-adjusted power estimates might aid our understanding of the test performance.

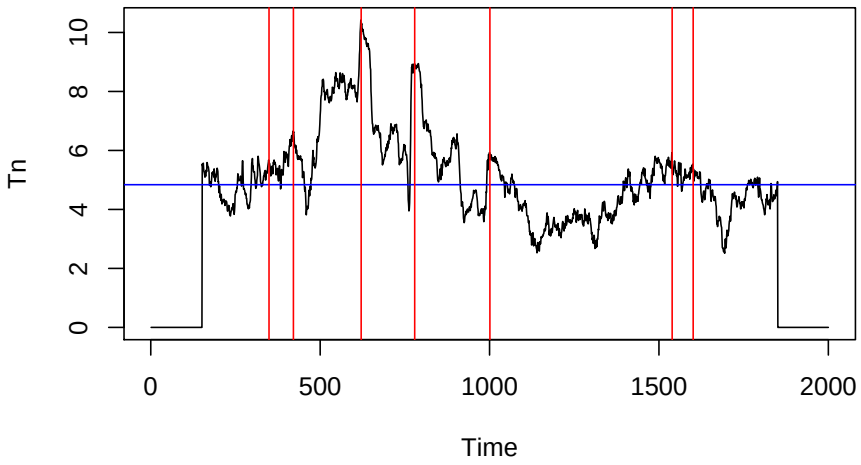


Figure 5.1: Example simulation 1 test result: Score Statistic under alternative, Diag-C estimator, $G = 150$

For the Wald procedure, with results in table 5.2, we can see that the all three estimators do a good job of controlling size and power. DIAG-C does very well in estimation, though all three under-estimate the number of changes. Note the behaviour around the last change point, where DIAG-H and FULL-H seem unable to detect the point at all; it is liekly the parameter change from regime 3 to regime 4 is not detectable at this resolution, with the specified threshold and ϵ criteria.

Table 5.2: Wald-type Statistic; d=4, p=1

	H_0 Empirical Size	H_1 Empirical Power	Estimated Number		Estimated Location		
			Mean	S.D.	$[500 \pm 40]$	$[1000 \pm 40]$	$[1500 \pm 40]$
DIAG-C							
G=100	0.61	1.00	2.92	1.05	0.86	0.42	0.83
G=150	0.20	1.00	2.72	0.67	0.95	0.51	0.93
G=200	0.09	1.00	2.79	0.52	0.96	0.60	0.95
DIAG-H							
G=100	0.31	0.83	1.34	0.90	0.57	0.39	0.00
G=150	0.06	0.85	1.30	0.77	0.66	0.40	0.00
G=200	0	0.93	1.53	0.67	0.78	0.56	0.00
FULL-H							
G=100	0.52	0.93	1.55	0.85	0.72	0.41	0.01
G=150	0.08	0.93	1.46	0.73	0.74	0.42	0.01
G=200	0.03	0.98	1.57	0.69	0.87	0.47	0.00

5.1.2 Simulation 2: $d = 3$, $p = 2$

We simulate a process with $d = 3$, $p = 2$. The dimensionality ($3 \times 7 = 21$) of this problem is also enough to cause instability.

$$
\begin{aligned}
\text{Regime 1} : \boldsymbol{A}_1 &= \begin{pmatrix} 0.5 & 0.1 & 0.1 \\ 0.1 & 0.5 & 0.1 \\ 0.1 & 0.1 & 0.5 \end{pmatrix} \boldsymbol{A}_2 = \begin{pmatrix} -0.2 & 0.2 & 0.2 \\ 0.2 & -0.2 & 0.2 \\ 0.2 & 0.2 & -0.2 \end{pmatrix} \\
\text{Regime 2} : \boldsymbol{A}_1 &= \begin{pmatrix} -0.2 & 0.2 & 0.2 \\ 0.2 & -0.2 & 0.2 \\ 0.2 & 0.2 & -0.2 \end{pmatrix} \boldsymbol{A}_2 = \begin{pmatrix} 0.5 & 0.1 & 0.1 \\ 0.1 & 0.5 & 0.1 \\ 0.1 & 0.1 & 0.5 \end{pmatrix} \\
\text{Regime 3} : \boldsymbol{A}_1 &= \begin{pmatrix} 0.4 & 0.4 & 0.0 \\ 0.4 & 0.4 & 0.4 \\ 0.0 & 0.4 & 0.4 \end{pmatrix} \boldsymbol{A}_2 = \begin{pmatrix} -0.2 & 0.0 & 0.0 \\ 0.0 & -0.2 & 0.0 \\ 0.0 & 0.0 & -0.2 \end{pmatrix} \\
\text{Regime 4} : \boldsymbol{A}_1 &= \begin{pmatrix} 0.5 & 0.1 & 0.1 \\ 0.1 & 0.5 & 0.1 \\ 0.1 & 0.1 & 0.5 \end{pmatrix} \boldsymbol{A}_2 = \begin{pmatrix} -0.2 & 0.2 & 0.2 \\ 0.2 & -0.2 & 0.2 \\ 0.2 & 0.2 & -0.2 \end{pmatrix}
\end{aligned}
\tag{5.2}
$$

In table 5.3, we again see that the Score DIAG-C estimator is not currently viable. DIAG-H and FULL-H control size and power well, but over-estimate the number of changes.

In table 5.4, the three estimators for the Wald test have empirical size very close to the

Table 5.3: Score-type Statistic; d=3, p=2

	H_0 Empirical Size	H_1 Empirical Power	Estimated Number		Estimated Location		
			Mean	S.D.	$[500 \pm 40]$	$[1000 \pm 40]$	$[1500 \pm 40]$
DIAG-C							
G=100	0.99	1.00	8.24	1.92	0.82	0.65	0.84
G=150	0.97	1.00	5.28	1.43	0.77	0.62	0.84
G=200	0.92	1.00	4.10	0.76	0.74	0.48	0.87
DIAG-H							
G=100	0.21	1.00	4.66	1.39	0.88	0.91	0.38
G=150	0.11	1.00	4.11	1.32	0.90	0.91	0.58
G=200	0.07	1.00	4.23	1.06	0.81	0.83	0.86
FULL-H							
G=100	0.35	1.00	5.33	1.76	0.95	0.96	0.47
G=150	0.14	1.00	4.34	1.53	0.89	0.89	0.72
G=200	0.13	1.00	4.34	1.12	0.88	0.74	0.80

Table 5.4: Wald-type Statistic; d=3, p=2

	H_0 Empirical Size	H_1 Empirical Power	Estimated Number		Estimated Location		
			Mean	S.D.	$[500 \pm 40]$	$[1000 \pm 40]$	$[1500 \pm 40]$
DIAG-C							
G=100	0.07	1.00	2.69	0.73	1.00	0.99	0.32
G=150	0.01	1.00	2.83	0.55	0.99	1.00	0.72
G=200	0.00	1.00	2.97	0.33	0.99	0.99	0.89
DIAG-H							
G=100	0.07	1.00	2.09	0.57	0.91	0.97	0.02
G=150	0.00	1.00	2.04	0.20	0.96	0.98	0.01
G=200	0.01	1.00	2.06	0.24	0.98	1.00	0.05
FULL-H							
G=100	0.16	1.00	2.27	0.60	0.95	1.00	0.02
G=150	0.03	1.00	2.13	0.42	0.97	0.98	0.03
G=200	0.00	1.00	2.05	0.22	0.96	0.97	0.04

theoretical size $\alpha = 0.05$. DIAG-C does well in estimation, though DIAG-H and FULL-H are again hindered by their inability to detect the final change. This is visible in figure 5.2 - the final change is clearly reflected by the shape of the statistic, but is not large enough to be detected. With a larger sample size, a different finite-sample threshold correction (as

discussed in 3.5) or a different value for ϵ, which tunes how we reject spurious positives, this would be detectable.

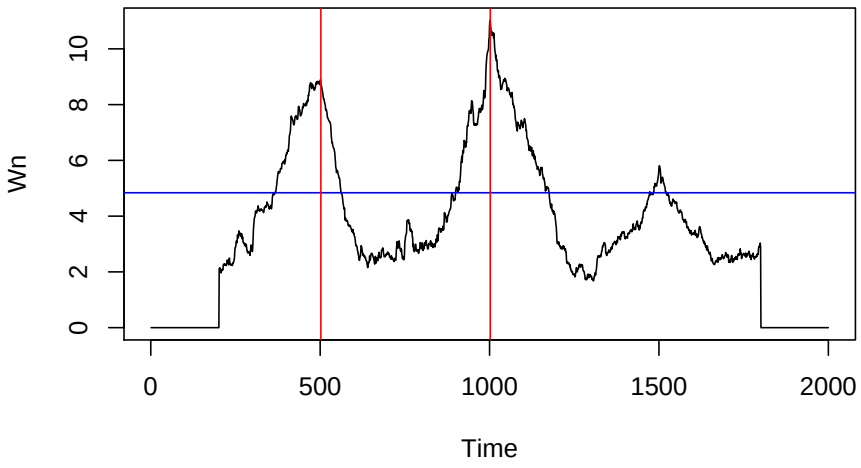


Figure 5.2: Example simulation 2 test result: Wald Statistic under alternative, Diag-H estimator, $G = 200$

5.1.3 Simulation 3: $d = 2$, $p = 1$

To investigate the phenomena seen for the previous two processes, namely size control for the Score test and the detectability of changes with H-based estimators in the Wald test, we consider a smaller model with $d = 2$, $p = 1$. The dimensionality ($2 \times 3 = 6$) of this is moderate and so no instability should follow, and the small-sample critical value D_n will be used as opposed the practical approximation. Under the null, the process is strongly autoregressive and the test is thus less affected by noise. Under the alternative, this process has large changes, which should be easy for both procedures to detect.

$$\begin{aligned}
\text{Regime 1} : \boldsymbol{A}_1 &= \begin{pmatrix} -0.75 & -0.75 \\ 0.75 & 0.75 \end{pmatrix} \\
\text{Regime 2} : \boldsymbol{A}_1 &= \begin{pmatrix} 0.25 & 0.25 \\ -0.25 & -0.25 \end{pmatrix} \\
\text{Regime 3} : \boldsymbol{A}_1 &= \begin{pmatrix} -0.25 & -0.25 \\ 0.25 & 0.25 \end{pmatrix} \\
\text{Regime 4} : \boldsymbol{A}_1 &= \begin{pmatrix} 0.75 & 0.75 \\ -0.75 & -0.75 \end{pmatrix}
\end{aligned} \tag{5.3}$$

Table 5.5: Score-type Statistic; d=2, p=1, Large Changes

	H_0 Empirical Size	H_1 Empirical Power	Estimated Number		Estimated Location		
			Mean	S.D.	[500 ± 40]	[1000 ± 40]	[1500 ± 40
DIAG-C							
G=100	0.01	1	4.42	0.9339825	1	0.41	1
G=150	0	1	4.95	0.4351941	1	0.93	1
G=200	0	1	4.96	0.2428784	1	1	1
DIAG-H							
G=100	0	1	2.06	0.3974667	0.95	0.01	0.98
G=150	0	1	3.15	0.8804843	1	0.26	0.99
G=200	0	1	4.53	0.6883592	1	0.86	1
FULL-H							
G=100	0	1	1.84	0.4431294	0.86	0.01	0.94
G=150	0	1	2.78	0.7046741	1	0.3	1
G=200	0	1	4.24	0.7401883	1	0.74	1

From the results in table 5.5, we can see that for this process the Score test is very liberal, with size zero for nearly every window size, and power instantly reaching one. This does a

very good job in estimating the locations, although the more conservative H-based estimators struggle to pick up the smaller second regime change.

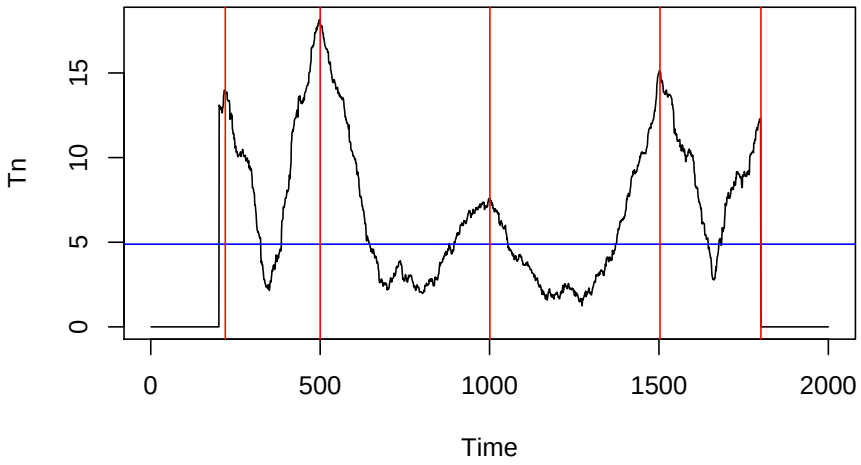


Figure 5.3: Example simulation 3 test result: Score statistic under alternative, Diag-C estimator, $G = 200$

For the Wald test, the results of which are in table 5.6, again the size is understated and the power reaches one very quickly, though we should expect this given the large relative size of the parameter changes. Examining behaviour with a smaller window size, or smaller significance level, might give us finer detail on how this is behaving. The Diag-C estimator does an incredible job for estimation, and was exactly right with $G = 200$ (confer figure 5.5). The H-based estimators had a harder time identifying the smaller second regime change, which is reflected by the smaller middle peak in 5.4.

Table 5.6: Wald-type Statistic; d=2, p=1, Large Changes

	H_0 Empirical Size	H_1 Empirical Power	Estimated Number		Estimated Location		
			Mean	S.D.	$[500 \pm 40]$	$[1000 \pm 40]$	$[1500 \pm 40$
DIAG-C							
G=100	0	1	2.22	0.4163332	1	0.22	1
G=150	0	1	2.85	0.3588703	1	0.84	1
G=200	0	1	3	0	1	1	1
DIAG-H							
G=100	0	0.95	1.65	0.5751592	0.81	0	0.84
G=150	0	1	2.06	0.277798	1	0.07	0.98
G=200	0	1	2.24	0.4292347	1	0.24	1
FULL-H							
G=100	0	0.84	1.24	0.7123726	0.67	0	0.57
G=150	0	1	1.95	0.2611165	0.98	0.01	0.96
G=200	0	1	2.21	0.4093602	1	0.21	1

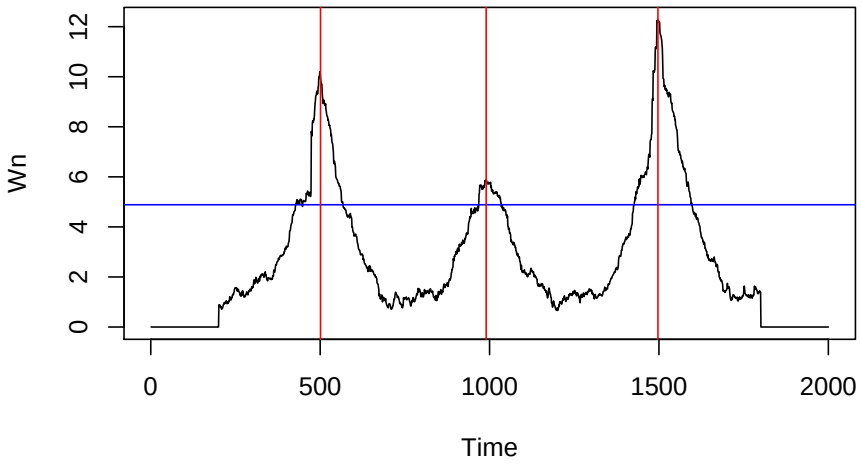


Figure 5.4: Example simulation 3 test result: Wald statistic under alternative, Diag-H estimator, $G = 200$

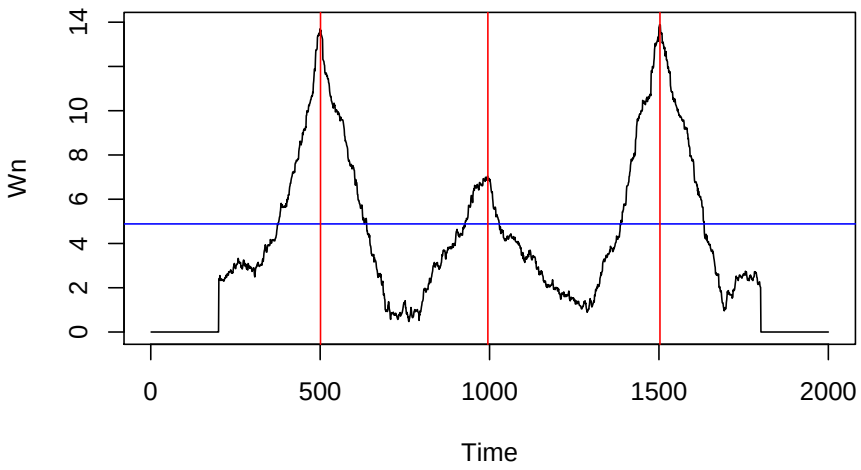


Figure 5.5: Example simulation 3 test result: Wald statistic under alternative, Diag-C estimator, $G = 200$

5.1.4 Simulation Summary

In summary, our method does has demonstrated it can do a very good job at detecting and locating changes in situations where the the dimensionality of the problem is small and the relative size of the changes is moderate. The Wald procedure with a DiagC estimator appears to be the best performing overall, and is recommended for use in practice. For situations in which testing is more important than estimation, we recommend using small ϵ values for better detection.

As the dimension grows, however, performance deteriorates at a non-negligible rate, meaning we cannot recommend the current procedure for large problems (even before taking into account the computational cost of the problem). Understanding how the nature of regime changes affects the ability of the procedures to detect and locate changes would help us explain what we have seen in this section.

5.2 Computational Considerations

The entire testing and estimation procedures have the potential to be computationally expensive even for small sample sizes n and moderate dimensions d and p. Here we give some consideration to how these methods scale in terms of time complexity and run time.

5.2.1 Score-type Procedure

We should expect the Score-type procedure to behave better with scale than the Wald-type procedure, given that the largest contributor to the number of operations to perform is the estimation of parameters, and this procedure conducts much estimation in a global manner as opposed to a MOSUM manner. In these calculations, we assume $G << n$.

Evaluating $\hat{\boldsymbol{a}}_{1,n}$ (3.39) amounts to solving d-many least-squares problems with n observations and $1 + dp$ regressors each, so has complexity $\mathcal{O}\big(nd(1+dp)^2\big) = \mathcal{O}\big(nd^3p^2\big)$.

Each channel-specific estimating function $\boldsymbol{H}_i$ has complexity $\mathcal{O}\big(d^2p^2\big)$,as determined by matrix multiplication, so finding $\boldsymbol{H}$ for all d-many channels has cost $\mathcal{O}\big(d^3p^2\big)$; calculating these for all time steps $k \in [G+p, n-G]$ costs $\mathcal{O}\big(nd^3p^2\big)$.

For Diagonal-C 3.40, evaluating the negative square root in the case of no prior test reduction (that is, $\mathbb{X}_{t-1}(i) = \mathbb{X}_{t-1} \forall i$) can be performed using one singular-value decomposition for $\widehat{\boldsymbol{C}}_{k-G+1,k+G}(i)$, so evaluating this has complexity $\mathcal{O}\big(d^3p^3\big)$. An estimator is calculated for each time step $k \in [G+p, n-G]$, so finding $\widehat{T}_n$ using this estimator has complexity $\mathcal{O}\big(nd^3p^3\big)$. Using the Woobury update formula 3.5.2, we invert $\widehat{\boldsymbol{\Sigma}}_{G+1}$ then update over $n - 2G$ time steps by multiplying matrices of dimension $(dp+1) \times (dp+1)$, so this has complexity $\mathcal{O}(d^3p^3 + nd^2p^2)$. This is hence cheaper when $n = \mathcal{O}(dp)$.

For Full-H 3.43 and Diagonal-H 3.44, the complexity will be again be dominated by matrix inversions. In computing Full-H, we invert a full-rank matrix of dimension $d(dp+1) \times d(dp+1)$, with complexity $\mathcal{O}\big((d(dp+1))^3\big) = \mathcal{O}\big(d^6p^3\big)$, meaning finding the maximum statistic has complexity $\mathcal{O}\big(nd^6p^3\big)$. Using the Woodbury update, by a similar calculation as above, this has complexity over all time steps of $\mathcal{O}(d^6p^3 + nd^4p^2)$. This is hence cheaper when $n = \mathcal{O}(d^2p)$.

In comparison, Diag-H has a block-diagonal form, and we can invert all d-many matrices with dimension $(dp+1) \times (dp+1)$ at cost $\mathcal{O}\big(d(dp+1)^3\big) = \mathcal{O}\big(d^4p^3\big)$, so evaluating the test statistic costs $\mathcal{O}\big(nd^4p^3\big)$. With a Woodbury update, this would cost $\mathcal{O}\big(d^4p^3 + nd^4p^2\big)$, which would be cheaper in the situation (impossible in reality) of $n = \mathcal{O}(p)$.

The model-fitting process is "embarrassingly parallel" in the sense that we can easily identify the evaluation of each $T_{k,n}, k \in [G+p, n-G]$ as a process to be assigned to a separate core,

so savings in run time are easy to obtain in an implementation.

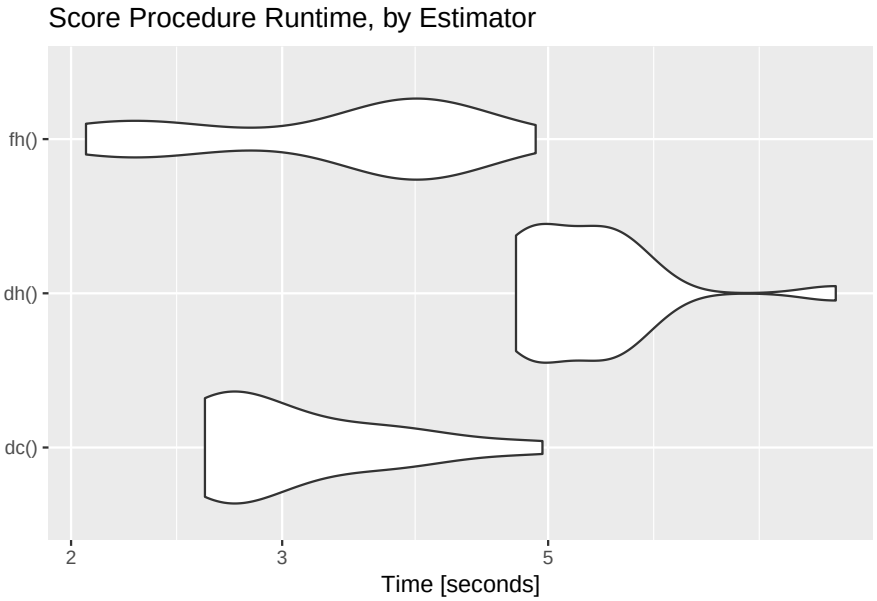


Figure 5.6: Score Procedure Runtime, $n = 2000, d = 5, p = 1$: fh = Full H, dh = Diagonal H, dc = Diagonal C

The runtime empirical distributions in 5.6, based on simulated data of size $n = 2000, d = 5$ modelled with order $p = 1$, show that surprisingly, DiagonalC and FullH are faster than DiagonalH, though it is not clear if this differences holds for larger dimensions. These are orders of magnitude faster than the runtimes for the Wald tests (see next subsection).

5.2.2 Wald-type Procedure

As mentioned in the last subsection, the Wald procedure requires evaluations of parameters and estimators at every time step. We see this leads to higher complexity.

Evaluating all parameter estimates $\hat{\boldsymbol{a}}_{l,u}$ (3.18) with bandwidth G, over n observations, has complexity $\mathcal{O}\big(nd^3p^2\big)$.

We compute $2G$-many evaluations of the estimating function $\boldsymbol{H}$ for each time step k, so computing all of these also has cost $\mathcal{O}\big(nd^3p^2\big)$.

Computing $\widehat{\boldsymbol{V}}_k$ only has cost $\mathcal{O}\big(d^2p^2\big)$ due to the blockwise-diagonal structure.

The root-inverted estimators Diagonal-C 3.40, Diagonal-H 3.56 and Full-H 3.57 for $\widehat{\boldsymbol{\Sigma}}_{n,k}$ have the same costs over all time steps as for the Wald statistic, respectively $\mathcal{O}\big(nd^3p^3\big)$, $\mathcal{O}\big(nd^4p^3\big)$

and $\mathcal{O}\big(nd^6p^3\big)$.

Likewise, the Wald-type procedure is trivial to parallelise over the calculations of $W_{k,n}, k \in [G+p, n-G]$. This will give even greater savings than the Score procedures.

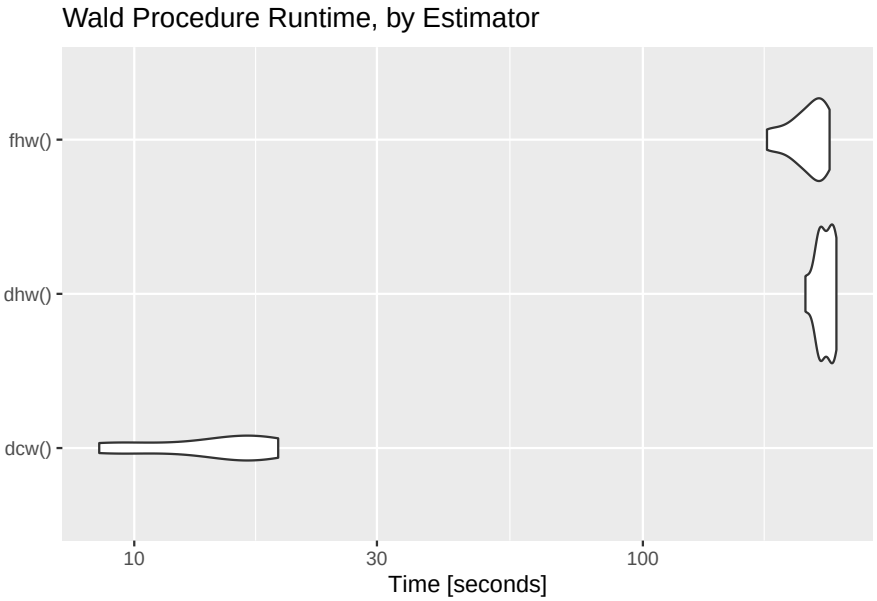


Figure 5.7: Score Procedure Runtime,$n = 2000, d = 5, p = 1$: fhw = Full H, dhw = Diagonal H, dcw = Diagonal C

The above runtime empirical distributions, based on simulated data of size $n = 2000, d = 5$ modelled with order $p = 1$, show that DiagonalC runs at around 10 times the speed of the H-based estimators, and is competitive in speed with the Score-type tests. Tests using DiagonalH and FullH took around 30 times as long as the Score-type test estimators (see previous subsection). Hence, a more efficient implementation in parallel and compiled language would be very useful for practice.

5.3 Data Analysis

Recall our motivation of applying causal network analysis to financial risk data, as outlined in 1.3. While we have data available for 30 banks, we know from our simulation study that the methods we have developed are most appropriate for data of moderate dimension, so we select a subset of three banks with minimal missing data.

By inspection of figure 5.8, the series appear to be integrated. Applying the augmented

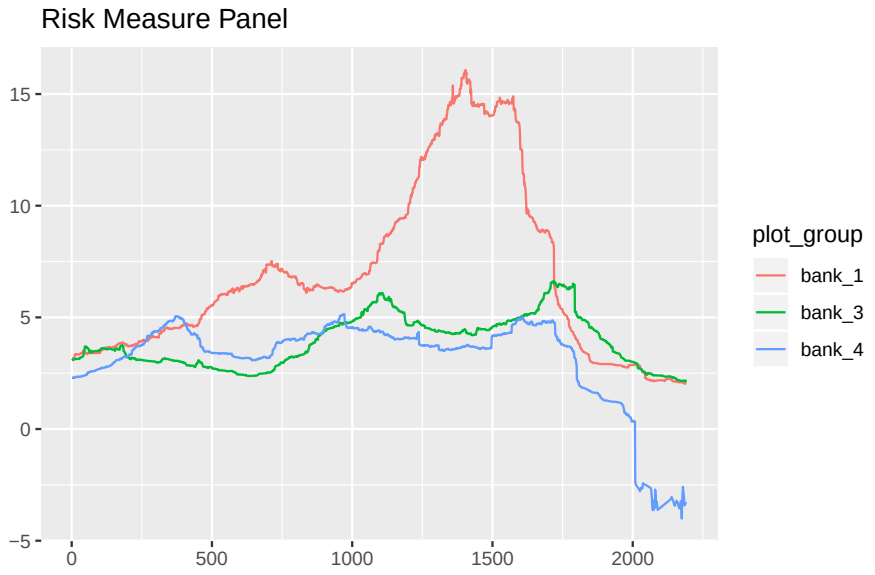


Figure 5.8: A panel of risk measures for three banks

Dickey-Fuller test for unit roots, with both intercept and trend, and where lags are selected by the Akaike Information Criterion (AIC), all three series fail to reject the null hypothesis of non-stationarity. With the aim of modelling these with a VAR model, we take the difference of each series (figure 5.9).

The AIC indicates a model of order $p = 2$ is most appropriate, giving total dimensionality of $3 \times 7 = 21$. Fitting this model to the data gives residuals with autocorrelation functions as shown in figure 5.10. Bank 1 seems to exhibit some small but significant correlation with its' own previous time lags, indicating our model may be misspecified, although no other series show significant autocorrelation.

We might take this as evidence that the original series are cointegrated, so a vector error correction model [Lüt05, Chapter 6] might perform better at modelling this series. Our method is designed to account for misspecification, however; the VAR model should still capture the second-order structure sufficiently to form a causal network.

Running each test with the DIAG-H estimator, significance level $\alpha = 0.05$, and bandwidth $G = 200$, we get that the Score test does not reject the null hypothesis of no changes in autoregression (figure 5.11) but the Wald test does reject the null hypothesis, locating a change-point at time $k = 1829$ (figure 5.12). A visual inspection of 1.3 does show a sharp dip in all three series after this time, though it is hard to verify if this is a change in autoregression. This estimated change point corresponds to the 1st January 1985; this

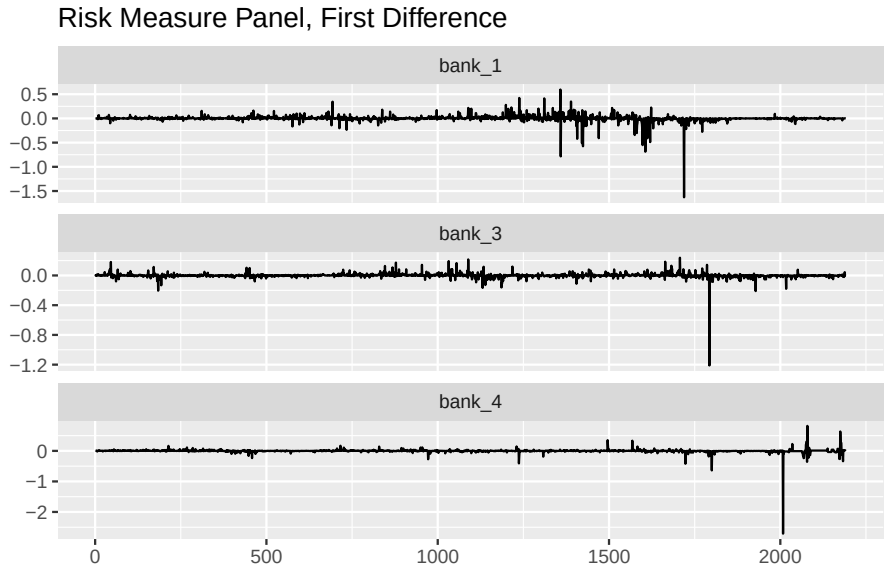


Figure 5.9: A panel of risk measures for three banks, first-order difference

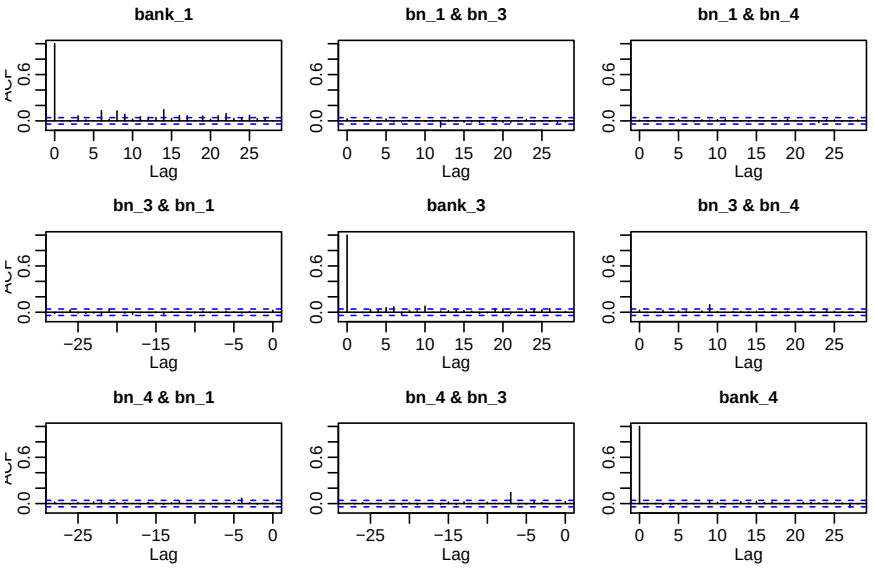


Figure 5.10: Autocorrelation of residuals, $p = 2$

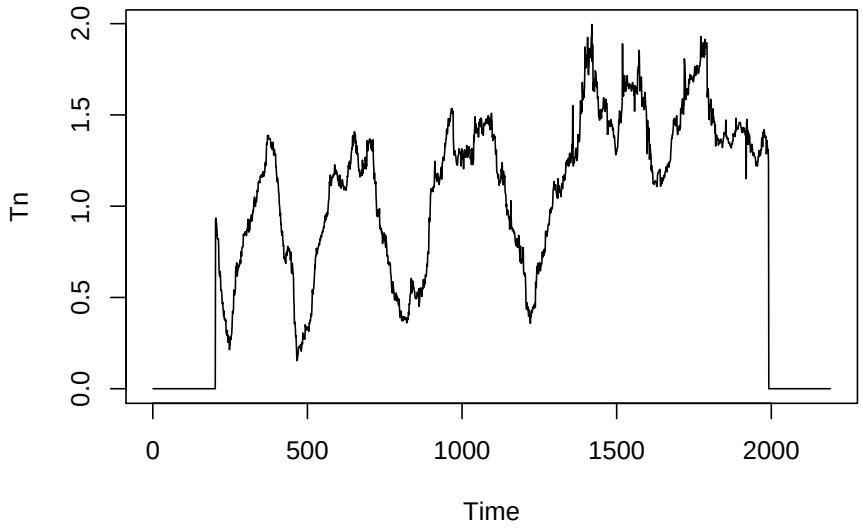


Figure 5.11: Score test with DIAG-H estimator, $\alpha = 0.05, G = 200$

may coincide with a policy intervention or macroeconomic event, though brief research has provided no explanation for this.

The two statistics seem to exhibit similar shape, but the Wald statistic spikes far higher around the estimated change point. The reason for this divergence in behaviour is unclear; it may be that the change is small with respect to the global parameter estimate, so is invisible to the Score test, but is large locally, causing the sharp growth in the Wald statistic.

We compare causal networks inferred from the series, using the truncating lasso from 4.10. We use error rates $\alpha = 0.3$ and $\beta = 0.05$ to encourage density, and over-state the dimensionality as $p = 4$, greater than that dictated by the information criterion ($p = 2$), to examine the truncation effect. First, we estimate a network using the entire dataset, as dictated by the lack of changes in the Score test. Figure 5.13 has many short-term dependencies over the first two lags, and dependency on the first channel from lags three and four; this reflects the autocorrelation in the residuals we saw in figure 5.10.

Then, we estimate networks using the data before and after the change point $k = 1829$, as dictated by the results of the Wald test. The network in the first regime (5.14) has dense causal links, over all four lags. The network in the second regime (5.15) has only three causal links, all over the first two lags. This indicates a change in structure from high interdependence to low interdependence. We should consider here, however, the dearth of data relative to the dimensionality of the parameter space in the second regime; the accuracy

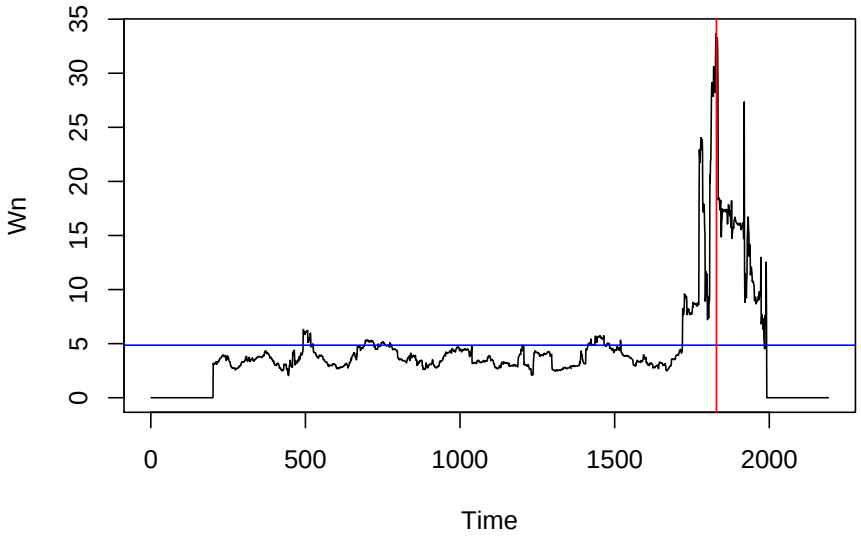


Figure 5.12: Wald test with DIAG-H estimator, $\alpha = 0.05, G = 200$

of our estimate is likely to be compromised as a result.

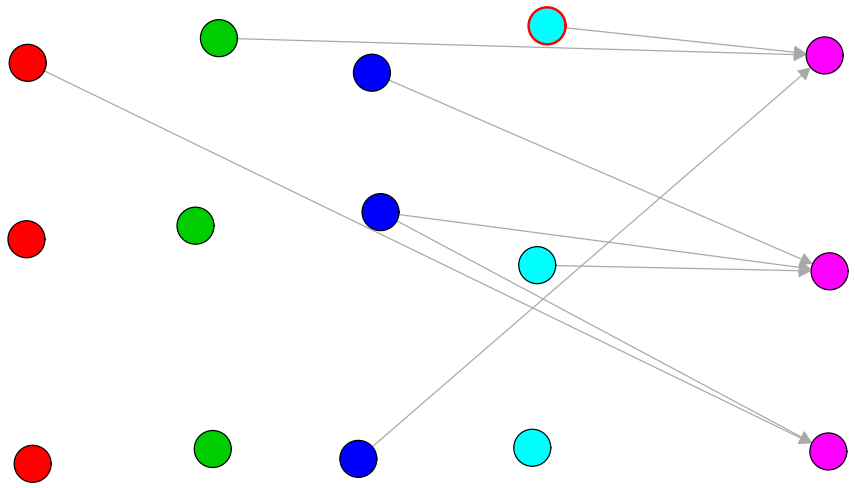


Figure 5.13: Network estimated over entire dataset

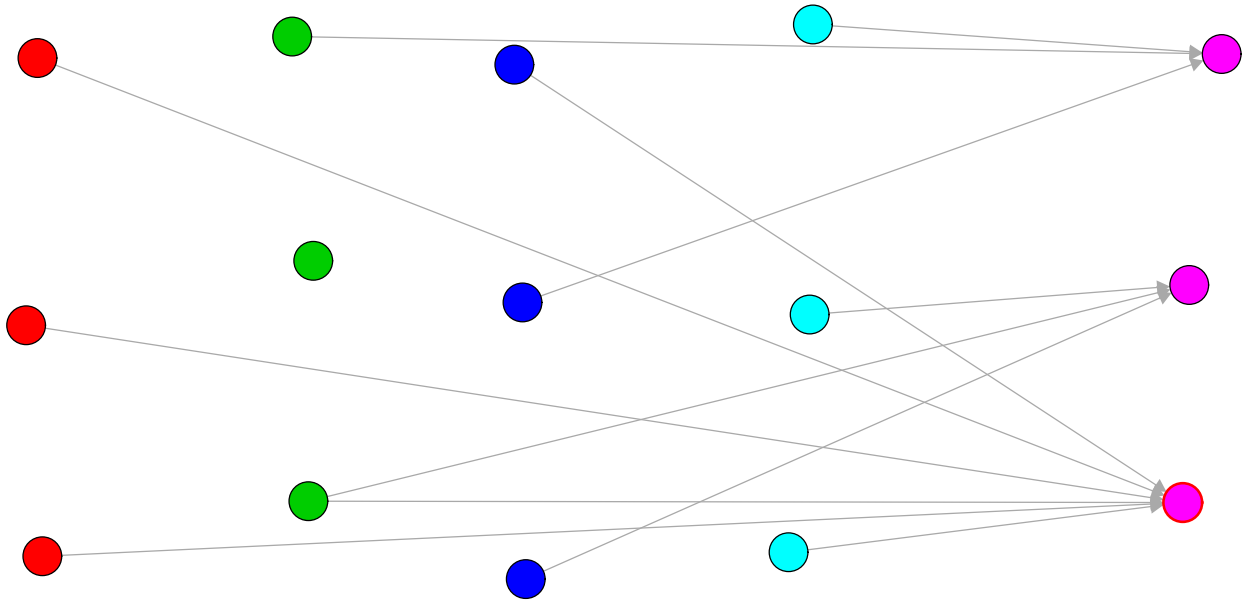


Figure 5.14: Network estimated before changepoint $k = 1829$

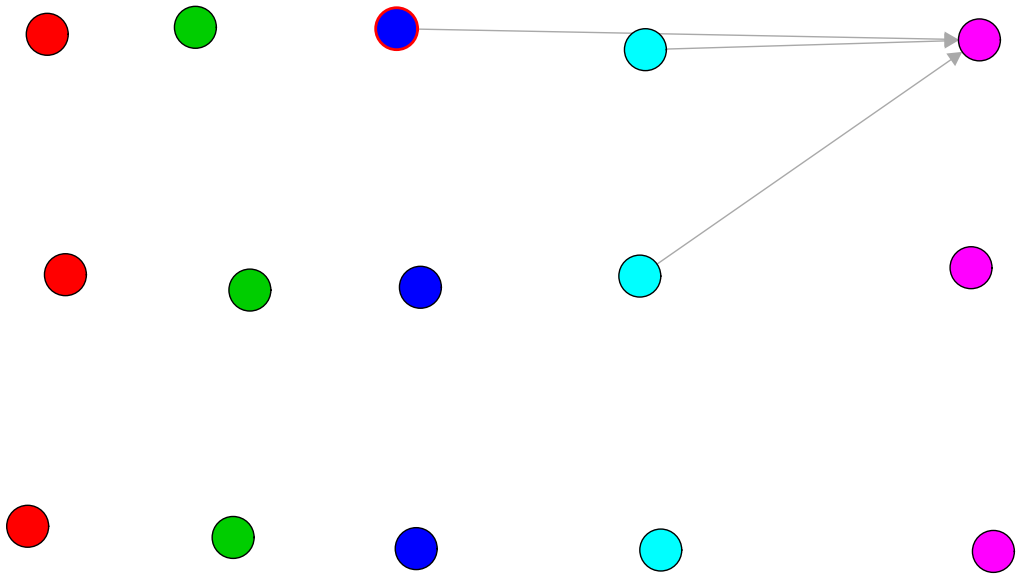


Figure 5.15: Networks estimated after changepoint $k = 1829$

Chapter 6

Conclusion

To conclude the report, we collect our findings and discuss future work.

6.1 Findings

We have derived a moving sum method for detecting and locating multiple change points in multivariate time series, as approximated by a vector autoregressive model. Theoretical results on asymptotic power and consistency in the number and location of changes, based on two types of statistic, were given in chapter 3. A method for inferring causal networks from stationary segments was discussed in chapter 4. Strong finite sample performance is demonstrated in low-dimensional monte-carlo simulations in chapter 5, and consideration was given to runtime and complexity. Analysis of financial risk data, given as our motivation, was carried out, demonstrating the ability of the method to capture changes in second-order structure.

6.2 Further Work

There are various directions this work can be taken further.

In [KMO15], some focus is given to how the model captures misspecification through bootstrap methods. It would be interesting to understand how ours captures misspecification - does ours account for simple generalisations such as changes in mean (step changes or deterministic/stochastic drift), and how does it perform for data from more complicated linear models such as VARMA? Simulation studies from these generating processes would be insightful.

Moreover, the data in chapter 5 might suggest that the series are cointegrated. We accounted for this by testing each for unit roots, differencing, and analysing the derivative series for change-points in autoregression. Could we extend the model to include the original cointegrated series, using error correction models? How would this compare to our current method in terms of describing structure?

We should urgently consider replacing the global variance estimator 3.42 used in 3.40, which does a poor job and almost certainly caused the results seen in 5.1. Adapting the LOCAL-1 estimator discussed in [Rec19] is an obvious solution to this.

An efficient implementation for the Wald procedure would make this competitive with other algorithms in run time, particularly for higher dimensions, and we should use the "embarassingly parallel" nature of the problem to do this. A full implementation as a package would aid practitioners.

Perhaps of most interest is to account for high dimensionality by proposing new estimators for Σ or Γ which do not require inversion, and to give theoretical results for this (especially in light of the slow convergence of the Gumbel distribution, as discussed in 3.5). This could involve assuming some low-dimensional structure.

A network estimation procedure that does not impose sparsity on the network might be more suitable for our motivating application; simulation studies for the truncating lasso method with dense networks would complement the sparse simulations in [SM10]. Truncation based on the ridge estimator is one possible way of doing this.

Finally, the data analysis in 5.3 should be extended to consider model misspecification, and to use all dimensions of the $d = 30$ dataset.

Chapter 7

Appendix

R code for the current implementation is available at https://github.com/Dom-Owens-UoB/VAR_MOSUM

Bibliography

[AC17] Samaneh Aminikhanghahi and Diane J Cook. A survey of methods for time series change point detection. *Knowledge and information systems*, 51(2):339–367, 2017.

[ADL18] Fumiya Akashi, Holger Dette, and Yan Liu. Change-point detection in autoregressive models with no moment assumptions. *Journal of Time Series Analysis*, 39(5):763–786, 2018.

[AG02] Elena Andreou and Eric Ghysels. Detecting multiple breaks in financial market volatility dynamics. *Journal of Applied Econometrics*, 17(5):579–600, 2002.

[AH13] Alexander Aue and Lajos Horváth. Structural breaks in time series. *Journal of Time Series Analysis*, 34(1):1–16, 2013.

[AMB+17] Siti Nur Afiqah Mohd Arif, Mohamad Farhan Mohamad Mohsin, Azuraliza Abu Bakar, Abdul Razak Hamdan, and Sharifah Mastura Syed Abdullah. Change point analysis: a statistical approach to detect potential abrupt change. *Jurnal Teknologi*, 79(5), 2017.

[ARK18] Syed Sazzad Ahmed, Swarup Roy, and Jugal K Kalita. Assessing the effectiveness of causality inference methods for gene regulatory networks. *IEEE/ACM transactions on computational biology and bioinformatics*, 2018.

[BA99] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *science*, 286(5439):509–512, 1999.

[Ber86] Ben S Bernanke. Alternative explanations of the money-income correlation, 1986.

[BGLP12] Monica Billio, Mila Getmansky, Andrew W Lo, and Loriana Pelizzon. Econometric measures of connectedness and systemic risk in the finance and insurance sectors. *Journal of financial economics*, 104(3):535–559, 2012.

[BH95] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

[BSM15] Sumanta Basu, Ali Shojaie, and George Michailidis. Network granger causality with inherent grouping structure. *The Journal of Machine Learning Research*, 16(1):417–453, 2015.

[C.08] Kirch C. Introduction to change-point analysis lecture notes, 2008.

[CF15] Haeran Cho and Piotr Fryzlewicz. 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, 2015.

[CH97] Miklos Csorgo and Lajos Horváth. *Limit theorems in change-point analysis*. John Wiley & Sons Chichester, 1997.

[E+89] JHJ Einmahl et al. On the standarized empirical process. *Statistica neerlandica*, 43(3):175–179, 1989.

[EK+18] Birte Eichinger, Claudia Kirch, et al. A mosum procedure for the estimation of multiple random change points. *Bernoulli*, 24(1):526–564, 2018.

[F+14] Piotr Fryzlewicz et al. Wild binary segmentation for multiple change-point detection. *The Annals of Statistics*, 42(6):2243–2281, 2014.

[FGP18] Lin Fan, Peter W Glynn, and Markus Pelger. Change-point testing and estimation for risk measures in time series. *arXiv preprint arXiv:1809.02303*, 2018.

[FL07] Paul Fearnhead and Zhen Liu. On-line inference for multiple changepoint problems. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(4):589–605, 2007.

[GLP17] Domenico Giannone, Michele Lenza, and Giorgio E Primiceri. Economic predictions with big data: The illusion of sparsity. 2017.

[Gra88] Clive WJ Granger. Some recent development in a concept of causality. *Journal of econometrics*, 39(1-2):199–211, 1988.

[HMPYP19] Oscar Hernan Madrid Padilla, Yi Yu, and Carey E Priebe. Change point localization in dependent dynamic nonparametric random dot product graphs. *arXiv*, pages arXiv–1911, 2019.

[IT94] Carla Inclan and George C Tiao. Use of cumulative sums of squares for retrospective detection of changes of variance. *Journal of the American Statistical Association*, 89(427):913–923, 1994.

[KFE12] Rebecca Killick, Paul Fearnhead, and Idris A Eckley. Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association*, 107(500):1590–1598, 2012.

[KLGS09] Gary Koop, Roberto Leon-Gonzalez, and Rodney W Strachan. On the evolution of the monetary policy transmission mechanism. *Journal of Economic Dynamics and Control*, 33(4):997–1017, 2009.

[KMO15] Claudia Kirch, Birte Muhsal, and Hernando Ombao. Detection of changes in multivariate time series with application to eeg data. *Journal of the American Statistical Association*, 110(511):1197–1216, 2015.

[KS09] Yoshinobu Kawahara and Masashi Sugiyama. Change-point detection in time-series data by direct density-ratio estimation. In *Proceedings of the 2009 SIAM International Conference on Data Mining*, pages 389–400. SIAM, 2009.

[KS12] Yoshinobu Kawahara and Masashi Sugiyama. Sequential change-point detection based on direct density-ratio estimation. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 5(2):114–127, 2012.

[LLR+09] Céline Lévy-Leduc, François Roueff, et al. Detection and localization of change-points in high-dimensional network traffic data. *The Annals of Applied Statistics*, 3(2):637–662, 2009.

[LM15] Thomas A Lubik and Christian Matthes. Time-varying parameter vector autoregressions: Specification, estimation, and an application. *Economic Quarterly*, (4Q):323–352, 2015.

[Lüt05] Helmut Lütkepohl. *New introduction to multiple time series analysis*. Springer Science & Business Media, 2005.

[LYCS13] Song Liu, Makoto Yamada, Nigel Collier, and Masashi Sugiyama. Change-point detection in time-series data by relative density-ratio estimation. *Neural Networks*, 43:72–83, 2013.

[MC07] Nitai D Mukhopadhyay and Snigdhansu Chatterjee. Causality and pathway search in microarray time series experiment. *Bioinformatics*, 23(4):442–449, 2007.

[MdB13] George Michailidis and Florence d'Alché Buc. Autoregressive models for gene regulatory network inference: Sparsity, stability and causality issues. *Mathematical biosciences*, 246(2):326–334, 2013.

[Muh13] Birte Chantal Simone Muhsal. *Change-point methods for multivariate autoregressive models and multiple structural breaks in the mean*. PhD thesis, Verlag nicht ermittelbar, 2013.

[NHZ16] Yue S Niu, Ning Hao, and Heping Zhang. Multiple change-point detection: A selective overview. *Statistical Science*, pages 611–623, 2016.

[PC06] Jianmin Pan and Jiahua Chen. Application of modified information criterion to multiple change point problems. *Journal of multivariate analysis*, 97(10):2221–2241, 2006.

[PDK+20] George P Papaioannou, Christos Dikaiakos, Christos Kaskouras, George Evangelidis, and Fotios Georgakis. Granger causality network methods for analyzing cross-border electricity trading between greece, italy, and bulgaria. *Energies*, 13(4):900, 2020.

[Per05] Roberto Perotti. Estimating the effects of fiscal policy in oecd countries. 2005.

[Pic85] Dominique Picard. Testing and estimating change-points in time series. *Advances in applied probability*, 17(4):841–867, 1985.

[PP15] Andrey Pepelyshev and Aleksey S Polunchenko. Real-time financial surveillance via quickest change-point detection methods. *arXiv preprint arXiv:1509.01570*, 2015.

[Prá18] Zuzana Prášková. Change point detection in vector autoregression. *Kybernetika*, 54(6):1122–1137, 2018.

[RAM17] Sandipan Roy, Yves Atchadé, and George Michailidis. Change point estimation in high dimensional markov random-field models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(4):1187–1206, 2017.

[Rec19] Kerstin Reckrühm. *Estimating multiple structural breaks in time series-a generalized MOSUM approach based on estimating functions*. PhD thesis, 2019.

[Set08] Anil K Seth. Causal networks in simulated neural systems. *Cognitive neurodynamics*, 2(1):49–64, 2008.

[She]

[SKTK19] Elsa Siggiridou, Christos Koutlis, Alkiviadis Tsimpiris, and Dimitris Kugiumtzis. Evaluation of granger causality measures for constructing networks from multivariate time series. *Entropy*, 21(11):1080, 2019.

[SM10] Ali Shojaie and George Michailidis. Discovering graphical granger causality using the truncating lasso penalty. *Bioinformatics*, 26(18):i517–i523, 2010.

[SS75] Ashish Sen and Muni S Srivastava. On tests for detecting change in mean. *The Annals of statistics*, pages 98–108, 1975.

[SS17] Abolfazl Safikhani and Ali Shojaie. Joint structural break detection and parameter estimation in high-dimensional non-stationary var models. *arXiv preprint arXiv:1711.07357*, 2017.

[STR10] Yunus Saatçi, Ryan D Turner, and Carl Edward Rasmussen. Gaussian process change point models. In *ICML*, pages 927–934, 2010.

[SW96] James H Stock and Mark W Watson. Evidence on structural instability in macroeconomic time series relations. *Journal of Business & Economic Statistics*, 14(1):11–30, 1996.

[SW01] James H Stock and Mark W Watson. Vector autoregressions. *Journal of Economic perspectives*, 15(4):101–115, 2001.

[TFP+16] Gaëtan Texier, Magnim Farouh, Liliane Pellegrin, Michael L Jackson, Jean-Baptiste Meynard, Xavier Deparis, and Hervé Chaudet. Outbreak definition by change point analysis: a tool for public health decision? *BMC medical informatics and decision making*, 16(1):33, 2016.

[TOV19] Charles Truong, Laurent Oudre, and Nicolas Vayatis. Selective review of offline change point detection methods. *Signal Processing*, page 107299, 2019.

[TRBK06] Alexander G Tartakovsky, Boris L Rozovskii, Rudolf B Blazek, and Hongjoong Kim. A novel approach to detection of intrusions in computer networks via adaptive sequential and batch-sequential change-point detection methods. *IEEE transactions on signal processing*, 54(9):3372–3382, 2006.

[VA16] Zeev Volkovich and Renata Avros. Text classification using a novel time series based methodology. In *KES*, pages 53–62, 2016.

[VdV00] Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

[Ven93] Ennapadam Seshan Venkatraman. *Consistency results in multiple change-point problems.* PhD thesis, 1993.

[Vos81] Lyudmila Yur'evna Vostrikova. Detecting "disorder" in multidimensional random processes. 259(2):270–274, 1981.

[Web17] S Weber. *Change-point procedures for multivariate dependent data.* PhD thesis, PhD thesis, Karlsruhe Institute of Technology (KIT), 2017. URN: urn: nbn: de . . . , 2017.

[WS18] Tengyao Wang and Richard J Samworth. High dimensional change point estimation via sparse projection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(1):57–83, 2018.

[WYR18] Daren Wang, Yi Yu, and Alessandro Rinaldo. Optimal change point detection and localization in sparse dynamic networks. *arXiv preprint arXiv:1809.09602*, 2018.

[WYRW19] Daren Wang, Yi Yu, Alessandro Rinaldo, and Rebecca Willett. Localizing changes in high-dimensional vector autoregressive processes. *arXiv preprint arXiv:1909.06359*, 2019.

[Yao88] Yi-Ching Yao. Estimating the number of change-points via schwarz'criterion. *Statistics & Probability Letters*, 6(3):181–189, 1988.