Novel Methods for Anomaly Detection

Alexander T. M. Fisch BA, MMath, MRes

Submitted for the degree of Doctor of

Philosophy at Lancaster University.

July 2020

STOR-i

excellence with impact

Abstract

Anomaly detection is of increasing importance in the data rich world of today. It can be applied to a broad range of challenges ranging from fault detection to fraud prevention and cyber-security. Many of these application require algorithms which are very scalable, as well as accurate, due to large data volumes and/or limited computational resources.

This thesis contributes three novel approaches to the field of anomaly detection. The first contribution, Collective And Point Anomalies (CAPA) detects and distinguishes between both collective and point anomalies in linear time. The second contribution, MultiVariate Collective And Point Anomalies (MVCAPA) extends CAPA to the multivariate setting. The third contribution is a novel particle based kalman filter which detects and distinguished between additive outliers and innovative outliers.

Acknowledgements

First and foremost, I would like to thank my Lancaster supervisors Paul Fearnhead and Idris Eckley for their help and guidance throughout my PhD. I am especially grateful for their willingness to wade through the rather long proofs. I would also like to thank Dan Grose, for helping with packaging up the various methods developed as part of this thesis. Thanks also to the wider time-series, changepoint, and computational stats community for stimulating discussions. Thanks also to my external examiner Jean-Philippe Vert and my internal examiner Azadeh Khaleghi for a very stimulating and thought provoking viva.

I am very grateful to the STOR-i CDT, EPSRC, and British Telecommunications PLC (BT) for funding this thesis. BT and its employees, especially Trevor Burbridge, Kjeld Jensen, and David Yearling, have also been very helpful by providing data examples and helpful discussions which have inspired much of this work. I am also very grateful for having been given the opportunity to visit their office on several occasions to get my hands dirty on real data.

On a more personal level, I would like to thank my wife, Solène, for bearing with me – and with the "brilliant ideas" that seem to come at 1am. I would also like to

thank the wider STOR-i and StatScale community for providing such a pleasant work environment. AMDG.

Declaration

I declare that the work in this thesis has been done by myself and has not been submitted elsewhere for the award of any other degree.

Chapter 3 is currently under review with Data Mining and Knowledge Discovery.

Chapter 4 is currently under review with the Journal of Computational and Graphical Statistics.

Chapter 5 is currently under review with IEEE Transactions on Signal Processing.

Alexander T. M. Fisch

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Anomaly detection is an area of considerable importance and has been subject to increasing attention in recent years. This is due to the wide range of applications the field lends itself to. Examples include fault detection (Theissler, 2017; Zhao et al., 2018), fraud prevention (Ahmed et al., 2016), and cyber security (Goh et al., 2017). The ubiquity of sensors and the emergence of the Internet of Things (IoT) has lead to the detection of anomalies in streaming data to emerge as a new and critical challenge.

One important aspect of anomalies is that they can come in different guises. One classification was offered by Chandola et al. (2009) who distinguish between global, contextual, and collective anomalies. Here global anomalies are single points which fall outside the general pattern of the data while contextual anomalies fall outside their local data pattern. Collective anomalies on the other hand are defined to be a sequence of observations which are not necessarily anomalous by themselves but together form an anomalous pattern. The Kalman filtering literature similarly distinguishes between punctual anomalies called additive outliers which affect the observations only and

persisting anomalies called innovative outliers which affect the system (Ruckdeschel et al., 2014).

This thesis introduces novel statistical methods for detecting and distinguishing between different types of anomalies in a computationally efficient manner. All algorithms have been inspired by anomalies observed in telecommunications network data.

The remainder of the thesis is organised as follows: We begin by reviewing relevant background in Chapter 2. The focus of this chapter will lie on robust statistics, Kalman filtering and changepoint methods.

In Chapter 3, we propose a new epidemic changepoint based algorithm which can detect and distinguish between collective an point anomalies in empirically linear time. We call the algorithm Collective And Point Anomalies (CAPA), theoretically prove its consistency, empirically evaluate it against competing methods, and apply it to monitoring machine temperature data and to exoplanet detection.

We propose extension of CAPA to the multivariate setting, which we call multivariate CAPA (MVCAPA), in Chapter 4. Crucially, the proposed methodology allows for related anomalies in different components to have imperfect alignment across time. We theoretically show MVCAPA's consistency and that it is able to optimally detect sparse anomalies, affecting only a few components, as well as dense anomalies, affecting a large subset of components.

A novel Kalman filter which is robust to both additive and innovative outliers is proposed in chapter 5. The proposed methodology, which we call Computationally Efficient Bayesian Anomaly detection by Sequential Sampling (CE-BASS) is fully

online, very scalable, and shown to compare favourably with other robust filters. CE-BASS is applied to both real router data and a benchmark dataset.

We discuss our contribution to the literature as well as potential areas of further research in Chapter 6.

Chapter 2

Background and Literature Review

In this chapter, we review some of the background literature relevant to this thesis. We will begin by reviewing the pertinent definitions and concepts in robust statistics in Section 2.1 before reviewing the robust Kalman Filter literature in Section 2.2 and the changepoint literature in 2.3. We purposefully omit reviewing the vast range of anomaly detection approaches proposed by the machine learning and computer science community and instead refer to the excellent reviews which can be found in Chandola et al. (2009) and Pimentel et al. (2014), as well as the more recent papers by Lavin and Ahmad (2015), Talagala et al. (2019), Ahmad et al. (2017), and references therein.

2.1 Robust Statistics

When the distribution of the typical data is known, detecting anomalies becomes almost trivial. However, inferring the distribution of the typical data from a data

set which is potentially polluted by anomalies is difficult as outliers can significantly affect the inference procedure. Take the sample mean for example: A single outlier has the potential of irreversibly polluting this statistic, hence making it useless for the purpose of detecting anomalies. Other commonly used statistics such as the sample variance, regression coefficients, sample covariance matrices, etc. are equally vulnerable to outliers.

Observations like the above have motivated the field of robust statistics. Robust statistics aim to bound the influence any single data point can have on the statistic while equally trying to achieve an efficiency which is close to that of maximum likelihood estimators. We will review the key concepts of the field in this section, as it provides important background to subsequent sections and chapters.

2.1.1 Definitions Around Robustness

The main concept in robustness is the influence function, first introduced by Hampel (1968). For a given statistic, $T()$, which is a functional mapping a cumulative density function $F()$ to a scalar, the influence function is defined via the Gateaux derivative

$$IF(x, F) = \lim_{\epsilon \to 0} \frac{T((1-\epsilon)F + \epsilon \triangle_x) - T(F)}{\epsilon}. \tag{2.1.1}$$

Here, $\triangle_x(y) = \mathbb{I}(y \geq x)$ denotes the CDF of point mass at x. It captures how much deviations from the assumed distribution can affect the statistic. A statistic is said to be robust if the influence function is bounded (Hampel et al., 1986).

For example, the mean is defined by the functional $T(G) = \int x dG(x)$. For a distribution F with mean μ the influence function is therefore given by $x - \mu$. Conversely,

the median is defined by the functional $T(G) = G^{-1}(1/2)$. Its influence function for a distribution with PDF f and median m is therefore

$$\frac{1}{f(m)}(\mathbb{I}(x < m) - \frac{1}{2}). \tag{2.1.2}$$

This confirms the intuition that the mean is not a robust statistic whilst the median is robust.

Another important metric in robust statistics is the breakdown point which measures the proportion of data that can be anomalous whilst guaranteeing that the difference between the statistic on the polluted data and the statistic on the unpolluted data is finite. Formally it can be defined as the largest γ such that

$$\sup_{G,\epsilon<\gamma} |T((1-\epsilon)F + \epsilon G) - T(F)| < \infty.$$

It can be shown to be 0 for the mean and $\frac{1}{2}$ for the median. The median therefore achieves the theoretical maximal breakdown point. Indeed, it can be shown that the breakdown point must be less or equal to $\frac{1}{2}$ – otherwise there would be identifiability issues (Hampel et al., 1986).

Finally, the asymptotic efficiency is an important metric. Robust estimators achieve robustness by ignoring or down-weighting parts of the data. They are therefore less efficient than the maximum likelihood estimator (MLE). This is captured by the asymptotic efficiency which is defined as the asymptotic variance of the MLE divided by that of the robust estimator. For instance, the median achieves and asymptotic efficiency of $2/\pi$ for normal data (Hampel et al., 1986).

2.1.2 M-Estimators

The median mentioned in the previous section is an example of an L-estimator, where L is short for location. Other examples of such L-estimators are the α-trimmed mean which neglects the $\alpha\%$ of lowest and highest observations. L-estimators for the variance also exist, such as the inter-quartile range, the α-trimmed variance, or the median absolute deviation (Hampel et al., 1986).

However, a different class of estimators called M-estimators, going back to Huber (1964b), tend to be preferred due to their better efficiency (Jurečková and Picek, 2005). M-estimators are obtained through a minimisation problem: For a suitable cost function $\rho(\cdot, \cdot)$, the M-estimator of a parameter θ is obtained from observations $x_1, ..., x_n$ by minimising

$$\left(\sum_{i=1}^{n} \rho(x_i, \theta) \right)$$

with respect to θ. This can be viewed as a generalisation of the MLE since taking $\rho(\cdot, \cdot)$ to be the negative log-likelihood recovers the MLE.

A range of cost functions $\rho()$ has been proposed with the aim of achieving robustness. One such function is Tukey's bi-weight loss (Tukey, 1960), defined as

$$\rho(x_i, \theta) = \begin{cases} (x_i - \theta)^2 & |x_i - \theta| < h, \\ h^2 & |x_i - \theta| \geq h, \end{cases}$$

which can be shown to bound the influence function at h. Here, the threshold parameter h governs the trade-off between robustness and efficiency – the higher h, the more efficient the estimate becomes, the lower h, the more robust.

Another robust loss function is Huber loss (Huber, 1964a) which is defined as:

$$\rho(x_i, \theta) = \begin{cases} (x_i - \theta)^2 & |x_i - \theta| < h, \\ 2h|x_i - \theta| - h^2 & |x_i - \theta| \geq h. \end{cases}$$

It can be shown to bound the influence function at h. Furthermore, it provides a trade-off between robustness and efficiency which is in some sense pareto-optimal. Indeed, Hampel (1968) showed that Huberisation, i.e. truncating the influence function of an MLE achieves maximal efficiency for a given breakdown point.

Other commonly used robust loss functions include the Dynamic Covariance Scaling (Agarwal et al., 2013) and the log-likelihood of the t-distribution distribution (Agamennoni et al., 2012).

2.2 Kalman Filtering Approaches

An important challenge in many anomaly detection applications is that sequential or online processing of time series data is often a requirement. The Kalman filter first proposed by Kalman (1960) uses a latent variable model which provides a convenient way of processing new observations at a fixed computational cost. Furthermore, having processed observations $\mathbf{Y}_1, ..., \mathbf{Y}_n$, the Kalman filter can return a mean and variance estimate for $\mathbf{Y}_{n+1}$, making it, in principle, very well suited for anomaly detection: A Mahalanobis distance can be computed and an anomaly declared if it exceeds a predefined quantile of the χ^2-distribution. However, the Gaussian noise model used by the classical Kalman filter makes it vulnerable to outliers. A range of outlier robust Kalman filters, more suitable to anomaly detection, has therefore been proposed.

We will begin this section by reviewing the Kalman filter and discuss the two types of anomalies which can affect it. We will then discuss the most common approaches aimed at robustifying the Kalman filter namely filters which use t-distributed noise in conjunction with Variational Bayes (VB), filters which use Huberisation, and filters which use heavy tailed noise in conjunction with other methods to maintain approximations to the posterior.

2.2.1 The Classical Kalman Filter

The Kalman filter goes back to the seminal paper Kalman (1960). It considers a model in which observations $\mathbf{Y}_1, ..., \mathbf{Y}_n$ are underpinned by hidden variables $\mathbf{X}_1, ..., \mathbf{X}_n$ in the following manner:

$$\mathbf{Y}_t = \mathbf{C}\mathbf{X}_t + \boldsymbol{\eta}_t \qquad \mathbf{X}_t = \mathbf{A}\mathbf{X}_{t-1} + \boldsymbol{\epsilon}_t.$$

Here the noise processes $\boldsymbol{\eta}_t \overset{i.i.d}{\sim} N(0, \mathbf{R})$ and $\boldsymbol{\epsilon}_t \overset{i.i.d}{\sim} N(0, \mathbf{Q})$ are independent for $t \geq 1$ and a prior $\mathbf{X}_0 \sim N(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$ is put on the initial state.

The main feature of the Kalman filter is that it allows for online updates of the hidden state. When $\mathbf{X}_t|\mathbf{Y}_t, ..., \mathbf{Y}_1 \sim N(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)$, it can be shown that $\mathbf{X}_{t+1}|\mathbf{Y}_{t+1}, ..., \mathbf{Y}_1 \sim N(\boldsymbol{\mu}_{t+1}, \boldsymbol{\Sigma}_{t+1})$, where the new mean and variance, $\boldsymbol{\mu}_{t+1}$ and $\boldsymbol{\Sigma}_{t+1}$ are obtained from

$\boldsymbol{\mu}_t$ and $\boldsymbol{\Sigma}_t$ through the update equations

$$\begin{aligned}
\hat{\boldsymbol{\mu}} &= \mathbf{A}\boldsymbol{\mu}_t \\
\hat{\boldsymbol{\Sigma}} &= \mathbf{A}\boldsymbol{\Sigma}_t\mathbf{A}^T + \mathbf{Q} \\
\mathbf{z} &= \mathbf{Y}_t - \mathbf{C}\hat{\boldsymbol{\mu}} \\
\mathbf{K} &= \left(\mathbf{C}\hat{\boldsymbol{\Sigma}}\mathbf{C}^T + \mathbf{R}\right)^{-1}\mathbf{C}\hat{\boldsymbol{\Sigma}} \\
\boldsymbol{\mu}_{t+1} &= \hat{\boldsymbol{\mu}} + \mathbf{K}^T\mathbf{z} \\
\boldsymbol{\Sigma}_{t+1} &= (\mathbf{I} - \mathbf{K}\mathbf{C}^T)\hat{\boldsymbol{\Sigma}}.
\end{aligned} \tag{2.2.1}$$

As mentioned in the introduction to this section, the Gaussian noise model used by the Kalman filter make it very vulnerable to outliers. One particular challenge is that two types of outliers can occur: additive outliers (Ruckdeschel et al., 2014), sometimes called observational outliers (Gandhi and Mili, 2009), affect the process $\boldsymbol{\eta}_t$. Their impact is limited to just one time point. Conversely, innovative (Ruckdeschel et al., 2014), or process (Huang et al., 2017) outliers, affect the process $\boldsymbol{\epsilon}_t$. Their impact can potentially affect many observations to come. Robustness against just one of these two types of outliers typically make the filter more vulnerable to the other. For instance, additive outlier robust filters tend to update the hidden variables even less than the classical Kalman filter when encountering an innovative outlier (Ruckdeschel et al., 2014).

2.2.2 Variational Bayes and t-Distributed Noise

A number of filters which achieve robustness to outliers by assuming t-distributed noise processes $\boldsymbol{\eta}_t$ and/or $\boldsymbol{\epsilon}_t$ has been proposed. For example, Agamennoni et al.

(2011) assumed t-distributed noise $\boldsymbol{\eta}_t$ to achieve robustness against additive outliers. The authors use the conditional noise model $\boldsymbol{\eta}_t \overset{i.i.d}{\sim} N(0, \mathbf{S}_t)$, where $\mathbf{S}_t^{-1} \sim \mathcal{W}(\mathbf{R}^{-1}/s, s)$. Here, $\mathcal{W}()$ denotes the Wishart distribution, which generalises the Gamma distribution to the multivariate setting. Whilst this model achieves robustness to additive outliers, there is no longer a tractable filter. Indeed, given a Gaussian prior $\mathbf{X}_t|\mathbf{Y}_t, ..., \mathbf{Y}_1 \sim N(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)$, the posterior $\mathbf{X}_{t+1}|\mathbf{Y}_{t+1}, ..., \mathbf{Y}_1$ is no longer Gaussian.

Variational Bayes is often used to obtain a Gaussian approximation to the posterior. It finds the Normal distribution which minimises the Kullback-Leibler divergence with the posterior distribution. In conjunction with t-distributed noise, this is often relatively easy to do. For example, in the model considered by Agamennoni et al. (2011), the posterior can be obtained by initialising $\mathbf{S} = \mathbf{R}$ and iterating the Kalman like equations

$$\begin{aligned}
\mathbf{K} &= \left(\mathbf{C}\hat{\boldsymbol{\Sigma}}\mathbf{C}^T + \mathbf{S}\right)^{-1}\mathbf{C}\hat{\boldsymbol{\Sigma}} \\
\boldsymbol{\mu}_{t+1} &= \hat{\boldsymbol{\mu}} + \mathbf{K}^T\mathbf{z} \\
\boldsymbol{\Sigma}_{t+1} &= \mathbf{K}\mathbf{S}\mathbf{K}^T + (\mathbf{I} - \mathbf{K}\mathbf{C}^T)\hat{\boldsymbol{\Sigma}}(\mathbf{I} - \mathbf{C}\mathbf{K}^T) \\
\mathbf{S} &= \frac{s\mathbf{R} + (\mathbf{Y}_t - \mathbf{C}\boldsymbol{\mu}_{t+1})(\mathbf{Y}_t - \mathbf{C}\boldsymbol{\mu}_{t+1})^T + \mathbf{C}\boldsymbol{\Sigma}_{t+1}\mathbf{C}^T}{s+1}
\end{aligned}$$

until convergence. Similar ideas can be found in the filters proposed by Ting et al. (2007), also robust to additive outliers and Huang et al. (2017, 2019) who propose filters which are robust against both additive and innovative outliers.

2.2.3 Huberisation and Robust Statistics

Approaches inspired by robust statistics can be used to truncate the effect of individual observations to achieve robustness against additive outliers or to truncate the effect of the prior state to achieve robustness against innovative outliers.

An example of such a filter is the additive outlier robust filter using Huberisation proposed by Ruckdeschel et al. (2014). The authors replace

$$\boldsymbol{\mu}_{t+1} = \hat{\boldsymbol{\mu}} + \mathbf{K}^T\mathbf{z}$$

in Equation (2.2.1) by

$$\boldsymbol{\mu}_{t+1} = \hat{\boldsymbol{\mu}} + H(\mathbf{K}^T\mathbf{z}, b).$$

Here $H(x, b)$ denotes x Huberised at level b, which is formally defined as

$$H(x, b) = x \min\left(1, \frac{b}{x}\right).$$

In the same paper, Ruckdeschel et al. (2014) also show that robustness against innovative outliers can be achieved if $\mathbf{C}$ is invertible by replacing

$$\boldsymbol{\mu}_{t+1} = \hat{\boldsymbol{\mu}} + \mathbf{K}^T\mathbf{z}$$

in Equation (2.2.1) by

$$\boldsymbol{\mu}_{t+1} = \hat{\boldsymbol{\mu}} + \mathbf{C}^{-1}\left(\mathbf{z} - H(\left(\mathbf{I} - \mathbf{C}\mathbf{K}^T\right)\mathbf{z}, b)\right).$$

Similar approaches, inspired by robust statistics were used by (Gandhi and Mili, 2009) for robustness against additive and innovative outliers and Chang et al. (2013) for robustness against additive outliers.

2.2.4 Other Approaches

Other approaches using heavy tailed noise and approximating the posterior have also been proposed. Kitagawa (1987) for instance proposed an approach which consists of using splines to approximate the posterior distribution at each time point. The proposed methodology is shown to be able to deal with both additive and innovative outliers but scales poorly with the dimensionality of the problem.

Particle filtering (Fearnhead and Künsch, 2018) to approximate the distribution of x_t has also been proposed. Gordon et al. (1993) proposed to sample from the noise at each time point and give particles weight proportional to the likelihood. However, such approaches are not computationally robust against outliers, as noted by Chang (2014): As outliers become stronger it is less and less likely that an appropriate particle will be sampled. Some particle filters offering computational robustness to specific models (Fearnhead and Clifford, 2003) and to additive outliers (Xu et al., 2013) have therefore been proposed. The filter by Harrison and Stevens (1976) is also often mentioned as being robust to both types of outliers. However, it uses a Gaussian mixture model and is therefore not robust to outliers.

2.3 Changepoint Approaches

Observing that our world is ever-changing, the ancient Greek philosopher Heraclitus claimed that "No man ever steps in the same river twice, for it's not the same river and he's not the same man". It can be assumed that Heraclitus would object to time series being modelled as stationary, on similar grounds. Indeed, data generating

mechanisms often change. One approach of modelling this non-stationarity is via changepoints.

The literature considers two main types of changepoint models: Classical changepoint models, typically only referred to as changepoint models, and epidemic changepoint models. The classical changepoint model, first considered by Page (1954), assumes that there exists a set of time-points at which the data-generating mechanism changes. The epidemic changepoint model, going back to Levin and Kline (1985) according to Yao (1993), assumes that the data follows some typical distribution for most of the time except during certain windows in which it behaves differently. These epidemic changes provide a natural model for collective anomalies.

In what follows, we will begin by reviewing the classical changepoint model as well as some some of the main approaches for changepoint inference in Section 2.3.1. This is mostly for background as the main focus of this section lies on the closely related epidemic changepoint detection problem. In Section 2.3.2, we will then review current approaches for the detection of univariate epidemic changepoints. This will be followed by a discussion of multivariate approaches in Section 2.3.3. For simplicity of exposition, we will focus on the change in mean setting, the frameworks being more general.

2.3.1 Univariate Changepoint Models

Consider a univariate time series $x_1, ..., x_n$. It is said to obey the classical changepoint model if there exists a set of changes $\tau = \{t_0, ..., t_{K+1}\}$, where $0 = t_0 < t_1 ... < t_K \leq$

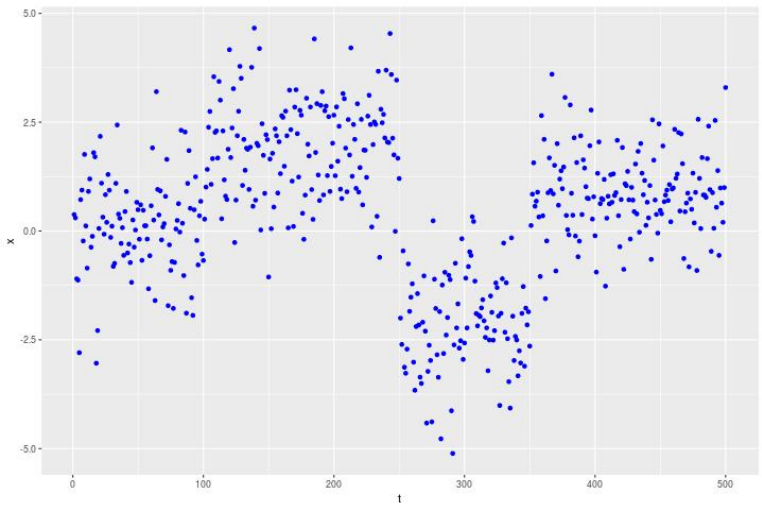


Figure 2.3.1: An example time series with $K = 3$ changes in mean. The first change occurs at $t_1 = 100$, the second at $t_2 = 250$, and the third at $t_3 = 350$.

$t_{K+1} = n$, such that

$$x_t \sim \begin{cases} \mathcal{M}_0 & t_0 < t \leq t_1, \\ \vdots & \qquad\qquad 1 \leq t \leq n. \\ \mathcal{M}_K & t_K < t \leq t_{K+1}, \end{cases}$$

Here $\mathcal{M}_k$ denotes the model obeyed by the data before the kth changepoint. Setting $\mathcal{M}_k = N(\mu_k, \sigma^2)$ gives rise to the change in mean problem with Gaussian noise. This setting has received a considerable amount of attention (Killick et al., 2012; Fryzlewicz, 2014) and will be the main focus of the remainder of this section for simplicity of exposition.

Detecting a Single Change

We will begin by reviewing the case in which at most one change (AMOC) is present. This is because approaches for the AMOC setting can be extended to multiple changes

using the Binary Segmentation algorithm described in the next section. In the AMOC setting, it is of interest to test the null hypothesis of a stationary mean

$$H_0 : \mu_1 = ... = \mu_n,$$

against the alternative hypothesis

$$H_1 : \exists T : 0 \leq T \leq n,\ \mu_1 = ... = \mu_T \neq \mu_{T+1} = ... = \mu_n,$$

which states that a change is present at some time T.

If a single time point, T, had to be investigated, using a log-likelihood ratio statistic would be a natural choice. It can be shown that this statistic is the square of the following cumulative sum (CUSUM) statistic

$$S_T = \left| \sqrt{\frac{n-T}{nT}} \sum_{t=1}^{T} x_t - \sqrt{\frac{T}{n(n-T)}} \sum_{t=T+1}^{n} x_t \right|.$$

It is therefore natural to compute this statistic for all candidate integers $1 \leq T \leq n-1$. If $\max(S_T^2)$ then exceeds a threshold λ a change at time $\operatorname{argmax}(S_T^2)$ is inferred. Otherwise, no changepoint is returned. Suitable choices for λ are discussed in Section 2.3.1. It should be noted that this approach is not restricted to the change in mean setting. It can be extended to any type of change provided an appropriate likelihood ratio tests exists.

Binary Segmentation Approaches

Binary Segmentation, introduced by Scott and Knott (1974) can be used to extend any AMOC changepoint method to infer multiple changepoints. The idea behind Binary Segmentation consists of repeatedly applying an AMOC procedure to segments between inferred changepoints. The algorithm can be summarised as follows:

1. Apply the AMOC procedure to the data x_1,...,x_n and obtain candidate changepoint points T.

2. If the statistic does not exceed the threshold λ stop. Otherwise consider T to be a changepoint

3. Repeat the above procedure on the sequences x_1,...,x_T and x_{T+1},...,x_n

Binary segmentation is computationally very efficient. Indeed, its computational complexity is $O(n\log(n))$. However, it has been shown to yield unsatisfactory results even in the absence of any noise in certain settings (Fryzlewicz, 2014). This has lead to the development of derived methods such as Wild Binary Segmentation (Fryzlewicz, 2014).

Penalised Cost Approaches

An alternative approach to Binary Segmentation consists of minimising a penalised cost (Killick et al., 2012; Jackson et al., 2005). In this approach, each segment of data between two inferred changepoints is allocated a cost, such as twice the negative log-likelihood evaluated at the segment's MLE. Every additional changepoint introduced typically reduces the total cost and therefore incurs a penalty β to avoid over-fitting. The number of changepoints $\hat{K}$ and changepoints $t_1, ..., t_{\hat{K}}$ are then inferred by minimising the penalised cost:

$$\sum_{i=0}^{\hat{K}} \mathcal{C}\left(x_{(t_i+1):t_{i+1}}\right) + \hat{K}\beta$$

The cost function $\mathcal{C}$ is chosen in the light of the model considered. For the change-in mean case, for example, it is natural to use the residual sum of squares, i.e.

$$\mathcal{C}\left(x_{a:b}\right) = \min_{\mu}\left(\sum_{i=a}^{b}(x_i-\mu)^2\right) = \sum_{i=a}^{b} x_i^2 - (b-a+1)\left(\bar{x}_{a:b}\right)^2.$$

It should be noted that for $a \leq b < c$

$$\mathcal{C}\left(x_{a:c}\right) - \left(\mathcal{C}\left(x_{a:b}\right) + \mathcal{C}\left(x_{(b+1):c}\right)\right) = \left(\sqrt{\frac{c-b}{(b-a+1)(c-a+1)}}\sum_{t=1}^{T} x_t - \sqrt{\frac{c-a+1}{(b-a+1)(c-b)}}\sum_{t=b+1}^{c} x_t\right)^2$$

holds, i.e. that the reduction in cost obtained by splitting a segment is equal to the CUSUM statistic. Consequently, Binary Segmentation can be viewed as a greedy heuristic for minimising the penalised cost.

Efficient Inference for Penalised Cost Approaches

The penalised cost introduced in the previous section can be minimised exactly via Optimal Partitioning (OP) introduced by Jackson et al. (2005). To this end, $C(m)$ is defined to be the cost of the optimal partition of all observations up to and including the mth one. The following recursive relationship then holds:

$$C(m) = \min_{1\leq k\leq m}\left(C(k-1) + \mathcal{C}\left(x_{k:m}\right) + \beta\right).$$

In the above, the optimal k represents the optimal changepoint preceding m, conditional on m being a changepoint. Solving the above dynamic programme therefore returns the optimal partition.

It can be shown that solving the full dynamic programme is at least $O(n^2)$ (Killick et al., 2012). However, Killick et al. (2012) showed that the solution space of the dynamic programme can be pruned thereby reducing the computational cost. Indeed

the authors showed that if the cost function is such that

$$\mathcal{C}\left(x_{a:c}\right) \geq \mathcal{C}\left(x_{a:b}\right) + \mathcal{C}\left(x_{(b+1):c}\right), \quad \forall a \leq b < c$$

i.e. such that adding an additional change does not reduce the cost, then if

$$C(k-1) + \mathcal{C}\left(x_{k:m}\right) > C(m) + \beta$$

holds for some $k \leq m$ then k can be disregarded for all future steps of the dynamic programme, without affecting the cost optimality of the returned partition. Killick et al. (2012) use this observation in their algorithm Pruned Exact Linear Time (PELT), to solve a pruned version of the dynamic programme considered by OP. The authors showed that PELT can be significantly faster than OP, especially when multiple changepoints are present and have a computational cost which is as low as $O(n)$ when the number of changes increases linearly in the number of observations, whilst still exactly minimising the penalised cost.

A different approach to this problem was proposed by Maidstone et al. (2017), who introduced Functional Pruning Optimal Partitioning (FPOP). Unlike PELT and OP, which condition on the location of the last changepoint, FPOP conditions on the last value of the parameter.

Choice of λ and β

The parameter λ and its analogue β are typically chosen in such a way that the number of false positives is controlled and that true changepoints are detected consistently.

For the change in mean case, the model

$$x_t \sim N(\mu_t, \sigma^2) \qquad \mu_t = \begin{cases} \mu_1 & t_0 < t \leq t_1 \\ ... & \\ \mu_{K+1} & t_K < t \leq t_{K+1} \end{cases}$$

with a fixed number of changepoints K is typically considered. The aim is then to show that under certain assumptions on the length of segments and the strength of the changes

$$\mathbb{P}\left(\hat{K} = K, |\hat{t}_i - t_i| < g(n) \quad 1 \leq i \leq p\right) \geq 1 - h(n)$$

holds for large enough n. Here, $g(n) = o(n)$, $h(n) = o(1)$, $\hat{K}$ denotes the inferred number of changes, and $\hat{t}_1, ..., \hat{t}_{\hat{K}}$ denote the inferred changes. The above statement therefore implies that the true number of changes as well as the true relative location t_i/n will be increasingly accurately estimated as the number of observations in between changes increases. Fryzlewicz (2014) showed that both Binary Segmentation and Wild Binary Segmentation are consistent if the threshold λ is set to $c\log(n)^{1+\alpha}$ for some positive c and α. Similarly, Tickle et al. (2018) showed that optimal partitioning is consistent provided that β is set to $(2+\epsilon)\log(n)$ for some $\epsilon > 0$.

2.3.2 Epidemic Changepoint Models for Univariate Data

The epidemic changepoint model assumes that data follows a typical distribution for most of the time but deviates from it during certain segments, the start and end point of which for epidemic changes. To formalise this, consider a univariate time series $x_1, ..., x_n$. It is said to obey the epidemic changepoint model if there exists a set

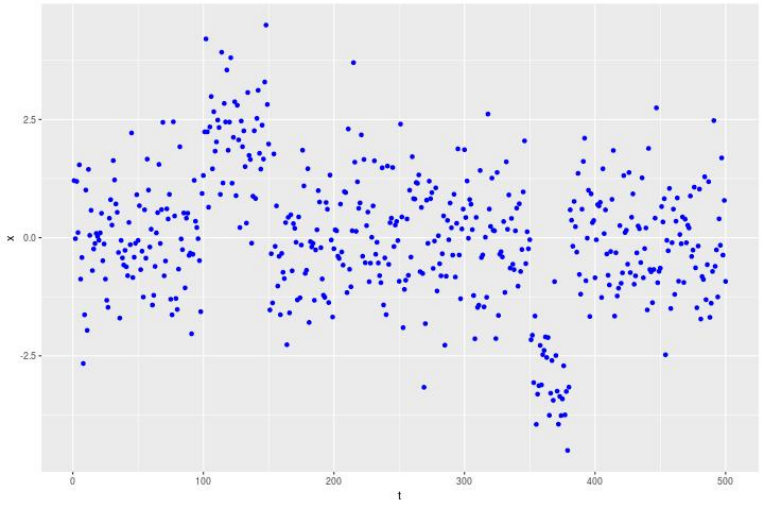


Figure 2.3.2: An example time series with $K = 2$ epidemic changes in mean. The first collective anomaly occurs between $s_1 = 100$ and $e_1 = 150$, and the second once occurs between $s_2 = 350$ and $e_2 = 381$.

of windows $\tau = \{(s_1, e_1), ..., (s_K, e_K)\}$, where $0 \leq s_1 < e_1 \leq s_2 < ... \leq s_K < e_K \leq n$, such that

$$x_t \sim \begin{cases} \mathcal{M}_1 & s_1 < t \leq e_1, \\ \vdots & \\ \mathcal{M}_K & s_K < t \leq e_K \\ \mathcal{M}_0 & \text{otherwise} \end{cases} \qquad 1 \leq t \leq n.$$

Here, $\mathcal{M}_k$ denotes the model obeyed by the data during the kth epidemic change. In between these epidemic changes it follows the null model $\mathcal{M}_0$. An example of such a timeseries can be found in Figure 2.3.2. As in the classical changepoint case, setting $\mathcal{M}_k = N(\mu_k, \sigma^2)$ gives rise to the change in mean problem with Gaussian noise. For simplicity, we will focus on this special case for the remainder of this section, pointing

to generalisations where appropriate.

Inferring at most one Single Change

The setting in which At Most One Change (AMOC) is present has received a significant amount of attention. This is because methodology detecting AMOC can naturally be extended to detect multiple changes, as we will show in Section 2.3.2. The problem of detecting a single epidemic change in mean can be formulated as the following hypothesis test (Yao, 1993): Consider data $x_1, ..., x_n$, where x_i $N(\mu_i, \sigma^2)$ and test the null hypothesis of a stationary mean.

$$H_0 : \mu_1 = ... = \mu_n,$$

against the alternative hypothesis that some segment (s, e) has a different mean

$$H_1 : \exists s, e : 0 \leq s < e \leq n,\ \mu_1 = ... = \mu_s = \mu_{e+1} = ... = \mu_n \neq \mu_{s+1} = ... = \mu_e.$$

This hypothesis testing framework can be extended to other types of epidemic changes, such as epidemic changes in variance, slope, mean and variance, etc.

The presence and location of epidemic changes are then typically inferred by using a likelihood ratio statistic (see, for example, Aston et al. (2012) and Yao (1993)). This likelihood ratio statistic is computed for all candidate start and end points. The pair of points for which this statistics is the largest is then declared a collective anomaly if the statistic exceeds a pre-defined threshold λ. Otherwise, the alternative hypothesis is rejected. The threshold λ is typically increased with the number of observations to account for multiple testing. (Yao, 1988)

Inferring Multiple Changes

Arguably the most commonly used method for the detection of multiple epidemic changes is circular binary segmentation (CBS) introduced by Olshen et al. (2004). Like Binary Segmentation for classical changepoints, CBS is capable of extending any AMOC test statistic to multiple epidemic changes by repeatedly applying the AMOC procedure to the parts of the data currently deemed typical. The algorithms can be summarised as follows:

1. Apply the AMOC procedure to the data x_1,...,x_n and obtain candidate start and end points s and e.

2. If the statistic does not exceed the threshold λ stop. Otherwise consider (s, e) to be an epidemic changepoint

3. Repeat the above procedure on the sequences x_1,...,x_s and x_{e+1},...,x_n

The computational cost of CBS is $O(n^2)$, when the whole data is searched. A faster approximation was proposed by Venkatraman and Olshen (2007). Another approach at speeding up CBS consists of imposing a maximum length m for epidemic changes, which reduces the computational cost to $O(mn)$.

On the choice of λ

The threshold λ is typically chosen in such a way that it controls the overall number of false positives at an acceptable level. Since $O(n^2)$ possible start and end points are investigated, this problem is closely linked to multiple testing. When trying to detect

epidemic changes in mean against a 0-background for example, as is the case in the CNV data, the log-likelihood ratio statistic for a segment s, e simplifies to

$$T(s,e) = (e-s)\left(\frac{1}{e-s}\sum_{t=s+1}^{e} x_t\right)^2,$$

assuming Gaussian noise. Yao (1988) showed that

$$\mathbb{P}\left(\max_{0\leq s<e\leq n} T(s,e) \leq (2+\epsilon)\sigma^2\log(n)\right) \to 1, \quad \text{as} \quad n\to\infty$$

under the null hypothesis for all $\epsilon > 0$. Here σ is the standard deviation of the noise. Consequently, setting $\lambda = (2+\epsilon)\sigma^2\log(n)$ asymptotically controls the number of false positives.

2.3.3 Epidemic Changepoint Models for Multivariate Data

Many time series are multivariate and epidemic changes can manifest themselves across multiple components. One commonly considered model for multivariate epidemic changes is the subset multivariate model, which assumes that components behave independently of each other, but that their anomalous time periods are linked. An example where this model is applicable can be found in the CNV data. Indeed, when no copy number variation is present, the data is independent across individuals. However, copy number variations can be shared by multiple subjects meaning that collective anomalies are likely to affect a subset of individuals at similar locations on the genome.

As with univariate epidemic changepoint detection, we can extend any AMOC procedure to multiple epidemic changes by circular binary segmentation. Consequently,

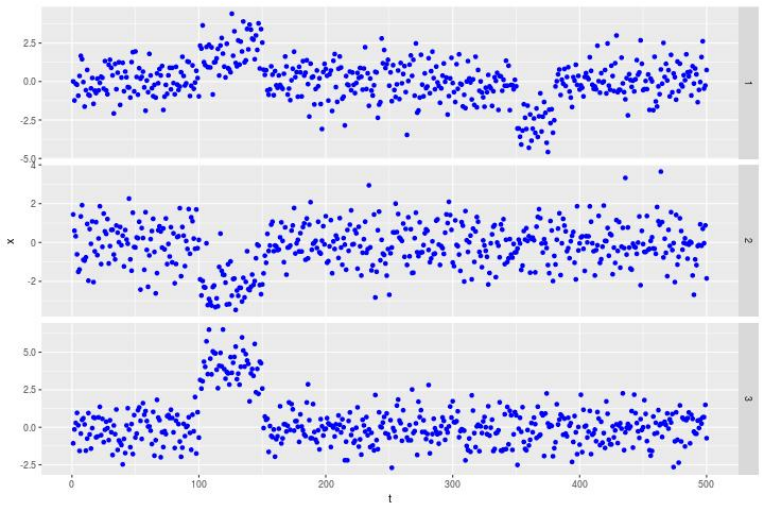


Figure 2.3.3: An example time series with $K = 2$ epidemic changes in mean. The first collective anomaly occurs between $s_1 = 100$ and $e_1 = 150$ and affects all components, i.e. $J_1 = \{1, 2, 3\}$. The second once occurs between $s_2 = 350$ and $e_2 = 381$ and affects only the first component, i.e. $J_2 = \{1\}$.

we will only review AMOC procedures in this section. In order to do so, we will begin by formalising the subset multivariate epidemic changepoint model before reviewing some theoretical results regarding sparse changes affecting few components strongly and dense changes which weakly affect a large number of components. This will be followed by a discussion of inference approaches.

Subset Multivariate Epidemic Changepoint Model

Consider multivariate data $\mathbf{x}_1, ..., \mathbf{x}_n \in \mathbb{R}^p$. This data is said to contain a subset multivariate epidemic changepoint if there exists a subset $\mathbf{J} \subset \{1, ..., p\}$ which exhibits a-typical behaviour during a time window (s, e). Formally we test the hypothesis

$$H_0 : \mathbf{x}_t^{(i)} \sim \mathcal{M}_0^{(i)} \quad 1 \leq t \leq n, \quad 1 \leq i \leq p,$$

which states that all n observations from the ith component follows the null model $\mathcal{M}_0^{(i)}$ for $1 \leq i \leq p$ against the alternative hypothesis

$$H_1 : \exists\, 0 \leq s < e \leq n, \mathbf{J} \subset \{1, ..., p\} : \ \mathbf{x}_t^{(i)} \sim \begin{cases} \mathcal{M}_1^{(i)} & s < t \leq e, \quad i \in \mathbf{J} \\ \mathcal{M}_0^{(i)} & \text{otherwises} \end{cases}$$

that an anomalous segment exists during which the ith component obeys a model $\mathcal{M}_1^{(i)}$ which is different from its typical model $\mathcal{M}_0^{(i)}$ for $i \in \mathbf{J}$.

Detectability Boundaries

Subset multivariate epidemic changes can be easier or harder to detect depending on how strong the changes are and how many components are affected. The literature (Jeng et al., 2012; Cai et al., 2011a) typically distinguishes between sparse changes, in which only a few components are affected by strong anomalies and dense changes which affect a large set of components, but potentially by very little. This has been formalised for the changes in mean and variance by Cai et al. (2011a) who for a segment s, e considered testing the null hypothesis of uncontaminated data

$$H_0 : \mathbf{x}_t^{(i)} \sim N(0, 1) \quad 1 \leq i \leq p, \quad s < t \leq e$$

against the alternative hypothesis of contaminated data

$$H_0 : \mathbf{x}_t^{(i)} \sim (1 - v_i)N(0, 1) + v_i N(\mu, 1 + \sigma^2) \quad 1 \leq i \leq p, \quad s < t \leq e$$

where the independent random variables $v_i \sim Ber(p^{-\xi})$ for $0 \leq \xi < 1$ determine whether the ith component exhibits anomalous behaviour or not. The parameter ξ then determines whether the regime is sparse ($\xi > \frac{1}{2}$) or dense ($\xi \leq \frac{1}{2}$). The strength

of the change in mean $(e-s)|\mu|$ can the be parametrised as

$$(e-s)|\mu| = \begin{cases} \sqrt{2r_p \log(p)} & \xi > \frac{1}{2} \\ p^{-r_p} & \xi \leq \frac{1}{2}, \end{cases}, \qquad r_p > 0$$

depending on whether the anomalous segment is dense or sparse. It is then called detectable if the exists a test whose type 1 and type 2 error both converge to 0 as $p \to \infty$. Jeng et al. (2012) showed that the detectability boundary for the case $\xi > \frac{1}{2}$ was

$$\rho^- = \min\left(\xi - \frac{1}{2} - \sigma^2, \left(1 - (1+\sigma^2)\sqrt{(1-\xi)}\right)^2\right)$$

and that such a test exists if $r_p > \rho^-$ and does not exist if $r_p < \rho^-$. Similarly the authors showed that the detectability boundary for the case $\xi \leq \frac{1}{2}$ was

$$\rho^- < \begin{cases} \frac{1}{2} - \xi & \sigma = 0 \\ \infty & \sigma > 0 \end{cases}$$

and that a consistent test exists if $r_p < \rho^-$ and does not exist if $r_p > \rho^-$. Boundary cases for the related problem of distinguishing between mixtures were treated in Cai et al. (2011a).

In particular these results mean that even the smallest changes in variance make any dense change detectable. Another point of interest is that the signal strength of a dense change can go to 0 as p goes to infinity without the change becoming undetectable.

Inference Approaches

A variety of methods has been proposed to detect epidemic changes, dense and/or sparse, in multivariate data. For example, Zhang et al. (2010) computes a likelihood ratio statistic for each of the p components individually and the sums these, hence obtaining a tests statistic for a given start and end point. This test statistic is good for dense changes but lacks power for detecting sparse ones. Conversely, the approach suggested by Jeng et al. (2010) consists of considering the largest test statistic only, an approach suitable only for sparse changes.

Approaches capable of detecting both sparse and dense epidemic changes have also been proposed. These approaches typically test each component individually thus obtaining p different p-values, which they then process to obtain a global significance value. The methods of both Zhang et al. (2010) and Jeng et al. (2010) fall into this framework thought they have good power only against certain types of epidemic changes. One example an approach suitable for both sparse and dense changes is proportion adaptive segment selection (PASS), introduced by Jeng et al. (2012), uses higher criticism (Donoho and Jin, 2004). Higher criticism builds on the fact that all p-values $q_1, ..., q_p$ are i.i.d. $U(0,1)$ distributed under the null hypothesis. Consequently, the ordered p-values $q_{(1)} \leq ... \leq q_{(p)}$ are all Beta distributed under the null hypothesis. This motivates the higher criticism statistic which is defined as:

$$HC_p^* = \max_{1 \leq i \leq p} \left(HC_{p,i} \right), \quad HC_{p,i} = \sqrt{p} \frac{\frac{i}{p} - q_{(i)}}{\sqrt{q_{(i)}(1 - q_{(i)})}}.$$

This statistic can be computed for all candidate start and end points. Jeng et al.

(2012) showed that a threshold value of

$$\frac{(1+\epsilon)\log(nm) + 2\log(\log(p))}{\sqrt{2\log(\log(p))}},$$

for some $\epsilon > 0$ asymptotically controls the number of false positives, as well as having power against all detectable sparse and most detectable dense alternatives. However, this asymptotic result is based on the asymptotic (and slow) convergence of the higher order statistic to a non-degenerate random variable (Shorack and Wellner, 2009). To improve finite sample performance, and reduce the number of false positives in particular, the authors suggest to consider $\max_{\alpha_0 \leq i \leq p}(HC_{p,i})$ instead, where $\alpha_0 \in \mathbb{N}$ is greater than 1. However, this is only advisable if no collective anomalies of interest affects fewer than α_0 components.

Other methods for combining the individual p-values have also been proposed. For example, the adaptive Fisher procedure Song et al. (2016), which exploits the fact that the differences between the logarithms of the ordered p-values is exponentially distributed under the null. Entirely different principal component analysis based approaches have also been proposed (Aston et al., 2012).

Chapter 3

Collective And Point Anomalies

3.1 Introduction

Anomaly detection is an area of considerable importance for many time series applications, such as fault detection or fraud prevention, and has been subject to increasing attention in recent years. See Chandola et al. (2009) and Pimentel et al. (2014) for comprehensive reviews of the area. As Chandola et al. (2009) highlight, anomalies can fall into one of three categories: global anomalies, contextual anomalies, or collective anomalies. Global anomalies and contextual anomalies are defined as single observations which are outliers with regards to the complete dataset and their local context respectively. Conversely, collective anomalies are defined as sequences of observations which are not anomalous when considered individually, but together form an anomalous pattern.

A number of different approaches can be taken to detect point (i.e. contextual and/or global) anomalies. These are observations that do not conform with the pat-

tern of the data. Hence, the problem of detecting point anomalies can be reformulated as inferring the general pattern of the data in a manner that is robust to anomalies. The field of robust statistics offers a wide range of methods aimed at this problem. For instance, Rousseeuw and Yohai (1984) proposed S-estimators to robustly estimate the mean and variance. These estimators were later extended to a multivariate setting by Rousseeuw (1985). A wide variety of robust time series models also exist. For example, Muler et al. (2009) proposed a robust ARMA model, Muler and Yohai (2002) a robust ARCH model, and Muler and Yohai (2008) a robust GARCH model. A robust non-parametric method, which decomposes time series into trend, seasonal component, and residual was proposed by Cleveland et al. (1990).

The machine learning community has also provided a rich corpus of work for the detection of point anomalies. Commonly used methods include nearest neighbour based approaches, such as the local outlier factor (Breunig et al., 2000), and information theoretical methods such as the one introduced by Guha et al. (2016). It is beyond the scope of this chapter to review them all. Instead we refer to excellent reviews that can be found in Chandola et al. (2009) and Pimentel et al. (2014). Lavin and Ahmad (2015), Talagala et al. (2019), Ahmad et al. (2017), and references therein provide examples of more recent developments in the area.

One common drawback of several point anomaly approaches is their limited ability to detect anomalous segments, or collective anomalies. Such features are of significance in many applications. One example is the analysis of brain imaging data, where periods in which the brain activity deviates from the pattern of the rest state have been associated with sudden shocks (Aston and Kirch, 2012). Another example is in

detecting regions of the genome with unusual copy number (Bardwell and Fearnhead, 2017; Siegmund et al., 2011; Zhang et al., 2010), with such copy number variation being associated with diseases such as cancer (Jeng et al., 2012).

The current statistical literature mostly uses hidden Markov models, smoothing based approaches or epidemic changepoint methods for the detection of collective anomalies. Hidden Markov models assume that a hidden state chain determines whether the data produced is anomalous or typical (Smyth, 1994). The underlying assumption that anomalous segments share one or multiple common behaviours is very attractive for some applications, such as brain imaging, where it can be assumed that there is a finite number of states, but can be a constraint in others. Hidden Markov models also suffer from the fact that they are not robust to global anomalies. Moreover, they tend to be slow to fit, which is an important disadvantage in many modern, big-data applications. This is in stark contrast with the typically very fast smoothing based approaches like the one proposed by Schwartzman et al. (2011). However, the smoothing step limits interpretability making the approach vulnerable to point anomalies and differentiating between point and collective anomalies nigh impossible. Furthermore, these methods achieve optimal power when the bandwidth of the smoothing kernel is of the same length-scale as the collective anomalies, meaning that they can struggle when anomalies are of very different lengths.

The epidemic changepoint model, first introduced by Levin and Kline (1985) assumes that there is a typical behaviour, from which the data deviates during collective anomalies. Epidemic changepoints can therefore be viewed as two classical (non-epidemic) changepoints: one away from and one back to the typical distribu-

tion. Thus, a simple approach to detecting collective anomalies would be to use one of the many methods for changepoint detection (e.g. Fearnhead and Rigaill, 2019a, Fryzlewicz, 2014, James et al., 2016, Killick et al., 2012, Ma and Yau, 2016, and references therein). However this does not exploit the fact that the behaviour of the segment before the start and after the end of an anomalous segment is the same. This reduces its statistical power, as can be seen in Section 3.5, which is a disadvantage, especially when faced with a weak signal.

The main corpus of work addressing the problem of detecting epidemic changes has been driven by the analysis of neuroimaging and genome data. An early application of epidemic changepoints to neuroimaging data can be found in Robinson et al. (2010), who use a hidden Markov model to detect epidemic changes in mean. This was later extended by Aston and Kirch (2012). Both methods are vulnerable to point anomalies, a shortcoming in some applications like the ones we consider in this chapter. Another limitation is that both approaches assume the presence of at most one change. Conversely, motivated by challenges arising in Genomics, a range of methods, both univariate and multivariate, have been proposed to detect epidemic changes in mean, mainly by considering sum of squares type test statistics (see Jeng et al., 2012; Siegmund et al., 2011; Cai et al., 2012), sometimes in combination with hidden states. They are therefore vulnerable to global anomalies. A more general Bayesian hidden state method for the detection of anomalous segments was proposed by Bardwell and Fearnhead (2017).

This article makes two main contributions. The first is the introduction of an estimation procedure that allows for the identification of **C**ollective **A**nd **P**oint **A**nomalies

(CAPA). Secondly, we establish finite sample consistency results not only for CAPA, but also for a commonly used penalised cost based method (Killick et al., 2012) aimed at detecting changes in mean and variance. This setting presents significant additional technical challenge compared to the change in mean setting, to which most existing theoretical results apply. Since the first version of this work appeared on arXiv, a similar algorithm has been independently proposed by Zhao and Yau (2019). However, the work of Zhao and Yau (2019) does not contain any consistency results and does not address the challenge of fitting point anomalies when using a data distribution with multiple parameters (e.g. changes in mean and variance).

The article is organised as follows: We begin by introducing a parametric model with epidemic changes in Section 3.2. This provides a general framework for collective anomalies, the location of which we infer by minimising a penalised cost. In Section 3.3, we introduce an algorithm which minimises an approximation to the penalised cost based on a robust estimate of the parameter of the typical distribution. This approximation can be minimised by a dynamic programme.

Section 3.4 presents a number of theoretical results. Specifically, we introduce a proof of consistency for the detection of joint classical changes in mean and variance using a penalised cost approach, which is of independent interest. We then prove that CAPA consistently estimates the number and location of collective anomalies, despite the simplicity of the approach used for the estimation of the parameters of the typical distribution. Section 3.5 concludes with a discussion of penalties. We then compare CAPA to other methods in a simulation study in Section 3.5 and show that it outperforms them, especially in the presence of point anomalies.

The chapter is concluded by applying CAPA to two real datasets in Section 3.6. The first dataset is lightcurve data gathered by the Kepler space telescope. There we show that CAPA can be used to detect Kepler 1132-b, an exoplanet which orbits the star Kepler 1132 (Morton et al., 2016). The second dataset is a machine temperature dataset obtained on an expensive industrial machine. There we show that CAPA can be used to detect both critical failures as well as early warning signs, highlighting the algorithms usefulness for predictive maintenance. The proofs of the theoretical results are all given in the appendix. CAPA has been implemented in the R package `anomaly` (Fisch et al., 2020) which is available from CRAN.

3.2 A Modelling Framework for Collective Anomalies

We assume that the data follow a parametric model where collective anomalies are epidemic changes in the model parameters. Whilst, in practice, it is unlikely that the distribution of the data in an anomalous segment will belong to the same family of distributions as the distribution of the typical data, it can nevertheless be expected that a set of parameters different from the typical distribution's will offer a better fit. We say that data $\mathbf{x}_1, ..., \mathbf{x}_n$ follow a parametric epidemic changepoint model if $\mathbf{x}_t$ has

probability density function $f(\mathbf{x}_t, \theta(t))$ and

$$\theta(t) = \begin{cases} \theta_1 & s_1 < t \leq e_1, \\ & \vdots \\ \theta_K & s_K < t \leq e_K, \\ \theta_0 & \text{otherwise,} \end{cases}$$

where θ_0 is the, usually unknown, parameter of the typical distribution, from which the model deviates during the K anomalous segments (s_1, e_1),...,(s_K, e_K) by adopting behaviours characterised by the parameters $\theta_1, ..., \theta_K$ all different from θ_0. We assume these windows do not overlap, i.e. $e_1 \leq s_2, ..., e_{K-1} \leq s_K$. Note that fitting an epidemic changepoint requires only one new set of parameters for θ, since the typical parameter is shared across the non-anomalous segments. This compares favourably with the two additional sets of parameters for θ introduced when an epidemic changepoint is fitted using two classical changepoints. We can therefore expect to gain statistical power, which is confirmed by the empirical results in Section 3.5.

It is possible to infer the number and location of epidemic changes by choosing $\tilde{K}$, $(\tilde{s}_1, \tilde{e}_1)$,...,$(\tilde{s}_{\tilde{K}}, \tilde{e}_{\tilde{K}})$, and $\tilde{\theta}_0$, which minimise the penalised cost

$$\sum_{t \notin \cup[\tilde{s}_i+1, \tilde{e}_i]} \mathcal{C}(\mathbf{x}_t, \tilde{\theta}_0) + \sum_{j=1}^{\hat{K}} \left[\min_{\tilde{\theta}_j} \left(\sum_{t=\tilde{s}_j+1}^{\tilde{e}_j} \mathcal{C}(\mathbf{x}_t, \tilde{\theta}_j) \right) + \beta \right], \tag{3.2.1}$$

subject to $e_i - s_i \geq \hat{l}$, where $\hat{l}$ is the minimum segment length for an appropriate cost function $\mathcal{C}(x, \theta)$ and a suitable penalty β. For example, $\mathcal{C}(x, \theta)$ could be defined as the negative log-likelihood of x under the parametric model using parameter θ. A common choice for the penalty β would then be $C \log(n)$ (Yao, 1988; Killick et al., 2012; Fryzlewicz, 2014), where the constant C depends on the model considered.

Using the formulation in (3.2.1), we can infer the location of joint epidemic changes in mean and variance by minimising the penalised cost related to the negative log-likelihood of Gaussian data. In this case $\theta = (\mu, \sigma^2)$ contains both the mean and variance and we estimate K, $s_1, ..., s_K$, and $e_1, ..., e_K$ by minimising

$$\sum_{t \notin \cup [\tilde{s}_i+1, \tilde{e}_i]} \left[\log(\sigma_0^2) + \left(\frac{x_t - \mu_0}{\sigma_0} \right)^2 \right] + \sum_{j=1}^{\tilde{K}} \left[(\tilde{e}_j - \tilde{s}_j) \left(\log \left(\frac{\sum_{t=\tilde{s}_j+1}^{\tilde{e}_j} (x_t - \bar{x}_{(\tilde{s}_j+1):\tilde{e}_j})^2}{(\tilde{e}_j - \tilde{s}_j)} \right) + 1 \right) + \beta \right], \tag{3.2.2}$$

subject to $\tilde{e}_i - \tilde{s}_i \geq 2$, i.e. a minimum segment length of 2, to account for the fact that θ is two dimensional.

It is well known that many changepoint detection methods struggle in the presence of point anomalies in the data and tend to fit two changepoints around each of them (Fearnhead and Rigaill, 2019a). An approach based on minimising the above cost function is not intrinsically immune to this phenomenon either. However, given that point outliers can naturally be viewed as single observations with a larger variance, we can incorporate them in the model as epidemic changes, in variance only, of length one. We therefore choose $\tilde{K}$, $(\tilde{s}_1, \tilde{e}_1)$,...,$(\tilde{s}_{\tilde{K}}, \tilde{e}_{\tilde{K}})$, μ_0, σ_0, as well as the set of point anomalies $O \subset \{1, ..., n\}$, which minimise the modified penalised cost

$$\begin{aligned} \sum_{t \notin \cup [\tilde{s}_i+1, \tilde{e}_i] \cup O} \left[\log(\sigma_0^2) + \left(\frac{x_t - \mu_0}{\sigma_0} \right)^2 \right] + \sum_{t \in O} \left[\log \left((x_t - \mu_0)^2 + \gamma \sigma_0^2 \right) + 1 + \beta' \right] + \\ \sum_{j=1}^{\hat{K}} \left[(\tilde{e}_j - \tilde{s}_j) \left(\log \left(\frac{\sum_{t=\tilde{s}_j+1}^{\tilde{e}_j} (x_t - \bar{x}_{(\tilde{s}_j+1):\tilde{e}_j})^2}{(\tilde{e}_j - \tilde{s}_j)} \right) + 1 \right) + \beta \right], \end{aligned} \tag{3.2.3}$$

where β' is a penalty smaller than β. This modification ensures that it is now cheaper to fit an outlier as an epidemic changepoint in variance only than as a full epidemic change. The constant, $\gamma > 0$ ensures that the argument of the logarithm will be larger than 0. We recommend setting γ to the level of precision of the observations, subject

to requiring $\gamma > \exp(-\beta')$ which ensures that no inliers are fitted as point anomalies, as shown by Proposition 3 in Section 3.4.

This modification has the added benefit that it allows the algorithm to distinguish between point and collective anomalies. This property is important for a range of applications in which collective and point anomalies have different interpretations (see Section 3.6.1 for an example).

3.3 Estimation of Collective and Point Anomalies

We now turn to consider the problem of minimising the penalised cost we introduced in the previous section. Unlike in the classical changepoint problem considered by Jackson et al. (2005), the penalised cost given by equation (3.2.1) can not be minimised using a dynamic programme, since the parameter θ_0 is shared across multiple segments and typically unknown. We therefore use robust statistics to obtain an estimate, $\hat{\theta}_0$, for θ_0. Such robust estimates can be obtained for a variety of models (Hampel et al., 1986; Jurečková and Picek, 2005). For example, the median, M-estimators, or the clipped mean can be used to robustly estimate the mean. The inter quartile range, the median absolute deviation, or the clipped standard deviation can be use to estimate the variance. Robust regression is available to estimate the parameters of AR models.

Having obtained $\hat{\theta}_0$, we then minimise

$$\sum_{t \notin \cup [\hat{s}_i+1, \hat{e}_i]} \mathcal{C}(\mathbf{x}_t, \hat{\theta}_0) + \sum_{j=1}^{\hat{K}} \left[\min_{\hat{\theta}_j} \left(\sum_{t=\hat{s}_j+1}^{\hat{e}_j} \mathcal{C}(\mathbf{x}_t, \hat{\theta}_j) \right) + \beta \right], \tag{3.3.1}$$

as an approximation to (3.2.1). Since it can be expected that most data belongs to the typical distribution, $\hat{\theta}_0$ should be close to θ_0. One might therefore expect that

using this estimate will have little impact on the performance of the method, which we also show theoretically for joint changes in mean and variance in Section 3.4.2.

The approximation to the penalised cost in (3.3.1) can be minimised exactly by solving the dynamic programme

$$C(m) = \min\left[C(m-1) + \mathcal{C}(\mathbf{x}_m, \hat{\theta}_0), \min_{0\leq k\leq m-\hat{l}}\left(C(k) + \min_{\hat{\theta}}\left(\sum_{t=k+1}^{m}\mathcal{C}(\mathbf{x}_t, \hat{\theta})\right) + \beta\right)\right], \tag{3.3.2}$$

where $C(m)$ is the cost of the most efficient partition of the first m observations and $C(0) = 0$. For example, solving the dynamic programme

$$C(m) = \min\left[C(m-1) + \log(\hat{\sigma}_0^2) + \left(\frac{x_m - \hat{\mu}_0}{\hat{\sigma}_0}\right)^2, \min_{0\leq k\leq m-2}\left(C(k) + (m-k)\left[\log\left(\frac{1}{m-k}\sum_{t=k+1}^{m}\left(x_t - \bar{x}_{(k+1):m}\right)^2\right) + 1\right] + \beta\right)\right],$$

minimises the approximate penalised cost for joint epidemic changes in mean and variance defined in equation (3.2.2), thus inferring the number and location of collective anomalies. Similarly, we can minimise the approximation to its point anomaly robust analogue in equation (3.2.3) by solving the dynamic programme

$$C(m) = \min\left[C(m-1) + \log(\hat{\sigma}_0^2) + \left(\frac{x_m - \hat{\mu}_0}{\hat{\sigma}_0}\right)^2, \min_{0\leq k\leq m-2}\left(C(k) + (m-k)\left[\log\left(\frac{1}{m-k}\sum_{t=k+1}^{m}\left(x_t - \bar{x}_{(k+1):m}\right)^2\right) + 1\right] + \beta\right), C(m-1) + 1 + \log\left(\gamma\hat{\sigma}_0^2 + (x_m - \hat{\mu}_0)^2\right) + \beta'\right],$$

thus also inferring the number and location of point anomalies, with a negligible increase in computational cost. We call this algorithm CAPA. Pseudocode for the full CAPA algorithm is given by Algorithm 1 in the appendix.

Solving the full dynamic program is at least $O(n^2)$. This is because it requires n steps over a solution space $0 \leq k \leq m - l$, which is of size $O(n)$ on average. However, we can reduce the size of the solution space by borrowing ideas on pruning from Killick et al. (2012), provided the loss function is such that adding a free changepoint will not increase the cost – a property which holds for many commonly used cost functions such as the negative log-likelihood. Indeed, the following proposition holds:

Proposition 1. *Let the cost function $\mathcal{C}(\cdot, \cdot)$ be such that*

$$\min_{\theta} \left(\sum_{t=a}^{c} \mathcal{C}(\boldsymbol{x}_t, \theta) \right) \geq \min_{\theta} \left(\sum_{t=a}^{b-1} \mathcal{C}(\boldsymbol{x}_t, \theta) \right) + \min_{\theta} \left(\sum_{t=b}^{c} \mathcal{C}(\boldsymbol{x}_t, \theta) \right)$$

holds for all a, b, and c such that $a + \hat{l} \leq b < c - \hat{l}$. Then, if

$$C(k) + \min_{\theta} \left(\sum_{t=k}^{m} \mathcal{C}(\boldsymbol{x}_t, \theta) \right) \geq C(m) \tag{3.3.3}$$

holds for some $k < m - \hat{l}$, we can exclude that k from the solution space for all future steps $m' \geq m + \hat{l}$ of the dynamic programme.

Thus we can keep track of the set of time indices that we need to minimise over in the dynamic programme. For each of these, we check, at each time step, if condition 3.3.3 holds, and, if it does, we remove the index from the set. See steps 15-22 of Algorithm 2 in the appendix for more detail. This can significantly reduce the computational cost. In practice, we found that it was close to $O(n)$ for the detection of joint epidemic changes in mean and variance when the number of true epidemic changes increased linearly with the number of observations. Note that the time after which a k can be discarded in the minimisation step also depends on the minimum segment length, something not considered by Killick et al. (2012).

3.4 Theory for Joint Changes in Mean and Variance

We now introduce some theoretical results for CAPA. In particular, we establish the consistency of CAPA at detecting collective anomalies and demonstrate that it can be viewed as a corollary of the consistency of a statistical procedure, such as optimal partitioning from Jackson et al. (2005) minimising a penalised cost function to detect classical (i.e. non-epidemic) changepoints. Consequently, we will begin by proving the consistency of a penalised cost method for the detection of changes in mean and variance in Section 3.4.1. To the best of our knowledge, no such result exists in the literature, which makes this proof of independent interest. We then proceed to proving the consistency of CAPA at detecting collective anomalies in Section 3.4.2. Like other cost function based approaches, CAPA is significantly affected by the choice of penalties. We therefore conclude this section by discussing suitable choices for this important hyper parameter in Section 3.4.3. The proofs of all theorems and propositions stated in this section can be found in the appendix.

3.4.1 Consistency of Classical Changepoint Detection

Consider the sequence $x_1, ..., x_n \in \mathbb{R}^n$ which is normally distributed with $K \in \mathbb{N}$ changepoints. The sequence therefore satisfies $x_t = \mu(t) + \sigma(t)\eta_t$ for $\eta_t \overset{i.i.d.}{\sim} N(0,1)$,

where

$$(\mu(t), \sigma(t)^2) = \begin{cases} (\mu_1, \sigma_1^2) & t_0 + 1 \leq t \leq t_1, \\ & \vdots \\ (\mu_{K+1}, \sigma_{K+1}^2) & t_K + 1 \leq t \leq t_{K+1}. \end{cases} \tag{3.4.1}$$

Here $0 = t_0 \leq ... \leq t_{K+1} = n$ denote the start of the series, the K changepoints, and the end of the series. We assume that the mean and/or variance changes at these changepoints, i.e. that $(\mu_k, \sigma_k^2) \neq (\mu_{k+1}, \sigma_{k+1}^2)$. These changes in mean and variance can be of varying strength. To quantify this, we define the signal strength $\triangle_{\sigma,k}$ of the change in variance at the kth changepoint to be

$$\triangle_{\sigma,k}^2 = \left(\sqrt{\frac{\sigma_k}{\sigma_{k+1}}} - \sqrt{\frac{\sigma_{k+1}}{\sigma_k}} \right)^2 = \frac{\sigma_k}{\sigma_{k+1}} + \frac{\sigma_{k+1}}{\sigma_k} - 2.$$

We note that $\triangle_{\sigma,k}^2$ is equal to 0 if, and only if, $\sigma_{k+1} = \sigma_k$. We also define the signal strength $\triangle_{\mu,k}$ of change in mean at the kth changepoint to be

$$\triangle_{\mu,k} = \frac{|\mu_k - \mu_{k+1}|}{\sqrt{\sigma_k \sigma_{k+1}}}.$$

Note that these two quantities can be combined into a global measure of signal strength

$$\triangle_k = \log \left(1 + \frac{1}{2} \triangle_{\sigma,k}^2 + \frac{1}{4} \triangle_{\mu,k}^2 \right)$$

for the kth change (see Lemma 7 in the appendix material for details).

We now define the penalised cost $\tilde{\mathcal{C}}(x_{i:j}, \tau', \alpha)$ of data $x_{i:j}$ under partition $\tau' = \{i - 1, \hat{t}'_1, ..., \hat{t}'_{\hat{K}'}, j\}$ to be

$$\tilde{\mathcal{C}}(x_{i:j}, \tau', \alpha) = \sum_{k=0}^{\hat{K}'} \tilde{\mathcal{C}}(x_{(\hat{t}_k+1):\hat{t}_{k+1}}) + \hat{K}' \alpha \log(n)^{1+\delta},$$

for $\delta, \alpha > 0$. Here $\alpha \log(n)^{1+\delta}$ corresponds to the commonly used strengthened SIC-type penalty (Fryzlewicz, 2014; Li et al., 2016) for introducing an additional changepoint. The cost of a segment $x_{a:b}$ is set to be

$$\tilde{\mathcal{C}}(x_{a:b}) = \tilde{\mathcal{C}}(x_{a:b}, \{a-1, b\}) = (b-a+1)\left(\log\left(\frac{\sum_a^b (x_t - \bar{x}_{a:b})^2}{b-a+1}\right) + 1\right),$$

which is the same as the segment cost used to infer the location of epidemic changes in mean and variance. Since this leaves two parameters to fit, we impose a minimum segment length of two for all partitions.

We also introduce the following assumption which ensures that the changepoints are sufficiently spaced apart to allow for their detection:

Assumption 1. *There exists some $\tilde{\delta} > 0$ such that for $1 \leq k \leq K+1$*

$$t_k - t_{k-1} \geq \frac{\log(n)^{1+\delta+\tilde{\delta}}}{\min(\triangle_k, \triangle_k^2, \triangle_{k-1}, \triangle_{k-1}^2)},$$

where $\triangle_0 = \triangle_{K+1} = \infty$ for convenience of notation.

Here the fact that $\min(\triangle_k, \triangle_k^2, \triangle_{k-1}, \triangle_{k-1}^2)$ is used instead of $\min(\triangle_k^2, \triangle_{k-1}^2)$ as in the change in mean setting is due to the fact that we have moved from a sub-Gaussian to a sub-exponential setting.

Then, the following consistency result holds:

Theorem 1. *Assume that observations $x_1, ..., x_n$ follow the distribution specified in Equation 3.4.1 and that Assumption 1 holds. Let $\hat{K}$ and $\hat{t}_1, ..., \hat{t}_{\hat{K}}$ be the number and locations of changepoints inferred by minimising the penalised cost function $\tilde{\mathcal{C}}(x_{1:n}, \tau, \alpha)$. Then there exists a constant C such that $\forall \epsilon > 0$ there exist constants $A(\alpha, \epsilon)$ and*

$B(\alpha, \tilde{\delta}, \delta, \epsilon)$ such that

$$\mathbb{P}\left(\hat{K} = K, \ \ |\hat{t}_k - t_k| < \frac{A(\alpha, \epsilon)}{\min(\triangle_k, \triangle_k^2)} \log(n)^{1+\delta} \quad 1 \leq k \leq K\right) \geq 1 - Cn^{-\epsilon}$$

holds for all $n \geq B(\alpha, \tilde{\delta}, \delta, \epsilon)$.

The proof of this result addresses several novel challenges such as obtaining upper and lower bounds on a range of transformations of the residual sum of squares arising from the fact that joint changes in mean and variance were considered instead of only changes in mean, as in pre-existing work (e.g. Fryzlewicz, 2014). In particular, the lower bound on the MGF of the weighted sum of χ^2-distribution in Lemma 12 is of independent interest. Note that the strengthened SIC was used as penalty in order to simplify the exposition of the proof. A very similar proof can be used to show that a $\alpha \log(n)$ type penalty for a sufficiently large α also achieves consistency.

Furthermore, the bounds on the accuracy of the detected changes can be tightened. For all $\delta > \delta_0$, it is possible to show that the detected changes must be within $\frac{A(\alpha,\epsilon)}{\min(\triangle_k, \triangle_k^2)} \log(n)^{1+\delta_0}$ of the true changes once $n \geq B(\alpha, \tilde{\delta}, \delta, \epsilon)$ and $n \geq B(\alpha, \tilde{\delta}, \delta_0, \epsilon)$. This is because Theorem 1 then also holds for the penalty $\alpha \log(n)^{1+\delta_0}$, which therefore guarantees that the fitted changes of the optimal partition with K fitted changes must be within $\frac{A(\alpha,\epsilon)}{\min(\triangle_k, \triangle_k^2)} \log(n)^{1+\delta_0}$ of the true changes.

3.4.2 Consistency of CAPA

The results we obtained in the previous section can be extended to prove the consistency of CAPA for the detection of joint epidemic changes in mean and variance. As in the previous section, consider data $x_1, ..., x_n$ which is of the form $x_t = \mu(t) + \sigma(t)\eta_t$,

where $\eta_t \sim N(0,1)$. Since we now assume epidemic changes, we have

$$(\mu(t), \sigma(t)^2) = \begin{cases} (\mu_1, \sigma_1^2) & s_1 < t \leq e_1, \\ & \vdots \\ (\mu_K, \sigma_K^2) & s_K < t \leq e_K, \\ (\mu_0, \sigma_0^2) & \text{otherwise.} \end{cases}$$

Here, μ_0 and σ_0^2 are the typical mean and variance respectively and K is the number of epidemic changepoints. The variables s_k and e_k denote the starting and end point of the kth anomalous window respectively. Treating the s_k and e_k like classical changepoints allows us to extend the definitions of $\triangle_\sigma$ and $\triangle_\mu$, and therefore $\triangle$, to the epidemic changepoint model by setting

$$\triangle_{\sigma,k}^2 = \frac{\sigma_k}{\sigma_0} + \frac{\sigma_0}{\sigma_k} - 2 \qquad \text{and} \qquad \triangle_{\mu,k} = \frac{|\mu_k - \mu_0|}{\sqrt{\sigma_k \sigma_0}}.$$

We make the following assumptions

Assumption 2. .

a) There exists some $\tilde{\delta} > 0$ *such that for* $1 \leq k \leq K$

$$e_k - s_k \geq \frac{\log(n)^{1+\delta+\tilde{\delta}}}{\min(\triangle_k, \triangle_k^2)},$$
$$s_{k+1} - e_k \geq \frac{\log(n)^{1+\delta+\tilde{\delta}}}{\min(\triangle_k, \triangle_k^2, \triangle_{k+1}, \triangle_{k+1}^2)}.$$

b) $e_k - s_k \leq \sqrt{n}$ *for* $1 \leq k \leq K$.

Assumption 2a is analogous to Assumption 1. Assumption 2b is only needed when the parameters of the typical distribution are unknown and guaranteess that

the robust estimates of the typical parameter will converges quickly enough to the ground truth. The following consistency result then holds for CAPA.

Theorem 2. *Let $(\hat{s}_1, \hat{e}_1, ..., \hat{s}_{\hat{K}}, \hat{e}_{\hat{K}})$ be the partition inferred by CAPA on observations $x_1, ..., x_n$ using a minimum segment length of 2 and $\alpha \log(n)^{1+\delta}$ as penalty for both point anomalies and epidemic changepoints. If $x_1, ..., x_n$ follow the distribution specified above and Assumption 2 holds, then there exists a constant C such that $\forall \epsilon > 0$ there exist constants $A(\alpha, \epsilon)$ and $B(\alpha, \tilde{\delta}, \delta, \epsilon)$ such that*

$$\mathbb{P}\left(\hat{K} = K,\ |\hat{e}_k - e_k| < \frac{A(\alpha,\epsilon)}{\min(\triangle_k, \triangle_k^2)} \log(n)^{1+\delta},\ |\hat{s}_k - s_k| < \frac{A(\alpha,\epsilon)}{\min(\triangle_k, \triangle_k^2)} \log(n)^{1+\delta} \quad 1 \leq k \leq K\right) \geq 1 - Cn^{-\epsilon}$$

holds for all $n \geq B(\alpha, \tilde{\delta}, \delta, \epsilon)$.

As in Theorem 1, the strengthened SIC was used as penalty to simplify the exposition of the proof. A very similar result could be derived to show that a $\alpha \log(n)$ type penalty for a sufficiently large α achieves consistency. This result suggests that CAPA, fits collective anomalies as such, despite being able to fit them as a mixture of point anomalies and data belonging to the typical distribution. It is possible to relax Assumption 2b to $e_k - s_k \leq D\sqrt{n}$ for some constant D, with the constants A, B, and C then also depending on D.

As in Theorem 1, the bounds on the accuracy of the detected epidemic changepoints can be tightened. For all $\delta > \delta_0$, it is possible to show that the detected epidemic changepoints must be within $\frac{A(\alpha,\epsilon)}{\min(\triangle_k, \triangle_k^2)} \log(n)^{1+\delta_0}$ of the true epidemic changepoints once $n \geq B(\alpha, \tilde{\delta}, \delta, \epsilon)$ and $n \geq B(\alpha, \tilde{\delta}, \delta_0, \epsilon)$. This is because Theorem 1 then also holds for the penalty $\alpha \log(n)^{1+\delta_0}$, which therefore guarantees that the fitted epidemic changepoints of the optimal partition with K fitted epidemic changepoints

must be within $\frac{A(\alpha,\epsilon)}{\min(\triangle_k,\triangle_k^2)}\log(n)^{1+\delta_0}$ of the true epidemic changepoints.

3.4.3 Penalties

We now turn to the problem of tuning the penalties β and β' introduced in Section 3.2. In the previous section, we showed that CAPA is consistent when using an $\alpha\log(n)^{1+\delta}$ penalty for both collective and point anomalies. The result can be tightened slightly to show that consistency is achieved by using an $\alpha\log(n)$ penalty for collective anomalies and an $\alpha'\log(n)$ penalty for point anomalies for sufficiently large constants α and α'. In practice, we recommend choosing α and α' with the aim of controlling the asymptotic rate of false positives under the null hypothesis that no anomalies – collective or point – are present.

Relatively tight results can be derived for the case in which the typical mean and variance are known and the observations are normally distributed. Indeed, the following proposition holds under these assumptions:

Proposition 2. *Let $x_1, ..., x_n$ be i.i.d. $N(\mu, \sigma^2)$ distributed, for known μ and σ. Then there exist constants C_1 and C_2 such that when the penalty for point anomalies, β', and the penalty for collective anomalies, β, satisfy $\beta' \geq 2\psi$ and $\beta \geq 2(2+2\psi+2\sqrt{2\psi})$, then*

$$\mathbb{P}\left(\hat{K} = 0,\ \hat{O} = \emptyset\right) \geq 1 - C_1 n e^{-\psi} - C_2\left(n e^{-\psi}\right)^2.$$

for all $\psi > 0$

The proof of this result relies on a bound of the MGF of the cost function under the data distribution and can therefore be extended to other cost functions and data

distributions and to the classical changepoint setting. As a consequence of Proposition 2, setting the penalty for collective anomalies to $(4+\epsilon)\log(n)$ and the penalty for point anomalies to $(2+\epsilon)\log(n)$, where $\epsilon > 0$ controls the asymptotic rate of false positives in this scenario, since the probability of observing false positives then tends to 0 as n tends to ∞.

Moreover, the following proposition holds under general (i.e. model-free) assumptions

Proposition 3. *Let $(\hat{\mu}, \hat{\sigma})$, $\hat{O}$, and β' correspond to the estimated parameters of the typical distribution, the inferred set of point anomalies, and the penalty for point anomalies used by CAPA respectively. Assume, moreover, that the constant γ, defined in Section 3.3, satisfies $1 \geq \gamma \geq \exp(-\beta')$. Then, there exists a constant $K(\beta') = \beta' + o(\beta')$ such that*

$$i \notin \hat{O} \implies (x_i - \hat{\mu})^2 < \hat{\sigma}^2 K(\beta')$$

and

$$(x_i - \hat{\mu})^2 > \hat{\sigma}^2 K(\beta') \implies i \in \hat{O} \vee \exists k \in \{1, ..., \hat{K}\} : i \in \{\hat{s}_k + 1, ..., \hat{e}_k\}$$

This proposition defines a threshold for $|x_i - \mu|/\sigma$ below which the ith observation will not be considered a point anomaly and above which the ith observation will not be considered typical. Furthermore, this detection threshold asymptotically behaves like β'. This provides a natural way of choosing the penalty β' for point anomalies based on how large outliers, as measured by $|x_i - \mu|/\sigma$ are assumed to be. For Gaussian data, extreme value theory places the threshold at $\sqrt{(2+\epsilon)\log(n)}$, which confirms the choice of β' derived from Proposition 2.

In practice, the typical parameters are often unknown and the constants α and α' should be slightly inflated to reflect this additional uncertainty. This effect on α, is offset by the fact that the Bonferroni correction used in the proof of Proposition 2 becomes loose when the minimum segment length exceeds 2 as it does not exploit correlations. We chose $\beta' = 3\log(n)$ and $\beta = 4\log(n)$ as default penalties for point and collective anomalies respectively in our software implementation and recommend inflating both penalties, whilst maintaining their ratio, when dealing with heavy tailed or autocorrelated data.

3.5 Simulation Study

To assess the potential of CAPA, we compare its performance to that of other popular anomaly and changepoint methods on simulated data. In particular, we compare with PELT as implemented in the changepoint package (Killick et al., 2018; Killick and Eckley, 2014), a commonly used changepoint detection method, luminol (Maheshwari et al., 2014), an algorithm developed by LinkedIn to detect segments of atypical behaviour, as well as BreakoutDetection (James et al., 2016) which was introduced by Twitter to detect changes in mean in a way which is robust to point anomalies.

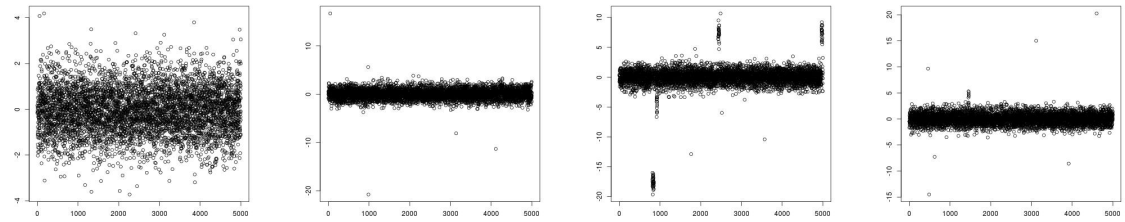
The simulation study was conducted over simulated time series each consisting of 5000 observations, for which the typical data follows a $N(0, 1)$ distribution. Epidemic changepoints occur at a rate of 0.0005 (corresponding to an average of about 2.5 epidemic changes in each series), with their length being i.i.d. $Pois(30)$-distributed. In each anomalous segment the data is again normally distributed, with the means

being i.i.d. $N(0, a^2)$ distributed and standard deviations i.i.d. $\Gamma(1/b, 1/b)$ distributed. We used $a = 1$, and $a = 10$ for weak and strong changes in mean respectively as well as $b = 1$ and $b = 10$ for weak and strong changes in variance respectively. Short anomalies and strong anomalies where also simulated using $a = b = 5$ and collective anomalies of i.i.d. $Pois(6)$-distributed length.

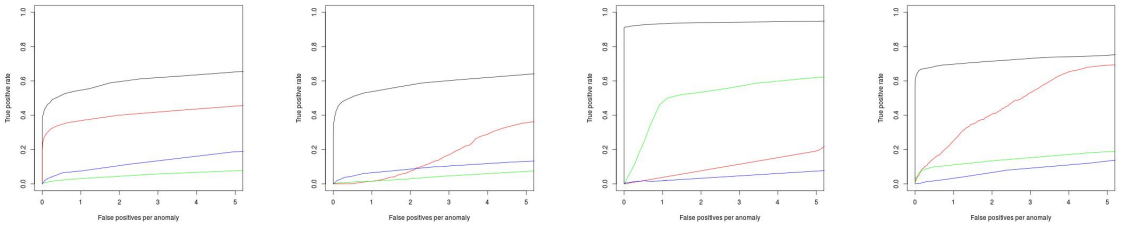
We compared the performance of the four methods in the presence of both strong and weak changes in mean and/or variance. We also repeated the analysis with 10 i.i.d. $N(0, 10^2)$ distributed point anomalies occurring at randomly sampled points in the typical data. Robustness to model misspecification was also investigated by using t_{10}-distributed noise and AR(1)-distributed noise with autocorrelation parameter $\rho = 0.3$. The comparison of these methods is made using the three different approaches we detail below.

3.5.1 ROC

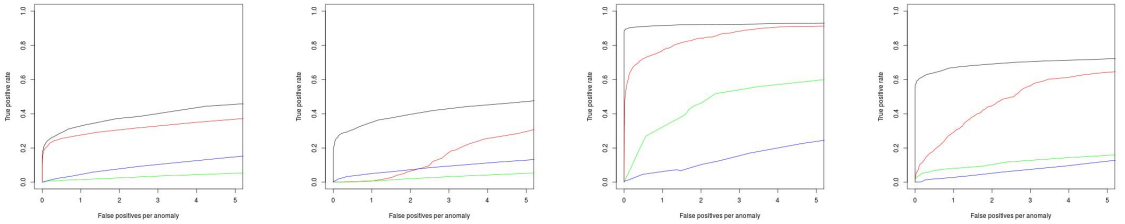
We obtained ROC curves for the four methods. For BreakoutDetection and PELT, we considered detected changes within 20 time points of true changes to be true positives and classified all other detected changes as false positives. For luminol and CAPA, we considered detected starting and end points of epidemic changes to be true positives if they were within 20 observations of a starting and end point respectively. The results regarding the precision of true positives in Section 3.5.2 suggest that the results in this section are robust with regard to the choice of error tolerance. We set the minimum segment length to ten for PELT, CAPA, and BreakoutDetection. To obtain the ROC curves we varied the penalty for collective anomalies β in CAPA, the penalty in PELT,



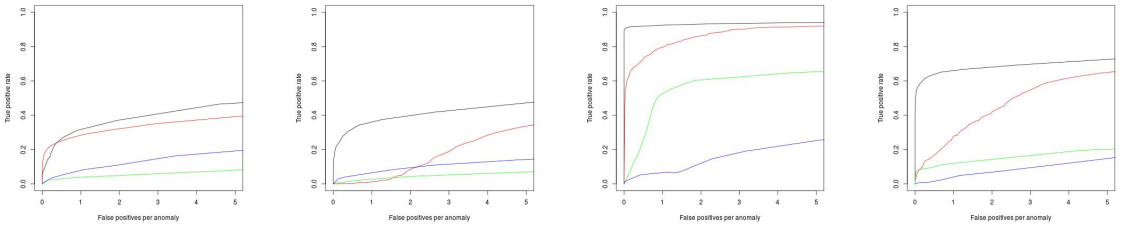
(a) Weak (b) Weak, PAs (c) Strong, PAs (d) Short, PAs



(e) Weak (f) Weak, PAs (g) Strong, PAs (h) Short, PAs



(i) Weak, AR (j) Weak, AR, PAs (k) Strong, AR, PAs (l) Short, AR, PAs



(m) Weak, T (n) Weak, T, PAs (o) Strong, T, PAs (p) Short, T, PAs

Figure 3.5.1: Data examples and ROC curves for changes in mean for CAPA (black), PELT (red), BreakoutDetection (green), and luminol (blue).

the threshold in luminol, and the beta parameter of BreakoutDetection.

The resulting ROC curves, as well as examples of realisations of the data for the scenario of weak and strong changes in mean can be found in Figure 3.5.1.The results for joint changes in mean and variance, as well as changes in variance can be found in the appendix. We see that CAPA generally outperforms PELT, even in the absence of point anomalies. This is due to it having more statistical power, by exploiting the epidemic nature of the change. This becomes particularly apparent when the changes are weak. CAPA also outperform BreakoutDetection and luminol for epidemic changes in mean, the scenario for which these methods were developed. Moreover, the performance of CAPA is barely affected by the presence of point anomalies, unlike that of the non-robust methods.

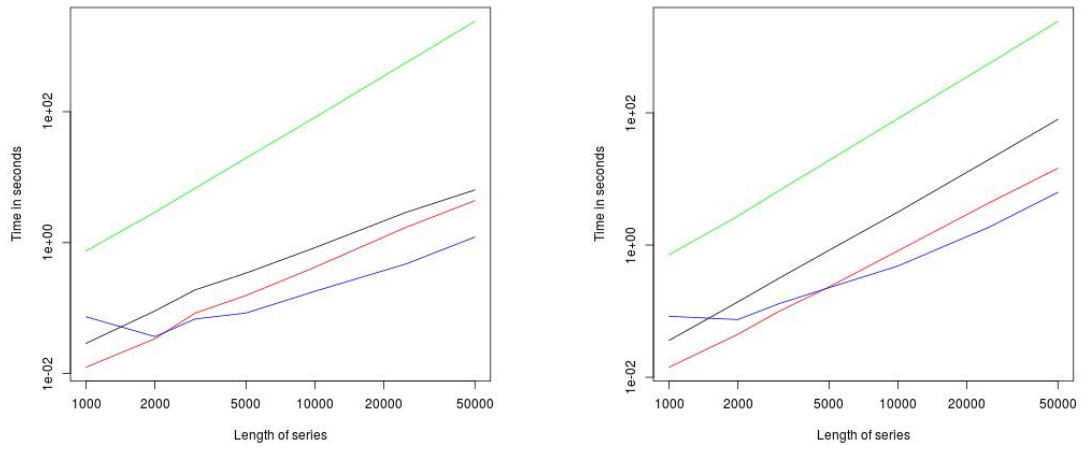
3.5.2 Precision

We also investigated the precision of the true positives for the four methods. We compared the mean absolute distance between detected changes (i.e. true changes which had a detected changes within 20 observations) and the nearest estimated change across all the 12 scenarios. We used the default penalties for all methods (i.e. the BIC for PELT) except BreakoutDetection, where we found that the default penalty returned no true positives at all for many scenarios. We therefore used the results we obtained when deriving the ROC curves to set the beta parameter to an appropriate level for each case.

The results of this analysis can be found in Table 3.5.1. We see that CAPA is generally the most precise one. Moreover, its precision is not too strongly affected by the

Mean	Variance	Point anomalies	CAPA	PELT	BreakoutDetection	luminol
weak	-	-	1.79	**1.50**	3.40	9.91
weak	-	10	**1.72**	2.27	3.75	10.70
strong	-	-	**0.16**	0.61	5.38	15.99
strong	-	10	**0.19**	0.67	4.68	15.60
-	weak	-	**1.41**	1.43	4.60	9.87
-	weak	10	**1.31**	1.89	4.49	10.76
-	strong	-	**0.33**	0.73	5.19	12.03
-	strong	10	**0.33**	0.79	5.17	11.29
weak	weak	-	**1.16**	1.30	4.00	11.40
weak	weak	10	**1.22**	1.63	4.00	11.30
strong	strong	-	**0.09**	0.56	3.78	16.31
strong	strong	10	**0.09**	0.58	3.77	15.71

Table 3.5.1: Precision of true positives measured in mean absolute distance for CAPA, PELT, luminol, and BreakoutDetection.



(a) With epidemic changes.

(b) Stationary data.

Figure 3.5.2: log-log-plot of the runtime of CAPA (black), PELT (red), BreakoutDetection (green), and luminol (blue).

presence of point anomalies, unlike that of PELT, whose performance is significantly deteriorated by anomalies, especially when the signal is weak. The reason for this is that PELT fits additional changes in the presence of anomalies, which results in shorter segments. This leads to less accurate parameter estimates, and consequently poorer estimates for the location of the changepoint. CAPA does not face this problem since the parameter of the typical distribution is shared across all segments.

3.5.3 Runtime

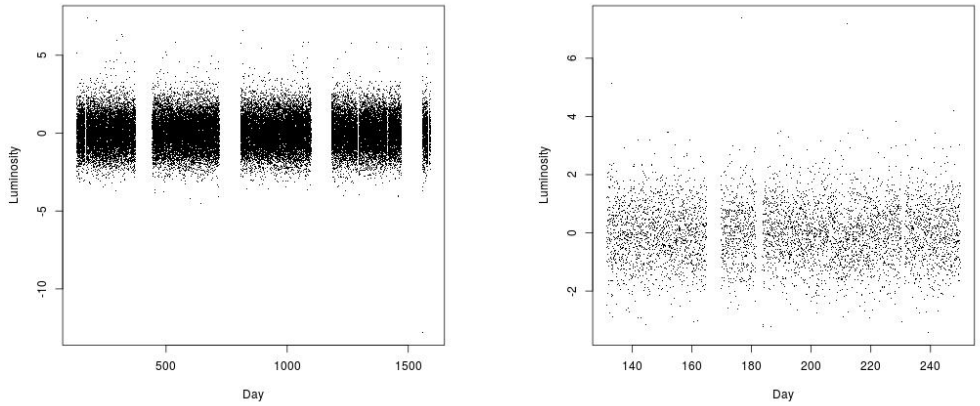
Finally, we investigated the relationship between the runtime of the 4 methods and the number of observations. Our comparison is based on data following a distribution identical to the one we used in Sections 3.5.1 and 3.5.2. Since this type of data

favours PELT and CAPA, because the expected number of changes increases with the number of observations, we also compared the runtime of the four methods on stationary $N(0, 1)$ data, which represents the worst case scenario for these methods.

Figure 3.5.2 displays the average speed over 50 repetitions for the two cases. When comparing the slopes between 10000 and 50000 datapoints we note that the slope is very close to 2 for BreakoutDetection in both cases as well as CAPA and PELT for stationary data, suggesting quadratic scaling. In the presence of epidemic changes however, that slope is 1.26 for CAPA – 1.14 even between 25000 and 50000 datapoints – thus suggesting near linear runtime.

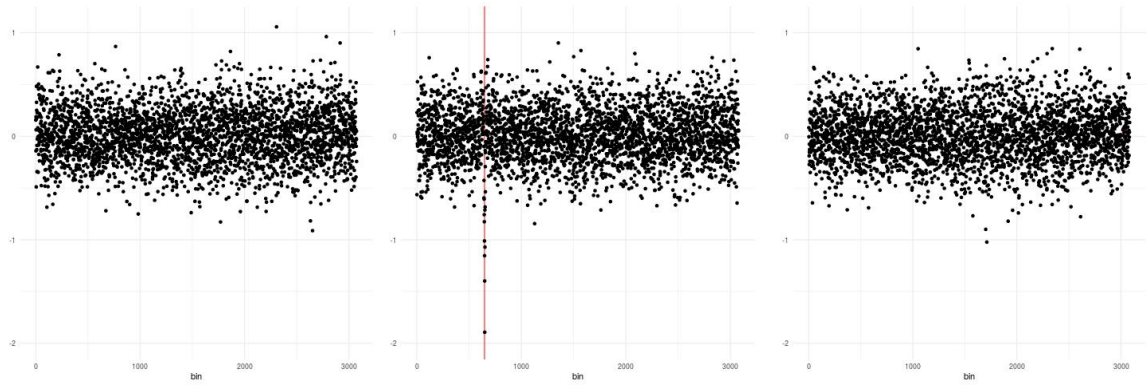
3.6 Applications

We now turn to applying CAPA to Kepler lightcurve data and a machine temperature dataset. These two applications are illustrative of the two main flavours of anomaly detection: In the Kepler lightcurve data, anomalies can point to the presence of an exoplanet, i.e. a potentially exciting scientific discovery. This application is therefore an example of novelty detection, similar in spirit to other applications such as the detection of copy number variations (Bardwell and Fearnhead, 2017) or the analysis of fMRI data (Aston and Kirch, 2012). Conversely, in the machine temperature dataset anomalies can point towards a potentially critical problem. It therefore falls under the same category as fraud and other fault detection procedures.



(a) Full data. (b) Subset (120 days).

Figure 3.6.1: Light curve of Kepler 1132, obtained at approximately 30 minute intervals. Missing values are due to periods in which the star was not observed. Note the presence of a point anomaly on day 1550 – with luminosity -13 – and the fact that no transit signature is apparent to the eye. This remains true after zooming in on a 120 day subset of the data, despite the known presence of Kepler 1132-b, an exoplanet orbiting this star every 62.9 days.



(a) 62.8 days. (b) 62.9 days. (c) 63.0 days.

Figure 3.6.2: CAPA applied with default penalties to the light curve of Kepler 1132 preprocessed using different periods.

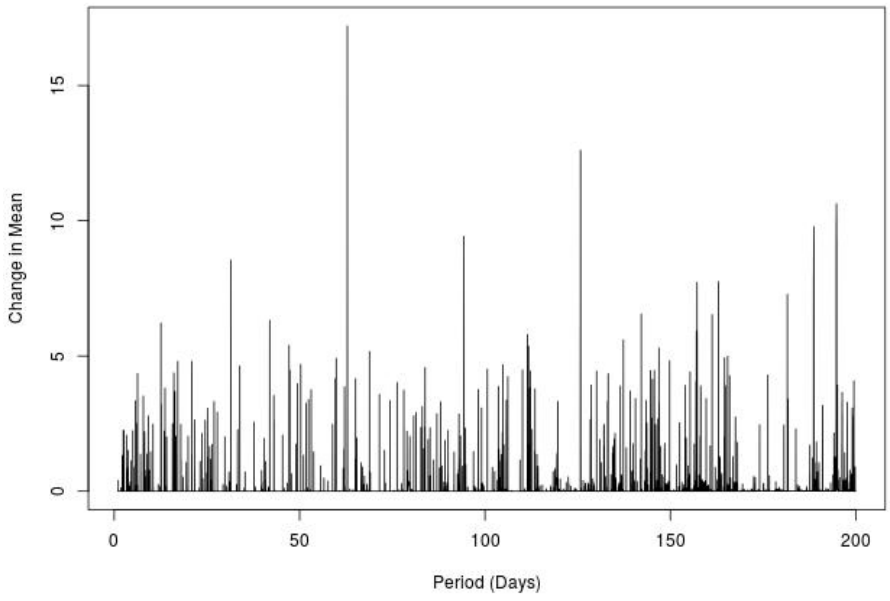


Figure 3.6.3: The strongest change in mean, as measured by $\max_k(\triangle_{\mu,k})$, detected by CAPA for the lightcurve of Kepler 1132. All periods from 1 to 200 days at 0.01 day increment were examined.

3.6.1 Kepler Light Curve Data

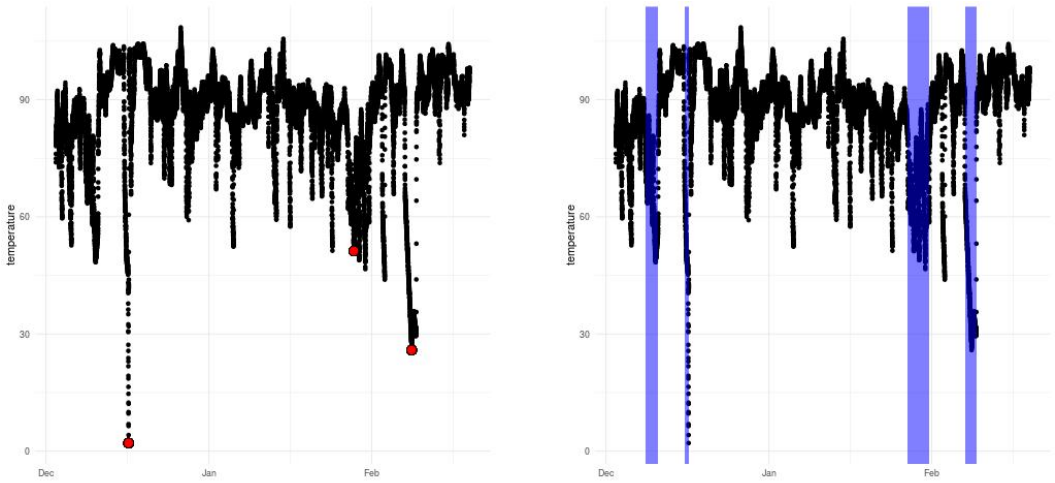
CAPA can be applied to the Kepler light curve data to detect exoplanets via the so called transit method (Sartoretti and Schneider, 1999) first proposed by Struve (1952). This method consist of measuring the luminosity of a star at regular intervals, with the aim of detecting segments of reduced luminosity. These indicate the transit of a planet (Sartoretti and Schneider, 1999) past the star, as in an eclipse, and can naturally be interpreted as collective anomalies: a short period of reduced luminosity followed by a return to the baseline level. The light curves are typically preprocessed (Mullally, 2016) and both the raw and whitened light curves for over 40 million stars can be accessed online. We have included the whitened light curve of the star Kepler 1132 in Figure 3.6.1 to illustrate the nature of this type of data. We note the presence of a global anomaly on day 1550 and the noisy nature of the data. These make the detection of transits challenging given the weak signal induced by planetary transits. Indeed, even the transit of Jupiter past the sun reduces the latter's luminosity by only 1% (Sartoretti and Schneider, 1999). The ability to automate the analysis of light curves would be beneficial, given the large number of light curves that have been gathered and need analysing.

Since the reduction in luminosity caused by transiting planets is known to be weak, we amplify the signal to noise ratio by exploiting the periodic nature of the transit signal. If the period of an orbiting planet were known, the signal of its transit could be strengthened by considering all data points to have been gathered at their measurement time modulo that period. We would thus obtain an irregularly sampled

time series which we can transform into a regularly sampled time series by binning the data into equally sized bins and taking the average within each bin. We could then apply CAPA to this preprocessed data, which would exhibit a stronger mean anomaly for any planet with the associated period. The results obtained by applying this method to the light curve of Kepler 1132 using a period of 62.8, 62.9, and 63.0 days can be found in Figure 3.6.2. We note that using a period of 62.9 days results in a promising dip, which is not present when using 62.8 or 63 days as period.

Given a light curve, the periods of exoplanets orbiting the corresponding star, if any are present, are obviously not known *a priori*. However, we can use the approach described in the previous paragraph to test a range of candidate periods: Since transits appear as periods of reduced mean, we simply record the strength of the strongest change in mean, as defined by $\max_k (\triangle_{\mu,k})$ (see Section 3.4), of any detected collective anomalies for each candidate period, expecting it to be largest for the orbital period of exoplanets. This approach can be compared with a spectral method. We tried all periods from 1 day to 200 days with increments of 0.01 days for the light curve of Kepler 1132. The result of this analysis can be found in Figure 3.6.3. Note that the largest change in mean is recorded at a period of 62.89 days. This result is consistent with the existing literature, which considers Kepler 1132 to be orbited approximately every 62.892 days by the exoplanet Kepler 1132-b, whose radius is about 2.5 times that of the earth (Morton et al., 2016). As with spectral methods, we also observe resonance of the main signal at integer fractions of that period.

Furthermore, recalling that there are currently over 40 million recorded light curves for different starts, we require a fast automated procedure. The cost of running CAPA



(a) Raw data with labels.

(b) CAPA output.

Figure 3.6.4: Machine temperature data with anonymised y-axis. The labels were provided by an engineer.

for one putative orbital period is small relative to the cost of binning the data for that period. Analysis of the Kepler 1132 data for all 20,000 orbital periods considered took less than 5 minutes on a standard laptop.

We also applied CAPA to the light curves of further stars with confirmed exoplanets and were able to detect their transit signal at the right period. A more detailed exposition of these results can be found in the appendix.

3.6.2 Machine Temperature Data

We now turn to analysing the machine temperature data taken from the Numenta Anomaly Benchmark Ahmad et al. (2017) and included, with permission, in the `anomaly` package. The data, displayed in Figure 3.6.4a, were obtained from a heat sensor on a large industrial machine over the course of approximately 3 months at a 5

minute sampling frequency. The data set, consisting of $n = 22695$ observations, was analysed by an engineer who identified three relevant events: A planned shutdown, a catastrophic system failure, as well as a period of anomalous behaviour preceding the failure which was, in hindsight, deemed to have been an early warning sign.

Given that there is a large amount of structure in the data, most notably auto-correlation, we used Minimum Covariance Determinant (MCD) covariance estimator (Rousseeuw, 1984) to robustly estimated the AR(1) coefficient $\rho = 0.98$ and, following Lavielle and Moulines (2000) inflated the default penalties by a factor of $\frac{1+\rho}{1-\rho}$. We then applied CAPA to the data using a maximum segment length $m = 1500$ to prevent the fitting of long anomalies which are merely a result of model mis-specification. The results can be seen in Figure 3.6.4b. Note that all anomalies flagged by the engineer where correctly detected. Furthermore, only a single additional collective anomaly was detected. We refrain from calling it a false positive as it is not possible to know what would have happened without the planned shut-down.

Chapter 4

Multivariate Collective And Point Anomalies

4.1 Introduction

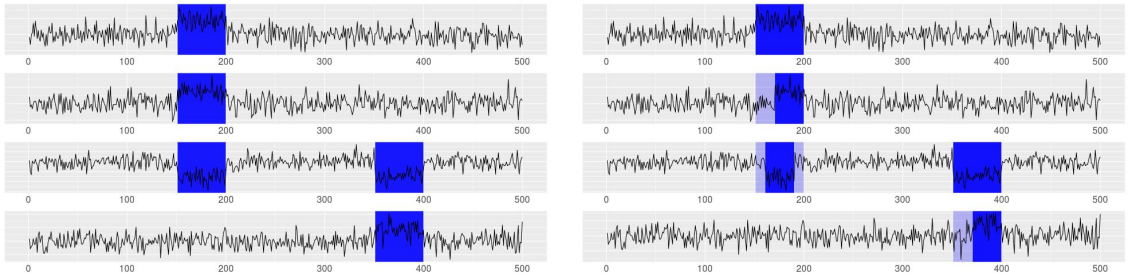
The field of anomaly detection has attracted considerable attention in recent years, in part due to an increasing need to automatically process large volumes of data gathered without human intervention. Comprehensive reviews of the area can be found in Chandola et al. (2009) and Pimentel et al. (2014). More recently the detection of anomalies in multivariate data has become more important (Boudt et al., 2020; Chen et al., 2020).

In this article, we focus on the following setting: we observe a multivariate time series $\mathbf{x}_1, ..., \mathbf{x}_n \in \mathbb{R}^p$ corresponding to n observations observed across p different components. Each component of the series has a typical behaviour, interspersed by windows of time where it behaves anomalously. In line with the definition in Chandola

et al. (2009), we call the behaviour within such a time window a collective anomaly. Often the underlying cause of such a collective anomaly will affect more than one, but not necessarily all, of the components. Our aim is to accurately estimate the location of these collective anomalies within the multivariate series, potentially in the presence of point anomalies.

Whilst it may be mathematically convenient to assume that anomalous structure occurs contemporaneously across all affected sequences, in practice one might expect some time delays (i.e. offsets or lags), as illustrated by Figure 4.1.1. In this article, we will consider two different scenarios for the alignment of related collective anomalies across different components. The first, is that concurrent collective anomalies perfectly align. That is, we can segment our time series into windows of typical and anomalous behaviour, with the latter affecting a subset of components. The second, and for many applications more realistic, setting considered in this chapter assumes that concurrent collective anomalies start and end at similar but not necessarily identical time points.

Current approaches aimed at detecting collective anomalies can broadly be divided into state space approaches and (epidemic) changepoint methods. State space models assume that a hidden state, which evolves following a Markov chain, determines whether the time series' behaviour is typical or anomalous. Examples of state space models for anomaly detection can be found in Bardwell and Fearnhead (2017) and Smyth (1994). These models have the advantage of providing interpretable output in the form of probabilities of certain segments being anomalous. However, they are often slow to fit and are sensitive to the choice of prior distributions, which can be



(a) No lags

(b) Lags are indicated in light blue.

Figure 4.1.1: A time series with $K = 2$ collective anomalies, highlighted in blue. Using notation from Section 4.2 these collective anomalies occur from times $s_1 = 150$ to $e_1 = 200$ and $s_2 = 350$ to $e_2 = 400$; the affected components are $\mathbf{J}_1 = \{1, 2, 3\}$ and $\mathbf{J}_2 = \{3, 4\}$.

difficult to specify.

The epidemic changepoint model (Levin and Kline, 1985) provides an alternative detection framework, built on an assumption that there is a typical behaviour from which the model deviates during certain windows. Each epidemic changepoint consists of two classical changepoints, one away from and one returning back to the typical distribution. Epidemic changepoints can be inferred by using classical changepoint methods (Killick et al., 2012; Fryzlewicz, 2014; Wang and Samworth, 2018). However, such approaches lead to sub-optimal power, as they do not exploit the fact that the typical parameter is shared across the non-anomalous segments.

Many epidemic changepoint methods are based on the popular circular binary segmentation algorithm (Olshen et al., 2004), an epidemic changepoint version of binary segmentation. For multivariate data, the key challenge for these methods is that theoretically detectable anomalies can either be sparse, with a few components

exhibiting strongly anomalous behaviour, or dense, with a large proportion of components exhibiting potentially very weak anomalous behaviour (Cai et al., 2011b; Jeng et al., 2012). A range of different epidemic changepoint methods have been proposed that use circular binary segmentation to deal with different types of anomalies: the methods of Zhang et al. (2010) for dense changes, LRS (Jeng et al., 2010) for sparse changes, and higher criticism (Donoho and Jin, 2004) based methods like PASS (Jeng et al., 2012) for both types of changes.

The approach we present in this chapter is fundamentally different from this earlier work. It builds on a penalised cost based test statistic to detecting collective anomalies, which we introduce in Section 4.2. As compared to earlier work, this approach is general, as we can choose different costs to detect different type of anomalies. It can also be model-based, as the cost is most naturally defined in terms of the negative log-likelihood of the data under an appropriate model for data in normal and anomalous segments. One of the challenges with implementing such an approach for multivariate data is how to choose the penalty, and in particular how the penalty varies as the number of anomalous series varies. By focussing on the case of at most one anomalous region we are able to propose appropriate penalties, and show that this choice both controls the false positive rate and has optimal power to detect both sparse and dense anomalies in the case of i.i.d. Gaussian data where anomalies correspond to changes in the mean of the data.

An advantage of this penalised cost approach is that we can extend the method from detecting a single anomaly to detecting multiple anomalies without having to resort to binary segmentation algorithms – instead we can exactly and efficiently opti-

mise the penalised cost criteria over the unknown number and position of the anomalies. We show how to extend this approach so as to allow for both point anomalies, and for the estimated anomalous segments to be misaligned across series. Furthermore we present a theoretical result that gives intuition regarding the trade-offs involved when specifying the maximum lag. The resulting algorithm is called **M**ulti-**V**ariate **C**ollective **A**nd **P**oint **A**nomalies (MVCAPA). MVCAPA has been implemented in the R package `anomaly` for detecting collective anomalies which correspond to changes in mean, or changes in mean and variance.

We give finite sample consistency results for MVCAPA for the problem of detecting anomalies that change the mean in Gaussian data in Section 4.5. Our main result shows the consistency of estimates of the number and location of the anomalous segments, albeit for the special case where we ignore point anomalies or any misalignment of the collective anomalies, and for a slightly different penalty regime from the one we recommend in practice. However we are unaware of any similar consistency result for detecting collective anomalies in multivariate data. Whilst there are similar consistency results for the related problem of detecting changes in multivariate data (Wang and Samworth, 2018; Cho and Fryzlewicz, 2015), these are for versions of methods that assume a minimum segment length, and under the condition that this increases with the amount of data. We do not assume MVCAPA imposes a minimum segment length in our proof. This greatly increases the technical challenge – as one of the main issues with dealing with multiple anomalies is showing that a single true anomaly is not fit using multiple collective anomalies, something that can be easily excluded if our inference procedure assumes a diverging minimum segment length.

Dealing with this challenge is particularly complicated in our setting, due to the possibility of fitting an anomaly as either sparse or dense.

4.2 Model and Inference for a Single Collective Anomaly

4.2.1 Penalised Cost Approach

We begin by focussing on the case where collective anomalies are perfectly aligned. We consider a p-dimensional data source for which n time-indexed observations are available. A general model for this situation is to assume that the observation $\mathbf{x}_t^{(i)} \in \mathbb{R}$, where $1 \leq t \leq n$ and $1 \leq i \leq p$ index time and components respectively, comes from a parametric family of distributions, which may depend on earlier observations of component i, and whose parameter, $\boldsymbol{\theta}^{(i)}(t) \in \mathbb{R}$, depends on whether the observation is associated with a period of typical behaviour or an anomalous window. Conditional on $\boldsymbol{\theta}^{(i)}(t)$, the series are assumed to be independent. We let $\boldsymbol{\theta}_{\mathbf{0}}^{(i)}$ denote the parameter associated with component i during its typical behaviour. Let K be the number of anomalous windows, with the k-th window starting at $s_k + 1$ and ending at e_k and affecting the subset of components denoted by the set $\mathbf{J}_k$. We assume that anomalous windows do not overlap, so $e_k \leq s_{k+1}$ for $k = 1, \ldots, K-1$. We let l denote the minimum length of a collective anomaly, and impose $e_k - s_k \geq l$ for each k; setting $l = 1$ imposes no minimum length. Our model then assumes that the parameter

associated with observation $\mathbf{x}_t^{(i)}$ is

$$\boldsymbol{\theta}^{(i)}(t) = \boldsymbol{\theta}_k^{(i)} \text{ if } s_k < t \leq e_k \text{ and } i \in \mathbf{J}_k \tag{4.2.1}$$

and $\boldsymbol{\theta}_0^{(i)}$ otherwise.

We start by considering the case where there is at most one collective anomaly, i.e. where $K \leq 1$, and introduce a test statistic to determine whether a collective anomaly is present and, if so, when it occurred and which components were affected. The methodology will be generalised to multiple collective anomalies in Section 4.3. Throughout we assume that the parameter, $\boldsymbol{\theta}_0$, representing the sequence's baseline structure, is known. In practice we can estimate $\boldsymbol{\theta}_0$ robustly over the whole data, as in Fisch et al. (2018a).

Given the start and end of a window, (s, e), and the set of components involved, $\mathbf{J}$, we can calculate the log-likelihood ratio statistic for the collective anomaly. To do so, we introduce a cost, which in the case of i.i.d. observations from a density $f(x, \boldsymbol{\theta})$ is

$$\mathcal{C}_i\left(\mathbf{x}_{s+1:e}^{(i)}, \boldsymbol{\theta}\right) = -2\sum_{t=s+1}^{e} \log f(\mathbf{x}_t^{(i)}, \boldsymbol{\theta}),$$

obtained as minus twice the log-likelihood of data $\mathbf{x}_{s+1:e}^{(i)}$ for parameter $\boldsymbol{\theta}$. For dependent data we can replace $f(\mathbf{x}_t^{(i)}, \boldsymbol{\theta})$ by the conditional density of $\mathbf{x}_t^{(i)}$ given $\mathbf{x}_{1:(t-1)}^{(i)}$. We can then quantify the saving obtained by fitting component i as anomalous for the window starting at $s+1$ and ending at e as

$$\mathcal{S}_i(s, e) = \mathcal{C}_i\left(\mathbf{x}_{(s+1):e}^{(i)}, \boldsymbol{\theta}_0^{(i)}\right) - \min_{\boldsymbol{\theta}}\left(\mathcal{C}_i\left(\mathbf{x}_{(s+1):e}^{(i)}, \boldsymbol{\theta}\right)\right).$$

For example, to detect anomalies that correspond to changes in the mean of the data we can use a cost based on a Gaussian model for the data, $\mathcal{C}_i\left(\mathbf{x}_{s+1:e}^{(i)}, \boldsymbol{\theta}\right) =$

$\sum_{i=s+1}^{e}((\mathbf{x}_t^{(i)} - \boldsymbol{\mu}_0^{(i)})/\boldsymbol{\sigma}_0^{(i)})^2$, where $\boldsymbol{\mu}_0^{(i)}$ and $\boldsymbol{\sigma}_0^{(i)}$ are the typical mean and standard deviation of the ith component respectively. This leads to a saving

$$\mathcal{S}_i(s,e) = \frac{(e-s)}{\left(\boldsymbol{\sigma}_0^{(i)}\right)^2}\left(\frac{1}{e-s}\sum_{t=s+1}^{e}\mathbf{x}_t^{(i)} - \boldsymbol{\mu}_0^{(i)}\right)^2,$$

If anomalies could correspond to changes in either or both of the mean and variance, we can again base the cost on a Gaussian model but allow both mean and variance to be estimated for an anomalous region. This leads to savings

$$\mathcal{S}_i(s,e) = \sum_{t=s+1}^{e}\left(\frac{\mathbf{x}_t^{(i)} - \boldsymbol{\mu}_0^{(i)}}{\boldsymbol{\sigma}_0^{(i)}}\right)^2 - (e-s-1)\left(\log\left(\frac{\sum_{t=s+1}^{e}\left(\mathbf{x}_t^{(i)} - \frac{1}{e-s}\sum_{t'=s+1}^{e}\mathbf{x}_{t'}^{(i)}\right)^2}{(e-s)\left(\boldsymbol{\sigma}_0^{(i)}\right)^2}\right) + 1\right).$$

Similarly, for count data, we could base our cost on a Poisson or negative-binomial model for the data.

Given a suitable cost function, the log-likelihood ratio statistic is $\sum_{i\in\mathbf{J}}\mathcal{S}_i(s,e)$. As the start or the end of the window, or the set of components affected, are unknown *a priori*, it is natural to maximise the log-likelihood ratio statistic over the range of possible values for these quantities. However, in doing so, we need to take account of the fact that different $\mathbf{J}$ will allow different numbers of components to be anomalous, and hence will allow maximising the log-likelihood, or equivalently minimising the cost, over differing numbers of parameters. This suggests penalising the log-likelihood ratio statistic differently, depending on the size of $\mathbf{J}$. That is we test the null hypothesis of there being no anomaly by calculating

$$\max_{\mathbf{J},s\leq e-l}\left[\sum_{i\in\mathbf{J}}\mathcal{S}_i(s,e) - P(|\mathbf{J}|)\right], \tag{4.2.2}$$

where $P(.)$ is a suitable positive penalty function of the number of components that change, and l is the minimum segment length. We will detect an anomaly if (4.2.2)

is positive, and estimate its location and the set of components that are anomalous based on the values of s, e, and $\mathbf{J}$ that give the maximum of (4.2.2).

To efficiently maximise (4.2.2), define positive constants α, $\beta_{1:p}$ with $P(1) = \alpha+\beta_1$, and, for $i = 2, \ldots, p$, $\beta_i = P(i) - P(i-1)$. So the β_is are the first differences of our penalty function $P(\cdot)$. Further, let the order statistics of $\mathcal{S}_1(s,e), \ldots, \mathcal{S}_p(s,e)$ be

$$\mathcal{S}_{(1)}(s,e) \geq ... \geq \mathcal{S}_{(p)}(s,e),$$

and define the penalised saving statistic of the segment $\mathbf{x}_{(s+1):e}$,

$$\mathcal{S}(s,e) = \max_k \left(\sum_{i=1}^{k} \left\{ \mathcal{S}_{(i)}(s,e) - \beta_i \right\} \right) - \alpha.$$

Then (4.2.2) is obtained by maximising $\mathcal{S}(s,e)$ over s and e, subject to $e - s \geq l$.

Clearly, α and β_1 are only well specified up to their sum and α can be absorbed into β_1 without altering the properties of our statistic. However, not doing so can have computational advantages: it removes the need of sorting if all the β_is are identical and equal to β, say, when

$$\mathcal{S}(s,e) = \sum_{i=1}^{p} \left(\mathcal{S}_i(s,e) - \beta \right)^+ - \alpha.$$

4.2.2 Choosing Appropriate Penalties

The choice of penalties will impact both the overall false error rate of the approach and how the power to detect an anomaly varies with the number of components that are affected. In particular, we want a penalty function, $P(\cdot)$, that allows us to match optimal power results for both sparse and dense anomalies, whilst having a false error rate that asymptotically tends to 0. In practice, we then suggest fixing the shape of the

penalty function and to use simulation from an appropriate model with no anomalies, to scale the penalty function to achieve a desired false error rate. Such a tuning of the penalty function is straightforward, as it involves tuning a single scaling factor, whilst making the choice of penalty robust to both deviations from assumptions and looseness in the bounds on the false error-rate.

The optimal power results correspond to models with a change in mean in Gaussian data – for which the savings using the square error loss, or equivalently the Gaussian log-likelihood, have a χ^2_1 distribution under the null. We thus derive penalty regimes under an assumption that the savings can be stochastically bounded by $a\chi^2_v$ under the null hypothesis that no anomalies are present for some positive integer v and some positive real number a. This bound also holds for a wide variety of other cost functions under a range of different assumptions. If the cost is based on twice the negative log-likelihood, the savings are equal to the deviance and, if standard regularity conditions hold, converge to a χ^2_v distribution as $e - s \to \infty$. Also, when the Gaussian log-likelihood is used to detect changes in mean the bound holds under a range of model mis-specifications, such as when the time series are i.i.d. sub-Gaussian with parameter a; or when data from each component follow an independent AR(1)-models with bounded positive auto-corrletion parameter (Lavielle and Moulines, 2000).

Our bounds on the false positive rate will be based on showing

$$\mathbb{P}\left(\hat{K} = 0\right) \geq 1 - Ae^{-\psi(p,n)}, \tag{4.2.3}$$

where A is a constant and $\psi := \psi(p, n)$ increases with n and/or p. The appropriate choice of ψ will depend on the setting. In panel data the number of time points n

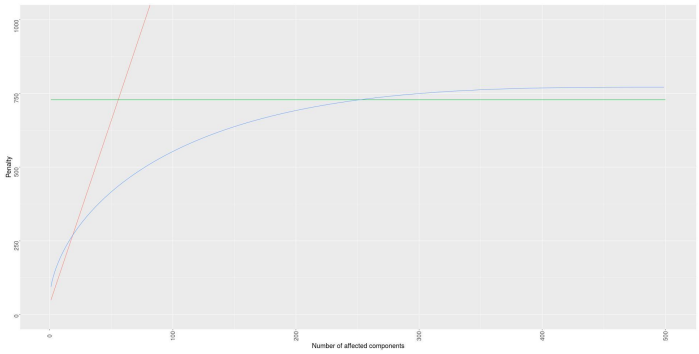


Figure 4.2.1: A comparison of the 3 penalty regimes for a χ^2_1-distributed saving when $p = 500$ and $\psi = 2\log(10000)$. Regime 1 is in green, regime 2 in red and regime 3 in blue.

may be small but we may have data from a large number of components, p. Setting $\psi(p, n) \propto \log(p)$ is therefore a natural choice so that the false positive probability tends to 0 as p increases. In a streaming data context, the number of sampled components p is typically fixed, while the number of observations n increases, so setting $\psi(p, n) \propto \log(n)$ is then natural.

We present three different penalty regimes (see Figure 4.2.1), each with power to detect anomalies with different proportions of anomalous segments. The regimes will be indexed by a parameter ψ which corresponds to the exponent of the probability bound, as defined in (4.2.3). We denote the penalty functions for each of these regimes by P_1, P_2 and P_3 respectively. The first penalty regime consists of just a single global penalty:

Penalty Regime 1: $P_1(j) = a\left(pv + 2\sqrt{pv\psi} + 2\psi\right)$, corresponding to setting $\beta_j = 0$ for $1 \leq j \leq p$ and $\alpha = a\left(pv + 2\sqrt{pv\psi} + 2\psi\right)$.

Under this penalty, we would infer that any detected anomaly region will affect

all components. This is inappropriate, and is likely to lead to a lack of power, if we have anomalous regions that only affect a small number of components. For such anomalies, the following regime offers a good alternative as it has a smaller penalty for fitting collective anomalies with few components:

Penalty Regime 2: $P_2(j) = 2(1+\epsilon)a\psi + 2(1+\epsilon)aj\log(p)$ which corresponds to setting $\alpha = 2(1+\epsilon)a\psi$ and $\beta_j = 2(1+\epsilon)a\log(p)$, for $1 \leq j \leq p$ and $\epsilon > 0$.

Comparing penalty regime 2 with penalty regime 1, we see that it has a lower penalty for small j, but a much higher penalty for $j >> \sqrt{p}/\log p$. As such it has higher power against collective anomalies affecting few components, but low power if the collective anomalies affect most components.

If $v \leq 2$, a third penalty regime can be derived:

Penalty Regime 3:

$$P_3(j) = a\left(2(\psi + \log(p)) + jv + 2pc_j f(c_j) + 2\sqrt{(jv + 2pc_j f(c_j))(\psi + \log(p))}\right),$$

where f is the PDF of the χ^2_v-distribution and c_j is defined via the implicit equation $\mathbb{P}(\chi^2_v > c_j) = j/p$.

As can be seen from Figure 4.2.1 for the special case of χ^2_1-distributed savings, this penalty regime provides a good alternative to the other penalty regimes, with lower penalties for intermediate values of $|\boldsymbol{J}|$.

All these regimes control the false positive rate, as shown in the following proposition.

Proposition 4. *Let the savings $S_i(s,e)$ be independent and stochastically bounded by $a\chi^2_v$ for $1 \leq i \leq p$ and let $\hat{K}$ denote the number of inferred collective anomalies. If we*

use penalty regime 1 or 2, or if $v \leq 2$ *and we use penalty regime 3, then, there exists a global constant* A *such that* $\mathbb{P}\{\hat{K} = 0\} \geq 1 - An^2 e^{-\psi}$.

Rather than choosing one penalty regime, we can maximise power against both sparse, intermediate and dense anomalies, by choosing $\alpha, \beta_1, ..., \beta_p$ so that the resulting penalty function $P(j)$, is the point-wise minimum of the penalty functions $P_1(j)$, $P_2(j)$, and, if available, $P_3(j)$. We call this the **composite regime**. It is a corollary from Proposition 4, that this composite penalty regime achieves $\mathbb{P}\{\hat{K} = 0\} \geq 1 - 3An^2 e^{-\psi}$ for the same global constant A as in Proposition 4, when $S_i(s, e)$ is stochastically bounded by $a\chi^2_v$.

4.2.3 Results on Power

For the case of a collective anomaly characterised by changes in the mean in a subset of the data's components, we can compare the power of our penalised saving statistic with established results regarding the optimal power of tests. Specifically, we examine behaviour under a large p regime. We follow the asymptotic parameterisation of Jeng et al. (2012) and therefore assume that the collective anomaly is of the form

$$\mathbf{x}_t^{(i)} = v^{(i)}\mu + \boldsymbol{\eta}_t^{(i)}, \quad v^{(i)} \sim \begin{cases} 0 & \text{with prob. } 1 - p^{-\xi}, \\ 1 & \text{with prob. } p^{-\xi}, \end{cases} \quad \text{and} \quad \boldsymbol{\eta}_t^{(i)} \overset{i.i.d.}{\sim} N(0,1), \quad \text{for} \quad s < t \leq e, \tag{4.2.4}$$

the noise $\boldsymbol{\eta}_t^{(1)}, ..., \boldsymbol{\eta}_t^{(p)}$ of the different series being independent.

Typically (Jeng et al., 2012), changes are characterised as either sparse or dense. In a sparse change, only a few components are affected. Such changes can be detected based on the saving of those few components being larger than expected after

accounting for multiple testing. The affected components therefore have to experience strong changes to be reliably detectable. On the other hand, a dense change is a change in which a large proportion of components exhibits anomalous behaviour. A well defined boundary between the two cases exists with $\xi \leq \frac{1}{2}$ corresponding to dense $\xi > \frac{1}{2}$ and corresponding to sparse changes (Jeng et al., 2012; Enikeeva and Harchaoui, 2019). Depending on the setting, the change in mean is parameterised by $r_p \in \mathbb{R}$ in the following manner:

$$(e-s)\mu^2 = \begin{cases} 2r_p \log(p) & \frac{1}{2} < \xi < 1, \\ p^{-2r_p} & 0 \leq \xi \leq \frac{1}{2}. \end{cases}$$

Both Jeng et al. (2012) and Cai et al. (2011b) derive detection boundaries for r_p, separating changes that are too weak to be detected from those changes strong enough to be detected. For the case in which the standard deviation in the anomalous segment is the same as the typical standard deviation, the detectability boundaries correspond to $\rho^- = (1 - \sqrt{1-\xi})^2$ if $3/4 < \xi < 1$, $\rho^- = \xi - 1/2$ if $1/2 < \xi \leq 3/4$ for the sparse case; and $\rho^+ = (1/2 - \xi)/2$ for the dense case ($0 \leq \xi \leq \frac{1}{2}$). The following proposition establishes that the penalised saving statistic has power against all sparse changes within the detection boundary, as well as against dense changes within the detection boundary

Proposition 5. *Let the typical mean be known and the series $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ contain an anomalous segment $\boldsymbol{x}_{s+1}, ..., \boldsymbol{x}_e$, which follows the model specified in (4.2.4). Let $r_p > \rho^-$ if $\frac{1}{2} < \xi < 1$ or $r_p < \rho^+$ if $0 \leq \xi \leq \frac{1}{2}$. Then the number of collective anomalies, $\hat{K}$, estimated by MVCAPA using the composite penalty with $a = 1$, $v = 1$ and $\psi(p, n) =$*

$2\log(n) + 2\log(\log(p))$ *on the data* $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$*, satisfies*

$$\mathbb{P}\left(\hat{K} \neq 0\right) \to 1 \quad as \quad p \to \infty$$

provided that $log(n) = o(log(p))$*.*

Rather than requiring μ_i to be 0, or a common value μ, it is trivial to extend the result to the case where $\mu_1, ..., \mu_p$ are i.i.d. random variables whose magnitude exceeds μ with probability $p^{-\xi}$. It is worth noticing that the third penalty regime is required to obtain optimal power against the intermediate sparse setting $\frac{1}{2} < \xi \leq \frac{3}{4}$.

4.3 Inference for Multiple Anomalies

A natural way of extending the methodology introduced in Section 4.2 to infer multiple collective anomalies, is to maximise the penalised saving jointly over the number and location of potentially multiple anomalous windows. That is we infer $\hat{K}$, $\left(\hat{s}_1, \hat{e}_1, \hat{\boldsymbol{J}}_1\right)$,..., $\left(\hat{s}_{\hat{K}}, \hat{e}_{\hat{K}}, \hat{\boldsymbol{J}}_{\hat{K}}\right)$ by directly maximising

$$\sum_{k=1}^{\hat{K}} \mathcal{S}\left(\hat{s}_k, \hat{e}_k\right), \tag{4.3.1}$$

subject to $\hat{e}_k - \hat{s}_k \geq l$ and $\hat{e}_k \leq \hat{s}_{k+1}$.

Such an approach may not be robust to point outliers, which could either be incorrectly inferred as anomalous segments or cause anomalous segments to be broken up, thus limiting interpretability. The occurrence of both these problems can be prevented by using bounded cost functions (Fearnhead and Rigaill, 2019b). For example, when looking for changes in mean using a square error loss function, we truncate the

loss at a value β' to obtain the cost $\mathcal{C}(x, \mu) = \min(\beta', (x-\mu)^2/\sigma^2)$, which is equivalent to using Tukey's biweight loss.

When only spurious detection of anomalous regions due to point outliers is to be avoided, just the cost used for the typical segments has to be truncated. To do this first we define $\mathcal{S}'(\mathbf{x}_t)$ to be the reduction in cost obtained by truncating the cost of a subset of the components of the observation $\mathbf{x}_t^{(i)}$. So, for example, if our anomalies correspond to changes in mean and we are using the square error loss, or equivalently the Gaussian log-likelihood base cost, then

$$\mathcal{S}'(\mathbf{x}_t) = \sum_{i=1}^{p} \max\left(\left(\frac{\mathbf{x}_t^{(i)} - \boldsymbol{\mu}_0^{(i)}}{\boldsymbol{\sigma}_0^{(i)}}\right)^2 - \beta', 0\right),$$

where as before $\boldsymbol{\mu}_0^{(i)}$ and $\boldsymbol{\sigma}_0^{(i)}$ are the mean and standard deviation of normal data for component i. Joint inference on collective and point anomalies is then performed by maximising the penalised saving

$$\sum_{k=1}^{\hat{K}} \mathcal{S}(\hat{s}_k, \hat{e}_k) + \sum_{t \in O} \mathcal{S}'(\mathbf{x}_t), \tag{4.3.2}$$

with respect to $\hat{K}$, $\left(\hat{s}_1, \hat{e}_1, \hat{\boldsymbol{J}}_1\right)$,..., $\left(\hat{s}_{\hat{K}}, \hat{e}_{\hat{K}}, \hat{\boldsymbol{J}}_{\hat{K}}\right)$, and the set of point anomalies O, subject to $\hat{e}_k - \hat{s}_k \geq l$, $\hat{e}_k < \hat{s}_{k+1}$ $(\cup_i [s_i + 1, e_i]) \cap O = \emptyset$.

Similarly, setting the cost $\mathcal{C}()$ of collective anomalies to the truncated loss prevents collective anomalies from being split up by point anomalies. This has been implemented for anomalies characterised by a change in mean, using Tukey's bi-weight loss, in the `anomaly` package. Using this cost function, however, makes computing the savings $\mathcal{S}()$ more computationally complex: it is $O(nM^2 p \log(p))$ when using a maximum segment length M.

The truncation threshold β' has to be chosen depending on whether it is of interest to detect point anomalies as such. When this is not the case, β' tunes the robustness of the approach – the lower it is, the more robust the approach becomes to outliers, whilst higher values of β' lead to more power. When point anomalies are of interest, β' can be chosen with the aim of controlling false positives under the null hypothesis, that no point anomalies are present. When Tukey's biweight loss is used for example, the following proposition holds for any penalty β':

Proposition 6. *Let $x_1^{(i)}, ..., x_n^{(i)}$ be i.i.d. sub-Gaussian(λ) with known mean μ_i. Let the series be independent for $1 \leq i \leq p$. Let $\hat{O}$ denote the set of point anomalies inferred by MVCAPA using cost $\mathcal{C}(x, \mu) = \min(\beta', (x - \mu)^2)$. Then, there exists a global constant A' such that*

$$\mathbb{P}\left(\hat{O} = \emptyset\right) \geq 1 - A'npe^{-\frac{1}{2\lambda}\beta'}.$$

This suggests setting $\beta' = 2\lambda \log(p) + 2\lambda\psi$, where ψ is as in Section 4.2.2.

4.4 Computation

The standard approach to extend a method for detecting an anomalous window to detecting multiple anomalous windows is through circular binary segmentation (CBS, Olshen et al. (2004)) – which repeatedly applies the method for detecting a single anomalous window or point anomaly. Such an approach is equivalent to using a greedy algorithm to approximately maximise the penalised saving and has computational cost of $O(Mn)$, where M is the maximal length of collective anomalies and n is the number

of observations. Consequently, the runtime of CBS is $O(n^2)$ if no restriction is placed on the length of collective anomalies. We will show in this section that we can directly maximise the penalised saving by using a pruned dynamic programme. This enables us to jointly estimate the anomalous windows, at the same or at a lower computational cost than CBS.

We will focus on the optimisation of the criteria that incorporates point anomalies (4.3.2), though a similar approach applies to optimising (4.3.1). Writing $S(m)$ for the largest penalised saving of all observations up to and including time m, it is straightforward to derive the recursion.

$$S(m) = \max\left(S(m-1) + \mathcal{S}'\left(\mathbf{x}_m\right), \max_{0 \leq t \leq m-l}\left(S(t) + \mathcal{S}\left(t, m\right)\right)\right)$$

with $S(0) = 0$. Calculating $\mathcal{S}\left(t, m\right)$ is, on average, an $O(p\log(p))$ operation, since it requires sorting the savings made from introducing a change in each component. This sorting is not required if the $\boldsymbol{\beta}_i$ are identical, whence the computational cost to $O(p)$. For a maximum segment length M, the computational cost of this dynamic programme approach scales like $O(Mn)$.

If no maximum segment length is specified, it scales quadratically in n. However, the solution space of the dynamic programme can be pruned in a fashion similar to Killick et al. (2012) and Fisch et al. (2018a) to reduced this computational cost. This is discussed in Section B.1.1 of the appendix. As a result of this pruning we found the runtime of MVCAPA to be close to linear in n, when the number of collective anomalies increased linearly with n.

4.5 Accuracy of Detecting and Locating Multiple Collective Anomalies

Whilst we have shown that MVCAPA has good properties when detecting a single anomalous window for the change in mean setting, it is natural to ask whether the extension to detecting multiple anomalous windows will be able to consistently infer the number of anomalous windows and accurately estimate their locations. Specifically, we will be considering the case of joint detection of sparse and dense collective anomalies in mean. Developing such results is notoriously challenging, as can be seen from the fact that previous work on this problem (Jeng et al., 2012) has not provided any such results. Another new feature of the following proof is that the results allow for the number of anomalous segments K to increase, whereas most results in the related changepoint literature (e.g. Fryzlewicz, 2014) assume K to be fixed. Our novel combinatorial arguments can be applied to other settings (e.g. mean and variance) within the penalised cost framework.

Consider a multivariate sequence $\mathbf{x}_1, ..., \mathbf{x}_n \in \mathbb{R}^p$, which is of the form $\mathbf{x}_t = \boldsymbol{\mu}(t) + \boldsymbol{\eta}_t$, where the mean $\boldsymbol{\mu}(t)$ follows a subset multivariate epidemic changepoint model with K epidemic changepoints in mean. For simplicity, we assume that within an anomalous window all affected components experience the same change in mean, and that the noise process is i.i.d. Gaussian, i.e. that for each component i,

$$\boldsymbol{\mu}^{(i)}(t) = \boldsymbol{\mu}_k \quad \text{if} \quad s_k < t \leq e_k \quad \text{and} \quad i \in \mathbf{J}_k \tag{4.5.1}$$

and 0 otherwise.

Consider also the following choice of penalty.

$$\sum_{i=1}^{j} \boldsymbol{\beta}_i = \begin{cases} C\psi + Cj\log(p) & \text{if} \quad j \leq k^*, \\ \\ p + C\psi + C\sqrt{p\psi} & \text{if} \quad j > k^*. \end{cases} \tag{4.5.2}$$

Here, C is a constant, $\psi := \psi(n, p)$ sets the rate of convergence and the threshold

$$k^* = p^{1/2}\frac{\psi}{\log(p)},$$

is defined as the threshold separating sparse changes from dense changes. This penalty regime is identical, up to $O(\epsilon)$, to the point-wise minimum between penalty regimes 1 and 2, when $C = 2 + \epsilon$.

Anomalous regions can be easier or harder to detect depending on the strength of the change in mean characterising them and the number of components ($|\mathbf{J}_k|$ for the kth anomaly) they affect. This intuition can be quantified by

$$\triangle_k^2 = \begin{cases} \dfrac{\boldsymbol{\mu}_k^2}{\log(p) + \psi|\mathbf{J}_k|^{-1}} & \text{if} \quad |\mathbf{J}_k| \leq k^*, \\ \\ \dfrac{\boldsymbol{\mu}_k^2}{\sqrt{p\psi}|\mathbf{J}_k|^{-1} + \psi|\mathbf{J}_k|^{-1}} & \text{if} \quad |\mathbf{J}_k| > k^*, \end{cases}$$

which we define to be the signal strength of the kth anomalous region. The following consistency result then holds

Theorem 3. *Let the typical means be known. There exists a global constants A and C_0 such that for all $C \geq C_0$ the inferred partition $\tau = \{(\hat{s}_1, \hat{e}_1, \hat{\mathbf{J}}_1), ..., (\hat{s}_{\hat{K}}, \hat{e}_{\hat{K}}, \hat{\mathbf{J}}_{\hat{K}})\}$ obtained by applying MVCAPA using the penalty regime specified in (4.5.2) and no minimum segment length, on data $\boldsymbol{x}$ which follows the distribution specified in (4.5.1) satisfies*

$$\mathbb{P}\left(\hat{K} = K, \ \ |\hat{s}_k - s_k| < \frac{10C}{\triangle_k^2}, \ \ |\hat{e}_k - e_k| < \frac{10C}{\triangle_k^2}\right) > 1 - An^3e^{-\psi}, \tag{4.5.3}$$

provided that

$$e_k - s_k \geq \frac{40C}{\triangle_k^2}, \qquad s_{k+1} - e_k \geq \frac{40C}{\triangle_k^2}, \qquad s_k - e_{k-1} \geq \frac{40C}{\triangle_k^2}$$

holds for $k = 1, ..., K$.

The result is proved in the appendix using combinatorial arguments. This finite sample result holds for a fixed C, which is independent of n, p, K, and/or $\triangle_k$. When $\psi = \log(p)$, the threshold k^* is identical to that in Jeng et al. (2012). However, if ψ is chosen to increase with $\log(n)$, so will k^*. This formalises the intuition that when $n >> p$, all detectable changes can be detected as sparse changes (see Liu et al., 2019, for similar results).

Theorem 3 can be extended to allow for both a minimum and maximum segment length. The proof of the theorem is based on partitioning all possible segmentations in to one of two classes, corresponding to those which are consistent with the event in the probability statement of (4.5.3) and those that are not. The proof then shows that conditional on a different event, whose probability is greater than $1 - An^3e^{-\psi}$, any segmentation in the latter class will have a lower penalised saving than at least one segmentation in the former class, and thus cannot be optimal under our criteria. This argument still works providing our choice of minimum or maximum segment lengths does not exclude any segmentations from the first class of segmentations, i.e. if

$$l \leq \min_k \left(e_k - s_k - \frac{20C}{\triangle_k^2} \right), \quad m \geq \max_k \left(e_k - s_k + \frac{20C}{\triangle_k^2} \right).$$

Theorem 3 has implications for the special case $p = 1$ of univariate collective mean anomaly detection. In this case, we detect the right number of collective anomalies,

each with a localisation error of $\frac{10C\psi}{\mu_k^2}$ with probability $1 - An^3e^{-\psi}$. Asymptotic consistency can therefore be obtained by setting $\psi = 4\log(n)$, which recovers standard $O(\log(n))$-localisation errors. As Theorem 3, the result is fully independent of the number of anomalies K and therefore improves on existing results (Fisch et al., 2018a) for the uni-variate setting.

4.6 Incorporating Lags

4.6.1 Extending the Test Statistic

So far, we have assumed that all anomalous windows are perfectly aligned. In some applications, such as the vibrations recorded by seismographs at different locations, certain components will start exhibiting atypical behaviour later and/or return to the typical behaviour earlier. The model in (4.2.1) can be extended to allow for lags in the start or end of each anomalous window. The parameter $\boldsymbol{\theta}^{(i)}(t)$ is then assumed to be

$$\boldsymbol{\theta}^{(i)}(t) = \boldsymbol{\theta}_k^{(i)} \quad \text{if} \quad s_k + \mathbf{d}_k^{(i)} < t \leq e_k - \mathbf{f}_k^{(i)} \quad \text{and} \quad i \in \mathbf{J}_k, \tag{4.6.1}$$

and $\boldsymbol{\theta}_0^{(i)}$ otherwise. Here the start and end lag of the ith component during the kth anomalous window are denoted, respectively, by $0 \leq \mathbf{d}_k^{(i)} \leq w$ and $0 \leq \mathbf{f}_k^{(i)} \leq w$, for some maximum lag-size, w, and satisfy $s_k + \mathbf{d}_k^{(i)} < e_k - \mathbf{f}_k^{(i)}$. The remaining notation is as before.

The statistic introduced in Section 4.2 can easily be extended to incorporate lags.

The only modification this requires is to re-define the saving $\mathcal{S}_i(s,e)$ to be

$$\max_{\substack{0\leq \mathbf{d}^{(i)},\mathbf{f}^{(i)}\leq w \\ e-s-\mathbf{d}^{(i)}-\mathbf{f}^{(i)}\geq l}} \left[\mathcal{C}_i\left(\mathbf{x}^{(i)}_{(s+1+\mathbf{d}^{(i)}):(e-\mathbf{f}^{(i)})}, \boldsymbol{\theta}_0^{(i)}\right) - \min_{\boldsymbol{\theta}}\left(\mathcal{C}_i\left(\mathbf{x}^{(i)}_{(s+1+\mathbf{d}^{(i)}):(e-\mathbf{f}^{(i)})}, \boldsymbol{\theta}\right)\right)\right], \quad (4.6.2)$$

where w is the maximal allowed lag. We then infer O, $\hat{K}$, $\left(\hat{s}_1, \hat{e}_1, \hat{\mathbf{d}}_1, \hat{\mathbf{f}}_1, \hat{\boldsymbol{J}}_1\right)$,..., $\left(\hat{s}_{\hat{K}}, \hat{e}_{\hat{K}}, \hat{\mathbf{d}}_{\hat{K}}, \hat{\mathbf{f}}_{\hat{K}}, \hat{\boldsymbol{J}}_{\hat{K}}\right)$ by directly maximising the penalised saving

$$\sum_{k=1}^{\hat{K}} \mathcal{S}\left(\hat{s}_k, \hat{e}_k\right) + \sum_{t\in O} \mathcal{S}'\left(\mathbf{x}_t\right), \quad (4.6.3)$$

with respect to $\hat{K}$, $\left(\hat{s}_1, \hat{e}_1, \hat{\mathbf{d}}_1, \hat{\mathbf{f}}_1, \hat{\boldsymbol{J}}_1\right)$,..., $\left(\hat{s}_{\hat{K}}, \hat{e}_{\hat{K}}, \hat{\mathbf{d}}_K, \hat{\mathbf{f}}_K, \hat{\boldsymbol{J}}_{\hat{K}}\right)$, and the set of point anomalies O, subject to $0 \leq \hat{\mathbf{d}}_k, \hat{\mathbf{f}}_k \leq w$, $(\hat{e}_k - \hat{\mathbf{f}}_k) - (\hat{s}_k + \hat{\mathbf{d}}_k) \geq l$ and $\hat{e}_k < \hat{s}_{k+1}$.

Introducing lags means searching over more possible start and end points for the anomalous segments in each series. Consequently, increased penalties are required to control the false error rate. A simple general way of doing is based on a Bonferonni correction to allow for the different start and end-points of anomalies in different series. It is shown in Section B.1.2 of the appendix that if we use the penalty regimes from Section 4.2.2 but inflate ψ by adding $4\log(w+1)$ we obtain the same bound on false positives.

When anomalies correspond to change in mean in Gaussian data, we obtain the following, stronger result.

Proposition 7. *Let $\boldsymbol{x}_1^{(i)}, ..., \boldsymbol{x}_n^{(i)}$ be i.i.d. $N(0,1)$ and independent for $1 \leq i \leq p$. Then, for all $\epsilon > 0$, $\mathcal{S}_i(s,e)$, as defined in (4.6.2) is stochastically bounded by*

$$(1+\epsilon)\chi_2^2 + 2(1+\epsilon)\left(\log(w+1) + \log(6) - \log(\log(1+\epsilon)) + \log(1+\log(w+1))\right), \quad (4.6.4)$$

when the cost function is $\mathcal{C}(x,\mu) = (x-\mu)^2$ and the typical mean known.

It should be noted that it is not possible to improve on the above result as the search space can contain $w+1$ independent savings when $e-s=w$. As a corollary of Proposition 7, the following modified version of penalty regime 2 controls the false positive probability:

Penalty Regime 2': $P_2'(j) = 2(1+\epsilon)\psi + 2(1+\epsilon)j\log(p)$ corresponding to $\alpha = 2(1+\epsilon)\psi$ and $\beta_j = 2(1+\epsilon)\log(p) + 2(1+\epsilon)\log(w+1)$, for $1 \leq j \leq p$.

4.6.2 Result on Power

Incorporating lags can improve power, especially when considering sparse collective anomalies. This becomes apparent when considering the following modification of the setting considered in Section 4.2.3. Let

$$\mathbf{x}_t^{(i)} = v^{(i)}\mathbb{I}\left(s+d_i < t \leq e-f_i\right)\mu + \boldsymbol{\eta}_t^{(i)}, \quad v^{(i)} \sim \begin{cases} 0 & \text{with prob. } 1-p^{-\xi}, \\ 1 & \text{with prob. } p^{-\xi}, \end{cases} \tag{4.6.5}$$

for $s < t \leq e$, where the noise $\boldsymbol{\eta}_t^{(1)}, ..., \boldsymbol{\eta}_t^{(p)}$ of the different series is independent and satisfies $\boldsymbol{\eta}_t^{(i)} \overset{i.i.d}{\sim} N(0,1)$ for $s < t \leq e$. Assume also that the start and end lags add up to w, i.e. that $d_i + f_i = w$ for $1 \leq i \leq p$. The following result holds:

Proposition 8. *Let* $\frac{3}{4} < \xi < 1$ *and* $(e-s-w)\mu^2 = 2r_p\log(p(w+1))$. *MVCAPA with a maximum lag of* w *using penalty regime 2' is able to detect the segment* $\boldsymbol{x}_{(s+1):e}$ *defined in* (4.6.5) *as being anomalous with probability going to 1, whilst controlling false positives as* $p \to \infty$ *if* $r_p > \left(1-\sqrt{1-\xi}\right)^2$ *and* $\log(n) = o(\log(p))$.

Conversely, it is possible to bound the power of any approach not considering lags using the following corollary of Theorem 1 in Cai et al. (2011b).

Proposition 9. *Let* $\frac{3}{4} < \xi < 1$ *and* $\frac{e-s-w}{e-s}(e-s-w)\mu^2 = 2r_p \log(p)$. *The sum of type I and type II error of any test of the alternative hypothesis*

$$H_1 : \sqrt{e-s}\bar{x}^{(i)}_{(s+1):e} \overset{i.i.d}{\sim} \begin{cases} N(0,1) & \text{with prob. } 1-p^{-\xi}, \\ N(\frac{e-s-w}{e-s}\mu, 1) & \text{with prob. } p^{-\xi}, \end{cases}$$

against the null hypothesis

$$H_0 : \sqrt{e-s}\bar{x}^{(i)}_{(s+1):e} \overset{i.i.d}{\sim} N(0,1)$$

converges to 1 as $p \to \infty$ *if* $r_p < \left(1-\sqrt{1-\xi}\right)^2$.

Thus for this setting, including lags, modifies the detectability boundary for μ^2 by a factor of

$$\frac{e-s-w}{e-s}\frac{\log((w+1)p)}{\log(p)}.$$

This shows that the gain from including lags is especially significant when the lags and segment lengths are on a similar scale. Furthermore, at constant lag and anomaly length, the improvement becomes more significant with increasing dimension p. Another corollary of this result is that specifying a lag w' which is too large (i.e. greater than w) is advantageous provided that

$$\frac{e-s-w}{e-s}\frac{\log((w'+1)p)}{\log(p)} < 1,$$

which, *ceteris paribus*, is bound to hold as $p \to \infty$.

4.6.3 Computational Considerations

The dynamic programming approach described in Section 4.4 can also be used to minimise the penalised negative saving in Equation (4.6.3). Solving the dynamic

programme requires the computation of $\mathcal{S}_i(t, m)$ for $1 \leq i \leq p$ for all permissible t at each step of the dynamic programme. Computing these savings *ex nihilo* every time leads to the computational cost of the dynamic programme to scale quadratically in $(w + 1)$.

However, it is possible to reduce the computational cost of including lags by storing the savings

$$\mathcal{C}_i\left(\mathbf{x}^{(i)}_{(a+1):b}, \boldsymbol{\theta}^{(i)}_0\right) - \min_{\theta}\left[\mathcal{C}_i\left(\mathbf{x}^{(i)}_{(a+1):b}, \boldsymbol{\theta}\right)\right]$$

for $t - w \leq b \leq t$ and $0 \leq a \leq b - l$. These can then be updated in each step of the dynamic programme at a cost of at most $O(np)$. From these, it is possible to calculate all $\mathcal{S}_i(t, m)$ required for a step of the dynamic programme in just $O(np(w + 1))$ comparisons. This reduces the computational cost of each step of the dynamic programme to $O(pn(w + 1) + pn\log(p))$. Crucially, only the comparatively cheap operations of allocating memory and finding the maximum of two numbers increase with $w + 1$. Furthermore, it is possible to adapt the pruning rule for the dynamic programme to incorporate lags. The details for this and full pseudocode can be found in the appendix.

4.7 Simulation Study

We now compare the performance of MVCAPA to that of other popular methods. In particular, we compare ROC curves, precision, as well as the runtime with PASS (Jeng et al., 2012) and Inspect (Wang and Samworth, 2018, 2016). PASS (Jeng et al., 2012) uses higher criticism in conjunction with circular binary segmentation (Olshen

et al., 2004) to detect subset multivariate epidemic changepoints. Inspect (Wang and Samworth, 2018) uses projections to find sparse classical changepoints and therefore provides a benchmark for the detection approach consisting of modelling epidemic changes as two classical changepoints. For the purpose of this simulation study, we used the implementation of PASS available on the author's website and the Inspect implementation from the R package `InspectChangepoint`.

The comparison was carried out on simulated multivariate time series with $n = 5000$ observations for p components with i.i.d. $N(0, 1)$ noise, $AR(1)$ noise ($\rho = 0.3$), or t_{10}-distributed noise for a range of values of p. To these, collective anomalies affecting k components occurring at a geometric rate of 0.001 (leading to an average of about 5 collective anomalies per series) were added. The lengths of these collective anomalies are i.i.d. Poisson-distributed with mean 20. Within a collective anomaly, the start and end lags of each component are drawn uniformly from the set $\{0, ..., w\}$, subject to their sum being less than the length of the collective anomaly. Note that $w = 0$ implies the absence of lags. The means of the components during the collective anomaly are drawn from an $N(0, \sigma^2)$-distribution. In particular, we considered the following cases, emulating different detectable regimes introduced in Section 4.2.3.

1. The most sparse regime possible: a single component affected by a strong anomaly without lags, i.e. $\sigma = 2\log(p)$, $w = 0$, and $k = 1$.

2. The most dense regime possible: all components affected by weak anomalies without lags, i.e. $\sigma = p^{-1/4}$, $w = 0$, and $k = p$.

3. A regime close to the boundary between sparse and dense changes, i.e. $k = 2$

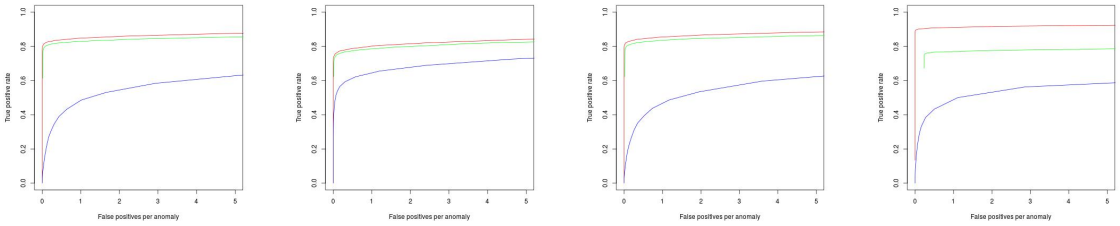
when $p = 10$ and $k = 6$ when $p = 100$ with $\sigma = \log(p)$ and $w = 0$.

4. A regime close to the boundary between sparse and dense changes, but with lagged collective anomalies, i.e. the same as 3 but with $w = 10$.

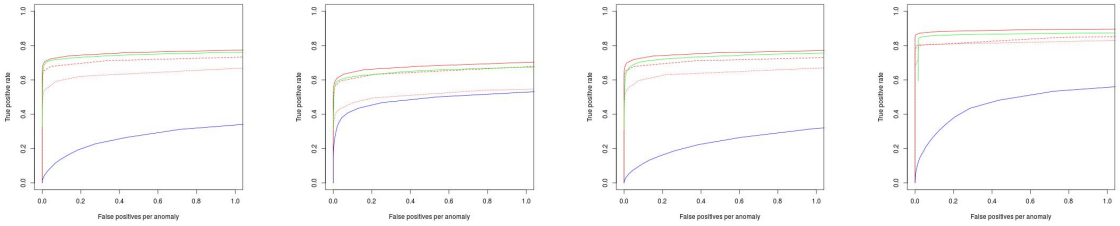
This analysis was repeated with 5 point anomalies distributed $N(0, 8\log(p))$. The $\log(p)$-scaling of the variance ensures that the point anomalies are anomalous even after correcting for multiple testing over the p different components.

4.7.1 ROC Curves

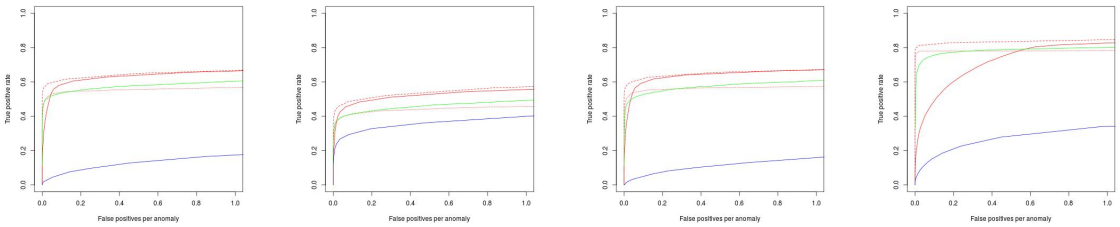
We use ROC curves to compare the methods. For our setting it is not clear how to define the number of true negatives, and thus we plot the true positive rate against the ratio of false positives to true positives. We obtained the curves by varying the threshold parameters of Inspect and PASS and by rescaling $\alpha, \beta', \beta_1, ..., \beta_p$ for MVCAPA. The curves were obtained over 1000 simulated datasets. For MVCAPA, we typically set $w = 0$, but also tried $w = 10$ and $w = 20$ for the third and fourth setting. The median and median absolute deviation were used to robustly estimate the mean and variance. Throughout the experiments, we used and rescaled the composite penalty regime (Section 4.2.2) for $w = 0$ and penalty regime 2' for $w > 0$. We also set the maximum segment lengths for both MVCAPA and PASS to 100 and the minimum segment length of MVCAPA to 2. The α_0 parameter of PASS, which excludes the $\alpha_0 - 1$ lowest p-values from the higher criticism statistic to obtain a better finite sample performance (see Jeng et al. (2012)) was set to k or 5, whichever was the smallest. For MVCAPA and PASS, we considered a detected segment to be a true



(a) p=10 (b) p=10, AR (c) p=10, PAs (d) p=100, PAs



(e) p=10 (f) p=10, AR (g) p=10, PAs (h) p=100, PAs



(i) p=10 (j) p=10, AR (k) p=10, PAs (l) p=100, PAs

Figure 4.7.1: Example series and ROC curves for setting 1, 3, and 4 (top row to bottom row). MVCAPA is in red, PASS in green, and Inspect in blue. The solid red line corresponds to $w = 0$, the dashed one to $w = 10$ and the dotted one to $w = 20$. The x-axis denotes the number of false discoveries normalised by the total number of real anomalies present in the data.

positive if its start and end point both lie within 20 observations of that of a true collective anomalies' start and end point respectively. For Inspect, we considered a detected change to be a true positive if it was within 20 observation of a true start or end point. When point anomalies were added to the data, we considered segments of length one returned by PASS to be point anomalies to make the comparison with MVCAPA fairer.

A subset of the results for three of the settings considered can be found in Figure 4.7.1. The full results for the four settings can be found in Figures B.4.1 to B.4.4 of the appendix. We can see that Inspect usually does worst. This is especially true when changes become dense, which is no surprise given the method was introduced to detect sparse changes. However it is also the case for very sparse changes – the setting for which Inspect has been designed, highlighting the advantage of treating epidemic changes as such. We additionally see that MVCAPA generally outperforms PASS. This advantage is particularly pronounced in the case in which exactly one component changes. This is a setting which PASS has difficulties dealing with due to the convergence properties of the higher criticism statistic at the lower tail (Jeng et al., 2012). PASS outperformed MVCAPA in the second setting for $p = 10$, when it was assisted by a large value of α_0, which considerably reduced the number of candidate collective anomalies it had to consider.

Figures 4.7.1e and 4.7.1i, show that MVCAPA performs best when the correct maximal lag is specified. They also demonstrate that specifying a lag and therefore overestimating the lag when no lag is present adversely affects performance of MVCAPA. However, when lags are present, over-estimating the maximal lag appears

preferable to underestimating it. Finally, the comparison between Figures 4.7.1i and 4.7.1k shows that the performance of MVCAPA is hardly affected by the presence of point anomalies.

We have also considered the case in which collective anomalies are characterised by joint changes in mean and variance. ROC curves for that setting can be found in the appendix.

4.7.2 Precision

We compared the precision of the three methods by measuring the accuracy (in mean absolute distance) of true positives. Only true positives detected by all methods were taken into account to avoid selection bias. We used the default parameters for MVCAPA and PASS, whilst we set the threshold for Inspect to a value leading to comparable number of true and false positives. To ensure a suitable number of true positives for Inspect we doubled σ in the second scenario. The results of this analysis can be found in Table 4.7.1 and show that MVCAPA is usually the most precise approach, exhibiting a significant gain in accuracy against PASS. Whilst we noted that erring on specifying too large a maximal lag was better in terms of power of the MVCAPA to detect collective anomalies, we see that it does have an adverse impact on the accuracy of their estimated locations.

Setting	p	Lag	PAs.	MVCAPA	MVCAPA, w=10	MVCAPA, w=20	Inspect	PASS
1	10	0	-	**0.09**	-	-	0.64	0.31
1	100	0	-	**0.02**	-	-	0.40	0.62
1	10	0	✓	**0.09**	-	-	0.62	0.38
1	100	0	✓	**0.03**	-	-	0.40	0.67
2	10	0	-	**0.09**	-	-	0.74	0.52
2	100	0	-	**0.01**	-	-	0.71	0.54
2	10	0	✓	**0.05**	-	-	0.69	0.46
2	100	0	✓	**0.01**	-	-	0.67	0.51
3	10	0	-	**0.11**	2.31	3.30	0.72	0.27
3	100	0	-	**0.01**	3.43	3.83	0.53	0.29
3	10	0	✓	**0.09**	2.23	3.26	0.69	0.22
3	100	0	✓	**0.01**	3.35	3.82	0.53	0.23
4	10	10	-	0.63	**0.46**	1.09	0.80	2.53
4	100	10	-	1.27	**0.18**	1.57	0.61	3.64
4	10	10	✓	0.72	**0.51**	1.22	0.83	2.60
4	100	10	✓	1.23	**0.21**	1.58	0.59	3.77

Table 4.7.1: Precision of true positives detected by all methods measured in mean absolute distance for MVCAPA, PASS, and Inspect.

4.8 Detecting Copy Number Variation

We now apply MVCAPA to extract copy number variations (CNVs) from genetics data. The data consists of a log-R ratio between observed and expected intensity (formally defined in Lin et al. (2013)) evaluated along the genome. The typical mean of this statistics is therefore equal to 0, whilst deviations from 0 correspond to CNVs. A multivariate approach to detecting CNVs is attractive because they are often shared across individuals. By borrowing signal across individuals we should gain power for detecting CNVs which have a weak signal. However, as we will become apparent from our results, shared variations do not always align perfectly across individuals.

In this section we re-use the design of Bardwell and Fearnhead (2017) to compare MVCAPA with PASS. We will therefore investigate the performance of both methods on two chromosomes (Chromosome 16 with $n = 59,590$ measurements and Chromosome 6 with $n = 126,695$ measurements) over 18 individuals, which we split into 3 folds of $p = 6$ individuals. We set the maximum segment length for MVCAPA and PASS to 100. To investigate the potential benefit of allowing for lags, we repeated the experiment for MVCAPA both with $w = 0$ (i.e. not allowing for lags) and $w = 40$. Since $n >> p$ in this application, we used the sparse penalty setting (Regime 2) for MVCAPA.

Whilst the exact ground truth is unknown, we can compare different methods by how accurately they detect known CNVs for a given test size. We used known CNVs from the HapMap project (Consortium, 2003) as true positives and tuned the penalties and thresholds in such a way that 4% of the genome was flagged up as anomalous

for all methods. For MVCAPA this involved scaling the penalties $\alpha, \beta_1, ..., \beta_p$ by a constant, as discussed in Section 4.2.2.

The results of this analysis on Chromosome 16 can be found in Table 4.8.1 while the reults for Chromosome 6 can be found in Table B.4.7 in the appendix. These tables show that MVCAPA shows much more consistency across folds than PASS. We can also see that allowing for lags generally led to a better performance of MVCAPA, thus suggesting non-perfect alignment of CNVs across individuals. Moreover, MVCAPA was very fast taking 5 seconds to analyse the longer genome on a standard laptop when we did not allow lags, and 10 seconds when we allowed for lags. The R implementation of PASS, on the other hand, took 17 minutes. No point anomalies were identified by MVCAPA.

Truth	PASS			MVCAPA ($w = 40$)			MVCAPA ($w = 0$)		
Start	Rep 1	Rep 2	Rep 3	Rep 1	Rep 2	Rep 3	Rep 1	Rep 2	Rep 3
2619669		✓							
2638575		✓							
21422575	✓	✓	✓	✓	✓	✓	✓	✓	✓
32165010	✓	✓	✓	✓	✓	✓	✓	✓	✓
34328205	✓	✓		✓	✓	✓	✓	✓	✓
54351338	✓	✓	✓	✓	✓	✓			
70644511				✓	✓	✓	✓	✓	✓

Table 4.8.1: A comparison between PASS, MVCAPA without lags, and MVCAPA with a lag of up to 40 for chromosome 16. Each row represents a known copy number variation, their starting point (as defined by the HapMap) being indicated in the first column. Successful detections are indicated by ticks.

Chapter 5

Innovative And Additive Outlier Robust Kalman Filtering

5.1 Introduction And Literature Review

Anomaly detection is an area of considerable importance and has been subject to increasing attention in recent years. Comprehensive reviews of the area can be found in Chandola et al. (2009); Pimentel et al. (2014). The field's growing importance arises from the increasing range of applications to which anomaly detection lends itself: from fraud prevention (Chandola et al., 2009; Pimentel et al., 2014), to fault detection (Chandola et al., 2009; Pimentel et al., 2014), and even the detection of exoplanets (Fisch et al., 2018b). More recently, the emergence of internet of things and the ubiquity of sensors has led to emergence of the online detection of anomalies as an important statistical challenge.

Kalman filters Kalman (1960) provide a convenient framework to detect anomalies

within a streaming data context. In particular, they can be updated in a fully online fashion at a fixed computational cost. At each time point, Kalman filters also provide an estimate both for the expectation and variance of the next observation. These can be used to determine whether that observation is anomalous or not. However, the major drawback of Kalman filters is their lack of robustness to outliers: once the filter has encountered an outlier, it will often produce inaccurate predictions for many future time points.

The anomaly detection literature distinguishes between two types of outliers. The first are additive outliers, sometimes referred to as observational outliers (Gandhi and Mili, 2009), which affect the observational noise only. The other type of outliers are the innovative, or process (Huang et al., 2017), outliers. These affect the updates of the hidden states. In practice, both have a similar effect on the next observation, but quite different effects on subsequent observations. Moreover, some innovative outliers cannot be detected immediately as their influence on the observations is only noticeable after, or over, a period of time.

A range of robust Kalman filters has been proposed to date. Many side-step the problem of distinguishing between the two outlier types. By far the largest class of filters aims to be robust against heavy tailed additive outliers. Examples of such filters include Ting et al. (2007); Agamennoni et al. (2011), which assume t-distributed additive noise and perform inference using variational Bayes, Ruckdeschel et al. (2014), who use Huberised residuals, and Chang (2014) inflate the noise covariance matrix whenever an outlier is encountered. A few filters have also been developed with the aim of achieving robustness against innovative outliers (Ruckdeschel et al., 2014).

The problem with such filters is that they exacerbate the shortcomings of the Kalman filter when they encounter the other type of anomaly: additive outlier robust Kalman filters, for example, update their hidden states even less than the classical Kalman filter when encountering innovative outliers.

In principle, it seems straightforward to combine the ideas of these two types of robust Kalman filter. One body of literature proposes to use Huberisation of both innovative and additive residuals (Gandhi and Mili, 2009; Chang, 2014). Others (Huang et al., 2017, 2019) have modelled both additive and innovative outliers using t-distributions, by imposing Wishart priors on the precision matrix of both the innovations and additions and maintaining the posterior by using variational Bayes approaches. The issue with these filters comes from how they approximate the filtering distribution of the state. Both return uni-modal posteriors after encountering an anomaly. This is a shortcoming given that the posterior after an anomaly is likely to be multi-modal: if the outlying observation was caused by an additive anomaly, the state will be close to the prior, whereas if it was caused by an innovative anomaly, the state would be far from it.

The ideal approach to constructing a robust filter would be to model the possibility of outliers in both the observation and system noise, and then use a filter algorithm that attempts to calculate, or approximate, the true filtering distribution for the model. An early attempt to do this was the spline based approach Kitagawa (1987), but the computational complexity increases very quickly with the number of dimensions and such a filter becomes impracticable when the state dimension is greater than 3. As a result we consider using particle filters (Gordon et al., 1993;

Fearnhead and Künsch, 2018). These are able to produce Monte Carlo approximations to the filtering distribution for an appropriate model that allows for outliers, and, in principle, can work even if the filtering distribution is multi-modal. However the Monte Carlo error of standard implementations of the particle can be prohibitively large (Chang, 2014).

In this chapter, we develop an efficient particle filter by using a combination of Rao-Blackwellisation and well-designed proposal distributions. The idea of Rao-Blackwellisation is to integrate out part of the state so that the particle filter approximates the filtering distribution of a lower-dimensional projection of the state. In our application this projection is whether each component of the additive and innovative noise is an outlier, and if it is how much the variance of the noise has been inflated. Conditional on this information, the state space model becomes linear-Gaussian and we can implement a Kalman Filter to calculate exactly the conditional filtering distribution, while being able to fully capture multi modal posteriors. This idea is similar to that which underpins the Mixture Kalman Filter (Chen and Liu, 2000).

Whilst Rao-Blackwellisation improves the Monte Carlo accuracy of the filter, such a filter can still have the shortcomings noted by Chang (2014) and perform poorly without good proposal distributions for the information we condition on. One of the main contributions of this work is a proposal distribution that accurately approximates the conditional distribution of the variance inflation for each component of the noise, and hence approximates the optimal proposal distribution (Pitt and Shephard, 1999). As a result of this proposal, we find that accurate results can be obtained even with only a few particles.

Another important challenge addressed by this chapter is that certain innovative outliers can not immediately be detected. An innovative outlier in a latent trend component for instance can cause a trend changes which may only become apparent – i.e. produce a visible outlier in the observations – many observations after the innovative outlier in the trend occurred. It is nevertheless important to capture such outliers as they can affect a potentially unlimited number of observations to come. The proposed particle filter includes the possibility to back-sample the variance inflation particles in light of more recent observations, which enables it to capture these important anomalies.

The remainder of this chapter is organised as follows: We discuss our robust noise model, consisting of a mixture distribution of Gaussian noise, representing typical behaviour, and heavy tailed noise, representing atypical behaviour, for both the additive (observational) and innovative (system) noise process in Section 5.2. The model is shown to be very similar to that considered by Huang et al. (2019). We then introduce the proposal distribution for the scale of the noise in Section 5.3, before extending it to anomalies which are not immediately identifiable in Section 5.4. The proposed filter is compared to others in Section 5.5 and applied to router data and a benchmark machine temperature data-set in Section 5.6. The proposed methodology, which we call Computationally Efficient Bayesian Anomaly detection by Sequential Sampling (CE-BASS) has been implemented in the the `R` package `RobKF` available from `https://github.com/Fisch-Alex/Robkf`. Derivations of theoretical results and complete pseudocode are available in the appendix.

5.2 Model And Examples

Throughout this chapter, we will consider inference about a latent state, $\mathbf{X}_t$, through partial observations, $\mathbf{Y}_t$, modelled as

$$\begin{aligned} \mathbf{Y}_t &= \mathbf{C}\mathbf{X}_t + \mathbf{V}_t^{\frac{1}{2}}\boldsymbol{\Sigma}_A^{\frac{1}{2}}\boldsymbol{\epsilon}_t, \\ \mathbf{X}_t &= \mathbf{A}\mathbf{X}_{t-1} + \mathbf{W}_t^{\frac{1}{2}}\boldsymbol{\Sigma}_I^{\frac{1}{2}}\boldsymbol{\nu}_t. \end{aligned} \tag{5.2.1}$$

Here the additive noise, $\boldsymbol{\epsilon}_t \in \mathbb{R}^p$, and the innovations $\boldsymbol{\nu}_t \in \mathbb{R}^q$ are both i.i.d. standard multivariate Gaussian. The diagonal matrices $\boldsymbol{\Sigma}_A$ and $\boldsymbol{\Sigma}_I$ denote the covariance of the additive and innovation noise respectively. The diagonal matrices $\mathbf{V}_t$ and $\mathbf{W}_t$ are used to capture additive and innovative outliers respectively, with large diagonal entries of $\mathbf{V}_t$ corresponding to additive outliers and large diagonal entries of $\mathbf{W}_t$ corresponding to innovative outliers. The classical Kalman model is recovered by setting $\mathbf{W}_t = \mathbf{I}$ and $\mathbf{V}_t = \mathbf{I}$ for all times t.

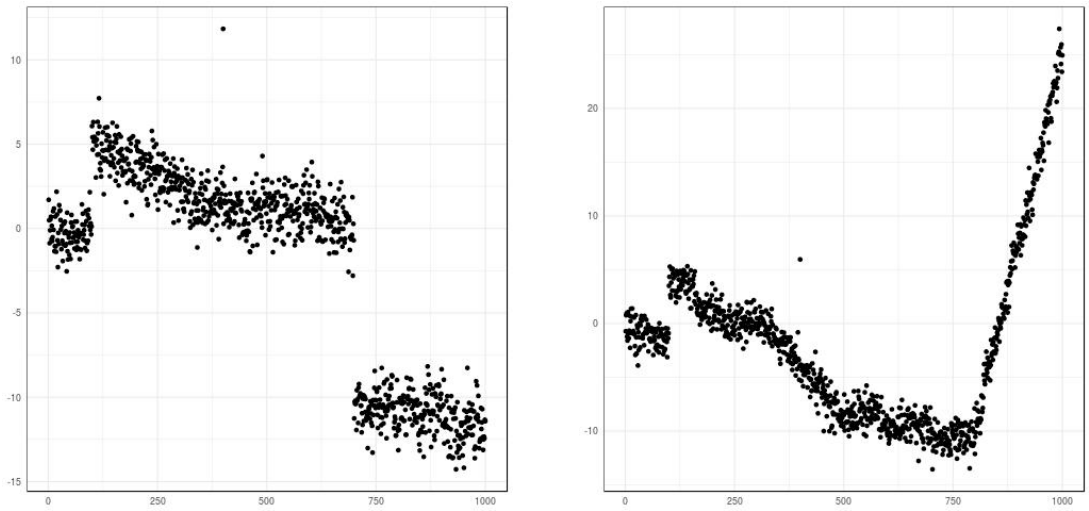
The model in Equation (5.2.1) can be used to model a range of time series behaviours. We will use the following two examples throughout the chapter:

Example 1: The random walk model with both changepoints and outliers, similar to the problem considered by Fearnhead and Rigaill (2019b). It can be formulated as

$$Y_t = X_t + V_t^{\frac{1}{2}}\sigma_A\epsilon_t, \qquad X_t = X_{t-1} + W_t^{\frac{1}{2}}\sigma_I\nu_t. \tag{5.2.2}$$

Here atypically large values of V_t correspond to outliers, whilst atypically large values of W_t correspond to changes. A realisation of this model can be found in Figure 5.2.1a.

Example 2: A time series with changes in trend, level shifts, as well as outliers, similar to the model considered by Maeng and Fryzlewicz (2019). It can be formulated



(a) Random walk

(b) Random walk with trends

Figure 5.2.1: Two examples of time series which are realisations of outlier infested Kalman models. (a) was simulated using the setup defined in Equation (5.2.2), with $\sigma_A = 1$, $\sigma_I = 0.1$, and outliers defined by $W_{100} = 3600$, $V_{400} = 100$, and $W_{700} = 10000$. Conversely (b) second example was simulated using the model defined in Equation (5.2.3) using $\sigma_A = 1$, $\sigma_I^{(1)} = 0.1$, $\sigma_I^{(2)} = 0.01$ and outliers defined by $W_{100}^{(1)} = 3600$, $V_{400} = 100$, and $W_{700}^{(2)} = 40000$.

as

$$\begin{aligned} Y_t &= X_t^{(1)} + V_t^{\frac{1}{2}} \sigma_A \epsilon_t \\ X_t^{(1)} &= X_{t-1}^{(1)} + X_{t-1}^{(2)} + \left(W_t^{(1)}\right)^{\frac{1}{2}} \sigma_I^{(1)} \nu_t^{(1)}, \\ X_t^{(2)} &= X_{t-1}^{(2)} + \left(W_t^{(2)}\right)^{\frac{1}{2}} \sigma_I^{(2)} \nu_t^{(2)}, \end{aligned} \tag{5.2.3}$$

with the first component of the hidden state denoting the current position and the second indicating the trend. Here, outliers are modelled by large values of V_t whilst level shift and changes in trend are modelled by atypically large values of $W_t^{(1)}$ and $W_t^{(2)}$ respectively. A realisation of this model can be found in Figure 5.2.1b.

A key feature of this second model is that an outlier in the trend component, $X_t^{(2)}$, may only become detectable many observations after the outlier – this challenging issue mentioned in the introduction is addressed via the methods in Section 5.4. A wide rage of other commonly used time series features, such as auto-correlation, moving averages, etc. can be incorporated in the model.

To infer the locations of anomalies we use the model

$$\mathbf{V}_t^{(i,i)} = 1 + \lambda_t^{(i)} \frac{1}{\tilde{\mathbf{V}}_t^{(i,i)}} \qquad \mathbf{W}_t^{(j,j)} = 1 + \gamma_t^{(j)} \frac{1}{\tilde{\mathbf{W}}_t^{(j,j)}} \tag{5.2.4}$$

for $1 \leq i \leq p$ and $1 \leq j \leq q$. The random variables $\lambda_t^{(i)} \sim Ber(r_i)$ and $\gamma_t^{(j)} \sim Ber(s_j)$ are indicators that determine whether an anomaly is present or not for $1 \leq i \leq p$ and $1 \leq j \leq q$ respectively. For additional interpretability, we impose that at most one anomaly is present at any given time t, and define r_i and s_j to be the probabilities that $\lambda_t^{(i)} = 1$ and $\gamma_t^{(j)} = 1$ respectively. The inverse scale, or precision, of an anomaly (if present) is given by the random variables $\tilde{\mathbf{V}}_t^{(i,i)} \sim \tilde{\sigma}_i \Gamma(a_i, a_i)$ and $\tilde{\mathbf{W}}_t^{(j,j)} \sim \hat{\sigma}_j \Gamma(b_j, b_j)$ for $1 \leq i \leq p$ and $1 \leq j \leq q$ respectively.

The proposed model bears similarities to the model used by Huang et al. (2019). Both use a mixture of Gaussian and heavy tailed noise. The main difference is that the anomalous behaviour is characterised by noise which is the sum of a Gaussian and a t-distribution in our model as opposed to just a t-distribution in the model used by Huang et al. (2019). This ensures that anomalies coincide with strictly greater noise and makes the result more interpretable. In practice, however, the noise distribution considered in this chapter and in Huang et al. (2019) are likely to be of very similar shape.

5.3 Particle Filter

We now turn to filtering the model defined by Equations (5.2.1) and (5.2.4). The main feature we exploit is the fact that if we knew the value of $(\mathbf{V}_t, \mathbf{W}_t)$ at all times t, we could just run the classical Kalman filter over the data. Consequently, our approach will consist of sampling particles for $(\mathbf{V}_t, \mathbf{W}_t)$, conditional on which the classical Kalman update equations for the hidden state $\mathbf{x}_t$ can be used. This approach, very similar to the mixture Kalman filter (Chen and Liu, 2000; Fearnhead and Clifford, 2003) is summarised by the pseudocode in Algorithm 1.

For each time, t, the code loops over the existing particles, $(\mathbf{V}_t, \mathbf{W}_t)$, and simulates M' descendants for each of them in step 4. They are stored in a set of candidate particles. If we have N particles at time t, keeping all candidates would produce NM' particles at time $t+1$. To avoid growing the number of particles exponentially with t, Step 7 resamples the candidates to keep just N particles. The filtering distribution

for each of these particles is then calculated using the Kalman Filter updates in step 10.

Algorithm 1 Basic Particle Filter (No Back-sampling)

Input: An initial state estimate $(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$

A number of descendants, $M' = M(p+q)+1$

A number of particles to be maintained, N.

A stream of observations $\mathbf{Y}_1, \mathbf{Y}_2, ...$

Initialise: Set $Particles(0) = \{(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)\}$

1: **for** $t \in \mathbb{N}^+$ **do**

2: $Candidates \leftarrow \{\}$

3: **for** $(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \in Particles(t-1)$ **do**

4: $(\mathbf{V}, \mathbf{W}, prob) \leftarrow$ Sample_Particles$(M', \boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{Y}_t, \mathbf{A}, \mathbf{C}, \boldsymbol{\Sigma}_A, \boldsymbol{\Sigma}_I)$

5: $Candidates \leftarrow Candidates \cup \{(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{V}, \mathbf{W}, prob)\}$

6: **end for**

7: $Descendants \leftarrow$ Subsample$(N, Candidates)$

8: $Particles(t) \leftarrow \{\}$

9: **for** $(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{V}, \mathbf{W}, prob) \in Descendants$ **do**

10: $(\boldsymbol{\mu}_{new}, \boldsymbol{\Sigma}_{new}) \leftarrow$ KF_Upd$(\mathbf{Y}_t, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{C}, \mathbf{A}, \mathbf{V}^{1/2}\boldsymbol{\Sigma}_A, \mathbf{W}^{1/2}\boldsymbol{\Sigma}_I)$

11: $Particles(t) \leftarrow Particles(t) \cup \{(\boldsymbol{\mu}_{new}, \boldsymbol{\Sigma}_{new})\}$

12: **end for**

13: **end for**

The main challenge in the above approach consists of selecting a good sampling procedure for the particles. Whilst it may be a natural choice to sample particles $(\mathbf{V}_{t+1}, \mathbf{W}_{t+1})$ from their prior distribution, this is not suitable for the problem considered in this chapter. In particular, this sampling procedure would not be robust to outliers: the stronger an anomaly was, the less likely we would be to sample a particle

with an appropriate value of $(\mathbf{V}_{t+1}, \mathbf{W}_{t+1})$, as discussed by Chang (2014).

Adopting ideas from Pitt and Shephard (1999) and Arulampalam et al. (2002), we overcome the above challenge by sampling particles from an approximation to the conditional distribution of $(\mathbf{V}_{t+1}, \mathbf{W}_{t+1})$ given observation $\mathbf{Y}_{t+1}$. Denote the model's prior distribution for $(\mathbf{V}_{t+1}, \mathbf{W}_{t+1})$ in (5.2.4) by $\pi_0(\cdot)$. The conditional distribution $\pi(\mathbf{W}_{t+1}, \mathbf{V}_{t+1}|\mathbf{Y}_{t+1})$ for the descendants of a particle whose filtering distribution for $\mathbf{x}_t$ is $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is then proportional to

$$\pi_0(\mathbf{W}, \mathbf{V})\mathcal{L}\left(\mathbf{Y}, \mathbf{CA}, \mathbf{CA}\boldsymbol{\Sigma}\mathbf{A}^T\mathbf{C}^T + \boldsymbol{\Sigma}_A\mathbf{V} + \mathbf{C}\boldsymbol{\Sigma}_I\mathbf{W}\mathbf{C}^T\right).$$

Here we have dropped time indices for convenience, and $\mathcal{L}(\mathbf{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the likelihood of an observation $\mathbf{x}$ under a $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$-model. Since at most one component is anomalous, we can re-write this as a sum over which, if any, component is anomalous

$$\mathbb{I}_{\{\mathbf{W}=\mathbf{I},\mathbf{V}=\mathbf{I}\}}\pi(\mathbf{I}, \mathbf{I}|\mathbf{Y}) + \sum_{j=1}^{q}\mathbb{I}_{\left\{\mathbf{W}=\mathbf{I}+\frac{\mathbf{I}^{(j)}}{\tilde{\mathbf{W}}^{(j,j)}},\mathbf{V}=\mathbf{I}\right\}}\hat{\pi}_j\left(\tilde{\mathbf{W}}^{(j,j)}\right) + \sum_{i=1}^{p}\mathbb{I}_{\left\{\mathbf{W}=\mathbf{I},\mathbf{V}=\mathbf{I}+\frac{\mathbf{I}^{(i)}}{\tilde{\mathbf{V}}^{(i,i)}}\right\}}\tilde{\pi}_i\left(\tilde{\mathbf{V}}^{(i,i)}\right).$$

Here, we use the shorthand

$$\tilde{\pi}_i\left(\tilde{\mathbf{V}}^{(i,i)}\right) = \pi\left(\mathbf{I}, \mathbf{I} + \frac{\mathbf{I}^{(i)}}{\tilde{\mathbf{V}}^{(i,i)}}|\mathbf{Y}\right)$$

and

$$\hat{\pi}_j\left(\tilde{\mathbf{W}}^{(j,j)}\right) = \pi\left(\mathbf{I} + \frac{\mathbf{I}^{(j)}}{\tilde{\mathbf{W}}^{(j,j)}}, \mathbf{I}|\mathbf{Y}\right).$$

Since the target distribution $\pi(\mathbf{W}, \mathbf{V}|\mathbf{Y})$ is intractable, we construct an approximation to it, which we denote $q(\mathbf{W}, \mathbf{V}|\mathbf{Y})$, and use this as our proposal distribution. This proposal is proportional to

$$\mathbb{I}_{\{\mathbf{W}=\mathbf{I},\mathbf{V}=\mathbf{I}\}}\beta_0 + \sum_{j=1}^{q}\mathbb{I}_{\left\{\mathbf{W}=\mathbf{I}+\frac{\mathbf{I}^{(j)}}{\tilde{\mathbf{W}}^{(j,j)}},\mathbf{V}=\mathbf{I}\right\}}\hat{\beta}_j\hat{q}_j\left(\tilde{\mathbf{W}}^{(j,j)}\right) + \sum_{i=1}^{p}\mathbb{I}_{\left\{\mathbf{W}=\mathbf{I},\mathbf{V}=\mathbf{I}+\frac{\mathbf{I}^{(i)}}{\tilde{\mathbf{V}}^{(i,i)}}\right\}}\tilde{\beta}_i\tilde{q}_i\left(\tilde{\mathbf{V}}^{(i,i)}\right).$$

Clearly, there is no benefit in simulating multiple identical descendants, so we wish to sample precisely one dependent that corresponds to no outliers. To do this, and also

to have the same number of descendant particles for each possible type of outlier, we set $\beta_0 = \frac{1}{1+M(p+q)}$, $\tilde{\beta}_i = \frac{M}{1+M(p+q)}$, and $\hat{\beta}_j = \frac{M}{1+M(p+q)}$, and use stratified subsampling as in Fearnhead and Clifford (2003). This leads to $M' = M(p+q)+1$ total descendants per particle, M for each of the p additive and q innovative outliers, and one for no outlier. Each of these particles is then given a weight proportional to

$$\frac{\pi(\mathbf{W}_{t+1}, \mathbf{V}_{t+1}|\mathbf{Y}_{t+1})}{q(\mathbf{W}_{t+1}, \mathbf{V}_{t+1}|\mathbf{Y}_{t+1})}.$$

The main challenge now consists of obtaining proposal distributions $\tilde{q}_i(\cdot)$ for $1 \leq i \leq p$ and $\hat{q}_j(\cdot)$ for $1 \leq j \leq q$ that provide good approximations to the conditional posteriors which are proportional to $\tilde{\pi}_i(\cdot)$ and $\hat{\pi}_j(\cdot)$ respectively. In the next subsection, we therefore derive proposal distributions that provide leading order approximations to the conditional posteriors. To simplify notation, we define the predictive variance $\hat{\mathbf{\Sigma}} = \mathbf{C}\mathbf{A}\mathbf{\Sigma}\mathbf{A}^T\mathbf{C}^T + \mathbf{\Sigma}_A + \mathbf{C}\mathbf{\Sigma}_I\mathbf{C}^T$ and use it throughout the remainder of this chapter. We also begin by assuming that $\mathbf{C}$ contains no 0-columns. The proposal introduced in the following subsection also forms the basis of back-sampling introduced in Section 5.4, which allows to relax this on $\mathbf{C}$.

5.3.1 Proposal Distributions

For $1 \leq i \leq p$, we would like the proposal distribution $\tilde{q}_i\left(\tilde{\mathbf{V}}^{(i,i)}\right)$ for the precision, $\tilde{\mathbf{V}}^{(i,i)}$, to be as close as possible to $\tilde{\pi}_i\left(\tilde{\mathbf{V}}^{(i,i)}\right)$ or, equivalently, proportional to

$$f_i\left(\tilde{\mathbf{V}}^{(i,i)}\right) \frac{\exp\left(-\frac{1}{2}\left(\mathbf{Y} - \mathbf{C}\mathbf{A}\boldsymbol{\mu}\right)^T \left(\hat{\mathbf{\Sigma}} + \frac{\mathbf{\Sigma}_A^{(i,i)}}{\tilde{\mathbf{V}}^{(i,i)}}\mathbf{I}^{(i)}\right)^{-1} \left(\mathbf{Y} - \mathbf{C}\mathbf{A}\boldsymbol{\mu}\right)\right)}{\sqrt{\left|\hat{\mathbf{\Sigma}} + \frac{\mathbf{\Sigma}_A^{(i,i)}}{\tilde{\mathbf{V}}^{(i,i)}}\mathbf{I}^{(i)}\right|}},$$

where $f_i()$ denotes the PDF of the $\tilde{\sigma}_i\Gamma(a_i, a_i)$-distributed prior of $\tilde{\mathbf{V}}^{(i,i)}$.

It should be noted that the intractable terms,

$$\left|\hat{\mathbf{\Sigma}} + \frac{\mathbf{\Sigma}_A^{(i,i)}}{\tilde{\mathbf{V}}^{(i,i)}}\mathbf{I}^{(i)}\right| \qquad \text{and} \qquad \left(\hat{\mathbf{\Sigma}} + \frac{\mathbf{\Sigma}_A^{(i,i)}}{\tilde{\mathbf{V}}^{(i,i)}}\mathbf{I}^{(i)}\right)^{-1} \tag{5.3.1}$$

can both be expanded using the matrix determinant lemma and the Sherman Morrison formula respectively, as they are rank 1 updates of a determinant and inverse respectively. Indeed, by the matrix determinant lemma,

$$\left|\hat{\mathbf{\Sigma}} + \frac{\mathbf{\Sigma}_A^{(i,i)}}{\tilde{\mathbf{V}}^{(i,i)}}\mathbf{I}^{(i)}\right| = \frac{\left|\hat{\mathbf{\Sigma}}\right|}{\tilde{\mathbf{V}}^{(i,i)}}\left(1 + \mathbf{\Sigma}_A^{(i,i)}\left(\hat{\mathbf{\Sigma}}^{-1}\right)^{(i,i)} + O\left(\tilde{\mathbf{V}}^{(i,i)}\right)\right),$$

the leading order term is conjugate to the prior of $\tilde{\mathbf{V}}^{(i,i)}$. Moreover, by the Sherman Morrison formula the second term in Equation (5.3.1) is equal to

$$\hat{\mathbf{\Sigma}}^{-1} - \hat{\mathbf{\Sigma}}^{-1}\mathbf{I}^{(i)}\hat{\mathbf{\Sigma}}^{-1}\left[\frac{1}{\left(\hat{\mathbf{\Sigma}}^{-1}\right)^{(i,i)}} - \left(\frac{1}{\left(\hat{\mathbf{\Sigma}}^{-1}\right)^{(i,i)}}\right)^2 \frac{\tilde{\mathbf{V}}^{(i,i)}}{\mathbf{\Sigma}_A^{(i,i)}}\right],$$

up to $O\left(\left(\tilde{\mathbf{V}}^{(i,i)}\right)^2\right)$. Crucially, the first two terms are constant in $\tilde{\mathbf{V}}^{(i,i)}$, while the third is linear in $\tilde{\mathbf{V}}^{(i,i)}$ and therefore returns a term which is conjugate to the prior of $\tilde{\mathbf{V}}^{(i,i)}$. Furthermore, we are most concerned about accurately sampling the particle when an anomaly occurs in the ith component, which happens when the precision, $\tilde{\mathbf{V}}^{(i,i)}$, and the higher order terms, become small.

Keeping only the leading order terms in the determinant and the exponential term results in the proposal distribution

$$\tilde{\mathbf{V}}^{(i,i)} \sim \tilde{\sigma}_i\Gamma\left(a_i + \frac{1}{2}, a_i + \frac{\tilde{\sigma}_i}{2\mathbf{\Sigma}_A^{(i,i)}}\left(\frac{\left(\hat{\mathbf{\Sigma}}^{-1}\right)^{(i,:)}(\mathbf{Y} - \mathbf{CA}\boldsymbol{\mu})}{\left(\hat{\mathbf{\Sigma}}^{-1}\right)^{(i,i)}}\right)^2\right)$$

for $\tilde{\mathbf{V}}^{(i,i)}$. More detailed derivations, including the associated weight are given by Theorem 1 in the appendix. This proposal has the property that as the observed

anomaly in the ith component becomes larger, i.e. as

$$\frac{1}{\mathbf{\Sigma}_A^{(i,i)}}\left(\frac{\left(\hat{\mathbf{\Sigma}}^{-1}\right)^{(i,:)}(\mathbf{Y}-\mathbf{CA}\boldsymbol{\mu})}{\left(\hat{\mathbf{\Sigma}}^{-1}\right)^{(i,i)}}\right)^2$$

increases, the mean of the proposal for $\tilde{\mathbf{V}}^{(i,i)}$ diverges from the prior mean and behaves asymptotically like

$$(2a_i+1)\mathbf{\Sigma}_A^{(i,i)}\left(\frac{\left(\hat{\mathbf{\Sigma}}^{-1}\right)^{(i,i)}}{\left(\hat{\mathbf{\Sigma}}^{-1}\right)^{(i,:)}(\mathbf{Y}-\mathbf{CA}\boldsymbol{\mu})}\right)^2.$$

Consequently, the variance and the squared residual will be on the same scale, thus achieving computational robustness.

A very similar approach can be used to obtain a proposal distribution $\hat{q}_j\left(\tilde{\mathbf{W}}^{(j,j)}\right)$ which provides a leading order approximation for the distribution proportional to $\pi\left(\mathbf{I}+\frac{1}{\tilde{\mathbf{W}}^{(j,j)}}\mathbf{I}^{(j)},\mathbf{I}|\mathbf{Y}\right)$. The proposal consists of sampling

$$\tilde{\mathbf{W}}^{(j,j)}\sim\hat{\sigma_j}\Gamma\left(b_j+\frac{1}{2},b_j+\frac{\hat{\sigma_i}}{2\mathbf{\Sigma}_I^{(j,j)}}\left(\frac{\left(\mathbf{C}^T\right)^{(j,:)}\hat{\mathbf{\Sigma}}^{-1}(\mathbf{Y}-\mathbf{CA}\boldsymbol{\mu})}{\left(\mathbf{C}^T\hat{\mathbf{\Sigma}}^{-1}\mathbf{C}\right)^{(j,j)}}\right)^2\right)$$

and is of very similar form to the proposal distribution for particles with an additive outlier and well defined if $\mathbf{C}$ has no $\mathbf{0}$-columns. Further details, including the associated weight, are given in Theorem 2 in the appendix. Like the proposal distribution for particles with an additive anomaly this proposal is computationally robust: it ensures that the squared residual and the variance will be on the same scale as the anomaly in the jth innovative component becomes stronger.

Finally, the "proposal" for particles without anomalies consists of deterministically setting $\mathbf{V}=\mathbf{I}$ and $\mathbf{W}=\mathbf{I}$. The weight associated with this particle is proportional to the likelihood, the closed form of which is given in Theorem 3 in the appendix.

5.3.2 Choices of Parameters

The choice of hyper-parameters, particularly $\hat{\sigma}_i$ and $\tilde{\sigma}_i$, has a significant effect of the performance of the proposed filter. One reason for this is that an outlier observation could be the result of either an additive or an innovative outlier. It may be that the root cause can only be determined after further observations are made. Thus, we wish to choose hyper-parameters in such a way as to ensure that observed anomalies, which are equally well explained by different classes of anomalies, are given similar importance weights. The following result describes such a choice:

Theorem 4. *Let the prior for the hidden state $\boldsymbol{X}_t$ be $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and an observation $\boldsymbol{Y}_{t+1} := \boldsymbol{Y}$ be available. When*

$$\tilde{\sigma}_i = \Sigma_A^{(i,i)} \left(\hat{\boldsymbol{\Sigma}}^{-1}\right)^{(i,i)} \quad \textit{and} \quad \hat{\sigma}_j = \Sigma_I^{(j,j)} \left(\boldsymbol{C}^T \hat{\boldsymbol{\Sigma}}^{-1} \boldsymbol{C}\right)^{(j,j)},$$

and $a_1 = ... = a_p = b_1 = ... = b_q = c$, the weights of additive and innovative anomalies are asymptotically proportional to

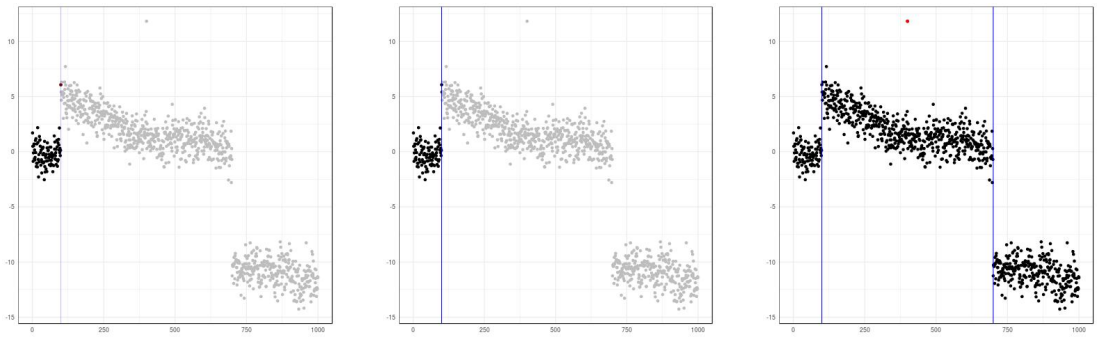
$$\frac{c^c \frac{1}{M} r_i \frac{\Gamma(c+\frac{1}{2})}{\Gamma(c)} \exp\left(\frac{1}{2}\delta^2\right)}{\left(\frac{\delta^2}{2}\right)^c} \quad \textit{and} \quad \frac{c^c \frac{1}{M} s_j \frac{\Gamma(c+\frac{1}{2})}{\Gamma(c)} \exp\left(\frac{1}{2}\delta^2\right)}{\left(\frac{\delta^2}{2}\right)^c}$$

when

$$\boldsymbol{Y} - \boldsymbol{CA\mu} = \frac{\delta \boldsymbol{e}_i}{\sqrt{\left(\hat{\boldsymbol{\Sigma}}^{-1}\right)^{(i,i)}}} \quad \textit{and} \quad \boldsymbol{Y} - \boldsymbol{CA\mu} = \frac{\delta \boldsymbol{C}^{(:,j)}}{\sqrt{\left(\boldsymbol{C}^T \hat{\boldsymbol{\Sigma}}^{-1} \boldsymbol{C}\right)^{(j,j)}}},$$

respectively, as $\delta \to \infty$

The above choice of hyper-parameters therefore leads to all components being given equal asymptotic importance weight under an anomaly they are able to account for. I.e. one which satisfies $\frac{\mathbf{C}^{(:,j)}}{\sqrt{\left(\mathbf{C}^T \hat{\boldsymbol{\Sigma}}^{-1} \mathbf{C}\right)^{(j,j)}}} \delta = \mathbf{Y} - \mathbf{CA\mu} = \frac{\delta \mathbf{e}_i}{\sqrt{\left(\hat{\boldsymbol{\Sigma}}^{-1}\right)^{(i,i)}}}$. Setting all the



(a) t=100 (b) t=101 (c) Full data

Figure 5.3.1: Robust particle filter output at various times. Additive anomalies are denoted by red points, innovative anomalies by blue lines. Grey observations are yet to be observed.

a_is and b_js to the same constant is advisable due to the fact that the convolution of two t-distributions whose means drift further and further apart yields two stable, i.e. non-vanishing modes if and only if they have the same scale parameter.

While, $\hat{\mathbf{\Sigma}}^{-1}$ is not fixed but time dependent, it nevertheless converges to a limit under an observable Kalman filter model. In practice, we therefore use this limit to set $\tilde{\sigma}_i$ and $\hat{\sigma}_j$.

5.3.3 Example 1 - revisited

The proposed filter can be applied to the data displayed in Figure 5.2.1a to detect anomalies in an online fashion. It is worth pointing out that the filter re-evaluates past anomalies as more data becomes available. This can be seen in Figure 5.3.1: When initially encountering the anomaly at time $t = 100$ the filter gives approximately equal weight to the possibility of it being an additive outlier and to it being an innovative

one. It is only when the next observation becomes available, that the filter (correctly) classifies it as an innovative anomaly. Note that only $N = 20$ particles were used and only $M = 1$ descendent of each anomaly type was sampled per particle.

5.4 Particle Filter With Back-Sampling – CE-BASS

As mentioned in the introduction, it is possible that innovative outliers may not immediately be observed. One such example are innovative outliers in the trend component of the model described in (5.2.3). The filter as described in Algorithm 1 can not deal with such anomalies as it only inflates the variance of the innovative process at time t when there is evidence in the observation at the same time t that an outlier occurred. This can be remedied by back-sampling particles representing innovative outliers at a later time, $t+k$, once more observations and therefore evidence for an anomaly are available. This can be done using nearly identical approximation strategies as used in the previous section and allows to relax the assumptions made in the previous section from $\mathbf{C}$ not having any $\mathbf{0}$-columns to requiring that the system be observable.

5.4.1 Back-Sampling Particles Using the Last $k+1$ Observations

The proposed back-sampling strategy at time t consists of sampling particles for $(\mathbf{V}_{t+1-k}, ...\mathbf{V}_{t+1}, \mathbf{W}_{t+1-k}, ..., \mathbf{W}_{t+1})$ given a $N(\boldsymbol{\mu}_{t-k}, \boldsymbol{\Sigma}_{t-k})$ filtering distribution for $\mathbf{x}_{t-k}$ and observations $\mathbf{Y}_{t-k+1}, ..., \mathbf{Y}_{t-k}$. Specifically, we sample particles with a inno-

vative single anomaly in $\mathbf{W}_{t+1-k}$ assuming no other innovative anomalies or additive anomalies. Conditional on these augmented particles classical Kalman updates can once more be used as shown in Algorithm 2. It should be noted that Algorithm 1 is a special case of Algorithm 2 which arises from setting $\mathcal{B}_1 = ... = \mathcal{B}_q = \{1\}$.

To sample a particle with an innovative anomaly in the jth component of $\mathbf{W}_{t+1-k}$, we define an augmented observation vector $\tilde{\mathbf{Y}}_{t+1-k}^{(k)} = (\mathbf{Y}_{t+1-k}^T, ..., \mathbf{Y}_{t+1}^T)^T$. This is normally distributed with mean $\tilde{\mathbf{C}}^{(k)}\mathbf{A}\boldsymbol{\mu}_{t-k}$ and variance

$$\tilde{\mathbf{C}}^{(k)}\left(\mathbf{A}\boldsymbol{\Sigma}_{t-k}\mathbf{A}^T + \tilde{\mathbf{Q}}^{(k)}\right)\left(\tilde{\mathbf{C}}^{(k)}\right)^T + \tilde{\mathbf{R}}^{(k)},$$

where $\tilde{\mathbf{C}}^{(k)} = \mathbf{C}\left(\left(\mathbf{A}^0\right)^T, ..., \left(\mathbf{A}^k\right)^T\right)^T$ denotes the augmented matrix mapping the hidden states to the observations,

$$\tilde{\mathbf{R}}^{(k)} = \begin{bmatrix} \mathbf{V}_{t+1-k}^{-1}\boldsymbol{\Sigma}_A & 0 & \ddots \\ 0 & \ddots & 0 \\ \ddots & 0 & \mathbf{V}_{t+1}^{-1}\boldsymbol{\Sigma}_A \end{bmatrix}$$

and

$$\tilde{\mathbf{Q}}^{(k)} = \begin{bmatrix} \mathbf{W}_{t+1-k}^{-1}\boldsymbol{\Sigma}_I & 0 & \ddots \\ 0 & \ddots & 0 \\ \ddots & 0 & \mathbf{W}_{t+1}^{-1}\boldsymbol{\Sigma}_I \end{bmatrix}$$

In a similar spirit, we define the augmented predictive variance to be

$$\hat{\boldsymbol{\Sigma}}^{(k)} = \tilde{\mathbf{C}}^{(k)}\left(\mathbf{A}\boldsymbol{\Sigma}_{t-k}\mathbf{A}^T + \mathbf{I}_{k+1} \otimes \boldsymbol{\Sigma}_I\right)\left(\tilde{\mathbf{C}}^{(k)}\right)^T + \mathbf{I}_{k+1} \otimes \boldsymbol{\Sigma}_A.$$

As a result of this reformulation, we retrieve update equations consisting of a single Kalman step, albeit with slightly different dimensions of the observation, $(k+1)p$ instead of p. It is therefore possible to use the sampling procedure for innovative outliers introduced in Section 5.3.1. This consists of sampling particles for $\tilde{\mathbf{W}}_{t+1-k}^{(j,j)}$ from

$$\hat{\sigma_j}\Gamma\left(b_j + \frac{1}{2}, b_j + \frac{\hat{\sigma_j}}{2\boldsymbol{\Sigma}_I^{(j,j)}}\left(\frac{\left(\left(\tilde{\mathbf{C}}^{(k)}\right)^T\right)^{(j,:)}\left(\hat{\boldsymbol{\Sigma}}^{(k)}\right)^{-1}\tilde{\mathbf{z}}_{t+1-k}^{(k)}}{\left(\left(\tilde{\mathbf{C}}^{(k)}\right)^T\left(\hat{\boldsymbol{\Sigma}}^{(k)}\right)^{-1}\tilde{\mathbf{C}}^{(k)}\right)^{(j,j)}}\right)^2\right).$$

Algorithm 2 Particle Filter (With Back Sampling) – CE-BASS

Input: An initial state estimate $(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$.

A number of descendants, $M' = M(p+q)+1$.

A number of particles to be maintained, N.

A stream of observations $\mathbf{Y}_1, \mathbf{Y}_2, ...$

Initialise: Set $Particles(0) = \{(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0, 1)\}$

Set $max\_horizon = \max\left(\cup_{i=1}^{q} \mathcal{B}_i\right)$

for $t \in \mathbb{N}^+$ **do**

  $Cand \leftarrow \{\}$ ▷ To Store Candidates

  for $(\boldsymbol{\mu}, \boldsymbol{\Sigma}, prob_{prev}) \in Particles(t-1)$ **do**

    $(\mathbf{V}, \mathbf{W}, prob) \leftarrow$ Sample_typical$(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{Y}_t, \mathbf{A}, \mathbf{C}, \boldsymbol{\Sigma}_A, \boldsymbol{\Sigma}_I)$

    $Cand \leftarrow Cand \cup \{(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{V}, \mathbf{W}, prob \cdot prob_{prev}, 1)\}$

    $Add\_Des \leftarrow$ Sample_add$(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{Y}_t, \mathbf{A}, \mathbf{C}, \boldsymbol{\Sigma}_A, \boldsymbol{\Sigma}_I, M)$

    for $(\mathbf{V}, \mathbf{W}, prob) \in Add\_Des$ **do**

      $Cand \leftarrow Cand \cup \{(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{V}, \mathbf{W}, prob \cdot prob_{prev}, 1)\}$

    end for

  end for

  for $hor \in \{1, ..., max\_horizon\}$ **do**

    for $(\boldsymbol{\mu}, \boldsymbol{\Sigma}, prob_{prev}) \in Particles(t-hor)$ **do**

      $\tilde{\mathbf{Y}} \leftarrow \left[\mathbf{Y}_{t-hor+1}^T, ..., \mathbf{Y}_t^T\right]^T$

      $Inn\_Des \leftarrow$ BS_inn$(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \tilde{\mathbf{Y}}, \mathbf{A}, \mathbf{C}, \boldsymbol{\Sigma}_A, \boldsymbol{\Sigma}_I, M, hor)$

      for $(\mathbf{V}, \mathbf{W}, prob) \in Inn\_Des$ **do**

        $Cand \leftarrow Cand \cup \{(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{V}, \mathbf{W}, prob \cdot prob_{prev}, hor)\}$

      end for

    end for

  end for

continues on next page

20: $Desc \leftarrow \text{Subsample}(N, Cand)$ ▷ Sampling proportional to $prob$

21: $Particles(t) \leftarrow \{\}$

22: **for** $(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{V}, \mathbf{W}, prob, hor) \in Desc$ **do**

23: $(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \leftarrow \text{KF\_Upd}(\mathbf{Y}_{t+1-hor}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{C}, \mathbf{A}, \mathbf{V}^{1/2}\boldsymbol{\Sigma}_A, \mathbf{W}^{1/2}\boldsymbol{\Sigma}_I)$

24: **if** $hor > 1$ **then**

25: **for** $i \in \{2, ..., hor\}$ **do**

26: $(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \leftarrow \text{KF\_Upd}(\mathbf{Y}_{t+i-hor}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{C}, \mathbf{A}, \boldsymbol{\Sigma}_A, \boldsymbol{\Sigma}_I)$

27: **end for**

28: **end if**

29: $Particles(t) \leftarrow Particles(t) \cup \{(\boldsymbol{\mu}, \boldsymbol{\Sigma}, prob \cdot \frac{|Cand|}{|Desc|})\}$

30: **end for**

31: **end for**

for the residual $\tilde{\mathbf{z}}_{t+1-k}^{(k)} \tilde{\mathbf{Y}}_{t+1-k}^{(k)} - \tilde{\mathbf{C}}^{(k)} \mathbf{A}\boldsymbol{\mu}_{t-k}$. The associated weight is given in Theorem 5 in the appendix.

As in Section 5.3.2, we want to give different particles equal weights if they explain anomalies equally well. In particular, we therefore want to balance out the weights given to the back-sampled particles and the descendants of particles with an anomaly sampled at time $t-k+1$ using just $\mathbf{Y}_{t+1-k}$. In order to do so, consider observations $\mathbf{Y}_{t+1}, ..., \mathbf{Y}_{t+1-k}$ which are such that they perfectly fit an innovative outlier in the ith innovative component at time $t-k+1$, i.e.

$$\tilde{\mathbf{Y}}_{t+1-k}^{(k)} - \left(\tilde{\mathbf{C}}^{(k)}\right) \mathbf{A}\boldsymbol{\mu}_{t-k} = \frac{\left(\tilde{\mathbf{C}}^{(k)}\right)^{(:,j)}}{\sqrt{\left(\left(\tilde{\mathbf{C}}^{(k)}\right)^T \left(\hat{\boldsymbol{\Sigma}}^{(k)}\right)^{-1} \left(\tilde{\mathbf{C}}^{(k)}\right)\right)^{(j,j)}}} \delta.$$

As δ grows, the importance weight behaves as

$$\frac{b_j^{b_j}\frac{1}{M}s_j\frac{\Gamma(b_j+\frac{1}{2})}{\Gamma(b_j)}\exp\left(-\delta^2\right)}{\left(\frac{\hat{\sigma}_j}{2\boldsymbol{\Sigma}_I^{(j,j)}\left(\left(\tilde{\mathbf{C}}^{(k)}\right)^T\left(\hat{\boldsymbol{\Sigma}}^{(k)}\right)^{-1}\left(\tilde{\mathbf{C}}^{(k)}\right)\right)^{(j,j)}}\delta^2\right)^{b_j}},$$

up to the likelihood term and the $\left(1-\sum_{i=1}^p r_i-\sum_{j=1}^q s_j\right)^k$ factor. However, these terms are also present in the weights of the descendants of the particles sampled at $t+1-k$ if no further anomaly was sampled at times $t+2-k,...,t+1$. Therefore, setting

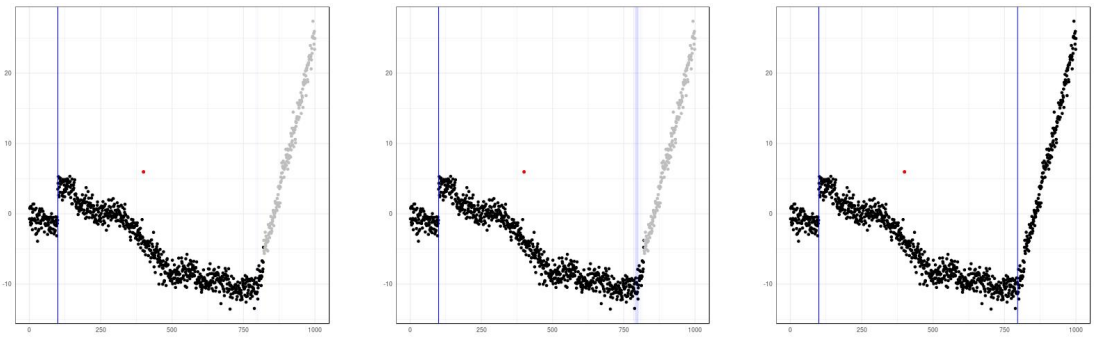
$$\hat{\sigma}_j=\boldsymbol{\Sigma}_I^{(j,j)}\left(\left(\tilde{\mathbf{C}}^{(k)}\right)^T\left(\hat{\boldsymbol{\Sigma}}^{(k)}\right)^{-1}\left(\tilde{\mathbf{C}}^{(k)}\right)\right)^{(j,j)}$$

results in the same asymptotic probabilities as the one obtained in Section 5.3.2. Given $\hat{\sigma}_j$ can only take a single value we set

$$\hat{\sigma}_j=\max_{k\in\mathcal{B}_j}\left(\boldsymbol{\Sigma}_I^{(j,j)}\left(\left(\tilde{\mathbf{C}}^{(k)}\right)^T\left(\hat{\boldsymbol{\Sigma}}^{(k)}\right)^{-1}\left(\tilde{\mathbf{C}}^{(k)}\right)\right)^{(j,j)}\right),$$

where $\mathcal{B}_j\subset\mathbb{N}$ denotes the set of horizons used to back-sample the jth component of the $\mathbf{W}_t$.

A range of observations guide the choice of the sets $\mathcal{B}_j$ for $1\leq j\leq q$. We assume that the Kalman model is observable, i.e. that there exists a k such that the matrix $\left[(\mathbf{C})^T,(\mathbf{CA})^T,...,\left(\mathbf{CA}^k\right)^T\right]$ has full column rank. Let k^* denote the lowest such k. It is advisable to choose the set $\mathcal{B}_j$ such that it contains at least one element greater or equal to k^*. The reason for this being that any innovative anomaly capable of eventually influencing the observations must do so within k^* observations from occurring. It should also be noted that a horizon h can only be in the set $\mathcal{B}_j$ if the jth column of the augmented mapping from the hidden states to



(a) t=820 (b) t=821 (c) Full data

Figure 5.4.1: Robust particle filter output at various times. Additive anomalies are denoted by red points, innovative anomalies by blue lines. Grey observations are yet to be observed.

the observations, $\tilde{\mathbf{C}}^{(h)}$, is non-zero as this is required by the proposal. Consequently, setting $\mathcal{B}_j = \left\{ k \in \{1, ..., k^*\} : \left(\tilde{\mathbf{C}}^{(k)}\right)^{(:,j)} \neq \mathbf{0} \right\}$ is a natural choice.

5.4.2 Example

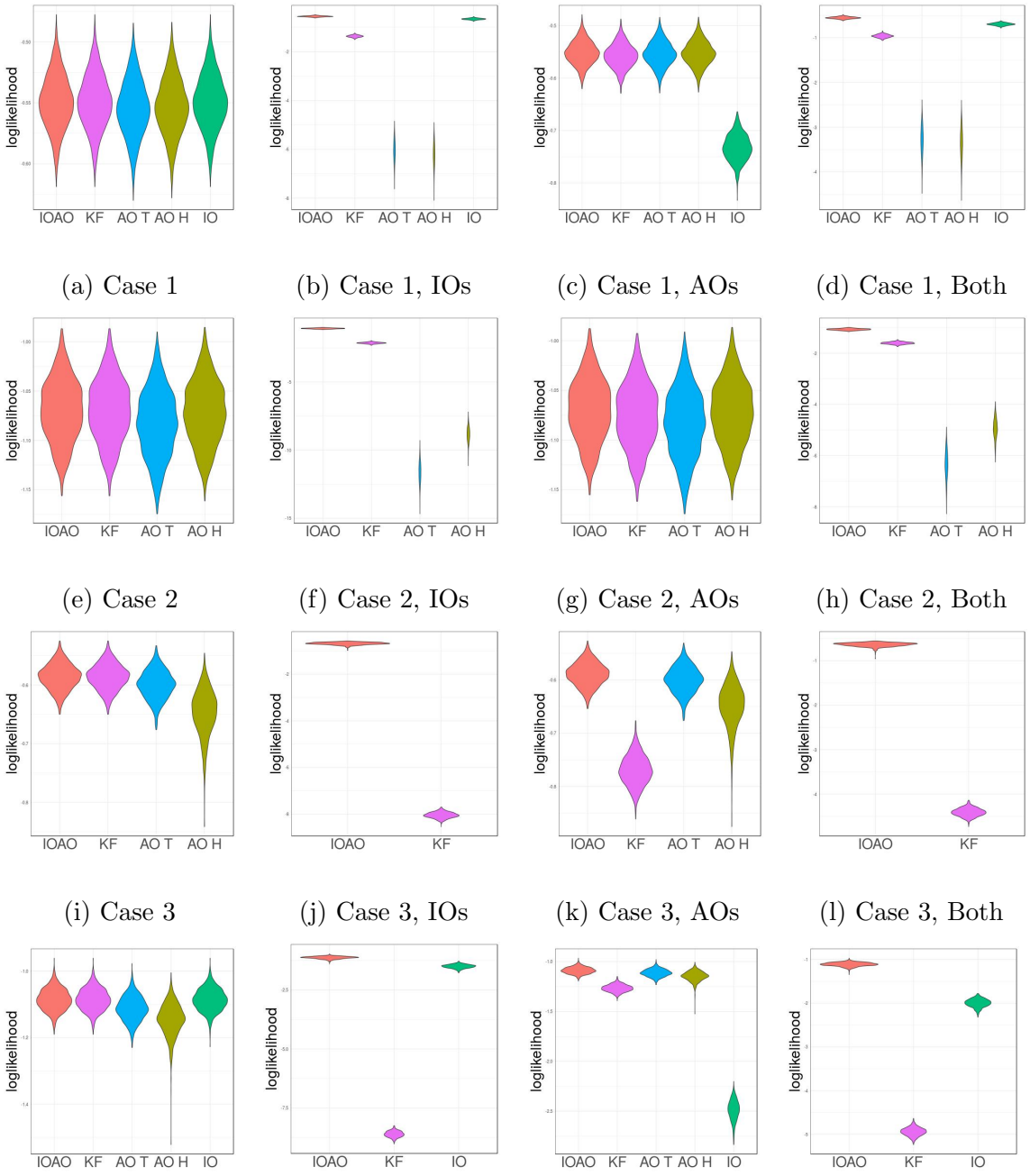
With back-sampling, we are now able to tackle the example from Figure 5.2.1b. We used $\mathcal{B}_1 = \{1, ..., 40\}$, $\mathcal{B}_2 = \{1, ..., 40\}$, to sample back up to 40 observations. We maintained $N = 40$ particles and sampled $M = 1$ descendants of each type. The output of the particle filter can be seen in Figure 5.4.1. As before, the filter updates its output as new observations become available. Whilst the trend innovation occurs at time $t = 800$, the anomaly is first detected around time $t = 820$. Even then, there is a large amount of uncertainty regarding the precise location of the anomaly which only gets resolved at a later time.

5.5 Simulations

We now turn to comparing CE-BASS against other methods. In particular, we compare against the t-distribution based additive outlier robust filter by Agamennoni et al. (2011), the Huberisation based additive outlier robust filter by Ruckdeschel et al. (2014), the Huberisation based innovative outlier robust filter by Ruckdeschel et al. (2014), and the classical Kalman Filter (Kalman, 1960). All these algorithms are implemented in the accompanying package.

We consider four different models and generate 1000 observations for each. For each of the four models, we consider a case in which no anomalies are present, a case in which only additive anomalies are present, a case in which only innovative anomalies are present, and a case in which both additive and innovative anomalies are present. When anomalies are added, they are added at times $t = 100$, $t = 300$, $t = 600$, and $t = 900$. Specifically we considered the following three models:

1. The model of Example 1 with $\sigma_A = 1$ and $\sigma_I = 0.1$. We consider a case with only additive outliers, a case with only innovative outliers, and a case where an additive outlier at $t = 100$, is followed by two innovative outliers at times $t = 300$ and $t = 600$, which were then followed by an additive outlier at time $t = 900$. To simulate additive anomalies, we set $V_t^{\frac{1}{2}}\sigma_A\epsilon_t = 10$ and to simulate the innovative outliers we set $W_t^{\frac{1}{2}}\sigma_I\nu_t = 10$.



(a) Case 1 (b) Case 1, IOs (c) Case 1, AOs (d) Case 1, Both

(e) Case 2 (f) Case 2, IOs (g) Case 2, AOs (h) Case 2, Both

(i) Case 3 (j) Case 3, IOs (k) Case 3, AOs (l) Case 3, Both

(m) Case 4 (n) Case 4, IOs (o) Case 4, AOs (p) Case 4, Both

Figure 5.5.1: Average predictive log-likelihood of the five filters (IOAO: CE-BASS, KF: The Kalman Filter, AO T: Agamennoni et al. (2011), AO H: Ruckdeschel et al. (2014), IO H: Ruckdeschel et al. (2014)) under a range of models. Higher values correspond to better performance. Methods are omitted if they can not be applied to the setting or if their performance is too poor.

2. The random walk model with two measurements

$$Y_t^{(1)} = X_t + \left(V_t^{(1)}\right)^{\frac{1}{2}} \sigma_A^{(1)} \epsilon_t^{(1)}, \qquad X_t = X_{t-1} + W_t^{\frac{1}{2}} \sigma_I \nu_t$$
$$Y_t^{(2)} = X_t + \left(V_t^{(2)}\right)^{\frac{1}{2}} \sigma_A^{(2)} \epsilon_t^{(2)},$$

where $\sigma_A^{(1)} = \sigma_A^{(2)} = 1$ for $i = 1, 2$ and $\sigma_I = 0.1$. We consider a case with only additive outliers (one in the first component, then two in the second, then one in the first), a case with only innovative outliers, and a case where an additive outlier in the first component at time $t = 100$ is followed by two innovative outliers at times $t = 300$ and $t = 600$, which are then followed by an additive outlier in the second component at time $t = 900$. For additive anomalies, we set $\left(V_t^{(1)}\right)^{\frac{1}{2}} \sigma_A^{(1)} \epsilon_t^{(1)} = 10$ or $\left(V_t^{(2)}\right)^{\frac{1}{2}} \sigma_A^{(2)} \epsilon_t^{(2)} = 10$ and for innovative outliers, we set $W_t^{\frac{1}{2}} \sigma_I \nu_t = 10$.

3. The model of Example 2 with $\sigma_A = 1$, $\sigma_I^{(1)} = 0.1$ and $\sigma_I^{(2)} = 0.01$. We consider a case with only additive outliers, a case with only innovative outliers (one in the second component, then one in the first, then one in the second, then one in the first), and a case with an additive outlier at $t = 100$, followed by an innovative outlier affecting the first component of the hidden state at times $t = 300$, followed by an innovative outlier affecting the second component of the hidden state at times $t = 600$, followed by an additive outlier at time $t = 900$. The additive anomalies were instances where we set $V_t^{\frac{1}{2}} \epsilon_t = 30$ and the innovative outliers were instances where we set $\left(W_t^{(1)}\right)^{\frac{1}{2}} \eta_t^{(1)} = 100$ or $\left(W_t^{(2)}\right)^{\frac{1}{2}} \eta_t^{(2)} = 500$.

4. An extension of Example 2 where the position is also observed. The equations governing the hidden state are as before whilst the equations governing the

observations are

$$Y_t^{(1)} = X_t^{(1)} + \left(V_t^{(1)}\right)^{\frac{1}{2}} \sigma_A^{(1)} \epsilon_t^{(1)},$$
$$Y_t^{(2)} = X_t^{(2)} + \left(V_t^{(2)}\right)^{\frac{1}{2}} \sigma_A^{(2)} \epsilon_t^{(2)},$$

where $\sigma_A^{(1)} = \sigma_A^{(2)} = 1$. We consider a case with only additive outliers (in the first component only), a case with only innovative outliers (one in the second component, then one in the first, then one in the second, then one in the first), and a case with an additive outlier at time $t = 100$, followed by an innovative outlier affecting the first component of the hidden state at time $t = 300$, followed by an innovative outlier affecting the second component of the hidden state at time $t = 600$, followed by an additive outlier at time $t = 900$. For additive anomalies, we set $\left(V_t^{(1)}\right)^{\frac{1}{2}} \sigma_A^{(1)} \epsilon_t^{(1)} = 30$ and for innovative outliers, we set $\left(W_t^{(1)}\right)^{\frac{1}{2}} \sigma_I^{(1)} \eta_t^{(1)} = 100$ or $\left(W_t^{(2)}\right)^{\frac{1}{2}} \sigma_I^{(2)} \eta_t^{(2)} = 500$.

We evaluate the different methods based on average predictive log-likelihood and average predictive mean squared error. We exclude all observations corresponding to anomalies from the calculation of these averages since the filters can not be expected to predict them. When calculating the average mean squared error we additionally remove one observation after the anomaly in the first setting and two observations in the third setting from the performance metric. This is to give the filter enough information to determine which type of anomaly the outlier corresponds to and return to a unimodal posterior, as the MSE is only an appropriate metric for unimodal posteriors.

The average log-likelihoods across all models can be found in Figure 5.5.1, while

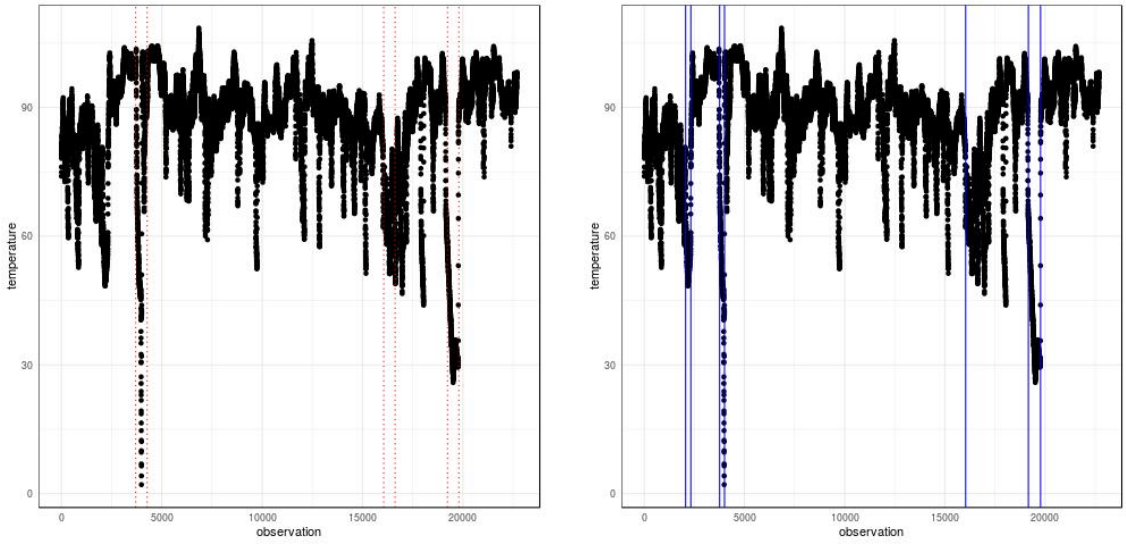
the qualitatively very similar results for the mean squared error can be found in the appendix. We see that the performance of CE-BASS compares favourably with that of the competing methods. In particular it is as accurate as the Kalman filter in the absence of anomalies and is more accurate than the additive outlier and innovative outlier robust filters even when only additive or innovative outliers are present, i.e. the settings for which these algorithms were designed.

5.6 Application

In this section, we apply CE-BASS to two real datasets. We will use different types of models for the two applications to illustrate the way in which CE-BASS can be used. The first dataset is a labelled benchmark dataset which consists of temperature readings on a large industrial machine. Here, we will use a model which considerably restricts the movements of the hidden states when no anomalies are present, and thus emulates a changepoint model. The second is an unlabelled dataset which consist of repeated throughput measurements on a router. For that application we will use a model which has a considerable amount of flexibility and where the hidden states tend to follow the observations and therefore detect localised anomalies.

5.6.1 Machine Temperature Data

We now apply CE-BASS to the machine temperature data taken from the Numenta Anomaly Benchmark (NAB, Lavin and Ahmad (2015)) which can be accessed at *https://github.com/numenta/NAB*. The data consists of over 20000 readings from a



(a) Raw data with labels

(b) CE-BASS output

Figure 5.6.1: Machine temperature dataset. The labelled anomalies are: a planned shutdown, an early warning sign of a problem, and the catastrophic system failure caused by the problem.

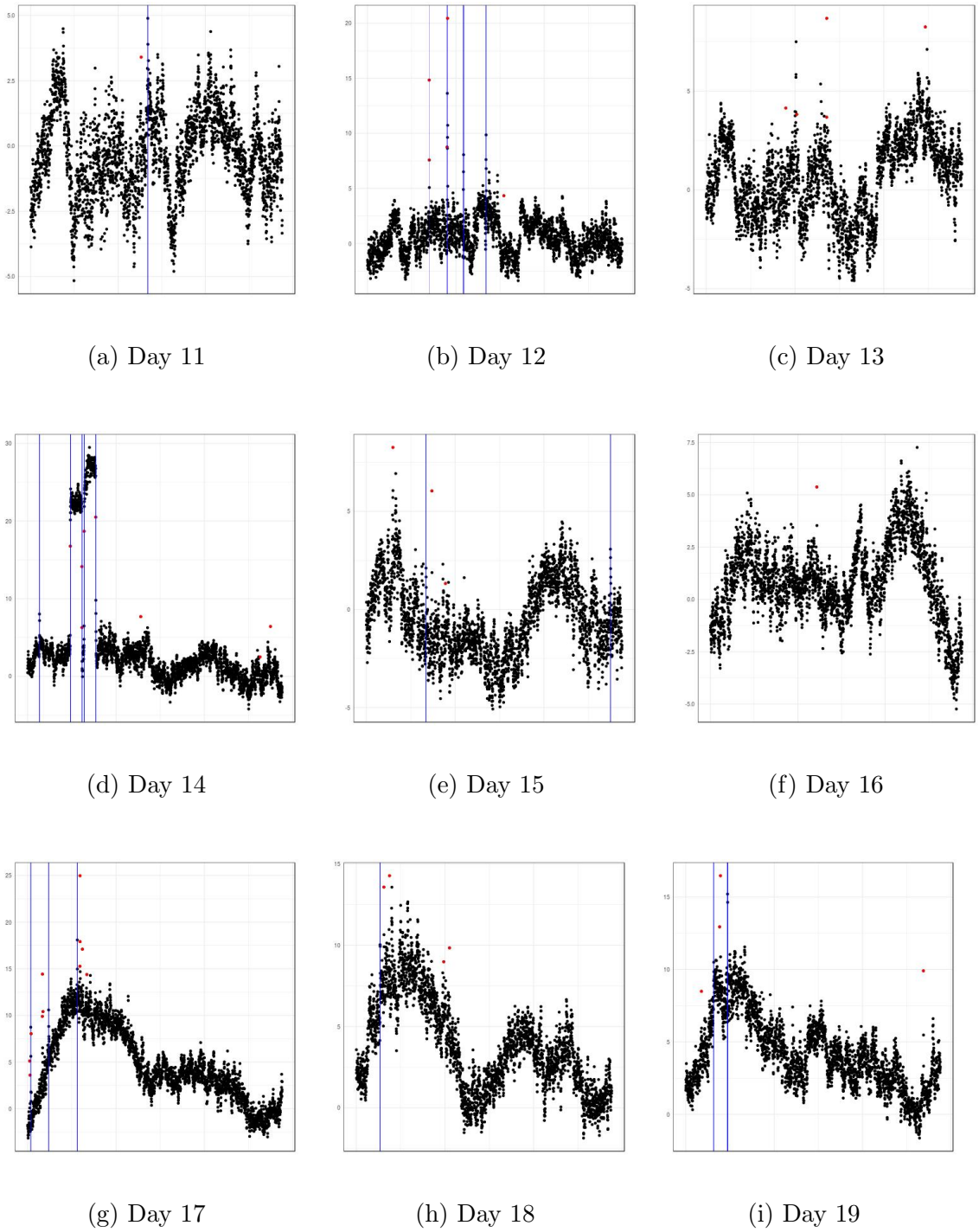
temperature sensor on a large industrial machine and is displayed in Figure 5.6.1a along the three periods of anomalous behaviour labelled by an engineer. The first corresponds to a planned shutdown and the second to an early warning sign of the third anomaly – a catastrophic failure.

In order to do so, we use the random walk model from Example 1 with the aim of detecting persistent changes in mean. We therefore use a maximum backsampling horizon of 250 by setting $\mathcal{B}_1 = \{1, 5, 10, 20, 40, 80, 150, 250\}$ and fix $\sigma_I = 1/10000\sigma_A$ to ensure that long and weak anomalies will not be interpreted as a persistent shift in the typical state. We use the first 15% of the data, marked by Lavin and Ahmad (2015) as train data, to estimate the standard deviation σ_A as well as the initial mean μ_0 using the median absolute deviation and the median respectively. Using robust covariance methods we also detect very strong auto-correlation ($\rho = 0.99$) and therefore took the default probabilities for anomalies to the power of $\frac{1}{1-\rho}$.

The results of this analysis can be seen in Figure 5.6.1b. We note that all anomalies flagged by the engineer are also being detected by CE-BASS. Two additional innovative anomalies around a prolonged drop which preceded the planned shutdown are also detected. They could be a false positive or an early warning sign of an anomaly prevented by the shutdown which has not been noticed by the engineer.

5.6.2 Router Data

The online analysis of aggregated traffic data on servers is an important challenge in both predictive maintenance and cyber security. This is because anomalies in throughput can point towards problems in the network such as malfunctions or ma-



(a) Day 11 (b) Day 12 (c) Day 13

(d) Day 14 (e) Day 15 (f) Day 16

(g) Day 17 (h) Day 18 (i) Day 19

Figure 5.6.2: CE-BASS applied to 9 days of de-seasonalised router data. Lines correspond to innovative anomalies, i.e. spikes or level shifts.

licious behaviour. Detecting anomalies as soon as possible therefore means that the root cause can be addressed more quickly – potentially even before user experience is affected or harm caused.

In this section, we consider 19 days worth of data from a network IP router which has been gathered at a frequency of one observation every 30 seconds. To preserve confidentiality, we de-seasonalised the data for days 11 to 19 using a seasonality model trained on days 1 to 10 and, for the purpose of this chapter, consider only the de-seasonalised data for days 11 to 19 which can be found in Figures 5.6.2a to 5.6.2i. The main features apparent in the daily series are spikes, outliers, and changepoints. In order to capture these, we use an AR(1) model with slowly changing mean to model the observations Y_t. Formally, we used the model

$$Y_t = X_t^{(1)} + X_t^{(2)} + V_t\sigma_A\epsilon_t, \qquad X_t^{(1)} = X_{t-1}^{(1)} + W_t^{(1)}\sigma_I^{(1)}\eta_t^{(1)},$$
$$X_t^{(2)} = \rho X_{t-1}^{(2)} + W_t^{(2)}\sigma_I^{(2)}\eta_t^{(2)}.$$

Here, anomalies in ϵ_t correspond to isolated outliers, anomalies in $\eta_t^{(1)}$ correspond to level shifts and outliers in $\eta_t^{(2)}$ correspond to spikes.

We use the first 1000 observations of the first day, to obtain the estimates $\sigma_A = 0.0516$, $\sigma_I^{(1)} = 0.0157$, $\sigma_I^{(2)} = 0.516$, and $\rho = 0.815$. The result obtained from running CE-BASS with these parameters on the daily router data is displayed in Figures 5.6.2a to 5.6.2i. We note that very few of the anomalies returned can be classed as false positives. At the same time, a large number of anomalies are flagged, including a large number of outliers and spikes, but also some level shifts (Day 14). Discussion with engineers highlighted that the anomalies detected matched well with their knowledge

of the data. This shows CE-BASS's ability to return a large number of diverse features which can be used as inputs to a supervised algorithm should labels become available.

Chapter 6

Conclusions And Further Research

This thesis has contributed three new algorithms, CAPA, MVCAPA, and CE-BASS, to the anomaly detection literature.

CAPA introduced in Chapter 3 is, to the best of our knowledge, the first algorithm aiming to detect and distinguish between collective and point anomalies. We established that CAPA can consistently detect collective anomalies in mean and variance. No comparable consistency result for joint changes in mean and variance exists in the literature. A simulation study shows that CAPA compares favourably with other anomaly and changepoint detection methods.

One potentially interesting avenue of further research would be to apply CAPA recursively: The observation CAPA identifies as typical can be used to re-estimate the typical parameter. CAPA could then be re-run using this new estimate for the typical distribution. Whilst this approach is not guaranteed to converge, it would nevertheless be able to provide more accurate robust estimates in the presence of collective anomalies than classical robust estimates like the Huber loss or Tukey's

Bi-weight loss which implicitly assume the presence of point anomalies only.

MVCAPA, introduced in Chapter 4, extends CAPA to the multivariate setting and aims to detect point anomalies as well as collective anomalies which affect multiple components. The main novelty of MVCAPA is that it does not assume that the collective anomalies are perfectly aligned. The algorithm is proven to be consistent at detecting changes in mean and to have optimal power.

It would be interesting to see whether an approach allowing for lags could be developed for the detection of classical (non-epidemic) multivariate changepoints. Another avenue of further research would be to extend the intermediate penalty regime to the chi^2_v-case with $v \geq 2$ and to derive sharper penalties for other popular settings such as the detection of anomalies characterised by joint changes in mean and variance.

CE-BASS, introduced in Chapter 5, is a novel Kalman filter which is robust to both additive and innovative outliers whilst fully capturing the multimodality. The algorithm is shown to be able to deal with identifiability issues through back-sampling and applied to two real data applications.

Extending the algorithm to deal with anomalies affecting more than one component of either the observations or the hidden states at a given time t would be an interesting extension. Another avenue of further research would be to investigate ways to tune the transition and noise parameters for a given model and to perform model comparison. This is likely to be difficult given the non-convex nature of the likelihood.

The new algorithms have been implemented in the `R` packages `anomaly` and `RobKF`. Both packages are available on CRAN and/or GitHub.

Appendix A

CAPA

A.1 Pseudocode for CAPA

Algorithm 3 CAPA Algorithm (No Pruning)

Input: A set of observations of the form, $(x_1, x_2, \ldots, x_n)$ where $x_i \in \mathbb{R}$.

Penalty constants β and β' for the introduction of a collective or a point anomaly

A minimum segment length $l \geq 2$

Initialise: Set $C(0) = 0$, $Anom(0) = NULL$.

$\hat{\mu} \leftarrow MEDIAN(x_1, x_2, \ldots, x_n)$ ▷ Obtain robust estimates of the mean and variance

$\hat{\sigma} \leftarrow IQR(x_1, x_2, \ldots, x_n)$

for $i \in \{1, ..., n\}$ **do**

    $x_i \leftarrow \frac{x_i - \hat{\mu}}{\hat{\sigma}}$ ▷ Centralise the data

end for

for $m \in \{1, ..., n\}$ **do**

    $C_1(m) \leftarrow \min_{0 \leq k \leq m-l} \left[C(k) + (m-k) \left[\log \left(\frac{1}{m-k} \sum_{t=k+1}^{m} \left(x_t - \bar{x}_{(k+1):m} \right)^2 \right) + 1 \right] + \beta \right]$

continues on next page

8: $s \leftarrow \underset{0\leq k\leq m-l}{\operatorname{argmin}} \left[C(k) + (m-k)\left[\log\left(\frac{1}{m-k}\sum_{t=k+1}^{m}\left(x_t - \bar{x}_{(k+1):m}\right)^2\right) + 1\right] + \beta\right]$

9: $C_2(m) \leftarrow C(m-1) + x_m^2$ ▷ No Anomaly

10: $C_3(m) \leftarrow C(m-1) + 1 + \log\left(\gamma + x_m^2\right) + \beta'\Big]$, ▷ Point Anomaly

11: $C(m) \leftarrow \min\left[C_1(m), C_2(m), C_3(m)\right]$

12: **switch** $\arg\min\left[C_1(m), C_2(m), C_3(m)\right]$ **do** ▷ Select type of anomaly giving the lowest cost

13: **case** 1 : $Anom(m) \leftarrow [Anom(s), (s+1, m)]$

14: **case** 2 : $Anom(m) \leftarrow Anom(m-1)$

15: **case** 3 : $Anom(m) \leftarrow [Anom(m-1), (m)]$

16: **end for**

Output The points and segments recorded in $Anom(n)$

Algorithm 4 CAPA Algorithm (With Pruning)

Input: A set of observations of the form, $(x_1, x_2, \ldots, x_n)$ where $x_i \in \mathbb{R}$.

Penalty constants β and β' for the introduction of a collective or a point anomaly

A minimum segment length $l \geq 2$

Initialise: Set $C(0) = 0$, $Anom(0) = NULL$, $\tau = \{\}$

1: $\hat{\mu} \leftarrow MEDIAN(x_1, x_2, \ldots, x_n)$ ▷ Obtain robust estimates of the mean and variance

2: $\hat{\sigma} \leftarrow IQR(x_1, x_2, \ldots, x_n)$

3: **for** $i \in \{1, ..., n\}$ **do**

4: $x_i \leftarrow \frac{x_i - \hat{\mu}}{\hat{\sigma}}$ ▷ Centralise the data

5: **end for**

6: **for** $m \in \{1, ..., n\}$ **do**

7: **if** $m \geq l$ **then**

8: $\tau \leftarrow \tau \cup \{(m-l, n)\}$ ▷ A tuple containing an option for the DP and the removal time

9: **end if**

continues on next page

10: $C_1(m) \leftarrow \min_{k\in\tau}\Big[C(k[1]) + (m-k[1])\left[\log\left(\frac{1}{m-k[1]}\sum_{t=k[1]+1}^{m}\left(x_t - \bar{x}_{(k[1]+1):m}\right)^2\right)+1\right]+\beta\Big]$ ▷ Coll. Anom.

11: $s \leftarrow \operatorname{argmin}_{k\in\tau}\Big[C(k[1]) + (m-k[1])\left[\log\left(\frac{1}{m-k[1]}\sum_{t=k[1]+1}^{m}\left(x_t - \bar{x}_{(k[1]+1):m}\right)^2\right)+1\right]+\beta\Big]$

12: $C_2(m) \leftarrow C(m-1) + x_m^2$ ▷ No Anomaly

13: $C_3(m) \leftarrow C(m-1) + 1 + \log\left(\gamma + x_m^2\right) + \beta'\Big]$, ▷ Point Anomaly

14: $C(m) \leftarrow \min\left[C_1(m), C_2(m), C_3(m)\right]$

15: **for** $k \in \tau$ **do**

16: **if** $(k[2] = n) \wedge \Big(C(m) < C(k[1]) + (m-k[1])\left[\log\left(\frac{1}{m-k[1]}\sum_{t=k[1]+1}^{m}\left(x_t - \bar{x}_{(k[1]+1):m}\right)^2\right)+1\right]\Big)$ **then**

17: $\tau \leftarrow \tau \setminus \{k\} \cup \{(k[1], m+l-1)\}$ ▷ Set destruction time

18: **end if**

19: **if** $m \geq k[2]$ **then**

20: $\tau \leftarrow \tau \setminus \{k\}$ ▷ Remove from solution space once destruction time reached

21: **end if**

22: **end for**

23: **switch** $\arg\min\left[C_1(m), C_2(m), C_3(m)\right]$ **do** ▷ Select type of anomaly giving the lowest cost

24: **case** 1 : $Anom(m) \leftarrow [Anom(s[1]), (s[1]+1, m)]$

25: **case** 2 : $Anom(m) \leftarrow Anom(m-1)$

26: **case** 3 : $Anom(m) \leftarrow [Anom(m-1), (m)]$

27: **end for**

Output The points and segments recorded in $Anom(n)$

A.2 Proofs of Propositions and Theorems

This Appendix contains proofs for all the results in this papers. Proofs for Lemmata we use can be found in Appendix A.4.

A.2.1 Proof of Proposition 1

Let $m' \geq m + \hat{l}$. We have

$$C(m) + \min_{\theta} \left(\sum_{t=m+1}^{m'} \mathcal{C}(\mathbf{x}_t, \theta) \right) + \beta \leq C(k) + \min_{\theta} \left(\sum_{t=k}^{m} \mathcal{C}(\mathbf{x}_t, \theta) \right) + \min_{\theta} \left(\sum_{t=m+1}^{m'} \mathcal{C}(\mathbf{x}_t, \theta) \right) + \beta$$
$$\leq C(k) + \min_{\theta} \left(\sum_{t=k+1}^{m'} \mathcal{C}(\mathbf{x}_t, \theta) \right) + \beta,$$

which shows that the cost of choosing k will always be larger than that of choosing m. We can thus disregard k.

A.2.2 Proof of Proposition 2

Assume, without loss of generality, that $\mu = 0$ and $\sigma = 1$. This Proof has two parts. The first one consist of showing that $\mathbb{P}\left(\hat{O} = \emptyset\right) \geq 1 - C_1' n e^{-\psi}$, the second one consists of showing that $\mathbb{P}\left(\hat{K} = 0\right) \geq 1 - C_1' n e^{-\psi} - C_2' \left(n e^{-\psi}\right)^2$

Part 1: We begin by proving that $\mathbb{P}\left(\hat{O} = \emptyset\right) \geq 1 - C_1' n e^{-\psi}$. Note that $x_i^2 \sim \chi_1^2$. We define $a_+(\psi)$ and $a_-(\psi)$ via the equation

$$\mathbb{P}\left(\chi_1^2 > a_+(\psi)\right) = \mathbb{P}\left(\chi_1^2 < a_-(\psi)\right) = e^{-\psi}$$

for $\psi > \log(1/2)$ Therefore, by a Bonferroni correction,

$$\mathbb{P}\left(a_-(\psi) < x_i^2 < a_+(\psi)\right) \geq 1 - 2e^{-\psi}.$$

Note that

$$x_i^2 - \log\left(\gamma + x_i^2\right) - 1 < x_i^2 - \log\left(x_i^2\right) - 1$$

and that the function $f(x) = (x-1) - \log(x)$ is decreasing for $x \leq 1$ and increasing thereafter. Consequently

$$x_i^2 - \log\left(x_i^2\right) - 1 \leq \max(f(a_+(\psi)), f(a_-(\psi)))$$

with probability $1 - 2e^{-\psi}$. We also know that the Chernoff bounds

$$\mathbb{P}\left(\chi_1^2 > a\right) \leq \exp\left(\frac{-a + 1 + \log(a)}{2}\right) \qquad \equiv \qquad a - 1 - \log(a) \leq -2\log(\mathbb{P}\left(\chi_1^2 > a\right))$$

and

$$\mathbb{P}\left(\chi_1^2 < a\right) \leq \exp\left(\frac{-a + 1 + \log(a)}{2}\right) \qquad \equiv \qquad a - 1 - \log(a) \leq -2\log(\mathbb{P}\left(\chi_1^2 < a\right))$$

hold for $a \geq 1$ and $a \leq 1$ respectively. Hence, $\max(f(a_+), f(a_-)) \leq 2\psi$ and thus

$$x_i \notin \hat{O} \quad \equiv \quad x_i^2 - \log\left(\gamma + x_i^2\right) - 1 \leq 2\psi.$$

holds with probability with probability $1 - 2e^{-\psi}$. A Bonferroni correction over $x_1, ..., x_n$ then gives the result.

Part 2: We now prove that $\mathbb{P}\left(\hat{K} = \emptyset\right) \geq 1 - C_1' n e^{-\psi} - C_2'\left(ne^{-\psi}\right)^2$. First of all, note that this is equivalent to showing that

$$\mathbb{P}\left(\sum_{s=i}^{j} x_s^2 - (j-i+1)\left[1 + \log\left(\frac{\sum_{s=i}^{j}(x_s - \bar{x}_{i:j})^2}{(j-i+1)}\right)\right] < 4 + 4\psi + 4\sqrt{2\psi} \quad 1 \leq i \leq j - l + 1 < j \leq n\right)$$
$$\geq 1 - C_1' n e^{-\psi} - C_2'\left(ne^{-\psi}\right)^2.$$

For a fixed i and j, writing $a = j - i + 1$, we have that $\sum_{s=i}^{j} x_s^2 = \sum_{s=i}^{j}(x_s - \bar{x}_{i:j})^2 + a\left(\bar{x}_{i:j}\right)^2$. Note that $a\left(\bar{x}_{i:j}\right)^2 \sim \chi_1^2$ and $\sum_{s=i}^{j}\left(x_s - \bar{x}_{i:j}\right)^2 \sim \chi_{a-1}^2$. Moreover, these two

random variables are independent. Consequently, the MGF of

$$a\left(\bar{x}_{i:j}\right)^2 - a\log\left(\frac{\sum_{s=i}^{j}\left(x_s-\bar{x}_{i:j}\right)^2}{a}\right) + \sum_{s=i}^{j}\left(x_s-\bar{x}_{i:j}\right)^2 - a$$

is given by

$$\begin{aligned}
&\sqrt{\frac{1}{1-2\lambda}}\int_0^\infty \left(\frac{a}{x}\right)^{a\lambda} e^{\lambda x-\lambda a}\left(\Gamma\left(\frac{a-1}{2}\right)2^{\frac{a-1}{2}}\right)^{-1} x^{\frac{a-1}{2}-1}e^{-x/2}dx \\
&= \left(\frac{1}{1-2\lambda}\right)^{\frac{1}{2}a(1-2\lambda)} \frac{a^{\lambda a}}{e^{\lambda a}2^{\lambda a}}\frac{\Gamma\left(\frac{a-1}{2}-\lambda a\right)}{\Gamma\left(\frac{a-1}{2}\right)} = \left(\frac{1}{1-2\lambda}\right)^{\frac{1}{2}a(1-2\lambda)} \frac{a^{\lambda a}}{e^{\lambda a}2^{\lambda a}}\frac{\frac{a-1}{2}}{\frac{a-1}{2}-\lambda a}\frac{\Gamma\left(\frac{a+1}{2}-\lambda a\right)}{\Gamma\left(\frac{a+1}{2}\right)}
\end{aligned}$$

We now use the following Stirling bounds from Artin (2015)

$$1 \le \Gamma(x)\frac{e^x}{\sqrt{2\pi}x^{x-\frac{1}{2}}} \le e^{\frac{1}{12x}}$$

which imply that:

$$\frac{\Gamma\left(\frac{a+1}{2}-\lambda a\right)}{\Gamma\left(\frac{a+1}{2}\right)} \le e^{\frac{1}{12\left(\frac{a+1}{2}-\lambda a\right)}} e^{\lambda a}\frac{\left(\frac{a+1}{2}-\lambda a\right)^{\left(\frac{a}{2}-\lambda a\right)}}{\left(\frac{a+1}{2}\right)^{\left(\frac{a}{2}\right)}} \le e^{1/6}e^{\lambda a}\left(1-2\frac{a}{a+1}\lambda\right)^{a/2}\left(\frac{a+1}{2}-\lambda a\right)^{-\lambda a},$$

since $\lambda \le \frac{1}{2}$. Consequently, the MGF is bounded by:

$$\begin{aligned}
&e^{1/6}\left(\frac{1}{1-2\lambda}\right)^{\frac{1}{2}a(1-2\lambda)}\frac{a^{\lambda a}}{2^{\lambda a}}\frac{1}{1-2\frac{a}{a-1}\lambda}\left(1-2\frac{a}{a+1}\lambda\right)^{a/2}\left(\frac{a+1}{2}-\lambda a\right)^{-\lambda a} \\
&= e^{1/6}\frac{1}{1-2\frac{a}{a-1}\lambda}\left[\left(\frac{1-2\lambda}{\frac{a+1}{a}-2\lambda}\right)^{2\lambda}\left(\frac{1-2\frac{a}{a+1}\lambda}{1-2\lambda}\right)\right]^{a/2} \\
&= e^{1/6}\frac{1}{1-2\frac{a}{a-1}\lambda}\left[\left(\frac{1}{1+\frac{1}{a(1-2\lambda)}}\right)^{2\lambda}\left(1+\frac{2\frac{1}{a+1}\lambda}{1-2\lambda}\right)\right]^{a/2} \\
&\le e^{1/6}\frac{1}{1-2\frac{a}{a-1}\lambda}\left[\left(\frac{1}{1+\frac{1}{a(1-2\lambda)}}\right)^{2\lambda}\left(1+\frac{1}{a(1-2\lambda)}\right)\right]^{a/2} \\
&= e^{1/6}\frac{1}{1-2\frac{a}{a-1}\lambda}\left[1+\frac{1}{a(1-2\lambda)}\right]^{a(1-2\lambda)/2} \\
&\le e^{1/6}e^{1/2}\frac{1}{1-2\frac{a}{a-1}\lambda} \le e\frac{1}{1-2\frac{a}{a-1}\lambda}.
\end{aligned}$$

This implies the following Chernoff bound

$$\mathbb{P}\left(\sum_{s=i}^{j}x_s^2 - a\log\left(\frac{\sum_{s=i}^{j}\left(x_s-\bar{x}_{i:j}\right)^2}{a}\right) - a > \frac{a}{a-1}\left(2+2\psi+2\sqrt{2\psi}\right)\right) \le ee^{-\psi}$$

and therefore, as $a \geq l \geq 2$

$$\mathbb{P}\left(\sum_{s=i}^{j} x_s^2 - a\log\left(\frac{\sum_{s=i}^{j}(x_s - \bar{x}_{i:j})^2}{a}\right) - a > 2\cdot 2 + 2\frac{a}{a-1}\psi' + 2\sqrt{2\cdot 2\frac{a}{a-1}\psi'}\right) \leq ee^{-\psi'}.$$

substituting $\psi' = 2\frac{a-1}{a}\psi$ yields

$$\mathbb{P}\left(\sum_{s=i}^{j} x_s^2 - a\log\left(\frac{\sum_{s=i}^{j}(x_s - \bar{x}_{i:j})^2}{a}\right) - a > 2\cdot 2 + 2\cdot 2\psi + 2\sqrt{2\cdot 2\psi}\right) \leq ee^{-2\frac{a-1}{a}\psi}.$$

Consequently, by a Bonferroni correction,

$$\mathbb{P}\left(\exists i,j : \sum_{s=i}^{j} x_s^2 - (j-i+1)\left(\log\left(\frac{\sum_{s=i}^{j}(x_s - \bar{x}_{i:j})^2}{j-i+1}\right) - 1\right) > 4 + 4\psi + 4\sqrt{\psi}\right)$$
$$\leq \sum_{a=l}^{n} nee^{-2\frac{a-1}{a}\psi} \leq ne\sum_{a=2}^{n} e^{-2\frac{a-1}{a}\psi}.$$

Now,

$$\sum_{a=2}^{n} e^{-2\frac{a-1}{a}\psi} = e^{-\psi} + \sum_{a=3}^{n} e^{-2\frac{a-1}{a}\psi} \leq e^{-\psi} + \int_{a=2}^{n} e^{-2\frac{a-1}{a}\psi} da$$
$$= e^{-\psi} + e^{-2\psi}\int_{a=2}^{n} e^{\frac{2}{a}\psi} da = e^{-\psi} + e^{-2\psi}\int_{x=\frac{1}{n}}^{\frac{1}{2}} \frac{1}{x^2} e^{2\psi x} dx.$$

Next note that

$$e^{-2\psi}\int_{x=\frac{1}{n}}^{\frac{1}{2\psi}} \frac{1}{x^2} e^{2\psi x} dx \leq ee^{-2\psi}\int_{x=\frac{1}{n}}^{\infty} \frac{1}{x^2} e^{2\psi x} dx = ene^{-2\psi},$$

which proves the result if $\psi \leq 1$. If $\psi > 1$, the proof can be obtained by noting that

$$e^{-2\psi}\int_{x=\frac{1}{2\psi}}^{\frac{1}{2}} \frac{1}{x^2} e^{2\psi x} dx \leq e^{-2\psi}\frac{1}{2}\max_{\frac{1}{2\psi}\leq x\leq\frac{1}{2}}\left(\frac{1}{x^2}e^{2\psi x}\right) \leq e^{-2\psi}\frac{1}{2}4e^{\psi} = 2e^{-\psi}.$$

A.2.3 Proof of Proposition 3

This proposition follows from the fact that CAPA will not fit x_i as a point anomaly if

$$\left(\frac{x-\hat{\mu}}{\hat{\sigma}}\right)^2 - 1 - \log\left(\gamma + \left(\frac{x-\hat{\mu}}{\hat{\sigma}}\right)^2\right) < \beta'$$

and not fit not x_i as typical if

$$\left(\frac{x-\hat{\mu}}{\hat{\sigma}}\right)^2 - 1 - \log\left(\gamma + \left(\frac{x-\hat{\mu}}{\hat{\sigma}}\right)^2\right) > \beta'.$$

It can be show by differentiation that the function $f(y) = y - 1 - \log(y+\gamma)$ is decreasing from 0 to $(y+\gamma)^{-1}$ and increasing thereafter. Since $f(0) < \beta'$, by the lower bound on γ and since $f(y) \to \infty$ as $y \to \infty$, there exists a constant $K(\beta')$ solving the equation $f(K(\beta')) = \beta'$ such that $f(y) < \beta'$ if $y < K(\beta')$ and $f(y) > \beta'$ if $y > K(\beta')$. Next note that

$$f(\beta') = \beta' - 1 - \log(\gamma + \beta') < \beta' - \log\left(e^{-\beta'} + \beta'\right) < \beta' - \log(1 - \beta' + \beta') = \beta'. \tag{A.2.1}$$

Moreover, we can show that

$$f(1 + \beta' + \sqrt{2(\beta'+\gamma)}) - \beta' = \sqrt{2(\beta'+\gamma)} - \log(1 + (\beta'+\gamma) + \sqrt{2(\beta'+\gamma)})$$

is equal to 0 when $z = \beta' + \gamma = 0$. Moreover the derivative of the above expression with respect to z is given by

$$\frac{z}{1 + z + \sqrt{2z}}$$

which is positive for all $z > 0$. Since $z = \beta' + \gamma > 0$ the following result also holds:

$$f(1 + \beta' + \sqrt{2(\beta'+\gamma)}) > \beta'. \tag{A.2.2}$$

This can be deduced from the fact that equality holds $\beta' = 0$ and comparing the derivatives. Equations (A.2.1) and (A.2.2) show that

$$\beta' < K(\beta') < \beta' + 1 + \sqrt{2(\beta'+\gamma)},$$

which finishes the proof.

A.2.4 Proof of Theorem 1

Before proving this theorem, we introduce some notation. We define the cost of a segment $x_{i:j}$ under the true partition $\{0, t_1, ..., t_K, n\}$ and true parameters to be

$$\mathcal{C}(x_{i:j}) = \sum_{t=i}^{j} \log(\sigma(t)^2) + \sum_{t=i}^{j} \eta_t^2.$$

Note that this cost is additive, i.e. for $a < b-1 < b+1 < c$ we have $\mathcal{C}(x_{a:c}) = \mathcal{C}(x_{a:b}) + \mathcal{C}(x_{(b+1):c})$, whilst the fitted cost satisfies the inequality $\tilde{\mathcal{C}}(x_{a:c}) \geq \tilde{\mathcal{C}}(x_{a:b}) + \tilde{\mathcal{C}}(x_{(b+1):c})$.

We also define the residual sum of squares $Y_{i:j} = \sum_{k=i}^{j} (\eta_k - \bar{\eta}_{i:j})^2$. Finally, we will work on the event sets E_1, E_2, E_3, E_4, E_5, and E_6 which we define below using notation $a := a(i,j) = j - i + 1$

$$E_1 := \left\{ a\bar{\eta}_{i:j}^2 < 4(1+\epsilon)\log(n), \quad 1 \leq i \leq j \leq n \right\},$$

$$E_2 := \left\{ Y_{i:j} \leq a - 1 + 2\sqrt{(a-1)(2+\epsilon)\log(n)} + (4+2\epsilon)\log(n), \quad 1 \leq i \leq j \leq n \right\},$$

$$E_3 := \left\{ Y_{i:j} \geq c(a,n)(a-1), \quad 1 \leq i < j \leq n \right\},$$

$$E_4 := \left\{ \frac{\sum_{t=t_k-1}^{t_k+1} (x_t - \bar{x}_{(t_k-1):(t_k+1)})^2}{\sigma_k^{2/3}\sigma_{k-1}^{4/3}} > n^{-\epsilon}, \frac{\sum_{t=t_k}^{t_k+2} (x_t - \bar{x}_{t_k:(t_k+2)})^2}{\sigma_k^{4/3}\sigma_{k-1}^{2/3}} > n^{-\epsilon}, \quad 1 \leq k \leq K-1 \right\},$$

$$E_5 := \left\{ \frac{(x_{t_k} - x_{t_k+1})^2}{\sigma_k \sigma_{k-1}} > n^{-\epsilon}, \quad 1 \leq k \leq K \right\},$$

$$E_6 := \left\{ Y_{i:j} \geq a - 1 - 2\sqrt{(a-1)(2+\epsilon)\log(n)}, \quad 1 \leq i \leq j \leq n \right\},$$

where $c(a,n) < 1$ satisfies

$$\frac{a}{2} \cdot \frac{c(a,n) - 1 - \log(c(a,n))}{2} = (2+\epsilon)\log(n).$$

Note that $c(a,n)$ is guaranteed to exist by the intermediate value theorem. Indeed, the function $f(x) = x - 1 - \log(x)$ is continuous and satisfies $f(1) = 0$ and $f(x) \to \infty$ as $x \to 0^+$. The motivation for these events is as follows: E_1 bounds the error in the estimates of the mean, while E_2, E_3, and E_6 bound the error in the estimates of the variance. E_5 and E_4 are needed to prevent the existence of segments length

two and three respectively in which the observations lie to close to each other, which would encourage the algorithm to erroneously fit them in a short segment of low variance. We also define E_7 which guarantees that the signal strength of true changes is utilised:

$$E_7 = \left\{ \sum_{t_k-D+1}^{t_k+D} \frac{(x_t - \bar{x}_{(t_k-D+1):(t_k+D)})^2}{2D\sigma_k\sigma_{k+1}c(D,n)\exp(\triangle_k)} \geq 1, \;\; 1 \leq D \leq \min(t_{k+1}-t_k, t_k-t_{k-1}), \;\; 1 \leq k \leq K \right\},$$

where $c(D,n) < 1$ satisfies $D\left(c(D,n) - 1 - \log(c(D,n))\right) = 2(2+\epsilon)\log(n)$. We write $E = \cap E_i$ and now in a position to prove the following lemmata:

Lemma 1. *(Yao 1988)* $\mathbb{P}(E_1) > 1 - \tilde{K}_1 n^{-\epsilon}$, *for some constant* $\tilde{K}_1$.

Lemma 2. $\mathbb{P}(E_2) > 1 - \tilde{K}_2 n^{-\epsilon}$, $\mathbb{P}(E_3) > 1 - \tilde{K}_3 n^{-\epsilon}$, $\mathbb{P}(E_4) > 1 - \tilde{K}_4 n^{-\epsilon}$, $\mathbb{P}(E_5) > 1 - \tilde{K}_5 n^{-\epsilon}$, $\mathbb{P}(E_6) > 1 - \tilde{K}_6 n^{-\epsilon}$, *and* $\mathbb{P}(E_7) > 1 - \tilde{K}_7 n^{-\epsilon}$*for some constants* $\tilde{K}_2$,$\tilde{K}_3$,$\tilde{K}_4$, $\tilde{K}_5$, $\tilde{K}_6$, *and* $\tilde{K}_7$.

Lemma 3. *There exists a constant* $\tilde{C}_1$ *such that* $Y_{i:j} - a - a\log(Y_{i:j}/a) \leq \tilde{C}_1 \log(n)$ *holds on* E *for all* $1 \leq i < j \leq n$.

Lemma 4. *Let* i, j *be such that there exists some* k *such that* $t_{k-1} < i < j \leq t_k$. *The following holds given* E *:*

$$0 \leq \mathcal{C}(x_{i:j}) - \tilde{\mathcal{C}}(x_{i:j}) \leq \tilde{C}_2 \log(n).$$

Lemma 5. *Let* i, j *be such that* $\exists k$ *such that* $t_{k-1} = i < j \leq t_k$ *or* $t_{k-1} < i < j = t_k{+}1$. *The following then holds given* E

$$\mathcal{C}(x_{i:j}) - \tilde{\mathcal{C}}(x_{i:j}) \leq \tilde{C}_3 \log(n)$$

Lemma 6. *Let $a, b, c \in \tau$ for some partition τ of $x_{i,j}$ such that $\exists k$ such that $t_{k-1} < a < b < c \leq t_k$. Then,*

$$\tilde{\mathcal{C}}\left(x_{i:j}, \tau, \alpha\right) - \tilde{\mathcal{C}}\left(x_{i:j}, \tau_{-b}, \alpha\right) \geq \frac{3}{4}\alpha \log(n)^{1+\delta},$$

where $\tau_{-b} = \tau \setminus \{b\}$ holds on E for large enough n.

Lemma 7. *For all $\alpha > 0$, there exists a constant $\tilde{\kappa}(\alpha, \epsilon)$ such that $\tilde{\mathcal{C}}\left(x_{i:j}\right) - (\mathcal{C}\left(x_{i:t_k}\right) + \mathcal{C}\left(x_{(t_k+1):j}\right)) \geq \alpha \log(n)^{1+\delta}$ holds on E if*

$$j - t_k = t_k + 1 - i \geq \frac{\tilde{\kappa}(\alpha, \epsilon)}{\min(\triangle_k, \triangle_k^2)} \log(n)^{1+\delta}$$

and $j \leq t_{k+1}, i > t_{k-1}$ for all $n > 2$.

We now define

$$\tilde{\kappa}_k = 2\frac{\tilde{\kappa}(3\alpha, \epsilon)}{\min(\triangle_k, \triangle_k^2)}$$

and the set of partitions

$$\mathcal{B} := \left\{ \{0, t_1', t_2', ..., t_K', n\} \quad | \quad |t_k' - t_k| \leq \tilde{\kappa}_k \log(n)^{1+\delta} \quad 1 \leq k \leq K \right\},$$

which are within $\tilde{\kappa}_k \log(n)^{1+\delta}$ of the true partition.

We will show that, for large enough n, the optimal partition lies in $\mathcal{B}$ given the event set E. Given the probability of E, this proves Theorem 1. Our approach will consist of showing that the cost of a partition $\tau \notin \mathcal{B}$ is higher than that of the true partition with the true parameters (see Proposition 12). We will achieve this by adding free changes to τ thus splitting up the series into multiple sub-segments each containing a single true changepoint and $\tilde{\kappa}_k \log(n)^{1+\delta}$ points either side of it. This

also defines a projection of τ onto the partitions of the sub-segments. We define the set of partitions

$$\mathcal{B}_k := \left\{ \{i-1, t'_k, j\} \quad | \quad |t'_k - t_k| \leq \tilde{\kappa}_k \log(n)^{1+\delta} \right\}$$

for segments $x_{i:j}$ for which there exist a k such that: $t_{k-1}+1 \leq i \leq t_k - \tilde{\kappa}_k \log(n)^{1+\delta} < t_k + \tilde{\kappa}_k \log(n)^{1+\delta} \leq j \leq t_{k+1}$ as an analogue of $\mathcal{B}$ for the whole of x.

If $\tau \notin \mathcal{B}$, there must be at least one sub-segment for which the projection of τ does not lie in $\mathcal{B}_k$. We will show in Proposition 11, that the cost of the true partition using the true parameters is at least $O(\log(n)^{1+\delta})$ lower than that of the projection of τ on such a segment. We will also show in Proposition 10 that the projections of τ which are in $\mathcal{B}_k$ have a cost which is at most $O(\log(n))$ lower than that of the true partition with true parameters.

Proposition 10. *Let $i, j \in N$, be such that there exists a k such that: $t_{k-1}+1 \leq i < t_k < j \leq t_{k+1}$, then there exists a constant $\tilde{C}_4$ such that given E,*

$$\left[\mathcal{C}\left(x_{i:j}\right) + \alpha \log(n)^{1+\delta}\right] - \tilde{\mathcal{C}}(x_{i:j}, \tau, \alpha) \leq \tilde{C}_4 \log(n)$$

for all valid partitions τ of the form $\tau = \{i-1, \hat{t}, j\}$, if n is large enough.

Proof of Proposition 10: The following cases are possible:

Case 1: $\hat{t} = t_k$. Then:

$$\left[\mathcal{C}\left(x_{i:j}\right) + \alpha \log(n)^{1+\delta}\right] - \tilde{\mathcal{C}}(x_{i:j}, \{i-1, \hat{t}, j\}, \alpha) = \mathcal{C}\left(x_{i:j}\right) - \left[\tilde{\mathcal{C}}(x_{i:t_k}) + \tilde{\mathcal{C}}(x_{(t_k+1):j})\right] \leq 2\tilde{C}_2 \log(n),$$

where the inequality follows from Lemma 4.

Case 2: $\hat{t} = t_k + 1$. Then:

$$\begin{aligned}&\left[\mathcal{C}\left(x_{i:j}\right) + \alpha\log(n)^{1+\delta}\right] - \tilde{\mathcal{C}}(x_{i:j}, \{i-1, \hat{t}, j\}, \alpha)\\ &= \mathcal{C}\left(x_{i:j}\right) - \left[\tilde{\mathcal{C}}(x_{i:(t_k+1)}) + \tilde{\mathcal{C}}(x_{(t_k+2):j})\right] \leq (\tilde{C}_2 + \tilde{C}_3)\log(n),\end{aligned}$$

where the inequality follows from Lemmata 4 and 5.

Case 3: $\hat{t} > t_k + 1$. Then:

$$\begin{aligned}&\left[\mathcal{C}\left(x_{i:j}\right) + \alpha\log(n)^{1+\delta}\right] - \tilde{\mathcal{C}}(x_{i:j}, \{i-1, \hat{t}, j\})\\ &\leq \mathcal{C}\left(x_{i:j}\right) + 2\alpha\log(n)^{1+\delta} - \tilde{\mathcal{C}}(x_{i:j}, \{i-1, t_k, \hat{t}, j\}, \alpha)\\ &= \mathcal{C}\left(x_{i:j}\right) - \left[\tilde{\mathcal{C}}(x_{i:t_k}) + \tilde{\mathcal{C}}(x_{(t_k+1):\hat{t}}) + \tilde{\mathcal{C}}(x_{(\hat{t}+1):j})\right] \leq 3\tilde{C}_2\log(n),\end{aligned}$$

where the first inequality follows from the fact that introducing an unpenalised change-point reduces cost and the second is a consequence of Lemma 4.

Case 4: $\hat{t} = t_k - 1$. Symmetrical to case 2.

Case 5: $\hat{t} < t_k - 1$. Symmetrical to case 3.

This finishes our proof.

Proposition 11. *There exists a constant $n_4(\alpha, \delta, \epsilon)$, such that $\forall i, j$ for which $\exists k$ such that $t_{k-1} + 1 \leq i \leq t_k - \tilde{\kappa}_k\log(n)^{1+\delta} < t_k + \tilde{\kappa}_k\log(n)^{1+\delta} \leq j \leq t_{k+1}$*

$$\tilde{\mathcal{C}}(x_{i:j}, \tau, \alpha) - \left[\mathcal{C}\left(x_{i:j}\right) + \alpha\log(n)^{1+\delta}\right] \geq \frac{1}{3}\alpha\log(n)^{1+\delta}$$

holds for all $\tau \notin \mathcal{B}_k$ given E and $n > n_4(\alpha, \delta, \epsilon)$.

Proof of Proposition 11: Consider $\tau' \notin \mathcal{B}_k$. We consider the following three cases and denote $H := \lceil \frac{1}{2}\tilde{\kappa}_k\log(n)^{1+\delta} \rceil$, noting that it is larger than

$$\frac{\tilde{\kappa}(3\alpha, \epsilon)}{\min(\triangle_k, \triangle_k^2)}\log(n)^{1+\delta}$$

Case 1: $|\tau'| = 2$. We have $\tau' = \{i-1, j\}$. Hence:

$$\begin{aligned}
&\tilde{\mathcal{C}}(x_{i:j}, \tau', \alpha) \geq \tilde{\mathcal{C}}(x_{i:(t_k-H)}) + \tilde{\mathcal{C}}(x_{(t_k-H+1):(t_k+H)}) + \tilde{\mathcal{C}}(x_{(t_k+H+1):j}) \\
\geq &\tilde{\mathcal{C}}(x_{i:(t_k-H)}) + \mathcal{C}(x_{(t_k-H+1):(t_k+H)}) + 3\alpha \log(n)^{1+\delta} + \tilde{\mathcal{C}}(x_{(t_k+H+1):j}) \\
\geq &2\alpha \log(n)^{1+\delta} - 2\tilde{C}_2 \log(n) + \left[\mathcal{C}\left(x_{i:j}\right) + \alpha \log(n)^{1+\delta}\right],
\end{aligned}$$

where the second inequality follows from the definition of H and Lemma 7 and the third from Lemma 4.

Case 2: $|\tau'| = 3$. We have $\tau' = \{i-1, t_k + L, j\}$, where $|L| > \tilde{\kappa}_k \log(n)^{1+\delta}$. We assume $L > 0$, the other case being very similar. We have:

$$\begin{aligned}
&\tilde{\mathcal{C}}(x_{i:j}, \{i-1, t_k+L, j\}, \alpha) = \tilde{\mathcal{C}}(x_{i:(t_k+L)}) + \tilde{\mathcal{C}}(x_{(t_k+L+1):j}) + \alpha \log(n)^{1+\delta} \\
\geq &\tilde{\mathcal{C}}(x_{i:(t_k-H-1)}) + \tilde{\mathcal{C}}(x_{(t_k-H):(t_k+H)}) + \tilde{\mathcal{C}}(x_{(t_k+H+1):(t_k+L)}) - \tilde{C}_2 \log(n) + \mathcal{C}(x_{(t_k+L+1):j}) + \alpha \log(n)^{1+\delta} \\
\geq &3\alpha \log(n)^{1+\delta} - 3\tilde{C}_2 \log(n) + \left[\mathcal{C}\left(x_{i:j}\right) + \alpha \log(n)^{1+\delta}\right],
\end{aligned}$$

where the inequalities follow from of the definition of H as well as Lemmata 7 and 4.

Case 3: $|\tau'| \geq 4$. Let $\tau' = \{a_1, a_2, ..., a_{|\tau'|}\}$, where $a_1 = i-1$ and $a_{|\tau'|} = j$. There must exist a $l \in \{2, ..., |\tau'|-1\}$, such that $a_{l-1} < t_k$ and $a_{l+1} > t_k + 1$. We thus have:

$$\begin{aligned}
&\tilde{\mathcal{C}}(x_{i:j}, \tau', \alpha) \\
&= (|\tau'|-3)\alpha \log(n)^{1+\delta} + \left(\sum_{m=1}^{l-2} + \sum_{m=l+1}^{|\tau'|-1}\right)\left[\tilde{\mathcal{C}}(x_{a_m+1, a_{m+1}})\right] + \tilde{\mathcal{C}}(x_{(a_{l-1}+1):a_{l+1}}, \{a_{l-1}, a_l, a_{l+1}\}, \alpha) \\
&\geq (|\tau'|-2)\alpha \log(n)^{1+\delta} + \left(\sum_{m=1}^{l-2} + \sum_{m=l+1}^{|\tau'|-1}\right)\left[\mathcal{C}(x_{a_m+1, a_{m+1}})\right] + \mathcal{C}(x_{(a_{l-1}+1):a_{l+1}}) - ((|\tau'|-3)\tilde{C}_2 + \tilde{C}_4)\log(n) \\
&= \mathcal{C}(x_{i:j}, \tau, \alpha) + \alpha \log(n)^{1+\delta} + (|\tau'|-3)\alpha \log(n)^{1+\delta} - \left[(|\tau'|-3)\tilde{C}_2 + \tilde{C}_4\right]\log(n),
\end{aligned}$$

by Lemma 4 and Proposition 10. This finishes the proof.

Proposition 12. *There exists a constant $\tilde{n}_5(\alpha, \delta, \tilde{\delta}, \epsilon)$ such that given E, we have*

$$\tilde{\mathcal{C}}(x_{1:n}, \tau, \alpha) - \left[\mathcal{C}(x_{1,n}) + K\alpha \log(n)^{1+\delta}\right] \geq \frac{1}{4}\alpha \log(n)^{1+\delta}$$

for all $\tau \notin \mathcal{B}$ *if* $n \geq \tilde{n}_5(\alpha, \delta, \tilde{\delta}, \epsilon)$.

Proof of Proposition 12: First, consider the special case $K = 0$. For this case, $\tau \notin \mathcal{B}$ implies that $\hat{K} \geq 1$. We have

$$\tilde{\mathcal{C}}(x_{1:n}, \tau, \alpha) \geq \tilde{\mathcal{C}}(x_{1:n}, \{0, n\}, \alpha) + \hat{K}\frac{3}{4}\alpha \log(n)^{1+\delta} \geq \mathcal{C}(x_{1:n}) + \hat{K}\frac{3}{4}\alpha \log(n)^{1+\delta} - \tilde{C}_2 \log(n),$$

where the first inequality follows from Lemma 6 and the second from Lemma 4.

Next assume $K \geq 1$. Let $\tau \notin \mathcal{B}$. We now introduce free changepoints $l_0, l_1, ..., l_K$ to break up the series into multiple sub-series with one true changepoint each. We impose $l_0 = 0$, $l_K = n$, $|l_k - t_k| > 4\tilde{\kappa}_k \log(n)^{1+\delta}$ for $0 < k \leq K$ and $|l_k - t_{k+1}| > 4\tilde{\kappa}_k \log(n)^{1+\delta}$ for $0 \leq k < K$. We also require that $\tau \cup \{l_0, ..., l_K\}$ is a valid partition (i.e. one which has segments of length at least two) and that there exists a $\hat{k}$ such that $\tau_{\hat{k}} := \tau \cap \{l_{\hat{k}-1} + 1, l_{\hat{k}-1} + 2, ..., l_{\hat{k}}\} \notin \mathcal{B}_{\hat{k}}$. We are guaranteed to find such points $l_0, l_1, ..., l_K$ if n is such that

$$\frac{1}{\min(\triangle_k, \triangle_k^2)} \log(n)^{1+\delta+\tilde{\delta}} \geq 12\tilde{\kappa}_k \log(n)^{1+\delta},$$

which is satisfied if $n > \tilde{n}_5(\alpha, \tilde{\delta}, \epsilon)$. Indeed, we can choose points near the middle of the true segments which are not in τ, or by select points in τ if the former is impossible because there are too many point in τ near the middle of some segment.

Since introducing free changes reduces the cost we then have

$$\begin{aligned}\tilde{\mathcal{C}}(x_{1:n}, \tau, \alpha) &\geq \sum_{k=1}^{K} \tilde{\mathcal{C}}(x_{(l_{k-1}+1):l_k}, \tau_k, \alpha) = \tilde{\mathcal{C}}(x_{(l_{\hat{k}-1}+1):l_{\hat{k}}}, \tau_{\hat{k}}, \alpha) + \sum_{k \neq \hat{k}} \tilde{\mathcal{C}}(x_{(l_{k-1}+1):l_k}, \tau_k, \alpha) \\ &\geq \mathcal{C}(x_{1:n}, \tau, \alpha) + \frac{1}{3}\alpha \log(n)^{1+\delta} - (K-1)\tilde{C}_4 \log(n),\end{aligned}$$

where the second inequality follows from Propositions 10 and 11. This finishes the proof.

Proof of Theorem 1: $\mathcal{B}$ contains the true partition with fitted parameters which is cheaper than the true partition with true parameters. Proposition 12 shows that conditional on E the true partition with true parameters will be cheaper than all $\tau \notin \mathcal{B}$ fo $n > \tilde{n}_5(\alpha, \delta, \tilde{\delta}, \epsilon)$. The optimal partition must therefore be in $\mathcal{B}$, given event set E. This proves Theorem 1, since Lemmata 1 and 2 imply that $\mathbb{P}(E) \geq 1 - (\tilde{K}_1 + \tilde{K}_2 + \tilde{K}_3 + \tilde{K}_4 + \tilde{K}_5 + \tilde{K}_6 + \tilde{K}_7)n^{-\epsilon}$.

A.2.5 Proof of Theorem 2

In order to prove this result, we will use the following notation in this section: We define $\tilde{\mathcal{C}}_E\left(x_{1:n}, \tau_E, \alpha, \mu, \sigma\right)$ to be the cost of an epidemic partition $\tau_E = \{\hat{s}_1, \hat{e}_1, ...\hat{s}_{\hat{K}}, \hat{e}_{\hat{K}}\}$ under a penalty $\alpha \log(n)^{1+\delta}$ and inferred parameters of the typical distribution μ, σ. We define $\mathcal{C}_E\left(x_{1:n}, \alpha, \mu, \sigma\right)$, to be the cost under the true partition using the true parameters for the epidemic segments and μ, σ as estimates for the parameters of the typical distribution. We also define the set of epidemic partitions

$$\mathcal{B}_E = \left\{ \{\hat{s}_1, \hat{e}_1, ..., \hat{s}_K, \hat{e}_K\} \;\mid\; |\hat{e}_k - e_k| < \tilde{\kappa}_k \log(n)^{1+\delta}, |\hat{s}_k - s_k| < \tilde{\kappa}_k \log(n)^{1+\delta}, \quad 1 \leq k \leq K \right\}$$

as an epidemic equivalent of $\mathcal{B}$. Finally, we note that we can extend the definition of the event set E to epidemic changepoints by treating the s_k and e_k like classical changepoints.

We will begin by proving a simplified version of the theorem in which we run our epidemic changepoint detection algorithm without allowing for epidemic changes of length one in variance only and imposing that each segment of the data allocated to the typical distribution is of length at least two. The reason for this is that this allows us to define an equivalent non-epidemic partition, whose segments must be of

length at least two, for each epidemic partition. We also begin by assuming that the parameter of the typical distribution is known.

This simplified version captures the main ideas of the full proof. We will proceed to showing that the result also holds when the typical mean and variance are unknown. This will be followed by a proof of the full result by means of introducing and proving the consistency of a modified version of the classical changepoint detection algorithm described in the previous section which also allows for segments of length one.

For now, we assume that all segments are of length at least two and that the true parameters μ_0 and σ_0 are known. This allows us to use the fact that the cost of the true epidemic partition using the true parameters is exactly the same as the cost of the corresponding true non-epidemic partition using the true parameters with twice the penalty. We can therefore prove the following proposition, as a corollary of Proposition 12.

Proposition 13. *There exists a constant* $\tilde{n}_6(\alpha, \delta, \tilde{\delta}, \epsilon)$*, such that for all* $\tau'_E \notin \mathcal{B}_E$

$$\tilde{\mathcal{C}}_E\left(x_{1:n}, \tau'_E, \alpha, \mu_0, \sigma_0\right) - \mathcal{C}_E\left(x_{1:n}, \alpha, \mu_0, \sigma_0\right) \geq \frac{1}{5}\alpha \log(n)^{1+\delta/2}$$

holds on E *for* $n > \tilde{n}_6(\alpha, \delta, \tilde{\delta}, \epsilon)$*.*

Proof of Proposition 13: We note that

$$\tilde{\mathcal{C}}_E\left(x_{1:n}, \tau'_E, \alpha, \mu_0, \sigma_0\right) \geq \tilde{\mathcal{C}}\left(x_{1:n}, \tau'_E \cup \{0, n\}, \frac{1}{2}\alpha\right) + \frac{\alpha}{2}\sum_{k=2}^{\hat{K}} \mathbb{I}\{s_k = e_{k-1}\} \log(n)^{1+\delta},$$

because using fitted parameters instead of μ_0 and σ_0 for segments allocated to the typical distribution under τ'_E can only reduce the cost. Additionally, two epidemic changes correspond to three classical changepoints if their end and starting points

coincide. Moreover,

$$\mathcal{C}_E\left(x_{1:n}, \alpha, \mu_0, \sigma_0\right) = \mathcal{C}\left(x_{1:n}\right) + K\alpha\log(n)^{1+\delta}.$$

Therefore:

$$\tilde{\mathcal{C}}_E\left(x_{1:n}, \tau_E', \alpha, \mu_0, \sigma_0\right) - \mathcal{C}_E\left(x_{1:n}, \alpha, \mu_0, \sigma_0\right)$$
$$\geq \tilde{\mathcal{C}}\left(x_{1:n}, \tau_E' \cup \{0, n\}, \frac{1}{2}\alpha\right) + \frac{\alpha}{2}\sum_{k=2}^{\hat{K}} \mathbb{I}\{s_k = e_{k-1}\}\log(n)^{1+\delta} - \left[\mathcal{C}\left(x_{1:n}\right) + 2K\frac{\alpha}{2}\log(n)\right].$$

This leaves two possibilities. If $\tau_E' \cup \{0, n\} \notin \mathcal{B}$ then the above will exceed

$$\frac{1}{4}\alpha\log(n)^{1+\delta},$$

by proposition 12. Since $\tau_E' \notin \mathcal{B}_E$, the only way we can have $\tau_E' \cup \{0, n\} \in \mathcal{B}$ is if there exists a k such that $s_k = e_{k-1}$. In that case the difference will exceed

$$\frac{1}{2}\alpha\log(n)^{1+\delta} - (2K+1)\tilde{C}_4\log(n),$$

by Proposition 10. This finishes the proof.

We can now use this proposition to prove Theorem 2 in the same way we used 12 to prove Theorem 1.

Proof of Theorem 2: Proposition 13 proves Theorem 2 as Lemmata 1 and 2 imply that $\mathbb{P}(E) \geq 1 - (\tilde{K}_1 + \tilde{K}_2 + \tilde{K}_3 + \tilde{K}_4 + \tilde{K}_5 + \tilde{K}_6 + \tilde{K}_7)n^{-\epsilon}$.

We now introduce the following lemma about the distribution of the median and inter-quantile range. It will allow us to prove Theorem 2 when the true parameters are unknown.

Lemma 8. *There exists a constants $\tilde{K}_8$, D_1, and D_2 such that for large enough n*

$$\mathbb{P}\left(|\hat{\mu}-\mu_0| \leq D_1\sigma_0\sqrt{\frac{\log(n)}{n}}, \left|\frac{\hat{\sigma}^2}{\sigma_0^2}-1\right| \leq D_2\sqrt{\frac{\log(n)}{n}}\right) \geq 1-\tilde{K}_8 n^{-\epsilon}$$

We can use this Lemma above to introduce a new event E_8 stating that the estimated parameters are close to the true parameters.

$$E_8 := \left\{|\hat{\mu}-\mu_0| \leq D_1\sigma_0\sqrt{\frac{\log(n)}{n}}, \left|\frac{\hat{\sigma}^2}{\sigma_0^2}-1\right| \leq D_2\sqrt{\frac{\log(n)}{n}}\right\}.$$

This event bounds the effect of using the estimated typical parameters instead of the true parameters for the cost of the true distribution with true non-typical parameters. Indeed, the following lemma holds:

Lemma 9. *There exists a constant $\tilde{C}_7$ such that given E and E_8 and n large enough we have:*

$$\tilde{\mathcal{C}}_E\left(x_{1:n}, \alpha, \hat{\mu}, \hat{\sigma}\right) - \mathcal{C}_E\left(x_{1:n}, \alpha, \mu_0, \sigma_0\right) \leq \tilde{C}_7 \log(n).$$

We can use this lemma to prove the following extension of Proposition 13 to the case when the typical parameters are inferred.

Proposition 14. *There exists a constant $\tilde{n}_7(\alpha, \delta, \tilde{\delta}, \epsilon)$ such that for all $\tau'_E \notin \mathcal{B}_E$*

$$\tilde{\mathcal{C}}_E\left(x_{1:n}, \tau'_E, \alpha, \hat{\mu}, \hat{\sigma}\right) - \mathcal{C}_E\left(x_{1:n}, \alpha, \hat{\mu}, \hat{\sigma}\right) \geq \frac{1}{5}\alpha \log(n)^{1+\delta/2}$$

holds on $E \cap E_8$ for $n > \tilde{n}_7(\alpha, \delta, \tilde{\delta}, \epsilon)$.

Proof of Proposition 14: We note that, as before,

$$\tilde{\mathcal{C}}_E\left(x_{1:n}, \tau'_E, \alpha, \hat{\mu}, \hat{\sigma}\right) \geq \tilde{\mathcal{C}}\left(x_{1:n}, \tau'_E \cup \{0, n\}, \frac{1}{2}\alpha\right) + \frac{\alpha}{2}\sum_{k=2}^{\hat{K}} \mathbb{I}\{s_k = e_{k-1}\} \log(n)^{1+\delta}$$

$$\mathcal{C}_E\left(x_{1:n}, \alpha, \mu, \sigma\right) = \mathcal{C}\left(x_{1:n}\right) + K\alpha \log(n)^{1+\delta},$$

Therefore we now have

$$\tilde{\mathcal{C}}_E\left(x_{1:n}, \tau'_E, \alpha, \hat{\mu}, \hat{\sigma}\right) - \mathcal{C}_E\left(x_{1:n}, \alpha, \hat{\mu}, \hat{\sigma}\right)$$
$$\geq \tilde{\mathcal{C}}\left(x_{1:n}, \tau'_E \cup \{0, n\}, \frac{1}{2}\alpha\right) + \frac{\alpha}{2}\sum_{k=2}^{\hat{K}} \mathbb{I}\{s_k = e_{k-1}\}\log(n)^{1+\delta} - \left[\mathcal{C}\left(x_{1:n}\right) + 2K\frac{\alpha}{2}\log(n)^{1+\delta}\right] - \tilde{C}_7\log(n),$$

by applying Lemma 9. The rest of the proof is identical to that of Proposition 13, with an added $O(\log(n))$ term.

In order to be able to extend Proposition 14 to the case in which we allow epidemic changes of length one in variance only, as well as segments of the typical distribution which are of length one, we will prove the consistency of the following adaptation of the algorithm detecting classical changepoints we introduced in the previous section. We now let the segment costs be

$$\tilde{\mathcal{C}}(x_{i:j}) = \tilde{\mathcal{C}}(x_{i:j}, \{i-1, j\}) = \begin{cases} (\hat{t}_{k+1} - \hat{t}_k)\left(\log\left(\frac{\sum_{\hat{t}_k+1}^{\hat{t}_{k+1}}(x_t - \bar{x}_{(\hat{t}_k+1):\hat{t}_{k+1}})^2}{(\hat{t}_{k+1} - \hat{t}_k)}\right) + 1\right) & i < j, \\ \min\left\{\log(\tilde{\sigma}^2) + \frac{(x_i - \tilde{\mu})^2}{\tilde{\sigma}^2}, 1 + \log(\gamma\tilde{\sigma}^2 + (x_t - \tilde{\mu})^2)\right\} & i = j, \end{cases}$$

where $|\tilde{\mu} - \mu_{k'}| \leq D_1\sigma_{k'}\sqrt{\frac{\log(n)}{n}}$ and $|\frac{\tilde{\sigma}^2}{\sigma^2_{k'}} - 1| < D_2\sqrt{\frac{\log(n)}{n}}$ for k' either $k-1$,k, or $k+1$, when i belongs to the kth segment. Given E_8 the range of allowed $\tilde{\sigma}^2$ and $\tilde{\mu}$ is therefore guaranteed to contain the estimated typical parameters $\hat{\sigma}^2$ and $\hat{\mu}$ when applied to x. The algorithm can obviously not be implemented in practice, as it requires knowledge of the true parameters. It is nevertheless a consistent method.

To prove this, we need to define a last event set E_9 which controls the newly introduced segments of length one:

$$E_9 := \left\{|x_t - \mu_{k+1}| \geq \sigma_k n^{-2+\epsilon}, \quad |x_t - \mu_{k-1}| \geq \sigma_k n^{-2+\epsilon}, \quad 1 \leq t \leq n\right\},$$

We can prove the following probability bounds

Lemma 10. *There exists a constant* $\tilde{K}_9$ *such that* $\mathbb{P}(E_9) \geq 1 - \tilde{K}_9 n^{-\epsilon}$

We can now prove the following proposition, adapted from Proposition 12 for this modified penalised cost approach:

Proposition 15. *There exists a constant* $\tilde{n}_7(\alpha, \delta, \tilde{\delta}, \epsilon)$ *such that given* $E \cap E_9$*, we have*

$$\tilde{\mathcal{C}}(x_{1:n}, \tau, \alpha) - \left[\mathcal{C}(x_{1,n}) + K\alpha \log(n)^{1+\delta}\right] \geq \frac{1}{5}\alpha \log(n)^{1+\delta}$$

for all $\tau \notin B$ *if* $n \geq \tilde{n}_7(\alpha, \delta, \tilde{\delta}, \epsilon)$

Proof of Proposition 15: Identical to the proof of Proposition 12. We just need to replace Lemma 4 by

Lemma 11. *There exists a constant* $\tilde{C}_2'$ *such that if* i, j *are such that there exists some* k *such that* $t_{k-1} < i \leq j \leq t_k$*, then given* $E \cap E_8$ *and* n *large enough*

$$\mathcal{C}\left(x_{i:j}\right) - \tilde{\mathcal{C}}\left(x_{i:j}\right) \leq \tilde{C}_2' \log(n).$$

to also account for the newly added segments of length one. We can now prove that

Proposition 16. *There exists a constant* $\tilde{n}_8(\alpha, \delta, \tilde{\delta}, \epsilon)$ *such that for all* $\tau_E' \notin \mathcal{B}_E$

$$\tilde{\mathcal{C}}_E\left(x_{1:n}, \tau_E', \alpha, \hat{\mu}, \hat{\sigma}\right) - \mathcal{C}_E\left(x_{1:n}, \alpha, \hat{\mu}, \hat{\sigma}\right) \geq \frac{1}{5}\alpha \log(n)^{1+\delta/2}$$

holds on $E \cap E_8 \cap E_9$ *for* $n > \tilde{n}_8(\alpha, \delta, \tilde{\delta}, \epsilon)$.

holds even when we allow for epidemic changes of length one in variance only and do not impose that segments allocated to the typical distribution have to be of length at least two.

Proof of Proposition 16: Identical to the proof of Proposition 14 using Proposition 15 instead of Proposition 12.

Proof of Theorem 2: Proposition 16 proves Theorem 2 since Lemmata 1, 2, 8, and 10 show that $\mathbb{P}(E \cap E_8 \cap E_9) \geq 1-(\tilde{K}_1+\tilde{K}_2+\tilde{K}_3+\tilde{K}_4+\tilde{K}_5+\tilde{K}_6+\tilde{K}_7+\tilde{K}_8+\tilde{K}_9)n^{-\epsilon}$.

A.3 Additional Lemmata

Lemma 12. *Let weights* $W_1, ..., W_p > 0$ *and* $A_1, ..., A_p > 0$. *Define* $\mu = \sum_1^p W_i A_i$. *Then*

$$\min_{\lambda<0} \left[\left(\prod_{i=1}^{p} \left(\frac{1}{1-A_i\lambda} \right)^{W_i} \right) e^{-\lambda c \mu} \right] \leq \exp\left(\left(\sum_{i=1}^{p} W_i \right) (\log(c) + 1 - c) \right)$$

holds for $0 < c < 1$

Proof of Lemma 12: This Lemma proves a multiplicative Chernoff lower bound for a weighted sum of chi-squared random variables. We define

$$Z(\lambda) = \left(\prod_{i=1}^{p} \left(\frac{1}{1-A_i\lambda} \right)^{W_i} \right) e^{-\lambda c \mu}$$

and note that it has derivative

$$\frac{d}{d\lambda}\left(Z(\lambda)\right) = \left(\sum_{i=1}^{p} \frac{A_i W_i}{1-A_i\lambda} - c\mu \right) Z(\lambda).$$

The minimise λ^* of $Z(\lambda)$ thus satisfies

$$\sum_{i=1}^{p} \frac{A_i W_i}{1-A_i\lambda^*} = c\mu.$$

Note that the LHS is strictly increasing from 0 to μ as λ^* increases from $-\infty$ to 0.

Consequently λ^* is well defined and unique. Moreover,

$$\begin{aligned} c\mu\lambda^* &= \sum_{i=1}^{p} \frac{A_i W_i \lambda^*}{1 - A_i\lambda^*} = -\sum_{i=1}^{p} W_i + \sum_{i=1}^{p} \frac{W_i}{1 - A_i\lambda^*} \\ &\geq -\sum_{i=1}^{p} W_i + \frac{\left(\sum_{i=1}^{p} W_i\right)^2}{\sum_{i=1}^{p} \left(W_i - W_i A_i \lambda^*\right)} = -\sum_{i=1}^{p} W_i + \frac{\left(\sum_{i=1}^{p} W_i\right)^2}{\sum_{i=1}^{p} W_i - \mu\lambda^*} \\ &= \left(\sum_{i=1}^{p} W_i\right) \frac{\mu\lambda^*}{\sum_{i=1}^{p} W_i - \mu\lambda^*}, \end{aligned}$$

with the inequality following from the fact that the arithmetic mean exceeds the harmonic mean (a special case of Jensen's inequality). This can be recomposed to yield

$$\lambda^* c\mu \leq (c-1) \sum_{i=1}^{p} W_i. \tag{A.3.1}$$

We can now use these results to bound $Z(\lambda^*)$ by noting

$$\begin{aligned} Z(\lambda^*) &= \left(\prod_{i=1}^{p} \left(\frac{1}{1 - A_i\lambda^*}\right)^{W_i}\right) e^{-\lambda * c\mu} \leq \left(\frac{1}{\sum_{i=1}^{p} W_i} \sum_{i=1}^{p} \frac{W_i}{1 - A_i\lambda^*}\right)^{\sum_{i=1}^{p} W_i} e^{-\lambda^* c\mu} \\ &= \left(1 + \frac{1}{\sum_{i=1}^{p} W_i} \sum_{i=1}^{p} \frac{W_i A_i \lambda^*}{1 - A_i\lambda^*}\right)^{\sum_{i=1}^{p} W_i} e^{-\lambda^* c\mu} = \left[\left(1 + \frac{\lambda^* c\mu}{\sum_{i=1}^{p} W_i}\right) e^{-\frac{\lambda^* c\mu}{\sum_{i=1}^{p} W_i}}\right]^{\sum_{i=1}^{p} W_i}, \end{aligned}$$

where the first inequality follows form The AMGM inequality. It can be shown by differentiation that the above bound is increasing in λ^*. Consequently, using (A.3.1) we have that

$$Z(\lambda^*) \leq \exp\left(\left(\sum_{i=1}^{p} W_i\right)(\log(c) + 1 - c)\right)$$

Lemma 13. *Let $c < 1$ solve the equation $c - 1 - \log(c) = t$ for some $t > 0$. Then $c - 1 \geq -\sqrt{2t}$*

Proof of Lemma 13: This Lemma helps bound $c(D, n)$ from the event set E_7.

$$t = c - 1 - \log(c) = c - 1 - \log(1 - (1 - c)) \geq c - 1 + (1 - c) + \frac{1}{2}(1 - c)^2 = \frac{1}{2}(1 - c)^2$$

which implies that $c - 1 \geq -\sqrt{2t}$. This finishes the proof.

A.4 Proofs of Main Lemmata

A.4.1 Proof of Lemma 1

See Yao (1988).

A.4.2 Proof of Lemma 2:

We note that $Y_{i:j} \sim \chi^2_{a-1}$. Laurent and Massart (2000) proved that

$$\mathbb{P}\left(-2\sqrt{kx} \leq \chi^2_k - k \leq 2\sqrt{kx} + 2x\right) \geq 1 - 2e^{-x}.$$

Therefore:

$$\mathbb{P}\left(-2\sqrt{(a-1)(2+\epsilon)\log(n)} \leq Y_{i:j} - (a-1) \leq 2\sqrt{(a-1)(2+\epsilon)\log(n)} + 2(2+\epsilon)\log(n)\right) \geq 1 - 2n^{-(2+\epsilon)}.$$

A Bonferroni correction therefore gives $\mathbb{P}(E_2 \cap E_6) > 1 - 2n^{-\epsilon}$.

We can derive the following Chernoff bound for $k \geq 1$ and $0 \leq \tilde{c} < 1$:

$$\begin{aligned}\mathbb{P}\left(\chi^2_k \leq k\tilde{c}\right) &= \mathbb{P}\left(\exp\left[\theta(\chi^2_k - k\tilde{c}\right)\right] \geq 1\right) \leq \mathbb{E}\left(\exp\left[\theta(\chi^2_k - k\tilde{c}\right)\right]\right) \\ &= e^{-k\tilde{c}\theta}\mathbb{E}\left(e^{\theta\chi^2_k}\right) = e^{-k\tilde{c}\theta}\left(\frac{1}{1-2\theta}\right)^{k/2},\end{aligned}$$

holds for all $\theta < 0$. Setting $\theta = \frac{1}{2}(1 - \frac{1}{\tilde{c}})$ we thus get

$$\mathbb{P}\left(\chi^2_k \leq k\tilde{c}\right) \leq \exp\left(-\frac{k}{2}\left(\tilde{c} - 1 - \log(\tilde{c})\right)\right).$$

Thus if we let $c(a,n) < 1$ be such that

$$\frac{a}{2} \cdot \frac{c(a,n) - 1 - \log(c(a,n))}{2} = (2+\epsilon)\log(n),$$

and write $c := c(a, n)$ for simplicity, we have

$$\mathbb{P}\left(Y_{i:j} \leq c(a-1)\right) \leq \exp\left(-\frac{a-1}{2}\left(c - 1 - \log(c)\right)\right) \leq \exp\left(-\frac{a}{4}\left(c - 1 - \log(c)\right)\right) = n^{-(2+\epsilon)},$$

for $a \geq 2$. A Bonferroni correction then gives $\mathbb{P}(E_3) > 1 - n^{-\epsilon}$.

Next we note that

$$\frac{\sigma_k \eta_{t_k+1} + \mu_{k+1} - \mu_k - \sigma_k \eta_{t_k}}{\sqrt{\sigma_{k+1}\sigma_k}} \sim N\left(\frac{\mu_{k+1} - \mu_k}{\sqrt{\sigma_{k+1}\sigma_k}}, \frac{\sigma_{k+1}^2 + \sigma_k^2}{\sigma_{k+1}\sigma_k}\right).$$

Consequently, we have that:

$$\mathbb{P}\left(\frac{|\sigma_{k+1}\eta_{t_k+1} + \mu_{k+1} - \mu_k - \sigma_k \eta_{t_k}|}{\sqrt{\sigma_{k+1}\sigma_k}} \leq n^{-\epsilon}\right) \leq \sqrt{\frac{2\sigma_{k+1}\sigma_k}{\pi(\sigma_{k+1}^2 + \sigma_{k-1}^2)}} n^{-\epsilon} \leq \sqrt{\frac{1}{\pi}} n^{-\epsilon}.$$

A Bonferroni correction then gives $\mathbb{P}(E_5) > 1 - K/\sqrt{\pi} n^{-\epsilon}$.

Finally, we have:

$$x_t - \bar{x}_{t_k:(t_k+2)} \sim \begin{cases} N\left(\frac{2\mu_k - 2\mu_{k+1}}{3}, \frac{4\sigma_k^2 + 2\sigma_{k+1}^2}{9}\right) & t = t_k, \\ N\left(\frac{\mu_{k+1} - \mu_k}{3}, \frac{2\sigma_k^2 + 5\sigma_{k+1}^2}{9}\right) & t = t_k + 1, t_k + 2, \end{cases}$$

which means that

$$\begin{aligned} (x_t - \bar{x}_{t_k:(t_k+2)})^2 &\geq \frac{4\sigma_k^2 + 2\sigma_{k+1}^2}{9} n^{-\epsilon}, \quad t = t_k, \\ (x_t - \bar{x}_{t_k:(t_k+2)})^2 &\geq \frac{1\sigma_k^2 + 5\sigma_{k+1}^2}{9} n^{-\epsilon}, \quad t = t_k + 1, t_k + 2, \end{aligned}$$

holds with probability exceeding $1 - 3n^{-\epsilon}$. Adding up the three inequalities then gives

$$\sum_{t=t_k}^{t_k+2} (x_t - \bar{x}_{t_k:(t_k+2)})^2 \geq \frac{12\sigma_{k+1}^2 + 6\sigma_k^2}{9} n^{-\epsilon} \geq 2\sigma_{k+1}^{4/3} \sigma_k^{2/3} n^{-\epsilon}.$$

By a similar argument,

$$\sum_{t=t_k-1}^{t_k+1} (x_t - \bar{x}_{(t_k-1):(t_k+1)})^2 \geq 2\sigma_{k+1}^{2/3} \sigma_k^{4/3} n^{-\epsilon}$$

must also hold with probability $1-3n^{-\epsilon}$. A Bonferroni correction then gives $\mathbb{P}(E_4) \geq 1-6K_4n^{-\epsilon}$.

Next, note that the MGF of

$$\sum_{t_k-D+1}^{t_k+D}\left(x_t-\bar{x}_{(t_k-D+1):(t_k+D)}\right)^2$$
$$=\sigma_k^2 Y_{(t_k-D+1):t_k}+\sigma_{k+1}^2 Y_{(t_k+1):(t_k+D)}+\frac{D}{2}\left(\mu_k+\sigma_k\bar{\eta}_{(t_k-D+1):t_k}-\mu_{k+1}-\sigma_{k+1}\bar{\eta}_{(t_k+1):(t_k+D)}\right)^2$$

is given by

$$\left(\frac{1}{1-2\sigma_k^2\lambda}\right)^{\frac{D-1}{2}}\left(\frac{1}{1-2\sigma_{k+1}^2\lambda}\right)^{\frac{D-1}{2}}\left(\frac{1}{1-(\sigma_k^2+\sigma_{k+1}^2)\lambda}\right)^{\frac{1}{2}}\exp\left(\frac{\frac{D}{2}(\mu_{k+1}-\mu_k)^2\lambda}{1-(\sigma_k^2+\sigma_{k+1}^2)\lambda}\right)$$

since $Y_{(t_k-D+1):t_k}$ and $Y_{(t_k+1):(t_k+D)}$ are independently χ^2_{D-1} distributed. Moreover, $\bar{\eta}_{(t_k-D+1):t_k}$ and $\bar{\eta}_{(t_k+1):(t_k+D)}$ are normally distributed and independent of the chi-squared random variables. The MGF is therefore the product of two chi-squared random variables and that of a noncentral chi-squared. Noting that

$$e^x=\sqrt{\frac{1}{e^{-2x}}}\leq\sqrt{\frac{1}{1-2x}}$$

holds for all $x<1/2$ shows that the MGF can be bounded by

$$\left(\frac{1}{1-2\sigma_k^2\lambda}\right)^{\frac{D-1}{2}}\left(\frac{1}{1-2\sigma_{k+1}^2\lambda}\right)^{\frac{D-1}{2}}\left(\frac{1}{1-2\left(\frac{\sigma_k^2+\sigma_{k+1}^2}{2}+\frac{D}{2}(\mu_{k+1}-\mu_k)^2\right)\lambda}\right)^{\frac{1}{2}}.$$

Lemma 12 then proves the following Chernoff bound for $0<c<1$:

$$\mathbb{P}\left(\sum_{t_k-D+1}^{t_k+D}(x_t-\bar{x}_{(t_k-D+1):(t_k+D)})^2<2D\sigma_{k+1}\sigma_k c\left(1+\left(1-\frac{1}{2D}\right)\triangle_{\sigma,k}^2+\frac{\triangle_{\mu,k}^2}{4}\right)\right)$$
$$<\exp\left(\frac{2D-1}{2}(\log(c)-1-c)\right).$$

Since $D \geq 1$ this implies,

$$\mathbb{P}\left(\sum_{t_k-D+1}^{t_k+D}(x_t - \bar{x}_{(t_k-D+1):(t_k+D)})^2 < 2D\sigma_{k+1}\sigma_k c\left(1+\frac{1}{2}\triangle_{\sigma,k}^2+\frac{1}{4}\triangle_{\mu,k}^2\right)\right) < \exp\left(\frac{D}{2}(\log(c)-1-c)\right).$$

and therefore,

$$\mathbb{P}\left(\sum_{t_k-D+1}^{t_k+D}(x_t - \bar{x}_{(t_k-D+1):(t_k+D)})^2 < 2D\sigma_{k+1}\sigma_k c(D,n)\exp\left(\triangle_k^2\right)\right) < \exp\left(\frac{D}{2}(\log(c(D,n))-1-c(D,n))\right) = n^{-2+\epsilon}.$$

A Bonferroni correction over all possible t_k and D, of which there are guaranteed to be fewer than n^2 gives $\mathbb{P}(E_7) \geq 1 - n^{-\epsilon}$. This finishes the proof.

A.4.3 Proof of Lemma 3

Consider the function $f(x) = x - a - a\log(x/a)$. This function decreases monotonically on $(0, a)$ and increases monotonically on (a, ∞). Since E_2 and E_3 bound $Y_{i:j}$ from above and below respectively we only have to show that $Y_{i:j} - a - a\log(Y_{i:j}/a) \leq \tilde{C}_1\log(n)$ holds for the bounds in order to prove the lemma.

Part 1: Upper bound: By E_2 there exist constants M and M' such that $Y_{i:j} \leq a + M\sqrt{a\log(n)} + M'\log(n)$. Substituting this upper bound for $Y_{i:j}$ gives:

$$Y_{i:j} - a - a\log\left(\frac{Y_{i:j}}{a}\right) \leq M\sqrt{a\log(n)} + M'\log(n) - a\log\left(1+\frac{M\sqrt{a\log(n)}+M'\log(n)}{a}\right). \tag{A.4.1}$$

Case 1: $a \leq \log(n)$. In that case we can bound equation (A.4.1) by

$$M\sqrt{a\log(n)} + M'\log(n) \leq (M+M')\log(n).$$

Case 2: $a \geq \log(n)$. We can use the fact that $\log(1+x) \geq x - x^2, \ \ \forall x > 0$ to bound equation (A.4.1) by

$$\frac{(M\sqrt{a\log(n)} + M'\log(n))^2}{a} \leq (M + M')^2 \log(n).$$

Part 2: Lower bound: E_3 implies that $Y_{i:j} \geq c(a,n)(a-1)$. Substituting this bound gives

$$\begin{aligned} Y_{i:j} - a - a\log\left(\frac{Y_{i:j}}{a}\right) &\leq a(c(a,n) - 1 - \log(c(a,n)) - c - a\log\left(\frac{a-1}{a}\right) \\ &\leq 4(4+\epsilon)\log(n) + a\log\left(\frac{a}{a-1}\right) \leq 4(4+\epsilon)\log(n) + \frac{a}{a-1} \leq 4(4+\epsilon)\log(n) + 2. \end{aligned}$$

This finishes the proof.

A.4.4 Proof of Lemma 4

This lemma bounds the reduction in cost we can obtain by using a mean and variance fitted to a segment rather than the true mean and variance of the segment. The left bound follows from the fact $\tilde{\mathcal{C}}(x_{i:j})$ fits the mean and variance to minimise the log likelihood on the segment $x_{i:j}$. The right bound follows from Lemma 3 and E_1. Indeed,

$$\begin{aligned} \mathcal{C}(x_{i:j}) - \tilde{\mathcal{C}}(x_{i:j}) &= (j-i+1)\log(\sigma_k^2) + \sum_{t=i}^{j} \eta_t^2 - (j-i+1)\left(\log\left(\frac{\sigma_k^2 Y_{i:j}}{j-i+1}\right) + 1\right) \\ &= a\bar{\eta_{i:j}}^2 + Y_{i:j} - a\log\left(\frac{Y_{i:j}}{a}\right) - a \leq \left(\tilde{C}_1 + 4 + \epsilon\right)\log(n), \end{aligned}$$

which finishes the proof.

A.4.5 Proof of Lemma 5

This Lemma is very similar to Lemma 4, except that we slightly relax the constraint that all the data has to be located between two changepoints. This is needed because of the minimum segment length of two. We will prove this lemma for the case where $t_{k-1} = i$, the other case being very similar. We consider 3 cases:

Case 1: $j = t_{k-1} + 1$. We have that:

$$\mathcal{C}\left(x_{i:j}\right) - \tilde{\mathcal{C}}\left(x_{i:j}\right) = \log(\sigma_k^2) + \log(\sigma_{k-1}^2) + \eta_{t_{k-1}}^2 + \eta_{t_{k-1}+1}^2 - 2\log\left(\frac{(x_{t_k+1} - x_{t_k})^2}{4}\right) - 2$$

$$\leq (8 + 2\epsilon)\log(n) - 2\log\left(\frac{(x_{t_k+1} - x_{t_k})^2}{4\sigma_{k-1}\sigma_k}\right) - 2 \leq (8 + 4\epsilon)\log(n) + 2\log(4) - 2,$$

where the first inequality follows from E_1 and the second from E_5.

Case 2: $j = t_{k-1} + 2$. We have:

$$\begin{aligned}
&\mathcal{C}\left(x_{i:j}\right) - \tilde{\mathcal{C}}\left(x_{i:j}\right) \\
&= 2\log(\sigma_k^2) + \log(\sigma_{k-1}^2) + \sum_{t=t_{k-1}}^{t_{k-1}+2} \eta_t^2 - 3\log\left(\frac{\sum_{t=t_{k-1}}^{t_{k-1}+2}\left(x_t - \bar{x}_{(t_{k-1}):(t_{k-1}+2)}\right)^2}{3}\right) - 3 \\
&\leq 2\log(\sigma_k^2) + \log(\sigma_{k-1}^2) + (12 + 3\epsilon)\log(n) - 3\log\left(\frac{n^{-\epsilon}\sigma_k^{4/3}\sigma_{k-1}^{2/3}}{3}\right) - 3 \\
&= (12 + 6\epsilon)\log(n) + 3\log(3) - 3,
\end{aligned}$$

where the inequality follows from E_1 and E_4.

Case 3: $j > t_{k-1} + 2$. We have:

$$\begin{aligned}
\mathcal{C}\left(x_{i:j}\right) - \tilde{\mathcal{C}}\left(x_{i:j}\right) &\leq \left[\mathcal{C}\left(x_{i:(i+1)}\right) - \tilde{\mathcal{C}}\left(x_{i:(i+1)}\right)\right] + \left[\mathcal{C}\left(x_{(i+2):j}\right) - \tilde{\mathcal{C}}\left(x_{(i+2):j}\right)\right] \\
&\leq (8 + 4\epsilon)\log(n) + 2\log(4) - 2 + \tilde{C}_2\log(n),
\end{aligned}$$

where the second inequality follows from case 1 and Lemma 4.

A.4.6 Proof of Lemma 6

This lemma applies Lemma 4 to show that removing false positives reduces the overall cost.

$$\tilde{\mathcal{C}}\left(x_{i:j},\tau,\alpha\right)-\tilde{\mathcal{C}}\left(x_{i:j},\tau_{-b},\alpha\right)=\tilde{\mathcal{C}}\left(x_{(a+1):b}\right)+\tilde{\mathcal{C}}\left(x_{(b+1):c}\right)-\tilde{\mathcal{C}}\left(x_{(a+1):c}\right)+\alpha\log(n)^{1+\delta}$$

$$\geq\mathcal{C}\left(x_{(a+1):b}\right)+\mathcal{C}\left(x_{(b+1):c}\right)-\mathcal{C}\left(x_{(a+1):c}\right)+\alpha\log(n)^{1+\delta}-2\tilde{C}_2\log(n)\geq\frac{3}{4}\alpha\log(n)^{1+\delta},$$

for large enough n.

A.4.7 Proof of Lemma 7

This lemma shows that not having an estimated changepoint near a true changepoint leads to high costs. Let $j-t_k=t_k+1-i=D$. We have

$$\tilde{\mathcal{C}}\left(x_{i:j}\right)-\left(\mathcal{C}\left(x_{i:t_k}\right)+\mathcal{C}\left(x_{(t_k+1):j}\right)\right)=2D\log\left(\frac{1}{2D\sigma_k\sigma_{k+1}}\sum_{t=i}^{j}(x_t-\bar{x}_{i:j})^2\right)+2D-Y_{i:j}-2D\bar{\eta}_{i:j}.$$

We note that E_1 and E_2 imply that

$$2D-Y_{i:j}-2D\bar{\eta}_{i:j}\geq 1-2\sqrt{2(2+\epsilon)D\log(n)}-(4+2\epsilon)\log(n)-(8+2\epsilon)\log(n).$$

This in conjunction with E_7 implies that $\tilde{\mathcal{C}}\left(x_{i:j}\right)-\left(\mathcal{C}\left(x_{i:t_k}\right)+\mathcal{C}\left(x_{(t_k+1):j}\right)\right)$ is bounded below by

$$2D\triangle_k+2D\log(c(D,n))-2\sqrt{2(2+\epsilon)D\log(n)}-(4+2\epsilon)\log(n)-(8+2\epsilon)\log(n)$$

$$\geq 2D\triangle_k-4(2+\epsilon)\log(n)+2D(c(D,n)-1)-2\sqrt{2(2+\epsilon)D\log(n)}-4(3+\epsilon)\log(n)$$

$$\geq 2D\triangle_k-4(5+2\epsilon)\log(n)-2\sqrt{2(2+\epsilon)D\log(n)}-2\sqrt{2(2+\epsilon)D\log(n)},$$

where the first inequality follows from E_7 and the second one from Lemma 13. Writing the above lower bound as

$$D\triangle_k-4(5+2\epsilon)\log(n)+\left(\sqrt{D\triangle_k^2}-4\sqrt{2(2+\epsilon)\log(n)}\right)\sqrt{D}$$

proves the result for

$$\tilde{\kappa}(\alpha, \epsilon) = a + 4\sqrt{2(2+\epsilon)}.$$

A.4.8 Proof of Lemma 8

Without loss of generality, we assume that $\mu_0 = 0$ and $\sigma_0 = 1$. Since $\hat{\mu}$ and $\hat{\sigma}$ only depend upon $x_{(0.25n)}$,$x_{(0.5n)}$, and $x_{(0.75n)}$ it is sufficient to show that there exists a constant D_3 such that

$$\mathbb{P}\left(|x_{(cn)} - q_c| < D_3\sqrt{\frac{\log(n)}{n}}\right) \geq 1 - n^{-\epsilon},$$

where q_c is the cth quantile of the normal, holds for $c = 0.25, 0.5, 0.75$.

In order to do so, we first define $y_{(i)}$ to be the ith largest observation belonging to the typical distribution. We note that $y_{(cn-m)} < x_{(cn)} < y_{(cn+m)}$, where $m = O(K\sqrt{n})$ is the number of points belonging to one of the anomalous windows. Since $q_{(cn\pm m)/(n-m)} - q_c = O(Kn^{-\frac{1}{2}})$, it is sufficient to show that there exists a constant D_4 such that

$$\mathbb{P}\left(|y_{(a(n-m))} - q_a| \leq D_4\sqrt{\frac{\log(n)}{n}}\right) \geq 1 - n^{-\epsilon}$$

for $a = (cn \pm m)/(n-m)$. We note that

$$y_{(a(n-m))} \sim \Phi^{-1}\Big(U_{(a(n-m)),(n-m)}\Big),$$

where Φ is the CDF of the normal distribution and $U_{s,t}$ the sth largest of t i.i.d. $U(0,1)$ random variables. The following concentration inequality (Reiss (2012)) applies to the uniform distribution

$$\mathbb{P}\left(\frac{\sqrt{n}}{v}\left|U_{r,n} - \frac{r}{n}\right| > t\right) \leq \exp\left(-\frac{t^2}{3(1+\frac{t}{v}\sqrt{n})}\right),$$

where $v^2 = (r/n)(1 - r/n) \leq 1/4$ by the AMGM inequality. This means that the event

$$\left\{ \left| U_{a(n-m),(n-m)} - a \right| \leq \sqrt{\epsilon} \sqrt{\frac{\log(n)}{n}} \right\}$$

for the six values of a which are of interest to us holds with probability at least

$$1 - 6 \exp\left(-\epsilon \log(n) \left(\frac{3}{4} + 3\sqrt{\frac{\epsilon \log(n)}{n}} \right)^{-1} \right),$$

by a Bonferroni correction, which is $1 - O(n^{-\epsilon})$. We note that this event implies that

$$\left| \Phi(y_{a(n-m)}) - a \right| = O\left(\sqrt{\frac{\log(n)}{n}} \right)$$

holds for all six a of interest, which will be confined to the interval $[0.1, 0.9]$ for large enough n. Hence we must also have

$$\left| \Phi^{-1}(\Phi(y_{a(n-m)})) - \Phi^{-1}(a) \right| = O\left(\sqrt{\frac{\log(n)}{n}} \right)$$

for large enough n. This finishes our proof.

A.4.9 Proof of Lemma 9

First of all we note that

$$\begin{aligned}
&\mathcal{C}_E\left(x_{1:n}, \alpha, \hat{\mu}, \hat{\sigma}\right) - \mathcal{C}_E\left(x_{1:n}, \alpha, \mu_0, \hat{\sigma}\right) = \frac{1}{\hat{\sigma}^2} \sum_{i=1}^{K+1} \sum_{t=e_i+1}^{s_{i+1}} \left[(x_t - \hat{\mu})^2 - (x_t - \mu_0)^2 \right] \\
&\leq \frac{1}{\hat{\sigma}^2} \sum_{i=1}^{K+1} \left[(s_{i+1} - e_i)(\hat{\mu} - \mu_0)^2 + 2\sigma_0 (s_{i+1} - e_i) |\bar{\eta}_{(s_{i+1}+1):e_i}| |(\hat{\mu} - \mu_0)| \right] \\
&\leq \frac{2}{\sigma_0^2} n (\hat{\mu} - \mu_0)^2 + \frac{2}{\sigma_0} |(\hat{\mu} - \mu_0)| \sum_{i=1}^{K+1} \sqrt{(s_{i+1} - e_i)(4 + \epsilon) \log(n)} \\
&\leq 2D_1^2 \log(n) + 2(2 + \epsilon)(K + 1) D_1 \log(n),
\end{aligned}$$

where the second inequality follows from E_1 and the third from E_8. Moreover,

$$\begin{aligned}
&\mathcal{C}_E\left(x_{1:n}, \alpha, \mu_0, \hat{\sigma}\right) - \mathcal{C}_E\left(x_{1:n}, \alpha, \mu_0, \sigma_0\right) = \sum_{i=1}^{K+1} \sum_{t=e_i+1}^{s_{i+1}} \left[\log(\hat{\sigma}^2) - \log(\sigma_0^2) + \left(\frac{1}{\hat{\sigma}^2} - \frac{1}{\sigma_0^2}\right)(x_t - \mu_0)^2\right] \\
&= \sum_{i=1}^{K+1} \left[-(s_{i+1} - e_i)\log\left(\frac{\sigma_0^2}{\hat{\sigma}^2}\right) + \left(\frac{\sigma_0^2}{\hat{\sigma}^2} - 1\right) \sum_{t=e_i+1}^{s_{i+1}} \eta_t^2\right] \\
&\leq \sum_{i=1}^{K+1} \left[-(s_{i+1} - e_i)\left[\left(\frac{\sigma_0^2}{\hat{\sigma}^2} - 1\right) + O\left(\left(\frac{\sigma_0^2}{\hat{\sigma}^2} - 1\right)^2\right)\right] + \left(\frac{\sigma_0^2}{\hat{\sigma}^2} - 1\right)\left(Y_{(e_i+1):(s_{i+1})} + (4+\epsilon)\log(n)\right)\right] \\
&\leq \sum_{i=1}^{K+1} \left[\left(\frac{\sigma_0^2}{\hat{\sigma}^2} - 1\right)\left(Y_{(e_i+1):(s_{i+1})} - (s_{i+1} - e_i)\right) + O(\log(n))\right] = O(K\log(n)),
\end{aligned}$$

where the first inequality follows from expanding $\log(x)$ around $x = 1$ and E_1, while the second inequality uses E_8 and E_2.

A.4.10 Proof of Lemma 10

Let $k' = k \pm 1$. Clearly, $x_t - \mu_{k'} \sim N(\mu_k - \mu_{k'}, \sigma_k^2)$. Consequently,

$$\mathbb{P}\left(|x_t - \mu_{k'}| < n^{-(2+\epsilon)}\sigma_k\right) \leq 2n^{-(2+\epsilon)}\sigma_k\sqrt{\frac{1}{2\pi\sigma_k^2}} = \sqrt{\frac{2}{\pi}}n^{-(2+\epsilon)}.$$

A Bonferroni correction therefore gives $\mathbb{P}(E_9) > 1 - \sqrt{\frac{8}{\pi}}n^{-\epsilon}$.

A.4.11 Proof of Lemma 11

We have to consider two cases:

Case 1: $i < j$. The result holds by Lemma 4.

Case 2: $i = j$, with the proxy for segments of length one. We have:

$$\mathcal{C}\left(x_{i:j}\right) - \tilde{\mathcal{C}}\left(x_{i:j}\right) = \log(\sigma_k^2) + \eta_i^2 - \log(\tilde{\sigma}^2) - \frac{(x_i - \tilde{\mu})^2}{\tilde{\sigma}^2} \leq (4+\epsilon)\log(n) + \log\left(\frac{\sigma_k^2}{\tilde{\sigma}^2}\right) - \frac{(x_i - \tilde{\mu})^2}{\tilde{\sigma}^2},$$

where the inequality follows from E_1. We now bound the above for all choices of $\tilde{\mu}$ and $\tilde{\sigma}$. First of all we consider the case $|\tilde{\mu} - \mu_k| < D_1\sqrt{\frac{\log(n)}{n}}\sigma_k$ and $|\frac{\tilde{\sigma}^2}{\sigma_k^2} - 1| < D_2\sqrt{\frac{\log(n)}{n}}$.

Then for large enough n:

$$\log\left(\frac{\sigma_k^2}{\tilde{\sigma}^2}\right) - \frac{(x_i-\tilde{\mu})^2}{\tilde{\sigma}^2} \leq \log\left(1 + 2D_2\sqrt{\frac{\log(n)}{n}}\right) \leq 2D_2\sqrt{\frac{\log(n)}{n}}.$$

Next we consider the cases $|\tilde{\mu} - \mu_{k'}| < D_1\sqrt{\frac{\log(n)}{n}}\sigma_{k'}$ and $\left|\frac{\tilde{\sigma}^2}{\sigma_{k'}^2} - 1\right| < D_2\sqrt{\frac{\log(n)}{n}}$, where $k' = k+1$ or $k' = k-1$. We have:

$$\log\left(\frac{\sigma_k^2}{\tilde{\sigma}^2}\right) - \frac{(x_i-\tilde{\mu})^2}{\tilde{\sigma}^2} \leq 2D_2\sqrt{\frac{\log(n)}{n}} + \log\left(\frac{\sigma_k^2}{\sigma_{k'}^2}\right) - \frac{(x_i-\tilde{\mu})^2}{2\sigma_{k'}^2}$$

for large enough n. If $\frac{\sigma_k^2}{\sigma_{k'}^2} < n^4$ the above is bounded by $5\log(n)$ for large enough n. Otherwise we have:

$$\begin{aligned}
&\log\left(\frac{\sigma_k^2}{\sigma_{k'}^2}\right) - \frac{(x_i-\tilde{\mu})^2}{2\sigma_{k'}^2} \leq \log\left(\frac{\sigma_k^2}{\sigma_{k'}^2}\right) + \frac{(|x_i-\mu_{k'}| - |\mu_{k'}-\mu|)_0^2}{2\sigma_{k'}^2} \\
&\leq \log\left(\frac{\sigma_k^2}{\sigma_{k'}^2}\right) - \frac{1}{2}\left(\frac{\sigma_k}{\sigma_{k'}}n^{-(2+\epsilon)} - D_1\sqrt{\frac{\log(n)}{n}}\right)_0^2 \leq \log\left(\frac{\sigma_k^2}{\sigma_{k'}^2}\right) - \frac{1}{8}\left(\frac{\sigma_k}{\sigma_{k'}}n^{-(2+\epsilon)}\right)^2 \\
&\leq \log(8n^{4+2\epsilon}) - 1 = (4+2\epsilon)\log(n) + \log(8) - 1.
\end{aligned}$$

Case 3: $i = j$, with the proxy for epidemic changes. We have:

$$\begin{aligned}
\mathcal{C}(x_{i:j}) - \tilde{\mathcal{C}}(x_{i:j}) &= \log(\sigma_k^2) + \eta_i^2 - \log(\tilde{\sigma}^2\gamma + (x_i-\tilde{\mu})^2) - 1 \\
&\leq (4+\epsilon)\log(n) - \log\left(\frac{\tilde{\sigma}^2}{\sigma_k^2}\gamma + \frac{(x_i-\tilde{\mu})^2}{\sigma_k^2}\right),
\end{aligned}$$

by E_1. We again bound the above for all choices of $\tilde{\mu}$ and $\tilde{\sigma}$. First of all we consider the case $|\tilde{\mu} - \mu_k| < D_1\sqrt{\frac{\log(n)}{n}}\sigma_k$ and $\left|\frac{\tilde{\sigma}^2}{\sigma_k^2} - 1\right| < D_2\sqrt{\frac{\log(n)}{n}}$. Then, for large enough n

$$-\log\left(\frac{\tilde{\sigma}^2}{\sigma_k^2}\gamma + \frac{(x_i-\tilde{\mu})^2}{\sigma_k^2}\right) \leq -\log\left(\left[1 - D_2\sqrt{\frac{\log(n)}{n}}\right]\gamma\right) \leq -\log(\gamma) + 2D_2\sqrt{\frac{\log(n)}{n}}$$

Next, we consider the cases $|\tilde{\mu} - \mu_{k'}| < D_1\sqrt{\frac{\log(n)}{n}}\sigma_{k'}$ and $|\frac{\tilde{\sigma}^2}{\sigma_{k'}^2} - 1| < D_2\sqrt{\frac{\log(n)}{n}}$,

where $k' = k+1$ or $k' = k-1$. We have

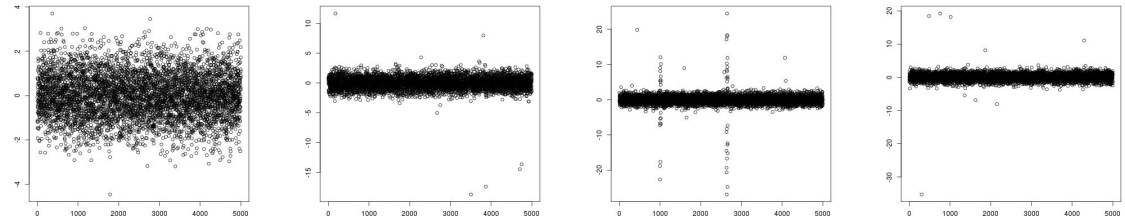
$$
\begin{aligned}
-\log\left(\frac{\tilde{\sigma}^2}{\sigma_k^2}\gamma + \frac{(x_i-\tilde{\mu})^2}{\sigma_k^2}\right) &\leq -\log\left(\frac{\tilde{\sigma}^2}{\sigma_{k'}^2}\right) - \log\left(\frac{\sigma_{k'}^2}{\sigma_k^2}\gamma + \frac{\sigma_{k'}^2}{\tilde{\sigma}^2}\frac{(x_i-\tilde{\mu})^2}{\sigma_k^2}\right) \\
&\leq 2D_2\sqrt{\frac{\log(n)}{n}} - \log\left(\frac{\sigma_{k'}^2}{\sigma_k^2}\gamma + \frac{1}{2}\frac{(|x_i-\mu_{k'}| - |\mu_{k'}-\mu|)^2}{\sigma_k^2}\right)
\end{aligned}
$$

for large enough n. If $\frac{\sigma_k^2}{\sigma_{k'}^2} < n^4$ the above is bounded by $4\log(n) - \log(\gamma)$. Otherwise we have:

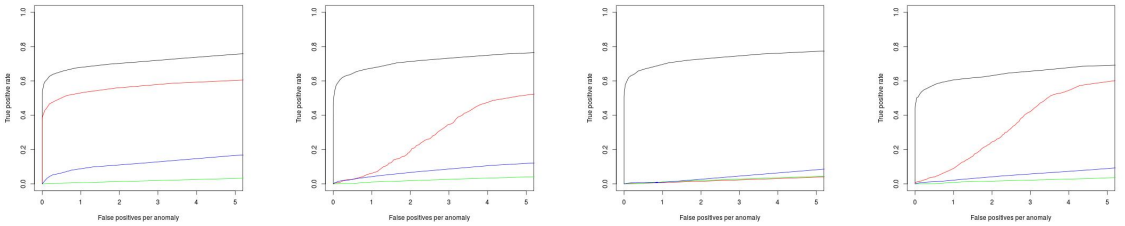
$$
\begin{aligned}
&-\log\left(\frac{\sigma_{k'}^2}{\sigma_k^2}\gamma + \frac{1}{2}\frac{(|x_i-\mu_{k'}| - |\mu_{k'}-\mu|)_0^2}{\sigma_k^2}\right) \leq -\log\left(\frac{\sigma_{k'}^2}{\sigma_k^2}\gamma + \frac{1}{2}\left(n^{-2+\epsilon} - D_1\sqrt{\frac{\log(n)}{n}}\frac{\sigma_{k'}}{\sigma_k}\right)_0^2\right) \\
&\leq -\log\left(\frac{1}{2}\left(n^{-2+\epsilon} - D_1\sqrt{\frac{\log(n)}{n}}n^{-2}\right)_0^2\right) \leq -\log\left(\frac{1}{8}\left(n^{-2+\epsilon}\right)^2\right) = \log(8) + (4+2\epsilon)\log(n),
\end{aligned}
$$

for large enough n. This finishes the proof.

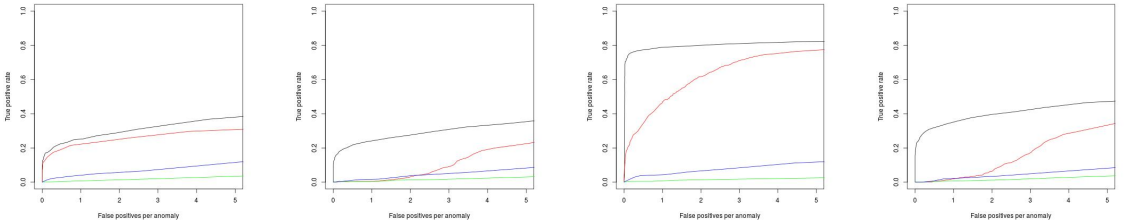
A.5 Further Simulation Study Results



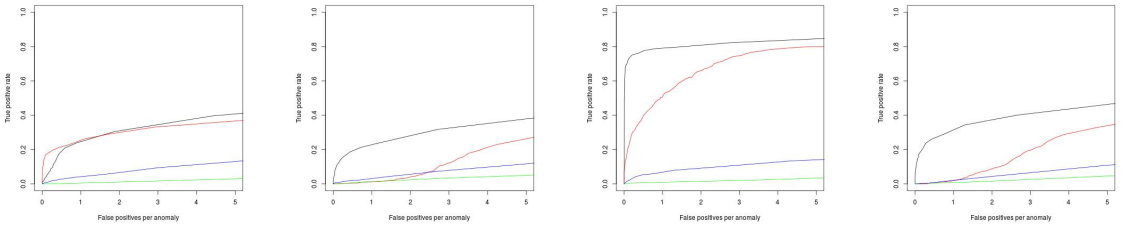
(a) Weak (b) Weak, PAs (c) Strong, PAs (d) Short, PAs



(e) Weak (f) Weak, PAs (g) Strong, PAs (h) Short, PAs

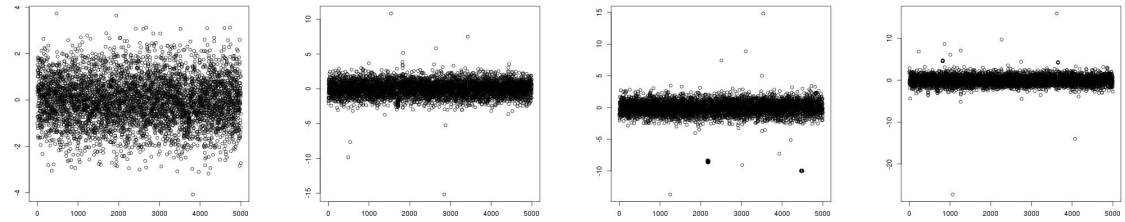


(i) Weak, AR (j) Weak, AR, PAs (k) Strong, AR, PAs (l) Short, AR, PAs

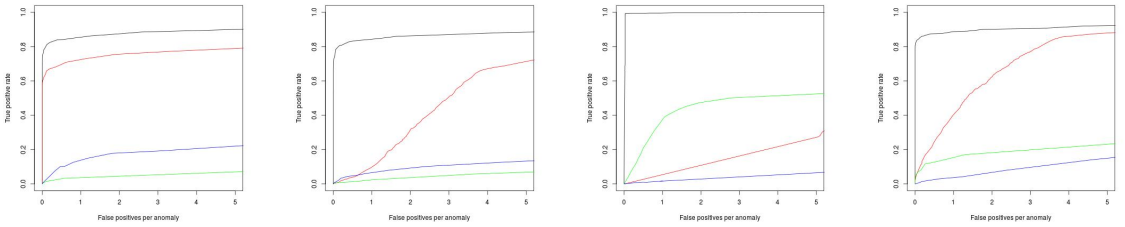


(m) Weak, T (n) Weak, T, PAs (o) Strong, T, PAs (p) Short, T, PAs

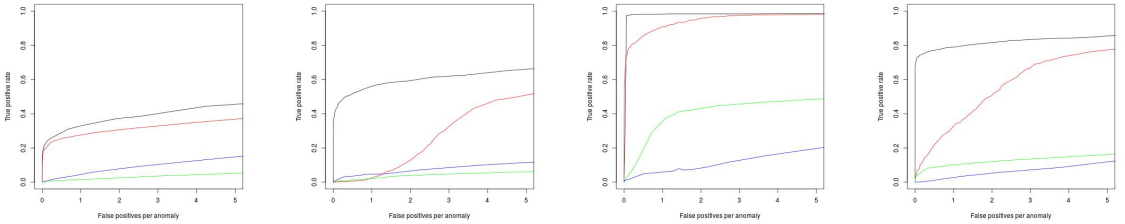
Figure A.5.1: Data examples and ROC curves for changes in variance for CAPA (black), PELT (red), BreakoutDetection (green), and luminol (blue).



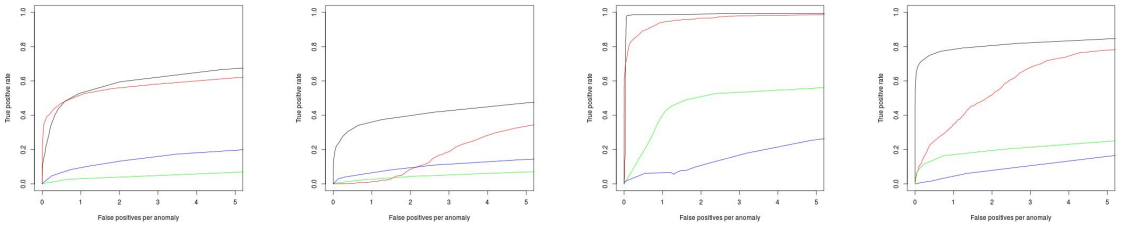
(a) Weak (b) Weak, PAs (c) Strong, PAs (d) Short, PAs



(e) Weak (f) Weak, PAs (g) Strong, PAs (h) Short, PAs



(i) Weak, AR (j) Weak, AR, PAs (k) Strong, AR, PAs (l) Short, AR, PAs

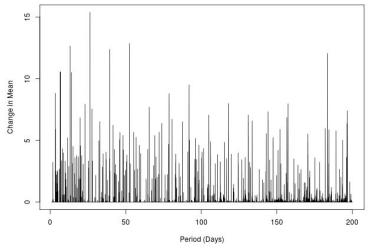


(m) Weak, T (n) Weak, T, PAs (o) Strong, T, PAs (p) Short, T, PAs

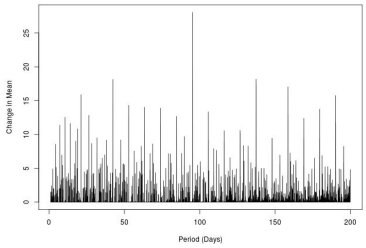
Figure A.5.2: Data examples and ROC curves for joint changes in mean and variance for CAPA (black), PELT (red), BreakoutDetection (green), and luminol (blue).

A.6 Application of CAPA to Further Stars

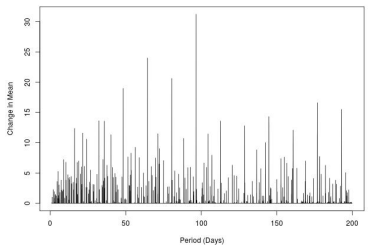
We applied the approach detailed in Section 6 to the light curves of five further stars with known exoplanets (Morton et al. (2016)). Figure A.6.1 depicts the largest detected change in mean as measured by $\max_k (\triangle_{\mu,k})$ per period for the five stars. We found that the 20 periods exhibiting the largest change in mean correspond to integer fractions of the orbital period of a known exoplanet in all cases. We thus observed no false positives. The results are summarised in Figure A.6.2. We note that not all planets appear in the 20 periods with largest change in mean. This is due to the fact that their signal is weaker than the resonance of the signal of larger planets. CAPA can nevertheless detect the transit signal of the missing planet at their orbital period, with the exception of Kepler 454-c. This planet however was discovered by a different method than the transit method.



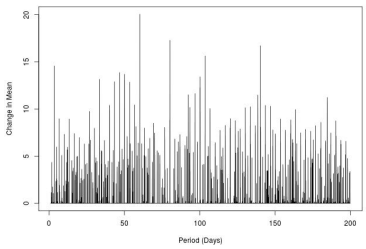
(a) Kepler 356.



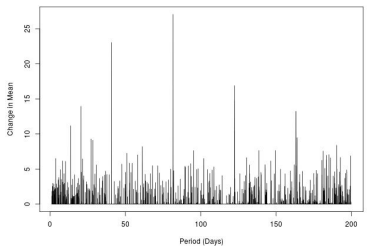
(b) Kepler 454.



(c) Kepler 275.



(d) Kepler 235.



(e) Kepler 264.

Figure A.6.1: The strongest change in mean, as measured by $\max_k(\triangle_{\mu,k})$, detected by CAPA for the lightcurves of five stars with known exoplanets. All periods from 1 to 200 days at 0.01 day increment were examined.

Star	Planet	Period	Period (or integer fraction) in top 20 modes
Kep. 275	Kep.275-b	10.3007	No
Kep. 275	Kep.275-c	16.0881	Yes
Kep. 275	Kep.275-d	35.6761	Yes
Kep. 264	Kep.264-b	40.806	Yes
Kep. 264	Kep. 264-c	140.101261	No
Kep. 356	Kep. 356-b	13.1216	Yes
Kep. 356	Kep. 356-c	4.6127	No
Kep. 454	Kep. 454-b	10.5738	Yes
Kep. 454	Kep. 454-c	523.90	No
Kep. 235	Kep. 235-b	3.340	Yes
Kep. 235	Kep. 235-c	7.824	No
Kep. 235	Kep. 235-d	20.0605	Yes
Kep. 235	Kep. 235-e	46.1836	Yes

Figure A.6.2: Five stars orbited by known exoplanets and whether their period or an integer fraction thereof was in the 20 periods with strongest change in mean according to CAPA.

Appendix B

MVCAPA

B.1 Additional Theoretical Results

B.1.1 Pruning Without Lags

The following proposition holds:

Proposition 17. *Let the costs $\mathcal{C}_i(,)$ be such that*

$$\min_{\boldsymbol{\theta}}\left(\sum_{t=a+1}^{c}\mathcal{C}_i\left(\boldsymbol{x}_t,\boldsymbol{\theta}\right)\right)\geq\min_{\boldsymbol{\theta}}\left(\sum_{t=a+1}^{b}\mathcal{C}_i\left(\boldsymbol{x}_t,\boldsymbol{\theta}\right)\right)+\min_{\boldsymbol{\theta}}\left(\sum_{t=b+1}^{c}\mathcal{C}_i\left(\boldsymbol{x}_t,\boldsymbol{\theta}\right)\right)$$

holds for all $\boldsymbol{x}$ and a,b,c such that $b-a\geq l$ and $c-b\geq l$. Then, if for some t there exists an $m\geq t-l$ such that

$$S(m)-\alpha-\sum_{1}^{p}\boldsymbol{\beta}_i>S(t)+\mathcal{S}\left(t,m\right),$$

then, for all $m'\geq m+l$,

$$S(m')>S(t)+\mathcal{S}\left(t,m'\right).$$

A wide range of cost functions (see Killick et al., 2012) satisfy the condition required by the above proposition. The proposition implies that if for some t there exists an $m \geq t - l$ such that

$$S(m) - \alpha - \sum_{1}^{p} \boldsymbol{\beta}_i > S(t) + \mathcal{S}(t, m)$$

holds, t can be dropped as an option from the dynamic programme for all steps after step $m + l$, thus reducing the cost of the algorithm.

B.1.2 Bounds on Lagged Savings

The following result provides a general way to extend the stochastic bound (and thus the penalties) from the lagged free to the lagged setting:

Proposition 18. *Let the cost function $\mathcal{C}_i()$ be such that the un-lagged saving*

$$\mathcal{C}_i\left(\boldsymbol{x}^{(i)}_{(s+1):e}, \boldsymbol{\theta}^{(i)}_0\right) - \min_{\boldsymbol{\theta}}\left(\mathcal{C}_i\left(\boldsymbol{x}^{(i)}_{(s+1):e}, \boldsymbol{\theta}\right)\right)$$

be stochastically bounded by $a\chi^2_v$, then the saving $\mathcal{S}_i(s, e)$ as defined in (4.6.2) *satisfies*

$$\mathbb{P}\left(\mathcal{S}_i(s, e) > x\right) \leq (w+1)^2 \mathbb{P}\left(a\chi^2_v > x\right).$$

consequently; replacing ψ by $\psi + 2\log(w+1)$ when going from the perfectly aligned case to the lagged case achieves at least the same error control.

B.1.3 Pruning the Dynamic Programme in the Presence of Lags

Even when lags are included in the model, the solution space of the dynamic programme can still be pruned in a fashion similar to Killick et al. (2012) and Fisch et al.

(2018a). Indeed, the following generalisation of Proposition 17 holds:

Proposition 19. *Let the costs $\mathcal{C}_i(,)$ be such that*

$$\min_{\boldsymbol{\theta}}\left(\sum_{t=a+1}^{c}\mathcal{C}_i\left(\boldsymbol{x}_t,\boldsymbol{\theta}\right)\right)\geq\min_{\boldsymbol{\theta}}\left(\sum_{t=a+1}^{b}\mathcal{C}_i\left(\boldsymbol{x}_t,\boldsymbol{\theta}\right)\right)+\min_{\boldsymbol{\theta}}\left(\sum_{t=b+1}^{c}\mathcal{C}_i\left(\boldsymbol{x}_t,\boldsymbol{\theta}\right)\right)$$

holds for all $\boldsymbol{x}$ and a,b,c such that $b-a\geq l$ and $c-b\geq l$. Then, if for some t there exists an $m\geq t-l-w$ such that

$$S(m)-\alpha-\sum_{1}^{p}\boldsymbol{\beta}_i>S(t)+\mathcal{S}\left(t,m\right)$$

holds,

$$S(m')>S(t)+\mathcal{S}\left(t,m'\right)$$

must also holds for all $m'\geq m+l+w$.

It implies that if for some t there exists an $m\geq t-l-w$ such that

$$C(m)-\alpha-\sum_{1}^{p}\boldsymbol{\beta}_i>C(t)+\mathcal{S}\left(t,m\right)$$

holds, t can be dropped as an option from the dynamic programme for all steps after step $m+l+w$, thus reducing the cost of the algorithm. Moreover, we only need to maintain the savings $S(a,b)$ for all a exceeding the smallest option not yet dropped from the dynamic programme, which further reduces the computational cost. As a result of this pruning we found the runtime of MVCAPA to be close to linear in n, when the number of anomalies increased linearly with n.

B.2 Proofs for Theorems and Propositions

B.2.1 Proof of Proposition 4

We will prove the existence of such a constant for each the three penalty regimes. The result follows from this.

Regime 1

Let $0 \leq s < e \leq n$. The probability that the segment $(s+1, e)$ is not flagged up as anomalous is given by

$$\begin{aligned}
&\mathbb{P}\left(\sum_{c \in J_m} \mathcal{S}_c(s,e) < a\left(pv + 2\sqrt{pv\psi} + 2\psi\right), \quad \forall S_m \subset \{1,...,p\} : |J_m| = m, \;\; 1 \leq m \leq p\right) \\
&= \mathbb{P}\left(\sum_{c=1}^{p} \mathcal{S}_c(s,e) < a\left(pv + 2\sqrt{pv\psi} + 2\psi\right)\right) \\
&\geq \mathbb{P}\left(\chi^2_{pv} < pv + 2\psi + 2\sqrt{pv\psi}\right) \geq 1 - e^{-\psi},
\end{aligned}$$

where the first inequality follows from the stochastic bound on $\mathcal{S}_c(s,e)$ and the second inequality follows from the bounds on the chi-squared distribution in Laurent and Massart (2000). A Bonferroni correction over all possible pairs s, e then finishes the proof.

Regime 2

Let $1 \leq s \leq e \leq n$. For this pair (s,e) define $Y_c = \mathcal{S}_c(s+1, e)$, noting that they are all independent and stochastically bounded by aZ_c where $Z_1, ..., Z_p$ are *i.i.d.* χ^2_v random

variables. The probability that the segment s, e will not be considered anomalous is

$$\begin{aligned}
&\mathbb{P}\left(\sum_{c\in S_m} Y_c < 2(1+\epsilon)a\psi + 2ma(1+\epsilon)\log(p),\ \forall S_m \subset \{1,...,p\} : |S_m| = m,\ 1 \le m \le p\right) \\
&\ge \mathbb{P}\left(\sum_{i=1}^{p}\left(\frac{Y_i - 2(1+\epsilon)a\log(p)}{a}\right)^+ < 2(1+\epsilon)\psi\right) \\
&\ge \mathbb{P}\left(\sum_{i=1}^{p}\left(Z_i - 2(1+\epsilon)\log(p)\right)^+ < 2(1+\epsilon)\psi\right) \\
&\ge 1 - \mathbb{E}\left(e^{\lambda(Z_1 - 2(1+\epsilon)\log(p))^+}\right)^p e^{-2\lambda(1+\epsilon)\psi},
\end{aligned}$$

for all $\lambda > 0$, where the final inequality corresponds to a Chernoff bound. Next set $\lambda = \frac{1}{2}\frac{1}{1+\epsilon}$ and note that the following Lemma holds:

Lemma 14. *Let $X \sim \chi^2_v$. Then the MGF of $(X-c)^+$ is given by:*

$$\mathbb{P}\left(\chi^2_v < c\right) + \frac{e^{-\lambda c}}{(1-2\lambda)^{\frac{v}{2}}}\mathbb{P}\left(\chi^2_v > c(1-2\lambda)\right).$$

Consequently,

$$\begin{aligned}
&\mathbb{E}\left(e^{\lambda(Z_1 - 2(1+\epsilon)\log(p))^+}\right)^p e^{-\psi} \\
&= \left[\mathbb{P}\left(\chi^2_v < 2(1+\epsilon)\log(p)\right) + \frac{e^{-2(1+\epsilon)\lambda\log(p)}}{(1-2\lambda)^{\frac{v}{2}}}\mathbb{P}\left(\chi^2_v > 2(1+\epsilon)(1-2\lambda)\log(p)\right)\right]^p e^{-\psi} \\
&\le \left[1 + \frac{e^{-2(1+\epsilon)\lambda\log(p)}}{(1-2\lambda)^{\frac{v}{2}}}\right]^p e^{-\psi} = \left[1 + \frac{1}{p}\left(\frac{1+\epsilon}{\epsilon}\right)^{\frac{v}{2}}\right]^p e^{-\psi} \le \exp\left(\left(\frac{1+\epsilon}{\epsilon}\right)^{\frac{v}{2}}\right)e^{-\psi}.
\end{aligned}$$

A Bonferroni correction over all possible pairs s, e then finishes the proof.

Regime 3

Let $1 \le s \le e \le n$. For this pair (s, e) define $Y_i = \mathcal{S}_c(s+1, e)$, noting that they are all independent and stochastically bounded by aZ_i where $Z_1, ..., Z_p$ are *i.i.d.* χ^2_v random variables. Next, define their order statistic $Z_{(1)} \ge ... \ge Z_{(p)}$ The probability that the

segment (s, e) is not flagged up as anomalous is given by

$$\begin{aligned}
&\mathbb{P}\left(\sum_{i=1}^{m} Y_{(i)} < a\left(2\psi' + mv + 2pc_m f(c_m) + 2\sqrt{(mv + 2pc_m f(c_m))\psi'}\right), \quad 1 \leq m \leq p\right) \\
&\geq 1 - \sum_{m=1}^{p} \mathbb{P}\left(\sum_{i=1}^{m}\left(\frac{Y_{(i)} - ac_m}{a}\right) > 2\psi' + m(v - c_m) + 2pc_m f(c_m) + 2\sqrt{(mv + 2pc_m f(c_m))\psi'}\right) \\
&\geq 1 - \sum_{m=1}^{p} \mathbb{P}\left(\sum_{i=1}^{m}\left(\frac{Y_{(i)} - ac_m}{a}\right)^{+} > 2\psi' + m(v - c_m) + 2pc_m f(c_m) + 2\sqrt{(vm + 2pc_m f(c_m))\psi'}\right) \\
&\geq 1 - \sum_{m=1}^{p} \mathbb{P}\left(\sum_{i=1}^{p}\left(\frac{Y_i - ac_m}{a}\right)^{+} < 2\psi' + m(v - c_m) + 2pc_m f(c_m) + 2\sqrt{(vm + 2pc_m f(c_m))\psi'}\right). \\
&\geq 1 - \sum_{m=1}^{p} \mathbb{P}\left(\sum_{i=1}^{p}(Z_i - c_m)^{+} < 2\psi' + m(v - c_m) + 2pc_m f(c_m) + 2\sqrt{(mv + 2pc_m f(c_m))\psi'}\right),
\end{aligned}$$

where we use the shorthand $\psi' := \psi + \log(p)$. We will now use the following lemma, which shows that $(Z_c - c_m)^{+}$ is sub-gamma.

Lemma 15. *Let $Z \sim \chi^2_v$ for $v \leq 2$. Then $(Z - c)^{+} - [2cf(c) + (v - c)\mathbb{P}(\chi^2_v > c)]$ is sub-gamma with scale parameter 2 and variance $V = 4cf(c) + 2v\mathbb{P}(\chi^2_v > c)$.*

Using Lemma 15 and the bounds on sub-gamma random-variables in Boucheron and Thomas (2012), we have that

$$\begin{aligned}
&\sum_{m=1}^{p} \mathbb{P}\left(\sum_{i=1}^{p}(Z_i - c_m)^{+} < 2\psi' + m(v - c_m) + 2pc_m f(c_m) + 2\sqrt{(mv + 2pc_m f(c_m))\psi'}\right) \\
&= \sum_{m=1}^{p} \mathbb{P}\left(\sum_{i=1}^{p}\left((Z_i - c_m)^{+} - (v - c_m)\mathbb{P}(\chi^2_v > c_m) + 2c_m f(c_m)\right) < 2\psi' + 2\sqrt{pv\left(\mathbb{P}(\chi^2_v > c_m) + 2c_m f(c_m)\right)\psi'}\right) \\
&\leq \sum_{m=1}^{p} e^{-\psi'} = \sum_{m=1}^{p} p^{-1}e^{-\psi} = e^{-\psi},
\end{aligned}$$

again using the shorthand $\psi' = \psi + \log(p)$. A Bonferroni correction over all possible pairs s, e then finishes the proof.

B.2.2 Proof of Proposition 5

Proof of Proposition 5: We will show that the penalised saving for the true anomalous segment is positive with probability converging to 1 as p increases. By the definition of signal strength, the distribution of the true anomalous segment's penalised

saving does not depend on the length, $s-e$, of the segment. Thus, we assume, without loss of generality, that $e = s+1$ and treat the cases $0 < \xi \leq \frac{1}{2}$, $\frac{1}{2} < \xi < \frac{3}{4}$, and $\frac{3}{4} < \xi < 1$, separately. We write $X_i := \mathbf{x}_e^{(i)}$, for $1 \leq i \leq p$.

Case 1: $0 < \xi \leq \frac{1}{2}$. Remember that the composite penalty used is the minimum between regimes 1, 2, and 3. It is therefore sufficient to show that the saving will exceed the penalty specified by one of these three regimes (regime 1 in this case) at some point. By definition, $X_i = (\epsilon_i + v_i\mu)^2$, where $\epsilon_1, ..., \epsilon_p$ are i.i.d. $N(0,1)$ and $v_1, ..., v_p$ are i.i.d. $Ber(p^{-\xi})$. Therefore

$$\begin{aligned}
&\mathbb{P}\left(\exists m : \sum_{i=1}^{m} X_{(i)} \geq \alpha + \sum_{i=1}^{m} \boldsymbol{\beta}_i\right) \geq \mathbb{P}\left(\sum_{i=1}^{p} X_i > p + 2\sqrt{p\psi} + 2\psi\right) \\
&= \mathbb{P}\left(\sum_{i=1}^{p} \epsilon_i^2 + \sum_{i=1}^{p} v_i\left(2\mu\epsilon_i + \mu^2\right) > p + 2\sqrt{p\psi} + 2\psi\right) \\
&\geq 1 - \mathbb{P}\left(\sum_{i=1}^{p} \epsilon_i^2 < p - 2\sqrt{p\psi}\right) - \mathbb{P}\left(\sum_{i=1}^{p} v_i\left(2\mu\epsilon_i + \mu^2\right) < 4\sqrt{p\psi} + 2\psi\right) \\
&\geq 1 - e^{-\psi} - \mathbb{P}\left(N\left(\mu^2\left(\sum_{i=1}^{p} v_i\right), 4\mu^2\left(\sum_{i=1}^{p} v_i\right)\right) < 4\sqrt{p\psi} + 2\psi\right)
\end{aligned}$$

Furthermore,

$$\mathbb{P}\left(N\left(\mu^2\sum_{i=1}^{p} v_i, 4\mu^2\sum_{i=1}^{p} v_i\right) > 4\sqrt{p\psi} + 2\psi\right) > \mathbb{P}\left(N\left(k\mu^2, 4k\mu^2\right) > 4\sqrt{p\psi} + 2\psi\right)\mathbb{P}\left(\sum_{i=1}^{p} v_i > k\right),$$

for all k such that $1 \leq k \leq p$. We therefore only have to show that there exists some sequence of integers k_p such that the right hand side converges to 1 as $p \to \infty$. Note that Hoeffding's inequality implies that

$$\mathbb{P}\left(\sum_{i=1}^{p} v_i > p^{1-\xi} - p^{\frac{1}{2}-\frac{1}{2}\xi}\sqrt{\log(p)}\right) \;\to\; 1 \quad as \quad p \;\to\; \infty$$

and therefore

$$\mathbb{P}\left(\sum_{i=1}^{p} v_i > \frac{1}{2}p^{1-\xi}\right) \;\rightarrow\; 1 \quad as \quad p \;\rightarrow\; \infty.$$

Setting $k_p = \lceil \frac{1}{2}p^{1-\xi} \rceil$, it is therefore sufficient to show that

$$\mathbb{P}\left(N(0,1) > \frac{4\sqrt{p\psi} + 2\psi - \frac{1}{2}\mu^2 p^{1-\xi}}{\sqrt{2\mu^2 p^{1-\xi}}}\right) = \mathbb{P}\left(N(0,1) > \frac{4\sqrt{p\psi} + 2\psi - \frac{1}{2}p^{1-\xi-2r_p}}{\sqrt{2\mu^2 p^{1-\xi-2r_p}}}\right)$$

converges to 1 as p tends to infinity. This is the case if $r_p < \frac{1}{2}(\frac{1}{2} - \xi)$, which finishes the proof.

Case 2: $\frac{3}{4} < \xi < 1$. By an argument similar to that made for case 1, it is sufficient to show that the saving will exceed the penalty specified by regime 2. We have that:

$$\mathbb{P}\left(\exists m : \sum_{i=1}^{m} X_{(i)} \geq \alpha + \sum_{i=1}^{m} \boldsymbol{\beta}_i\right) \geq \mathbb{P}\left(X_{(1)} > 2\psi + 2\log(p)\right) = 1 - \left(1 - \mathbb{P}\left(X_1 > 2\psi + 2\log(p)\right)\right)^p$$

By definition, $X_1 = (\mu v_1 + \epsilon_1)^2$, where $\epsilon_1 \sim N(0,1)$ and $v_1 \sim Ber(p^{-\xi})$. We can therefore bound the above by

$$\begin{aligned}
&1 - \left(1 - \mathbb{P}\left(X_1 > 2\psi + 2\log(p), v_1 = 1\right)\right)^p \\
&= 1 - \left(1 - p^{-\xi}\mathbb{P}\left(\left(\epsilon_1 + \sqrt{2r_p \log(p)}\right)^2 > 2\psi + 2\log(p)\right)\right)^p \\
&> 1 - \left(1 - p^{-\xi}\mathbb{P}\left(N(0,1) > \sqrt{2\psi + 2\log(p)} - \sqrt{2r_p \log(p)}\right)\right)^p \\
&\geq 1 - \exp\left(-p^{1-\xi}\mathbb{P}\left(N(0,1) > \sqrt{2\psi + 2\log(p)} - \sqrt{2r_p \log(p)}\right)\right),
\end{aligned}$$

where the second inequality follows from the fact that $1 - x \leq e^{-x}$. We consider separately the cases $\sqrt{2\psi + 2\log(p)} - \sqrt{2r_p \log(p)} > 1$ and $\sqrt{2\psi + 2\log(p)} - \sqrt{2r_p \log(p)} \leq 1$. In the latter case the above clearly converges to 1 as p goes to infinity. In the former case we can use the lower tail bound $\mathbb{P}\left(N(0,1) > x\right) > \frac{1}{\sqrt{2\pi}} \frac{x}{x^2+1} \exp\left(-\frac{x^2}{2}\right)$, for $x > 0$

to bound the above by

$$1-\exp\left(-\frac{1}{\sqrt{2\pi}}p^{1-\xi}\frac{p^{\left(\sqrt{1+\frac{2\psi}{2\log(p)}}-\sqrt{r_p}\right)^2}}{1+\left(\sqrt{2\psi+2\log(p)}-\sqrt{2r_p\log(p)}\right)^2}\right).$$

Thus, for a fixed $r_p > \left(1-\sqrt{1-\xi}\right)^2$ this converges to 1, as $\psi/\log(p)$ converges to 0.

Case 3: $\frac{1}{2} < \xi < \frac{3}{4}$. By an argument similar to that made for case 1, it is sufficient to show that the saving will exceed the penalty specified by regime 3. We assume, without loss of generality, that $\mu > 0$. If $r_p \geq \frac{1}{4}$. Our approach is to define a threshold, b, and a number of excesses, $\tilde{k}$, such that the number of savings in cost that exceed b will be great than $\tilde{k}$ with probability going to 1 as p increases. We then show that the overall sum of the $\tilde{k}$ largest savings will be greater than the penalty for fitting $\tilde{k}$ components as anomalous.

We introduce the following new random variable:

$$Y_i = \begin{cases} \left[(\mu+\epsilon_i)^+\right]^2 & \text{if} \quad v_i = 1 \\ \\ \epsilon_i^2 & \text{if} \quad v_i = 0, \end{cases}$$

where $(x)^+$ denotes the positive part of x. Note that $Y_i \leq X_i$. We also introduce the following four technical lemmata

Lemma 16. *Let $a > 0$ and let $Z \sim \chi_1^2$. Then, for all positive $x \in \mathbb{R}$*

$$\mathbb{P}\left(Y_i \geq a+x | Y_i \geq a\right) \geq \mathbb{P}\left(Z > a+x | Z \geq a\right).$$

Lemma 17. *Let $Z_i \overset{i.i.d.}{\sim} \chi_1^2$ for $1 \leq i \leq k$ and $a > 0$. Then for all $t \in R$*

$$\mathbb{P}\left(\sum_{i=1}^{k}(Z_i-a)|(Z_i>a) < k\mathbb{P}\left(\chi_1^2>a\right)^{-1}\mathbb{E}\left((Z-a)^+\right) - 2\sqrt{k\mathbb{P}\left(\chi_1^2>a\right)^{-1}\left(\mathbb{P}\left(\chi_1^2>a\right)+2af(a)\right)t}\right) < e^{-t}$$

Lemma 18. *Let a_k be defined implicitly as $\mathbb{P}(\chi_1^2 > a_k) = \frac{k}{p}$ and let $f(\cdot)$ denote the probability density function of the χ_1^2 distribution. Then*

$$p\mathbb{E}\left((\chi_1^2 - a_k)^+\right) + ka_k = k + 2pa_k f(a_k) \leq 2k + 2k\log(p/k)$$

Lemma 19. *For all $b > 0$:*

$$\mathbb{E}\left((\chi_1^2 - b)^+ | \chi_1^2 > b\right) > 1.$$

Next write $b = 8r_p \log(p)$ and let $\tilde{k}$ be an integer such that both $p\mathbb{P}(\chi_1^2 > b) \leq \tilde{k} \leq p$ and $a_{\tilde{k}} < b$. Note that since $r_p < \frac{1}{4}$, we have $b \leq 2\log(p)$ and such a $\tilde{k}$ is guaranteed to exist for sufficiently large values of p. For convenience, write $\tilde{\mu} = \mathbb{E}((\chi_1^2 - a_{\tilde{k}})^+)$. Using the shorthand $\psi' = \psi + \log(p)$, the following holds:

$$\mathbb{P}\left(\exists m : \sum_{i=1}^{m} X_{(i)} \geq \alpha + \sum_{i=1}^{m} \boldsymbol{\beta}_i\right) \geq \mathbb{P}\left(\sum_{i=1}^{\tilde{k}} Y_{(i)} \geq 2\psi' + \tilde{k}a_{\tilde{k}} + p\tilde{\mu} + 2\sqrt{(\tilde{k}a_{\tilde{k}} + p\tilde{\mu})\psi'}\right)$$
$$\geq \mathbb{P}\left(\sum_{i=1}^{\tilde{k}} Y_{(i)} \geq 2\psi' + \tilde{k}a_{\tilde{k}} + p\tilde{\mu} + 2\sqrt{(\tilde{k}a_{\tilde{k}} + p\tilde{\mu})\psi'} \middle| \sum_{i=1}^{p} \mathbb{I}(Y_i > b) \geq \tilde{k}\right) \mathbb{P}\left(\sum_{i=1}^{p} \mathbb{I}(Y_i > b) \geq \tilde{k}\right),$$

where the first inequality follows from substituting the third penalty regime (using the equality from Lemma 18) and the second inequality follows from conditioning on the number of Y_i exceeding b. Next note that

$$\mathbb{P}\left(\sum_{i=1}^{\tilde{k}} Y_{(i)} \geq 2\psi' + \tilde{k}a_{\tilde{k}} + p\tilde{\mu} + 2\sqrt{(\tilde{k}a_{\tilde{k}} + p\tilde{\mu})\psi'} \middle| \sum_{i=1}^{p} \mathbb{I}(Y_i > b) \geq \tilde{k}\right)$$
$$\geq \mathbb{P}\left(\sum_{i=1}^{\tilde{k}} Y_{(i)} \geq 2\psi' + \tilde{k}a_{\tilde{k}} + p\tilde{\mu} + 2\sqrt{(\tilde{k}a_{\tilde{k}} + p\tilde{\mu})\psi'} \middle| \sum_{i=1}^{p} \mathbb{I}(Y_i > b) = \tilde{k}\right)$$
$$= \mathbb{P}\left(\sum_{i=1}^{\tilde{k}} (Y_i - b)^+ \geq 2\psi' - \tilde{k}(b - a_{\tilde{k}}) + p\tilde{\mu} + 2\sqrt{(\tilde{k}a_{\tilde{k}} + p\tilde{\mu})\psi'} \middle| Y_1, ..., Y_{\tilde{k}} > b\right)$$

Let $Z_1, ..., Z_{\tilde{k}}$ be i.i.d. χ_1^2 distributed. Lemma 16 then implies that the above exceeds

$$\mathbb{P}\left(\sum_{i=1}^{\tilde{k}} (Z_i - b)^+ \geq 2\psi' - \tilde{k}(b - a_{\tilde{k}}) + p\tilde{\mu} + 2\sqrt{(\tilde{k}a_{\tilde{k}} + p\tilde{\mu})\psi'} \middle| Z_1, ..., Z_{\tilde{k}} > b\right).$$

Using the inequality in Lemma 18 and the fact that $\psi < \log(p)$ for sufficiently large values of p we can further bound the above by

$$\mathbb{P}\left(\sum_{i=1}^{\tilde{k}}(Z_i-b)^+ \geq 4\log(p)+p\tilde{\mu}-\tilde{k}\left(b-a_{\tilde{k}}\right)+2\sqrt{4(\tilde{k}+\tilde{k}\log(p/\tilde{k}))\log(p)}\middle| Z_1,...,Z_{\tilde{k}}>b\right).$$

Defining $W_i=(Z_i-b)\,|(Z_i>b)$, we can further bound the above by

$$\mathbb{P}\left(\sum_{i=1}^{\tilde{k}}W_i \geq p\tilde{\mu}-\tilde{k}\left(b-a_{\tilde{k}}\right)+8\sqrt{\tilde{k}\log(p)^2}\right), \tag{B.2.1}$$

provided p is large enough. Next, note that since $b \geq a_{\tilde{k}}$

$$\tilde{\mu}=\mathbb{E}\left[\left(\chi_1^2-a_{\tilde{k}}\right)^+\right] \leq \mathbb{E}\left[\left(\chi_1^2-b\right)^+\right]+\mathbb{P}\left(\chi_1^2>a_{\tilde{k}}\right)\left(b-a_{\tilde{k}}\right).$$

Consequently, we can bound (B.2.1) by

$$\begin{aligned}
&\mathbb{P}\left(\sum_{i=1}^{\tilde{k}}W_i \geq p\mathbb{E}\left[\left(\chi_1^2-b\right)^+\right]+8\sqrt{\tilde{k}\log(p)^2}\right)\\
&=\mathbb{P}\left(\sum_{i=1}^{\tilde{k}}\left[W_i-\mathbb{E}\left(\chi_1^2-b|\chi_1^2>b\right)\right] \geq \left(p\mathbb{P}\left(\chi_1^2>b\right)-\tilde{k}\right)\mathbb{E}\left(\chi_1^2-b|\chi_1^2>b\right)+8\sqrt{\tilde{k}\log(p)^2}\right)\\
&\geq 1-\exp\left(-\frac{\left[\left((\tilde{k}-p\mathbb{P}\left(\chi_1^2>b\right))\mathbb{E}\left((\chi_1^2-b)^+|\chi_1^2>b\right)-8\sqrt{\tilde{k}\log(p)^2}\right)^+\right]^2}{4\tilde{k}\mathbb{P}\left(\chi_1^2>b\right)^{-1}\left(\mathbb{P}\left(\chi_1^2>b\right)+2bf(b)\right)}\right),
\end{aligned}$$

where the inequality follows from Lemma 17. Given Lemma 19 and the fact that Lemma 18 implies that $\mathbb{P}\left(\chi_1^2>b\right)+2bf(b)<2\mathbb{P}\left(\chi_1^2>b\right)(1+\log(p))$, we can further bound the above by

$$1-\exp\left(-\frac{\left[\left((\tilde{k}-p\mathbb{P}\left(\chi_1^2>b\right))-8\sqrt{\tilde{k}}\log(p)\right)^+\right]^2}{8(\tilde{k}(1+\log(p)))}\right).$$

The arithmetic-mean-geometric-mean-inequality can be used to show that $((a-b)^+)^2 > \frac{\left((a)^+\right)^2}{2}-4b^2$. The above quantity is therefore bounded by

$$1-\exp\left(-\frac{1}{16(1+\log(p))}\left(\left[\frac{\tilde{k}-p\mathbb{P}\left(\chi_1^2>b\right)}{\sqrt{\tilde{k}}}\right]^+\right)^2+72\log(p)\right).$$

Note that if $\tilde{k} \geq p\mathbb{P}\left(\chi_1^2 > b\right) + p^{\frac{1}{2}-2r_p+\delta}$ for some $\delta > 0$, then

$$\frac{\tilde{k} - p\mathbb{P}\left(\chi_1^2 > b\right)}{\sqrt{\tilde{k}}} \geq \frac{p^{\frac{1}{2}-2r_p+\delta}}{\sqrt{p\mathbb{P}\left(\chi_1^2 > b\right) + p^{\frac{1}{2}-2r_p+\delta}}} \geq \frac{p^{\frac{1}{2}-2r_p+\delta}}{\sqrt{p^{1-4r_p} + p^{\frac{1}{2}-2r_p+\delta}}} \geq \frac{1}{2}p^{\frac{\delta}{2}},$$

with the first inequality following from the fact that the left-hand side is increasing in $\tilde{k}$, the second one following from the fact that $\mathbb{P}\left(\chi_1^2 > b\right) < \mathbb{P}\left(\chi_2^2 > b\right) = p^{-4r_p}$ and the last one following from the fact that $r_p < \frac{1}{4}$.

Consequently, it is sufficient to show that there exists a $\delta > 0$ such that

$$\mathbb{P}\left(\sum_{i=1}^{p} \mathbb{I}\left(Y_i \geq b\right) > p\mathbb{P}\left(\chi_1^2 > b\right) + p^{\frac{1}{2}-2r_p+\delta}\right) \to 1 \quad \text{as} \quad p \to \infty$$

This can be seen from the fact that $\sum_{i=1}^{p} \mathbb{I}\left(Y_i > b\right)$ is $Bin(p, q)$-distributed with

$$q = \mathbb{P}\left(Y_i > 8r_p \log(p)\right) = (1 - p^{-\xi})\mathbb{P}\left(\chi_1^2 > b\right) + p^{-\xi}\mathbb{P}\left(N(0,1) > \sqrt{2r_p \log(p)}\right).$$

Note that

$$q > \mathbb{P}\left(\chi_1^2 > b\right) - p^{-\xi-4r_p} + p^{-\xi}\mathbb{P}\left(N(0,1) > \sqrt{2r_p \log(p)}\right),$$

since $\mathbb{P}\left(\chi_1^2 > b\right) < p^{-4r_p}$. Moreover,

$$q < \mathbb{P}\left(\chi_1^2 > b\right) + p^{-\xi}\mathbb{P}\left(N(0,1) > \sqrt{2r_p \log(p)}\right) \leq p^{-4r_p} + p^{-\xi-r_p},$$

by standard tail bounds of the normal distribution and the definition of b. Standard Hoeffding bounds show that

$$\mathbb{P}\left(\sum_{i=1}^{p} \mathbb{I}\left(Y_i > b\right) > pq - \sqrt{pq \log(p)}\right) \to 1 \quad as \quad p \to \infty$$

Hence,

$$\mathbb{P}\left(\sum_{i=1}^{p} \mathbb{I}\left(Y_i > b\right) > p\mathbb{P}\left(\chi_1^2 > b\right) + p^{1-\xi}\mathbb{P}\left(N(0,1) > \sqrt{2r_p \log(p)}\right) - p^{1-\xi-4r_p} - \sqrt{p(p^{-4r_p} + p^{-\xi-r_p}) \log(p)}\right)$$

converges to 1 as $p \to \infty$. Note that

$$p^{1-\xi}\mathbb{P}\left(N(0,1) > \sqrt{2r_p \log(p)}\right) - p^{1-\xi-4r_p} - \sqrt{p(p^{-4r_p} + p^{-\xi-r_p})\log(p)} > p^{\frac{1}{2}-2r_p+\delta}$$

for all δ such that $r_p - \xi + \frac{1}{2} > \delta$, provided p is large enough. This follows from the fact that

$$p^{1-\xi}\mathbb{P}\left(N(0,1) > \sqrt{2r_p \log(p)}\right) > \frac{1}{\sqrt{2\pi}} p^{1-\xi-r_p} \frac{\sqrt{2r_p \log(p)}}{1 + 2r_p \log(p)},$$

by standard tail bounds on the normal distribution. Since $0 < r_p$, the the above dominates $p^{1-\xi-4r_p}$ as p increases. Similarly, because $r_p - \xi + \frac{1}{2} > 0$, it dominates $\sqrt{p^{1-4r_p}}$, since, $r_p < \frac{1}{4}$ and $\xi < \frac{3}{4}$, $1 - r_p - \xi > 0$ and the above therefore also dominates $\sqrt{p^{1-r_p-\xi}}$. Finally, if δ is such that $r_p - \xi + \frac{1}{2} > \delta$ it must also dominate $p^{\frac{1}{2}-2r_p+\delta}$. This finishes the proof.

B.2.3 Proof of Proposition 6

Standard tail bounds on the subgaussian distribution give that for all $\aleph > 0$

$$\mathbb{P}\left(\left(\mathbf{x}_t^{(i)-\mu_i}\right)^2 < 2\aleph\right) \geq 1 - Ae^{-\aleph/\lambda}$$

holds for a constant A under the null hypothesis. A Bonferroni correction therefore gives $\mathbb{P}\left(\hat{O} = \emptyset\right) \geq 1 - Anp\exp(-\frac{1}{2\lambda}\beta')$.

B.2.4 Proof of Propositions 17 and 19

We give the proof of Proposition 19, as Proposition 17 corresponds to as special case. We write

$$\mathcal{S}(s, e, \mathbf{d}, \mathbf{f}, \mathbf{J}) = \sum_{i \in \mathbf{J}} \left(\mathcal{C}_i\left(\mathbf{x}^{(i)}_{(s+1+\mathbf{d}^{(i)}):(e-\mathbf{f}^{(i)})}, \boldsymbol{\theta}_0^{(i)}\right) - \min_{\boldsymbol{\theta}}\left[\mathcal{C}_i\left(\mathbf{x}^{(i)}_{(s+1+\mathbf{d}^{(i)}):(e-\mathbf{f}^{(i)})}, \boldsymbol{\theta}\right)\right]\right) - \alpha - \sum_{i=1}^{|\mathbf{J}|} \boldsymbol{\beta}_i$$

and note that

$$\mathcal{S}(t,m) = \max_{\mathbf{d},\mathbf{f},\mathbf{J}:\ m-t-d-f\geq l} [\mathcal{S}(t,m,\mathbf{d},\mathbf{f},\mathbf{J})]$$

The proof of Proposition 19 is then a corollary of the observation that for all $\mathbf{d},\mathbf{f} < w$

$$\begin{aligned}
\mathcal{S}(t,m',\mathbf{d},\mathbf{f},\mathbf{J}) &= \sum_{i\in\mathbf{J}} \left(\mathcal{C}_i\left(\mathbf{x}^{(i)}_{(t+1+\mathbf{d}^{(i)}):(m'-\mathbf{f}^{(i)})}, \boldsymbol{\theta}_0^{(i)}\right) - \min_{\boldsymbol{\theta}} \left[\mathcal{C}_i\left(\mathbf{x}^{(i)}_{(t+1+\mathbf{d}^{(i)}):(m'-\mathbf{f}^{(i)})}, \boldsymbol{\theta}\right)\right]\right) - \alpha - \sum_{i=1}^{|\mathbf{J}|}\boldsymbol{\beta}_i \\
&\leq \sum_{i\in\mathbf{J}} \left(\mathcal{C}_i\left(\mathbf{x}^{(i)}_{(t+1+\mathbf{d}^{(i)}):(m'-\mathbf{f}^{(i)})}, \boldsymbol{\theta}_0^{(i)}\right) - \min_{\boldsymbol{\theta}} \left[\mathcal{C}_i\left(\mathbf{x}^{(i)}_{(t+1+\mathbf{d}^{(i)}):m}, \boldsymbol{\theta}\right)\right] - \min_{\boldsymbol{\theta}} \left[\mathcal{C}_i\left(\mathbf{x}^{(i)}_{(m+1):(m'-\mathbf{f}^{(i)})}, \boldsymbol{\theta}\right)\right]\right) \\
&\qquad\qquad - \alpha - 2\sum_{i=1}^{|\mathbf{J}|}\boldsymbol{\beta}_i + \sum_{i=1}^{p}\boldsymbol{\beta}_i \\
&= \mathcal{S}(t,m,\mathbf{d},0,\mathbf{J}) + \mathcal{S}(m,m',0,\mathbf{f},\mathbf{J}) + \sum_{i=1}^{p}\boldsymbol{\beta}_i + \alpha,
\end{aligned}$$

since $m-t-w \geq l$ and $m'-m-w \geq l$. Indeed, the above inequality shows that

$$\begin{aligned}
&S(t) + \max_{\mathbf{d}\leq w,\mathbf{f}\leq w,\mathbf{J}:\ m-t-d-f\geq l} [\mathcal{S}(t,m',\mathbf{d},\mathbf{f},\mathbf{J})] \\
&\leq S(t) + \max_{\mathbf{d}\leq w,\mathbf{f}\leq w,\mathbf{J}:\ m-t-d-f\geq l} [\mathcal{S}(t,m,\mathbf{d},\mathbf{0},\mathbf{J}) + \mathcal{S}(m,m',\mathbf{0},\mathbf{f},\mathbf{J})] + \alpha + \sum_{i=1}^{p}\boldsymbol{\beta}_i \\
&\leq S(t) + \max_{\mathbf{d}\leq w,\mathbf{J}:\ m-t-d\geq l} [\mathcal{S}(t,m,\mathbf{d},\mathbf{0},\mathbf{J})] + \max_{\mathbf{f}\leq w,\mathbf{J}:\ m-t-f\geq l} [\mathcal{S}(m,m',\mathbf{0},\mathbf{f},\mathbf{J})] + \alpha + \sum_{i=1}^{p}\boldsymbol{\beta}_i \\
&\leq S(t) + \max_{\mathbf{d}\leq w,\mathbf{f}\leq w,\mathbf{J}:\ m-t-d-f\geq l} [\mathcal{S}(t,m,\mathbf{d},\mathbf{f},\mathbf{J})] + \max_{\mathbf{d}\leq w,\mathbf{f}\leq w,\mathbf{J}:\ m-t-d-f\geq l} [\mathcal{S}(m,m',\mathbf{d},\mathbf{f},\mathbf{J})] + \alpha + \sum_{i=1}^{p}\boldsymbol{\beta}_i \\
&< S(m) + \max_{\mathbf{d},\mathbf{f},\mathbf{J}:\ m-t-d-f\geq l} [\mathcal{S}(m,m',\mathbf{d},\mathbf{f},\mathbf{J})] \leq S(m').
\end{aligned}$$

This finishes the proof.

B.2.5 Proof of Proposition 18

We prove the more general case of the savings being stochastically bounded by $a\chi^2_v + b$.

The result follows from a Bonferroni correction:

$$\mathbb{P}\left(\mathcal{S}_i\left(s,e\right)>y\right) \leq \sum_{d=0}^{w}\sum_{f=0}^{w}\mathbb{P}\left(\mathcal{C}_i\left(\mathbf{x}^{(i)}_{(s+1+d):(e-f},\boldsymbol{\theta}^{(i)}_0\right)-\min_{\boldsymbol{\theta}}\left(\mathcal{C}_i\left(\mathbf{x}^{(i)}_{(s+1+d):(e-f)},\boldsymbol{\theta}\right)\right)>y\right)$$
$$\leq \sum_{d=0}^{w}\sum_{f=0}^{w}\mathbb{P}\left(a\chi^2_v+b>y\right)=(w+1)^2\mathbb{P}\left(a\chi^2_v+b>y\right)$$

B.2.6 Proof of Proposition 7

The proof of Proposition 7 follows almost directly from the following Lemma:

Lemma 20. *Let* $\eta_t \overset{i.i.d.}{\sim} N(0,1)$ *for* $i \leq t \leq j$. *Then*

$$\mathbb{P}\left(\max_{0\leq j,d\leq w:j-f-d-i\geq 0}\left((j-f-d-i+1)\left(\bar{\boldsymbol{\eta}}^{(c)}_{(i+d):(j-f)}\right)^2\right)>u\right)\leq 6(w+1)\frac{1+\log(w+1)}{\log(b)}e^{-\frac{u}{2b}}.$$

for all $b\in\mathbb{R}$ *such that* $1<b\leq 2$.

Setting $b=1+\epsilon$, we can therefore see that

$$\max_{0\leq j,d\leq w:j-f-d-i\geq 0}\left((j-f-d-i+1)\left(\bar{\boldsymbol{\eta}}^{(c)}_{(i+d):(j-f)}\right)^2\right) \tag{B.2.2}$$

is stochastically bounded by

$$(1+\epsilon)\chi^2_2+2(1+\epsilon)\left(\log\left(w+1\right)+\log(6)-\log(\log(1+\epsilon))+\log(1+\log(w+1))\right)$$

B.2.7 Proof of Proposition 8

Let $\delta=r_p-(1-\sqrt{1-\xi})^2>0$. Then Penalty regime 2' with $\epsilon=\frac{\delta}{2}$ and $\psi=3\log(n)$ controls false positives. Given MVCAPA examines all possible lags up to w, we can bound the power by the probability that the test statistic for the true collective anomaly with true lags for each anomalous series is greater than the threshold for the test. Thus it is sufficient to show that

$$\mathbb{P}\left(\max_i\left(\left(\sqrt{e-s-w}\bar{\mathbf{x}}^{(i)}_{(s+1+d_i):(e-f_i)}\right)^2\right)>2(1+\epsilon)\psi+2(1+\epsilon)\log(p)+2(1+\epsilon)\log(w+1)\right)$$

goes to 1 as $p \to \infty$. This holds by a very similar argument as case 2 in the proof of Proposition 5 since

$$\sqrt{e-s-w}\bar{\mathbf{x}}^{(i)}_{(s+1+d_i):(e-f_i)} \overset{i.i.d.}{\sim} N(\sqrt{2r_p \log(p(w+1))}, 1).$$

B.2.8 Proof of Theorem 3

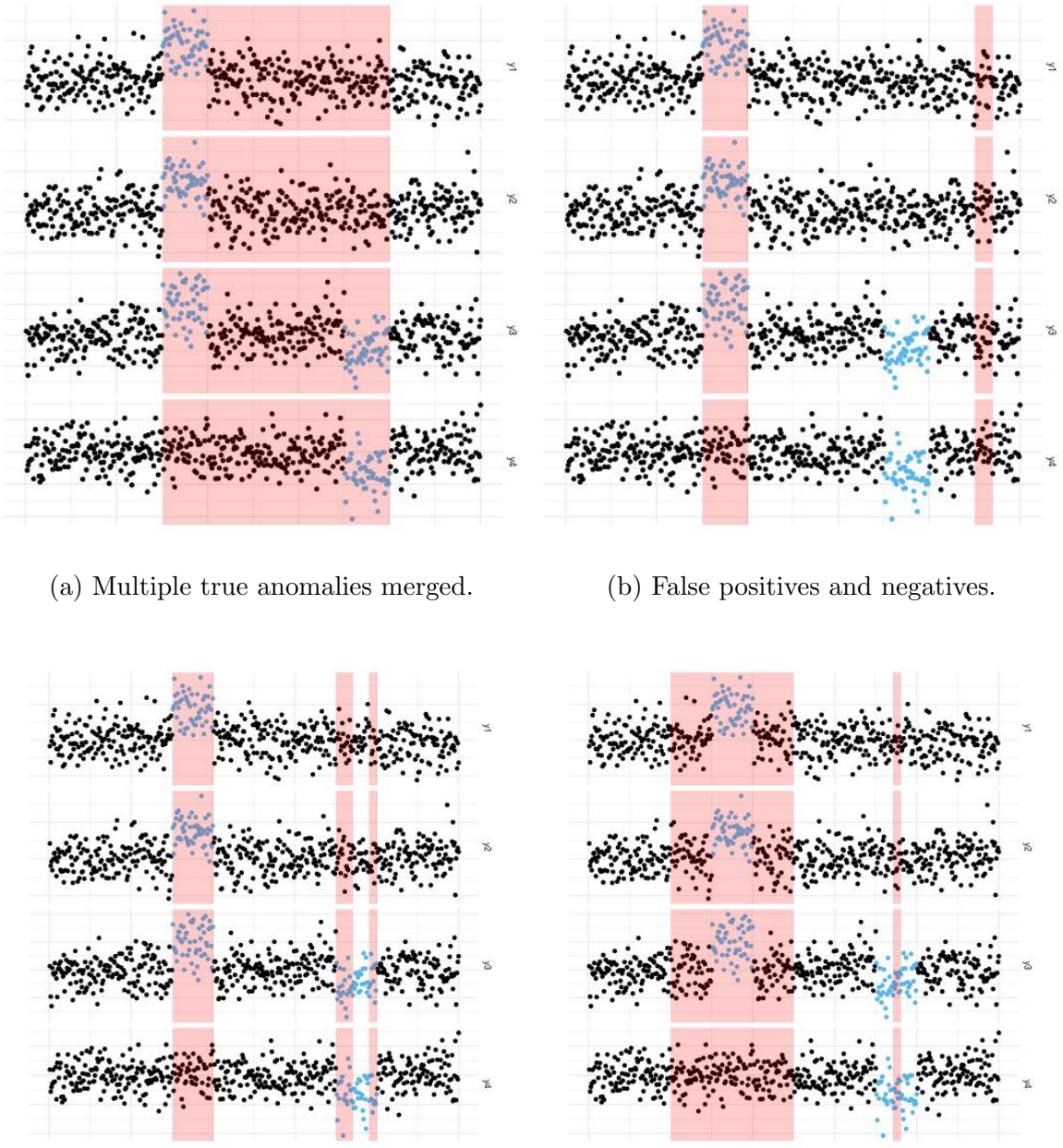
We define the penalised cost of a segment $\mathbf{x}_{i:j}$ under a partition $\tau = \{\hat{\tau}_1, ..., \hat{\tau}_{\hat{K}}\}$, where $\hat{\tau}_k = (\hat{s}_k, \hat{e}_k, \hat{\mathbf{J}}_k)$ to be

$$\mathcal{C}\left(\mathbf{x}_{i:j}, \hat{\tau}\right) = \sum_{k=1}^{\hat{K}} \left[\mathcal{C}(\mathbf{x}_{(\hat{s}_k+1):\hat{e}_k}, \hat{\mathbf{J}}_k)\right].$$

Here the penalised cost of introducing the kth anomalous window is

$$\begin{aligned}
&\mathcal{C}\left(\mathbf{x}_{(s+1):e}, \{(s, e, \mathbf{J})\}\right) \\
&= \mathcal{C}(\mathbf{x}_{(s+1):e}, \mathbf{J}) := -\mathcal{S}(\mathbf{x}_{(s+1):e}, \mathbf{J}) + \sum_{i=1}^{|\mathbf{J}|} \boldsymbol{\beta}_{\boldsymbol{i}\cdot} := -(e-s)\sum_{i \in \mathbf{J}} \mathcal{C}\left(\bar{\mathbf{x}}^{(i)}_{(s+1):e}\right)^2 + \sum_{i=1}^{|\mathbf{J}|} \boldsymbol{\beta}_{\boldsymbol{i}},
\end{aligned}$$

where $\mathcal{S}(\mathbf{x}_{(s+1):e}, \mathbf{J})$, is defined as the saving made by fitting the segment $\mathbf{x}_{(s+1):e}$ with $\mathbf{J}$ and $\bar{\mathbf{x}}^{(i)}_{(s+1):e} := (e-s)^{-1}\sum_{t=s+1}^{e} \mathbf{x}^i_t$ is defined as the arithmetic mean of the ith component from time $t = s+1$ to $t = e$. It should be noted that minimising the penalised cost, is equivalent to maximising the penalised saving. We call the partition which minimises the penalised cost, $\mathcal{C}\left(\mathbf{x}_{1:n}, \hat{\tau}\right)$, over all feasible partitions, $\hat{\tau}$, the optimal partition.



(a) Multiple true anomalies merged.

(b) False positives and negatives.

(c) Anomaly fitted using multiple segments

(d) Bad fit to true anomalies

Figure B.2.1: Examples of the four ways a fitted partition (in red) can be outside the set of good partitions, $\mathcal{B}_C$, defined in Equation (B.2.3). True anomalies are indicated in blue.

We also define the following event sets over all pairs i, j such that $1 \leq i \leq j \leq n$

$$E_1 := \left\{ \sum_{c \in S} (j-i+1) \left(\bar{\boldsymbol{\eta}}_{i:j}^{(c)} \right)^2 < 2\psi + 2|S| \log(p) \quad \forall S \subset \{1, ..., p\} \right\}$$

$$E_2 := \left\{ \sum_{c \in S} (j-i+1) \left(\bar{\boldsymbol{\eta}}_{i:j}^{(c)} \right)^2 < p + 2\psi + 2\sqrt{p\psi} \quad \forall S \subset \{1, ..., p\} \right\}$$

$$E_3 := \left\{ \sum_{c=1}^{p} (j-i+1) \left(\bar{\boldsymbol{\eta}}_{i:j}^{(c)} + \bar{\boldsymbol{\mu}}_{i:j}^{(c)} \right)^2 > p - 2\sqrt{p\psi} \right\}$$

$$E_4 := \left\{ \left| \sum_{c \in S} \sqrt{j-i+1} \bar{\boldsymbol{\eta}}_{i:j}^{(c)} \right| < \sqrt{2|S|\psi + 2|S|^2 \log(p)} \quad \forall S \subset \{1, ..., p\} \right\}$$

$$E_5 := \left\{ \sum_{c \notin S} (j-i+1) \left(\bar{\boldsymbol{\eta}}_{i:j}^{(c)} + \bar{\boldsymbol{\mu}}_{i:j}^{(c)} \right)^2 > p - 2\sqrt{p\psi} - 2\psi - 2|S| \log(p) \quad \forall S \subset \{1, ..., p\} \right\}$$

$$E_6 := \left\{ \frac{\left| \sum_{c \in S} \left(\sum_{t=i}^{j} \left(\boldsymbol{\mu}_t^{(c)} - \bar{\boldsymbol{\mu}}_{i:j}^{(c)} \right) \boldsymbol{\eta}_t^{(c)} \right) \right|}{\sqrt{\sum_{c \in S} \sum_{t=i}^{j} \left(\boldsymbol{\mu}_t^{(c)} - \bar{\boldsymbol{\mu}}_{i:j}^{(c)} \right)^2}} \leq \sqrt{2\psi + 2\left|S \cap W_{i,j}\right| \log(p)} \quad \forall S \subset \{1, ..., p\} \right\}$$

$$E_7 := \left\{ \sum_{c \in S} \frac{(j-j')(j'-i+1)}{j-i+1} \left(\bar{\boldsymbol{\eta}}_{i:j'}^{(c)} - \bar{\boldsymbol{\eta}}_{(j'+1):j}^{(c)} \right)^2 < 2\psi + 2|S| \log(p) \quad \forall S \subset \{1, ..., p\} \right\}$$

$$E_8 := \left\{ \sum_{c \in S} \frac{(j-j')(j'-i+1)}{j-i+1} \left(\bar{\boldsymbol{\eta}}_{i:j'}^{(c)} - \bar{\boldsymbol{\eta}}_{(j'+1):j}^{(c)} \right)^2 < p + 2\psi + 2\sqrt{p\psi} \quad \forall S \subset \{1, ..., p\} \right\}$$

$$E_9 := \left\{ \sum_{c=1}^{p} \frac{(j-j')(j'-i+1)}{j-i+1} \left(\bar{\mathbf{x}}_{i:j'}^{(c)} - \bar{\mathbf{x}}_{(j'+1):j}^{(c)} \right)^2 > p - 2\sqrt{p\psi} \right\}$$

$$E_{10} := \left\{ \left| \sum_{c \in \mathbf{J}_k} \sqrt{j-i+1} \bar{\boldsymbol{\eta}}_{i:j}^{(c)} \right| < \sqrt{2|\mathbf{J}_k|\psi} \quad \forall i, j : \exists k \in \{1, ..., K\} : e_{k-1} < i, \;\; j \leq s_{k+1} \right\}$$

$$E_{11} := \left\{ \frac{\left| \sum_{c \in \mathbf{J}_k} \sqrt{\frac{(j-e_k)(e_k-i+1)}{j-i+1}} \left(\bar{\boldsymbol{\eta}}_{i:e_k}^{(c)} - \bar{\boldsymbol{\eta}}_{(e_k+1):j}^{(c)} \right) \right|}{\sqrt{2|\mathbf{J}_k|\psi}} < 1 \quad \forall i, j : \exists k \in \{1, ..., K\} : e_{k-1} < i, \;\; j \leq s_{k+1} \right\},$$

where the set $W_{i,j}$ of components with non constant mean in the interval $[i, j]$ is defined as

$$W_{i,j} = \left\{ c \in \{1, ..., p\} : \exists t \in [i, j-1] : \boldsymbol{\mu}_t^{(c)} \neq \boldsymbol{\mu}_{t+1}^{(c)} \right\}.$$

The intuition behind these events is as follows: Events E_1 and E_2 bound the saving obtained from fitting an anomalous region on data belonging to the typical distribution and so ensure no false positives are fitted. Events E_7, E_8, E_9, and E_{11} provide bounds on the additional un-penalised cost of splitting a fitted segment in two or merging

two existing segments, assuring that anomalous regions are fitted by one rather than multiple adjacent segments. They are assisted by events E_3 and E_5 which bound the additional un-penalised cost incurred by fitting any given segment by a dense change, extending the result to showing the sub-optimality of a collective anomaly being fitted by multiple non-adjacent segments. Events E_4, E_6, and E_{10} bound the interaction between the signal and the noise thus ensuring that anomalous regions are detected. For brevity, we denote $E = \cap E_i$ and note that it occurs with high probability. Indeed, the following Lemma holds:

Lemma 21. *There exists a constant A such that*

$$\mathbb{P}(E) > 1 - An^3 e^{-\psi}$$

We now define the set of good partitions $\mathcal{B}_C$ to be

$$\mathcal{B}_C = \left\{ \tau : |\tau| = K, \ \ |\hat{s}_k - s_k| \leq \frac{10C}{\triangle_k^2} \ \ |\hat{e}_k - e_k| \leq \frac{10C}{\triangle_k^2} \right\}. \tag{B.2.3}$$

It is sufficient to prove the following proposition in order to prove Theorem 3

Proposition 20. *Let the assumptions of Theorem 1 hold. Given E holds and C exceeds a global constant, the partition τ_0 minimising the penalised cost $\mathcal{C}(\boldsymbol{x}_{1:n}, \tau)$ satisfies $\tau_0 \in \mathcal{B}_C$*

The main ideas of the proof of Proposition 20 are that given E:

I Each fitted anomalous segment overlaps with at most one true anomalous segment; this excludes the situation depicted in Figure B.2.1a.

II Each fitted anomalous segment overlaps with at least one true anomalous region; this excludes the situation depicted in Figure B.2.1b.

III Each true anomalous segment overlaps with at most one fitted anomalous region, i.e. there exists a bijection between fitted and true segments; this excludes the situation depicted in Figure B.2.1c.

IV Each fitted anomalous segment is close (in the sense of $\mathcal{B}_C$) to the true segment it fits; this excludes the situation depicted in Figure B.2.1d.

We will prove these properties which exclude the various types of poor partitions in Figure B.2.1 in the following order: First we will prove property II, then IV, then III, and then I. We will then use these to prove Proposition 20. In the subsequent proofs we will use a certain number of technical Lemmata, all proved in Section B.3.

Throughout these proofs we will use the following two lemmata. The first one describes the increase in un-penalised cost incurred by splitting a fitted segment into two fitted segments and the second one bounds this increase in penalised cost for splitting fitted dense collective anomalies.

Lemma 22. *Let $i \leq j' < j'+1 \leq j$. The following property is satisfied for all $\boldsymbol{J}$*

$$\mathcal{S}\left(\boldsymbol{x}_{i:j'}, \boldsymbol{J}\right) + \mathcal{S}\left(\boldsymbol{x}_{(j'+1):j}, \boldsymbol{J}\right) = \mathcal{S}\left(\boldsymbol{x}_{i:j}, \boldsymbol{J}\right) + \sum_{c \in \boldsymbol{J}} \left(\frac{(j'-i+1)(j-j')}{j-i+1} \left(\bar{\boldsymbol{x}}^{(c)}_{i:j'} - \bar{\boldsymbol{x}}^{(c)}_{(j'+1):j} \right)^2 \right)$$

Lemma 23. *Let $i \leq j' < j'+1 \leq j$ The following holds given E*

$$\mathcal{C}\left(\boldsymbol{x}_{i:j'}, \boldsymbol{1}\right) + \mathcal{C}\left(\boldsymbol{x}_{(j'+1):j}, \boldsymbol{1}\right) \leq \mathcal{C}\left(\boldsymbol{x}_{i:j}, \boldsymbol{1}\right) + C\psi + C\sqrt{p\psi} + 2\sqrt{p\psi},$$

provided C exceeds some global constant.

We will also use the following lemma which shows that merging two adjacent fitted collective anomalies which are both contained within a true anomalous segment reduces the penalised cost substantially.

Lemma 24. *Let i, j', and j be such that there exists a k such that $s_k < i \leq j' < j' + 1 \leq j \leq e_k$. Then,*

$$\mathcal{C}\left(x_{i:j}, \boldsymbol{J}_k\right) \leq \mathcal{C}\left(x_{i:j'}, \boldsymbol{J}_k\right) + \mathcal{C}\left(x_{(j'+1):j}, \boldsymbol{J}_k\right) - \frac{79}{80} C\left(\psi + |\boldsymbol{J}_k| \log(p)\right)$$

and

$$\mathcal{C}\left(x_{i:j}, \boldsymbol{1}\right) \leq \mathcal{C}\left(x_{i:j'}, \boldsymbol{1}\right) + \mathcal{C}\left(x_{(j'+1):j}, \boldsymbol{1}\right) - \frac{79}{80} C\left(\psi + \sqrt{p\psi}\right)$$

when $|\boldsymbol{J}_k| \leq k^*$ and $|\boldsymbol{J}_k| > k^*$ respectively , provided C exceeds some global constant and the event E holds.

The proof of part IV will mostly rely on the following three lemmata. The first one shows that fitting a true collective anomaly as anomalous reduces the penalised cost. The second and third one show that if a fitted sparse or dense collective anomaly contains a large number of observations both from a true anomalous segment and from a typical segment, then removing the typical data from the fitted anomaly reduces the penalised cost.

Lemma 25. *Let i and j be such that there exists a k such that $s_k < i \leq j \leq e_k$. Moreover assume that*

$$j - i + 1 \geq \frac{4C}{\triangle_k^2}.$$

Then given E

$$\mathcal{C}\left(\boldsymbol{x}_{i:j}, \boldsymbol{J}_k\right) < 0$$

holds if the kth anomalous window is sparse; i.e. if $|\boldsymbol{J}_k| \leq k^*$; and

$$\mathcal{C}\left(\boldsymbol{x}_{i:j}, \boldsymbol{1}\right) < 0$$

holds if the kth anomalous window is dense; i.e. if $|\boldsymbol{J}_k| > k^*$*; provided* C *exceeds some global constant and the event* E *holds.*

Lemma 26. *Let* i *and* j *be such that there exists a* k *such that either of the following holds:*

1. $s_k < i \leq j \leq s_{k+1}$ *and*

$$\min(e_k - i + 1, j - e_k) \geq \frac{10C}{\triangle_k^2}.$$

2. $e_{k-1} < i \leq j \leq e_k$ *and*

$$\min(s_k - i + 1, j - s_k) \geq \frac{10C}{\triangle_k^2}.$$

Then the corresponding holds given E

1. *if the kth anomalous window is sparse; i.e. if* $|\boldsymbol{J}_k| \leq k^*$

$$\mathcal{C}\left(x_{i:j}, \boldsymbol{J}_k\right) \geq \mathcal{C}\left(x_{i:e_k}, \boldsymbol{J}_k\right) + 6C(\psi + \log(p))$$

   if the kth anomalous window is dense; i.e. if $|\boldsymbol{J}_k| > k^*$

$$\mathcal{C}\left(x_{i:j}, \boldsymbol{1}\right) \geq \mathcal{C}\left(x_{i:e_k}, \boldsymbol{1}\right) + 6C(\psi + \sqrt{p\psi})$$

2. *if the kth anomalous window is sparse; i.e. if* $|\boldsymbol{J}_k| \leq k^*$

$$\mathcal{C}\left(x_{i:j}, \boldsymbol{J}_k\right) \geq \mathcal{C}\left(x_{(s_k+1):j}, \boldsymbol{J}_k\right) + 6C(\psi + \log(p))$$

   if the kth anomalous window is dense; i.e. if $|\boldsymbol{J}_k| > k^*$

$$\mathcal{C}\left(x_{i:j}, \boldsymbol{1}\right) \geq \mathcal{C}\left(x_{(s_k+1):j}, \boldsymbol{1}\right) + 6C(\psi + \sqrt{p\psi})$$

provided C exceeds some global constant and the event E holds.

Lemma 27. *Let i and j be such that there exists a k such that the kth anomalous window is dense, $|\boldsymbol{J}_k| > k^*$, and either of the following holds:*

1. $s_k < i \leq j \leq s_{k+1}$ *and*

$$\min(e_k - i + 1, j - e_k) \geq \frac{10C}{\triangle_k^2}.$$

2. $e_{k-1} < i \leq j \leq e_k$ *and*

$$\min(s_k - i + 1, j - s_k) \geq \frac{10C}{\triangle_k^2}.$$

Then the corresponding holds for all $\boldsymbol{J}$ given E

1.

$$\mathcal{C}\left(x_{i:j}, \boldsymbol{J}\right) \geq \mathcal{C}\left(x_{i:e_k}, \boldsymbol{1}\right) + 4C(\psi + \sqrt{p\psi})$$

2.

$$\mathcal{C}\left(x_{i:j}, \boldsymbol{J}\right) \geq \mathcal{C}\left(x_{(s_k+1):j}, \boldsymbol{1}\right) + 4C(\psi + \sqrt{p\psi})$$

provided C exceeds some global constant and the event E holds.

For Part II, we will require the following six lemmata. The first one proves that merging two fitted collective anomalies contained within a truly anomalous segment reduces the overall penalised cost substantially, even if they are non-adjacent. The second one shows that if a fitted collective anomaly contains both typical and atypical data, then the atypical data can be removed from the fitted collective anomaly without increasing the penalised cost too much. The remaining Lemmata are mostly used to

show that if a true anomaly has been fitted using the wrong set of components (i.e. fitting a sparse anomaly as a dense one, a dense anomaly as a sparse one, or a sparse anomaly as a sparse anomaly but not with the correct set of components), then it is possible to replace this fitted collective anomaly by one with the right components without increasing the overall penalised cost by too much.

Lemma 28. *Let i, j', and j be such that there exists a k such that $s_k < i \leq j' < j'' + 1 \leq j \leq e_k$. Then,*

$$\mathcal{C}\left(x_{i:j}, \boldsymbol{J}_k\right) \leq \mathcal{C}\left(x_{i:j'}, \boldsymbol{J}_k\right) + \mathcal{C}\left(x_{(j''+1):j}, \boldsymbol{J}_k\right) - \frac{19}{20} C\left(\psi + |\boldsymbol{J}_k| \log(p)\right)$$

and

$$\mathcal{C}\left(x_{i:j}, \boldsymbol{1}\right) \leq \mathcal{C}\left(x_{i:j'}, \boldsymbol{1}\right) + \mathcal{C}\left(x_{(j''+1):j}, \boldsymbol{1}\right) - \frac{19}{20} C\left(\psi + \sqrt{p\psi}\right)$$

when $|\boldsymbol{J}_k| \leq k^*$ and $|\boldsymbol{J}_k| > k^*$ respectively, provided C exceeds some global constant and the event E holds.

Lemma 29. *Let i, j be such that there exists a k such that $[s_k + 1, e_k] \cap [i, j] \neq \emptyset$, $e_{k-1} < i$, and $s_{k+1} \geq j$. Then,*

$$\mathcal{C}\left(\boldsymbol{x}_{i':j'}, \boldsymbol{J}\right) - \mathcal{C}\left(\boldsymbol{x}_{i:j}, \boldsymbol{J}\right) \leq 8\psi + 8|\boldsymbol{J}| \log(p)$$

for $|\boldsymbol{J}| \leq k^*$ and

$$\mathcal{C}\left(\boldsymbol{x}_{i':j'}, \boldsymbol{1}\right) - \mathcal{C}\left(\boldsymbol{x}_{i:j}, \boldsymbol{1}\right) \leq 8\psi + 8\sqrt{p\psi}$$

where $i' = \max(i, s_k + 1)$ and $j' = \min(j, e_k)$ both hold given E.

Lemma 30. *Let E hold and C exceed a global constant. Moreover, let i and j be such that there exists a k such that $s_k < i \leq j \leq e_k$. Then*

$$\mathcal{S}(\boldsymbol{x}_{i:j}, \boldsymbol{J}) \geq \alpha\left(C\psi + C|\boldsymbol{J}| \log(p)\right)$$

for some $\alpha > 0$ *implies that*

$$\sqrt{|\boldsymbol{J}|(j-i+1)\boldsymbol{\mu}_k^2} \geq \left(\sqrt{\alpha C} - \sqrt{2}\right)\sqrt{\psi + |\boldsymbol{J}|\log(p)}$$

for any sparse $\boldsymbol{J}$*.*

Lemma 31. *Let* i *and* j *be such that there exists a* k *such that* $s_k < i \leq j \leq e_k$*. If the* k*th anomalous window is sparse; i.e. if* $|\boldsymbol{J}_k| \leq k^*$*; and*

$$\mathcal{S}\left(\boldsymbol{x}_{i:j}, \boldsymbol{J}\right) \geq \frac{19}{20} C\left(|\boldsymbol{J}|\log(p) + \psi\right),$$

then

$$\mathcal{C}\left(\boldsymbol{x}_{i:j}, \boldsymbol{J}_k\right) - \mathcal{C}\left(\boldsymbol{x}_{i:j}, \boldsymbol{J}\right) \leq \frac{1}{10} C|\boldsymbol{J}_k|\log(p) + 2\psi$$

holds for all sparse $\boldsymbol{J}$*, i.e.* $\boldsymbol{J}$ *satisfying* $|\boldsymbol{J}| \leq k^*$*, if* C *is larger than some global constant and the event* E *holds.*

Lemma 32. *Let* i *and* j *be such that there exists a* k *such that* $s_k < i \leq j \leq e_k$*. If the* k*th anomalous window is dense; i.e. if* $|\boldsymbol{J}_k| > k^*$*; and*

$$\mathcal{S}\left(\boldsymbol{x}_{i:j}, \boldsymbol{J}\right) \geq \frac{19}{20} C\left(|\boldsymbol{J}|\log(p) + \psi\right),$$

then

$$\mathcal{C}\left(\boldsymbol{x}_{i:j}, \boldsymbol{1}\right) - \mathcal{C}\left(\boldsymbol{x}_{i:j}, \boldsymbol{J}\right) \leq \frac{1}{10} C\sqrt{p\psi} + 2\psi$$

holds for all sparse $\boldsymbol{J}$*, i.e.* $\boldsymbol{J}$ *satisfying* $|\boldsymbol{J}| \leq k^*$*, if* C *is larger than some global constant and the event* E *holds.*

Lemma 33. *Let the event* E *hold. Moreover, let* i *and* j *be such that there exists a* k *such that* $s_k < i \leq j \leq e_k$*. Then, if the* k*th anomalous window is sparse; i.e. if*

$|\boldsymbol{J}_k| \leq k^*$;

$$\mathcal{C}\left(\boldsymbol{x}_{i:j}, \boldsymbol{J}_k\right) - \mathcal{C}\left(\boldsymbol{x}_{i:j}, \boldsymbol{1}\right) \leq \frac{13}{20} C|\boldsymbol{J}_k| \log(p) - \frac{6}{10} C\sqrt{p\psi} + 2\psi \leq \frac{1}{10} C|\boldsymbol{J}_k| \log(p) - \frac{1}{20} C\sqrt{p\psi} + 2\psi$$

holds if C is larger than some global constant

For Part I we will then require the following lemmata, which are again concerned with bounding the increase in penalised cost for replacing fitted segments with the wrong number of components by fitted segments with the right number of components.

Lemma 34. *Let i and j be such that there exists a k such that $s_k < i \leq j \leq e_k$. If the kth anomalous window is dense; i.e. if $|\boldsymbol{J}_k| > k^*$; and*

$$\mathcal{S}\left(\boldsymbol{x}_{i:j}, \boldsymbol{J}\right) \geq \frac{3}{10} C\left(|\boldsymbol{J}| \log(p) + \psi\right),$$

then

$$\mathcal{C}\left(\boldsymbol{x}_{i:j}, \boldsymbol{1}\right) - \mathcal{C}\left(\boldsymbol{x}_{i:j}, \boldsymbol{J}\right) \leq \frac{8}{10} C\sqrt{p\psi} - \frac{6}{10} C|\boldsymbol{J}| \log(p) + 2\psi$$

holds for all sparse $\boldsymbol{J}$, i.e. $\boldsymbol{J}$ satisfying $|\boldsymbol{J}| \leq k^*$, if C is larger than some global constant and the event E holds..

Lemma 35. *Let i and j be such that there exists a k such that $s_k < i \leq j \leq e_k$. If the kth anomalous window is sparse; i.e. if $|\boldsymbol{J}| \leq k^*$; and*

$$\mathcal{S}\left(\boldsymbol{x}_{i:j}, \boldsymbol{J}\right) \geq \frac{3}{10} C\left(|\boldsymbol{J}| \log(p) + \psi\right),$$

then

$$\mathcal{C}\left(\boldsymbol{x}_{i:j}, \boldsymbol{1}\right) - \mathcal{C}\left(\boldsymbol{x}_{i:j}, \boldsymbol{J}\right) \leq \frac{8}{10} C|\boldsymbol{J}_k| \log(p) - \frac{6}{10} C|\boldsymbol{J}| \log(p) + 2\psi$$

holds for all sparse $\boldsymbol{J}$, i.e. $\boldsymbol{J}$ satisfying $|\boldsymbol{J}| \leq k^*$, if C is larger than some global constant and the event E holds.

Property III

We can prove that a fitted segment must overlap with at least one true anomalous segments:

Proposition 21. *Let the assumptions of Theorem 1 hold. Let τ be an optimal partition and E hold. Then, $\forall(s, e, \boldsymbol{J}) \in \tau \quad \exists k : [s+1, e] \cap [s_k + 1, e_k] \neq \emptyset$, provided $C > 2$.*

Proof of Proposition 21: By contradiction: If $(s, e, \mathbf{J})$ overlaps with no true anomalous it can be shown that the partition $\tau \setminus (s, e, \mathbf{J})$ has lower penalised cost than τ, because of E_1 if $\mathbf{J}$ is sparse and E_4 if $\mathbf{J}$ is dense.

Property IV

We now prove the following proposition, which shows that if each true anomalous region is fitted by exactly one segment, then the boundaries of that segment are close to the boundaries of the corresponding anomalous region. To this end, we define the set of partitions $\mathcal{T}_1$ as the set of all partitions fitting exactly K anomalous segments in such a way that each fitted anomalous segment overlaps with exactly one true anomalous region and each true anomalous region overlaps with exactly one fitted anomalous segment. More formally,

$$\mathcal{T}_1 = \{\tau : |\tau| = K \ \wedge (\forall(s, e, \mathbf{J}) \in \tau \ \ \exists k : s_{k+1} \geq e \ \ \wedge \ \ e_{k-1} \leq s \ \ \wedge \ \ [s+1, e] \cap [s_k + 1, e_k] \neq \emptyset)$$
$$\wedge (\forall k \ \ \exists(s, e, \mathbf{J}) \in \tau : [s+1, e] \cap [s_k + 1, e_k] \neq \emptyset)\}.$$

The following proposition then holds:

Proposition 22. *Let the assumptions of Theorem 1 hold. Given E, if a partition $\tau \in \mathcal{T}_1$ is optimal it must also satisfy $\tau \in \mathcal{B}_C$, if C exceeds a global constant.*

Proof of Proposition 22: Let τ be optimal. Consider the kth true anomalous segment $[s_k+1, e_k]$, which τ fits with the segment $(\hat{s}_k, \hat{e}_k, \hat{\mathbf{J}})$. We begin by showing this fitted segment needs to cover most of the true anomalous region, because otherwise adding an additional segment to τ would reduce the penalised cost.

Indeed, $\hat{s}_k \leq s_k + \frac{10C}{\triangle_k^2}$, as otherwise either the partition $\tau \cup (s_k, s_k + \lceil \frac{10C}{\triangle_k^2} \rceil, \mathbf{J}_k)$, if the kth anomalous segment is sparse, or the partition $\tau \cup (s_k, s_k + \lceil \frac{10C}{\triangle_k^2} \rceil, \mathbf{1})$, if the kth anomalous segment is dense, would have a lower overall penalised cost than τ by Lemma 25, which would contradict the optimality of τ. $\hat{e}_k \geq e_k - \frac{10C}{\triangle_k^2}$ holds by a similar argument.

The next step consists of showing that $(\hat{s}_k, \hat{e}_k, \hat{\mathbf{J}})$ does not extend too far beyond the kth anomalous region. Our approach consists of using Lemmata 26 and 27 to show that if this were to happen we could replace $(\hat{s}_k, \hat{e}_k, \hat{\mathbf{J}})$ by a different fitted segment in a way which reduces penalised cost. We will just show that $\hat{e}_k \leq e_k + \frac{10C}{\triangle_k^2}$, as a similar argument implies that $\hat{s}_k \geq s_k - \frac{10C}{\triangle_k^2}$.

We already know that $\hat{s}_k \leq s_k + \frac{10C}{\triangle_k^2}$. Thus, if $\hat{e}_k > e_k + \frac{10C}{\triangle_k^2}$, the segment $[\hat{s}_k+1, \hat{e}_k]$ would contain at least $\lceil \frac{10C}{\triangle_k^2} \rceil$ observations both from the typical distribution and the kth anomalous window. It is possible to show that this partition can be replaced by splitting up $[\hat{s}_k + 1, \hat{e}_k]$ in such a way that the penalised cost is reduced.

- If $\mathbf{J}_k$ is dense, we can replace $(\hat{s}_k, \hat{e}_k, \hat{\mathbf{J}})$ first with $(\hat{s}_k, e_k - \lfloor \frac{10C}{\triangle_k^2} \rfloor, \hat{\mathbf{J}})$ and $(e_k - \lfloor \frac{10C}{\triangle_k^2} \rfloor, \hat{e}_k, \hat{\mathbf{J}})$, increasing the penalised cost by no more than $C\psi + C|\mathbf{J}| \log(p)$ if

$\hat{\mathbf{J}}$ is sparse and $C\psi+(C+2)\sqrt{p\psi}$ if $\hat{\mathbf{J}}=\mathbf{1}$ (By event E_9). Lemma 27 then shows that replacing $(e_k-\lfloor\frac{10C}{\triangle_k^2}\rfloor,\hat{e}_k,\hat{\mathbf{J}})$ with $(e_k-\lfloor\frac{10C}{\triangle_k^2}\rfloor,e_k,\mathbf{1})$ reduces the penalised cost by at least $4C\psi+4C|\hat{\mathbf{J}}|\log(p)$ if $\hat{\mathbf{J}}$ is sparse and $4C\psi+4C\sqrt{p\psi}$ when $\hat{\mathbf{J}}=\mathbf{1}$ respectively. Chaining these two transformations therefore leads to a reduction in penalised cost contradicting optimality of τ.

- If $\mathbf{J}_k$ is sparse, the cases $\hat{\mathbf{J}}=\mathbf{1}$, and $|\hat{\mathbf{J}}|\leq k^*$ have to be considered separately. If $\hat{\mathbf{J}}=\mathbf{1}$,

$$\mathcal{C}\left(\mathbf{x}_{(\hat{s}_k+1):\hat{e}_k},\mathbf{1}\right)=\mathcal{C}\left(\mathbf{x}_{(\hat{s}_k+1):\hat{e}_k},\mathbf{J}_k\right)+p+C\sqrt{p\psi}-\sum_{c\notin\mathbf{J}_k}(\hat{e}_k-\hat{s}_k)\left(\bar{\boldsymbol{\eta}}_{(\hat{s}_k+1):\hat{e}_k}\right)^2+C|\mathbf{J}_k|\log(p)$$

$$\geq\mathcal{C}\left(\mathbf{x}_{(\hat{s}_k+1):\left(e_k-\lceil\frac{10C}{\triangle_k^2}\rceil\right)},\mathbf{J}_k\right)+\mathcal{C}\left(\mathbf{x}_{\left(e_k-\lfloor\frac{10C}{\triangle_k^2}\rfloor\right):\hat{e}_k},\mathbf{J}_k\right)-2C|\mathbf{J}_k|\log(p)-(C+2)\psi+(C-2)\sqrt{p\psi},$$

with the inequality following from E_2 and the fact that splitting a segment does not increase the un-penalised cost. Lemma 26, then shows that the above quantity exceeds

$$\mathcal{C}\left(\mathbf{x}_{(\hat{s}_k+1):\left(e_k-\lceil\frac{10C}{\triangle_k^2}\rceil\right)},\mathbf{J}_k\right)+\mathcal{C}\left(\mathbf{x}_{\left(e_k-\lfloor\frac{10C}{\triangle_k^2}\rfloor\right):e_k},\mathbf{J}_k\right)+6C\psi+4C|\mathbf{J}_k|\log(p)-(C+2)\psi+(C-2)\sqrt{p\psi},$$

which exceeds

$$\mathcal{C}\left(\mathbf{x}_{(\hat{s}_k+1):\left(e_k-\lceil\frac{10C}{\triangle_k^2}\rceil\right)},\mathbf{J}_k\right)+\mathcal{C}\left(\mathbf{x}_{\left(e_k-\lfloor\frac{10C}{\triangle_k^2}\rfloor\right):e_k},\mathbf{J}_k\right),$$

thus contradicting the optimality of τ. Similarly, if $|\hat{\mathbf{J}}|\leq k^*$,

$$\mathcal{C}\left(\mathbf{x}_{(\hat{s}_k+1):\hat{e}_k},\hat{\mathbf{J}}\right)\geq\mathcal{C}\left(\mathbf{x}_{(\hat{s}_k+1):\hat{e}_k},\mathbf{J}_k\right)-\sum_{c\in\hat{\mathbf{J}}\setminus\mathbf{J}_k}(\hat{e}_k-\hat{s}_k)\left(\bar{\boldsymbol{\eta}}_{(\hat{s}_k+1):\hat{e}_k}\right)^2+C(|\hat{\mathbf{J}}|-|\mathbf{J}_k|)\log(p)$$

$$\geq\mathcal{C}\left(\mathbf{x}_{(\hat{s}_k+1):\left(e_k-\lceil\frac{10C}{\triangle_k^2}\rceil\right)},\mathbf{J}_k\right)+\mathcal{C}\left(\mathbf{x}_{\left(e_k-\lfloor\frac{10C}{\triangle_k^2}\rfloor\right):\hat{e}_k},\mathbf{J}_k\right)-2C|\mathbf{J}_k|\log(p)-C\psi-2\psi$$

$$-2|\hat{\mathbf{J}}\setminus\mathbf{J}_k|\log(p)+C|\hat{\mathbf{J}}|\log(p)$$

$$\geq\mathcal{C}\left(\mathbf{x}_{(\hat{s}_k+1):\left(e_k-\lceil\frac{10C}{\triangle_k^2}\rceil\right)},\mathbf{J}_k\right)+\mathcal{C}\left(\mathbf{x}_{\left(e_k-\lfloor\frac{10C}{\triangle_k^2}\rfloor\right):e_k},\mathbf{J}_k\right)-2C\psi-2C|\mathbf{J}_k|\log(p),$$

with the second inequality following from E_1 and the fact that splitting a segment does not increase the un-penalised cost. The third equality holds for large enough values of C. As before, Lemma 26 shows that the above quantity exceeds

$$\mathcal{C}\left(\mathbf{x}_{(\hat{s}_k+1):\left(e_k-\lceil\frac{10C}{\triangle_k^2}\rceil\right)},\mathbf{J}_k\right)+\mathcal{C}\left(\mathbf{x}_{\left(e_k-\lfloor\frac{10C}{\triangle_k^2}\rfloor\right):e_k},\mathbf{J}_k\right),$$

thus contradicting the optimality of τ.

Property II

We now prove that if all fitted segments of the optimal partition overlap with at most one true anomalous segment, then each true anomalous segment must overlap with exactly one fitted segment. To this end, we now define $\mathcal{T}_2$ as the set of partitions in which each fitted anomalous segment overlaps with exactly one truly anomalous region. More formally we define

$$\mathcal{T}_2=\{\tau:\forall(s,e,\mathbf{J})\in\tau\ \ \exists k:s_{k+1}\geq e\ \ \wedge\ \ e_{k-1}\leq s\ \ \wedge\ \ [s+1,e]\cap[s_k+1,e_k]\neq\emptyset\}.$$

and note that $\mathcal{T}_1\subset\mathcal{T}_2$. The following proposition holds:

Proposition 23. *Let the assumptions of Theorem 1 hold. Given E, if a partition $\tau\in\mathcal{T}_2$ is optimal it must also satisfy $\tau\in\mathcal{T}_1$ if C exceeds a global constant.*

Proof of Proposition 23: The proof has two parts:

1. We need to show that the optimality of τ implies that each true anomalous segment overlaps with at least one fitted segment in τ.

2. We need to show that the optimality of τ implies that each true anomalous segment overlaps with at most one fitted segment in τ.

We prove both statements by contradiction: First assume that τ is optimal but that there exists a k such that $[s_k + 1, e_k]$ is not covered at all by any fitted segment in τ. Then by Lemma 25, the partition $\tau \cup (s_k, e_k, \mathbf{J}_k)$, if the kth change is sparse, or $\tau \cup (s_k, e_k, \mathbf{1})$, if the kth change is dense, has a lower penalised cost than τ, so contradicting its optimality.

Now assume that there exists a k such that τ contains two or more fitted segments overlapping with $[s_k + 1, e_k]$. We will show that it is possible to merge any two fitted segments (called $(a, b, \mathbf{J}_1), (c, d, \mathbf{J}_2)$, where $c \geq b$ without loss of generality) in a way which reduces the total penalised cost thereby contradicting the optimality of τ. In order to do so, we define $a' = \max(s_k, a)$ and $d' = \min(e_k, d)$. The following two cases have to be considered separately, but share in the following idea: Merging $(a, b, \mathbf{J}_1), (c, d, \mathbf{J}_2)$, into a single fitted segment increases the un-penalised cost by at most $O(\sqrt{C})$. At the same time merging reduces the penalty by $O(C)$. Hence, if C is large enough, merging reduces the overall penalised cost.

1. $\mathbf{J}_k$ is dense : We will show that replacing $(a, b, \mathbf{J}_1), (c, d, \mathbf{J}_2)$ with $(a', d', \mathbf{1})$ reduces the penalised cost. Lemma 28, implies that it is sufficient to show that

$$\mathcal{C}(\mathbf{x}_{a':b}, \mathbf{1}) - \mathcal{C}(\mathbf{x}_{a:b}, \mathbf{J}_1) \leq \frac{5}{40} C \left(\psi + \sqrt{p\psi} \right)$$

and

$$\mathcal{C}(\mathbf{x}_{c:d'}, \mathbf{1}) - \mathcal{C}(\mathbf{x}_{c:d}, \mathbf{J}_2) \leq \frac{5}{40} C \left(\psi + \sqrt{p\psi} \right).$$

We limit ourselves to proving the first statement, as the second one can be proven via a symmetrical argument. If $\mathbf{J}_1 = \mathbf{1}$, the statement follows directly from Lemma 29.

If $|\mathbf{J}_1| \leq k^*$, we first note that Lemma 29 implies that

$$\mathcal{C}(x_{a':b}, \mathbf{J}_1) \leq \mathcal{C}(\mathbf{x}_{a:b}, \mathbf{J}_1) + 8\psi + 8|\mathbf{J}_1| \log(p) \leq \mathcal{C}(\mathbf{x}_{a:b}, \mathbf{J}_1) + 8\psi + 8\sqrt{p\psi} \tag{B.2.4}$$

By optimality of τ, $\mathcal{C}(\mathbf{x}_{a:b}, \mathbf{J}_1) < 0$, must hold. This implies that

$$\mathcal{S}(x_{a':b}, \mathbf{J}_1) \geq \frac{19}{20} C \left(\psi + |\mathbf{J}| \log(p)\right).$$

Consequently, Lemma 32 shows that

$$\mathcal{C}(x_{a':b}, \mathbf{1}) \leq \mathcal{C}(x_{a':b}, \mathbf{J}_1) + \frac{1}{10} C \left(\psi + \sqrt{p\psi}\right). \tag{B.2.5}$$

Combining (B.2.4) and (B.2.5) finishes the proof.

2. $\mathbf{J}_k$ is sparse : We will show that replacing $(a, b, \mathbf{J}_1), (c, d, \mathbf{J}_2)$ with $(a', d', \mathbf{J}_k)$ reduces the penalised cost. Lemma 28, implies that it is sufficient to show that

$$\mathcal{C}(\mathbf{x}_{a':b}, \mathbf{J}_k) - \mathcal{C}(\mathbf{x}_{a:b}, \mathbf{J}_1) \leq \frac{5}{40} C \left(\psi + |\mathbf{J}_k| \log(p)\right)$$

and

$$\mathcal{C}(\mathbf{x}_{c:d'}, \mathbf{J}_k) - \mathcal{C}(\mathbf{x}_{c:d}, \mathbf{J}_2) \leq \frac{5}{40} C \left(\psi + |\mathbf{J}_k| \log(p)\right).$$

These proofs for both statements are symmetrical. We therefore only prove the first one. As before we begin by considering the case $\mathbf{J}_1 = \mathbf{1}$. We have

$$\begin{aligned}
\mathcal{C}(\mathbf{x}_{a':b}, \mathbf{J}_k) &= \mathcal{C}(\mathbf{x}_{a:b}, \mathbf{1}) + (\mathcal{C}(\mathbf{x}_{a':b}, \mathbf{1}) - \mathcal{C}(\mathbf{x}_{a:b}, \mathbf{1})) + (\mathcal{C}(\mathbf{x}_{a':b}, \mathbf{J}_k) - \mathcal{C}(\mathbf{x}_{a':b}, \mathbf{1})) \\
&\leq \mathcal{C}(\mathbf{x}_{a:b}, \mathbf{1}) + \left(8\psi + 8\sqrt{p\psi}\right) + \left(2\psi + \frac{1}{10} C|\mathbf{J}_k| \log(p) - \frac{1}{20} C\sqrt{p\psi}\right) \\
&\leq \mathcal{C}(\mathbf{x}_{a:b}, \mathbf{1}) + 10\psi + \frac{1}{10} C|\mathbf{J}_k| \log(p),
\end{aligned}$$

where the first inequality follows from Lemmata 29 and 33, while the second inequality golds if C exceeds a fixed constant. Turning to the case in which $|\mathbf{J}_1| \leq k^*$, we note

that the same strategy of proof used for the case in which $\mathbf{J}_k$ is dense can be reapplied, the only difference being that Lemma 31 has to be used instead of Lemma 32.

Property I

We will now prove that an optimal partition can not contain a fitted segment overlapping with more than one true anomalous segment. We formalise this in the following Proposition:

Proposition 24. *Let the assumptions of Theorem 1 hold. Let τ be an optimal partition. Then, $\tau \in \mathcal{T}_2$, given that the event E holds and that the constant C exceeds a global constant.*

Note that this result trivially holds when $K = 1$. In order to prove this proposition, we will use a variation of Proposition 23. For this we introduce the set of fitted sparse segments, which either begin or end at the start of a true anomalous segment and only contain a small fraction of the true anomalous segment

$$\mathcal{A}_1 = \left\{ (s, e, \mathbf{J}) : |\mathbf{J}| < k^* \ \wedge \ \exists k : \left(s = s_k \ \wedge \ e \leq s_k + \frac{10C}{\triangle_k^2} \right) \vee \left(e = e_k \ \wedge \ s \geq e_k - \frac{10C}{\triangle_k^2} \right) \right\},$$

as well as its analogue for dense changes

$$\mathcal{A}_2 = \left\{ (s, e, \mathbf{1}) : \exists k : \left(s = s_k \ \wedge \ e \leq s_k + \frac{10C}{\triangle_k^2} \right) \vee \left(e = e_k \ \wedge \ s \geq e_k - \frac{10C}{\triangle_k^2} \right) \right\}.$$

The following two propositions can then be proven

Proposition 25. *Let the assumptions of Theorem 1 hold. Let $\tau' \in \mathcal{T}_2$ and E hold true. Then there exists another partition $\tau'' \in \mathcal{T}_2$ such that*

$$\mathcal{C}\left(\boldsymbol{x}_{1:n}, \tau''\right) \leq \mathcal{C}\left(\boldsymbol{x}_{1:n}, \tau'\right) - \frac{6}{10} \left(\sum_{(s,e,\boldsymbol{J}) \in \tau' \cap \mathcal{A}_1} (C\psi + C|\boldsymbol{J}| \log(p)) + \sum_{(s,e,\boldsymbol{1}) \in \tau' \cap \mathcal{A}_2} \left(C\psi + C\sqrt{p\psi} \right) \right),$$

if C exceeds a global constant.

Proposition 26. *Let the assumptions of Theorem 1 hold. Let τ be an optimal partition and E hold true. Then, there exists a partition $\tau' \in \mathcal{T}_2$ such that*

$$\mathcal{C}\left(\boldsymbol{x}_{1:n}, \tau'\right) \leq \mathcal{C}\left(\boldsymbol{x}_{1:n}, \tau\right) + \frac{11}{20}\left(\sum_{(s,e,\boldsymbol{J})\in\tau\cap\mathcal{A}_1}(C\psi + C|\boldsymbol{J}|\log(p)) + \sum_{(s,e,\boldsymbol{1})\in\tau\cap\mathcal{A}_2}\left(C\psi + C\sqrt{p\psi}\right)\right),$$

with equality if and only if $\tau \in \mathcal{T}_2$, if C exceeds a global constant.

Note that Proposition 25 does not assume that τ' is optimal. Using these two propositions it is easy to derive the following:

Proof of Proposition 24: Assume that the optimal partition τ is such that $\tau \notin \mathcal{T}_2$. Then, by Proposition 26 there exists a partition $\tau' \in \mathcal{T}_2$ such that

$$\mathcal{C}\left(\mathbf{x}_{1:n}, \tau\right) > \mathcal{C}\left(\mathbf{x}_{1:n}, \tau'\right) - \frac{11}{20}\left(\sum_{(s,e,\mathbf{J})\in\tau\cap\mathcal{A}_1}(C\psi + C|\mathbf{J}|\log(p)) + \sum_{(s,e,\mathbf{1})\in\tau\cap\mathcal{A}_2}\left(C\psi + C\sqrt{p\psi}\right)\right),$$

Moreover, Proposition 25 implies that there exists another partition $\tau'' \in \mathcal{T}_2$ such that

$$\mathcal{C}\left(\mathbf{x}_{1:n}, \tau'\right) \geq \mathcal{C}\left(\mathbf{x}_{1:n}, \tau''\right) + \frac{6}{10}\left(\sum_{(s,e,\mathbf{J})\in\tau'\cap\mathcal{A}_1}(C\psi + C|\mathbf{J}|\log(p)) + \sum_{(s,e,\mathbf{1})\in\tau'\cap\mathcal{A}_2}\left(C\psi + C\sqrt{p\psi}\right)\right),$$

Consequently,

$$\mathcal{C}\left(\mathbf{x}_{1:n}, \tau\right) > \mathcal{C}\left(\mathbf{x}_{1:n}, \tau''\right),$$

which contradicts the optimality of τ.

Proof of Proposition 25: Proposition 23 shows that fitting an anomalous region with two segments, or with one very short segment leaving most of the anomalous region uncovered is sub-optimal. This proposition goes further by showing it is suboptimal by at least $O(\frac{6}{10}C)$. Crucially, this is larger than $O(\frac{1}{2}C)$ and will help us break up fitted segments spanning multiple anomalous regions. The proof of this Proposition is similar in flavour to the proof of the second part of Proposition 23. The main idea is that there are at most two fitted partitions $\in \tau' \cap (\mathcal{A}_1 \cup \mathcal{A}_2)$ overlapping with the kth true anomalous region. These partitions therefore leave at least $\frac{20C}{\triangle_k^2}$ of the

kth anomalous region uncovered. Therefore, if no other segment in τ' overlaps with the kth anomalous region, one can be added without increasing the penalised cost. It can then be merged with the fitted partitions in $\in \tau' \cap (\mathcal{A}_1 \cup \mathcal{A}_2)$ and overlap with the kth true anomalous region. This yields a new partition still in $\mathcal{T}_2$ with the claimed reduction in penalised cost.

Since $\tau' \in \mathcal{T}_2$, we can consider each of the K true anomalous regions separately. We define the set of fitted segments in τ' which overlap with the kth anomalous region to be

$$\tau'_k = \{(s, e, \mathbf{J}) \in \tau' : [s+1, e] \cap [s_k + 1, e_k] \neq \emptyset\}.$$

Proving the full result is therefore equivalent to proving the existence of a τ''_k which yields the required reduction in penalised cost. The following 3 cases are possible:

1. $|\tau'_k \cap (\mathcal{A}_1 \cup \mathcal{A}_2)| = 0$, which happens when τ' does not contain a short fitted segment at either the beginning or the end of the kth anomalous region. No further transformation is required in this case, i.e. $\tau''_k = \tau'_k$

2. $|\tau'_k \cap (\mathcal{A}_1 \cup \mathcal{A}_2)| = 1$.

3. $|\tau'_k \cap (\mathcal{A}_1 \cup \mathcal{A}_2)| = 2$.

We will only explicitly describe the transformation for the second case, as applying it twice yields a transformation for the third case. Without loss of generality we further assume that $\tau'_k \cap (\mathcal{A}_1 \cup \mathcal{A}_2) = (s, e_k, \mathbf{J})$, i.e. that the short fitted segment lies at the end of the kth anomalous window. A first special case can be treated very quickly. If $|\mathbf{J}| \leq k^*$ and $\mathcal{C}\left(\mathbf{x}_{(s+1):e_k}, \mathbf{J}\right) \geq \frac{6}{10} C(\psi + |\mathbf{J}| \log(p))$, removing $(s, e_k, \mathbf{J})$ from τ'_k is

sufficient. If $|\mathbf{J}| \leq k^*$ and $\mathcal{C}\left(\mathbf{x}_{(s+1):e_k}, \mathbf{J}\right) < \frac{6}{10}C(\psi + |\mathbf{J}|\log(p))$, we nevertheless have

$$\mathcal{C}\left(\mathbf{x}_{(s+1):e_k}, \mathbf{J}_k\right) \leq \mathcal{C}\left(\mathbf{x}_{(s+1):e_k}, \mathbf{J}\right) + \frac{8}{10}C|\mathbf{J}_k|\log(p) - \frac{6}{10}C|\mathbf{J}|\log(p) + 2\psi$$

if $\mathbf{J}_k$ is sparse and

$$\mathcal{C}\left(\mathbf{x}_{(s+1):e_k}, \mathbf{1}\right) \leq \mathcal{C}\left(\mathbf{x}_{(s+1):e_k}, \mathbf{J}\right) + \frac{8}{10}C\sqrt{p\psi} - \frac{6}{10}C|\mathbf{J}|\log(p) + 2\psi$$

if $\mathbf{J}_k$ is dense, by Lemmata 35 and 34 respectively. Similarly, if $\mathbf{J} = \mathbf{1}$ we have that

$$\mathcal{C}\left(\mathbf{x}_{(s+1):e_k}, \mathbf{J}_k\right) \leq \mathcal{C}\left(\mathbf{x}_{(s+1):e_k}, \mathbf{J}\right) + \frac{8}{10}C|\mathbf{J}_k|\log(p) - \frac{6}{10}C\sqrt{p\psi} + 2\psi$$

if $\mathbf{J}_k$ is sparse as a direct consequence of Lemma 33 and, trivially,

$$\mathcal{C}\left(\mathbf{x}_{(s+1):e_k}, \mathbf{1}\right) \leq \mathcal{C}\left(\mathbf{x}_{(s+1):e_k}, \mathbf{J}\right) + \frac{8}{10}C\sqrt{p\psi} - \frac{6}{10}C\sqrt{p\psi} + 2\psi$$

if $\mathbf{J}_k$ is dense.

Consequently, if the next fitted change in τ_k' to the left of $(s, e, \mathbf{J})$ is of the form $(\tilde{s}, \tilde{e}, \mathbf{J}_k)$, if $\mathbf{J}_k$ is sparse or $(\tilde{s}, \tilde{e}, \mathbf{1})$ if $\mathbf{J}_k$ is dense, for some $\tilde{s} \geq s_k$, Lemma 28 shows that the required reduction in penalised cost can be obtained by merging these two fitted segments. If there is no other fitted change in τ_k', or if the next fitted segment in τ_k' to the left of $(s, e, \mathbf{J})$ is $(\tilde{s}, \tilde{e}, \mathbf{J})$, where $\tilde{e}$ satisfies $s - \tilde{e} \geq \frac{10C}{\triangle_k^2}$, Lemma 25 implies that adding $(s - \lceil\frac{10C}{\triangle_k^2}\rceil, s, \mathbf{J}_k)$, if $\mathbf{J}_k$ is sparse or $(s - \lceil\frac{10C}{\triangle_k^2}\rceil, s, \mathbf{1})$ if $\mathbf{J}_k$ is dense, does not increase the penalised cost. Lemma 28 can then be applied as before to show that merging this new fitted segment with $(s, e, \mathbf{J})$ yields a new partition exhibiting the required reduction in penalised cost.

Hence, in order to finish proving the result we only need to show that any $(\tilde{s}, \tilde{e}, \mathbf{J}) \in \tau_k'$ can either be removed without increasing the penalised cost or replaced by (max($\tilde{s}$,

$s_k), \tilde{e}, \mathbf{J}_k)$ in a way which increases the penalised cost by at most $\frac{5}{40}C(|\mathbf{J}_k|\log(p)+\psi)$ if $\mathbf{J}_k$ is sparse or $(\max(\tilde{s}, s_k), \tilde{e}, \mathbf{1})$ in a way which increases the penalised cost by at most $\frac{5}{40}C\left(\sqrt{p\psi}+\psi\right)$ if $\mathbf{J}_k$ is dense. This however, was already shown in the proof of Proposition 23. This finishes the proof.

Proof of Proposition 26: If $\tau \in \mathcal{T}_2$, the result trivially holds. In order to prove the result when $\tau' \notin \mathcal{T}_2$, we consider all possible fitted segments $(s, e, \mathbf{J}) \in \tau \setminus \mathcal{T}_2$ which overlap with at least two anomalous regions and show that

1. No such segment can overlap a true fitted dense change, the k'th say, by more than $\frac{10C}{\triangle_k'^2}$ as this would contradict the optimality of τ.

2. All other fitted segments, overlapping with at least two anomalous regions, including, potentially, a certain number of sparse changes by more that $\frac{10C}{\triangle}$ can be replaced by fitted segments each overlapping with exactly one true anomalous segment in a way which strictly bounds the increase in penalised cost as stipulated by the proposition.

1) First of all we can show that the optimality of τ implies that no partition $(s, e, \mathbf{J}) \in \tau \setminus \mathcal{T}_2$ can overlap a dense change (the k'th change say) by more than $\frac{10C}{\triangle_k'^2}$. Otherwise, the interval $[s+1, e]$ would also contain at least $\frac{10C}{\triangle_k'^2}$ observations belonging to the typical distribution. We could therefore split it up into three segments (increasing the penalised cost by at most $2C\psi + 2C|\mathbf{J}|\log(p)$ or $2C\psi + 2(C+2)\sqrt{p\psi}$), one of which containing exactly $\lceil \frac{10C}{\triangle_k'^2} \rceil$ of observations belonging to the typical distribution and $\lceil \frac{10C}{\triangle_k'^2} \rceil$ of observations belonging to the k'th anomalous window. Lemma 27 shows that such a segment can be replaced in a way which reduces the penalised cost by

at least $4C\psi + 4C\sqrt{p\psi}$. Overall, we would thus obtain a new partition with a lower penalised cost than τ contradicting the optimality of τ.

2) Consider now, a segment $(s, e, \mathbf{J}) \in \tau \setminus \mathcal{T}_2$ not overlapping with any dense changes by more than $\frac{10C}{\triangle_k^2}$. For this segment define the set of true anomalous segments it overlaps by more than $\frac{10C}{\triangle_k^2}$ to be

$$\mathcal{D}_{e,s} := \left\{ k : |[s+1, e] \cap [s_k + 1, e_k + 1]| \geq \frac{10C}{\triangle_k^2} \right\}.$$

and note that $|\mathbf{J}_k|$ is sparse if $k \in \mathcal{D}_{s,e}$ for some $(s, e, \mathbf{J}) \in \tau \setminus \mathcal{T}_2$. We have to consider the following 4 scenarios

1. The beginning of the fitted segment $(s, e, \mathbf{J}) \in \tau \setminus \mathcal{T}_2$ overlaps with a true anomalous region $[s_{k'} + 1, e_{k'}]$, but does so by less than $\frac{10C}{\triangle_{k'}^2}$. i.e. $\exists k' : e_{k'} - \frac{10C}{\triangle_{k'}^2} \leq s + 1 \leq e_{k'}$.

2. The end of the fitted segment $(s, e, \mathbf{J}) \in \tau \setminus \mathcal{T}_2$ overlaps with a true anomalous region $[s_{k''} + 1, e_{k''}]$, but does so by less than $\frac{10C}{\triangle_{k''}^2}$. i.e. $\exists k'' : s_{k''} + 1 + \frac{10C}{\triangle_{k''}^2} \geq e \geq s_{k''} + 1$.

3. Both apply

4. None of 1 and 2 apply. Note that this allows for the beginning and or the end of $(s, e, \mathbf{J}) \in \tau \setminus \mathcal{T}_2$ to lie in a truly anomalous region provided the overlap with that region exceeds the critical threshold of $\frac{10C}{\triangle^2}$.

We then replace $(s, e, \mathbf{J})$ in τ to obtain a new partition $\tilde{\tau}$. depending on the cases above we define $\tilde{\tau}$ to be

1. $$(\tau \setminus \{(s, e, \mathbf{J})\}) \cup \{(s, e_{k'}, \mathbf{J})\} \cup \left(\bigcup_{k \in \mathcal{D}_{e,s}} \{(s_k, e_k, \mathbf{J}_k)\} \right)$$

2. $$(\tau \setminus \{(s, e, \mathbf{J})\}) \cup \left(\bigcup_{k \in \mathcal{D}_{e,s}} \{(s_k, e_k, \mathbf{J}_k)\} \right) \cup \{(s_{k''}, e, \mathbf{J})\}$$

3. $$(\tau \setminus \{(s, e, \mathbf{J})\}) \cup \{(s, e_{k'}, \mathbf{J})\} \cup \left(\bigcup_{k \in \mathcal{D}_{e,s}} \{(s_k, e_k, \mathbf{J}_k)\} \right) \cup \{(s_{k''}, e, \mathbf{J})\}$$

4. $$(\tau \setminus \{(s, e, \mathbf{J})\}) \cup \bigcup_{k \in \mathcal{D}_{e,s}} \{(s_k, e_k, \mathbf{J}_k)\}$$

depending on which case applies. The main effect of this transformation is the same across all cases: It results in all true anomalous regions contained in $(s, e, \mathbf{J})$ to be fitted separately and according to the ground truth. Only the number of fitted segments belonging to $\mathcal{A}_1$ and/or $\mathcal{A}_2$ depends on the case. Since applying this transformation for all $(s, e, \mathbf{J}) \in \tau \setminus \mathcal{T}_2$ leads to a new partition τ' which is contained in $\mathcal{T}_2$, it is sufficient to prove that each transformation individually increases the penalised cost by strictly less than

1. $\frac{11}{20}C\left(\psi + |\mathbf{J}|\log(p)\right)$ if $\mathbf{J}$ is sparse or $\frac{11}{20}C\left(\psi + \sqrt{p\psi}\right)$ if $\mathbf{J}$ is dense.

2. $\frac{11}{20}C\left(\psi + |\mathbf{J}|\log(p)\right)$ if $\mathbf{J}$ is sparse or $\frac{11}{20}C\left(\psi + \sqrt{p\psi}\right)$ if $\mathbf{J}$ is dense.

3. $\frac{22}{20}C\left(\psi + |\mathbf{J}|\log(p)\right)$ if $\mathbf{J}$ is sparse or $\frac{22}{20}C\left(\psi + \sqrt{p\psi}\right)$ if $\mathbf{J}$ is dense.

4. 0

depending on the case in order to prove the proposition. The fourth case follows directly from the following Lemma:

Lemma 36. *Let the event E hold and C exceed some global constant. Let s and e be such the fourth scenario applies, i.e.*

1. $\nexists k' : e_{k'} - \frac{10C}{\triangle_{k'}^2} \leq s+1 \leq e_{k'}$.

2. $\nexists k'' : s_{k''} + 1 + \frac{10C}{\triangle_{k''}^2} \geq e \geq s_{k''} + 1$

Then, the following holds true for all sparse $\boldsymbol{J}$

$$\mathcal{C}\left(\boldsymbol{x}_{s,e}, \boldsymbol{J}\right) \geq \frac{19}{20} C\left(\psi + |\boldsymbol{J}| \log(p)\right) + \sum_{k \in \mathcal{D}_{s,e}} \left(\mathcal{C}\left(\boldsymbol{x}_{(s_k+1):e_k}, \boldsymbol{J}_k\right)\right)$$

Moreover, the following statement is also true:

$$\mathcal{C}\left(\boldsymbol{x}_{s,e}, \boldsymbol{1}\right) \geq \frac{19}{20} C\left(\psi + \sqrt{p\psi}\right) + \sum_{k \in \mathcal{D}_{s,e}} \left(\mathcal{C}\left(\boldsymbol{x}_{(s_k+1):e_k}, \boldsymbol{J}_k\right)\right).$$

This Lemma can also be used to bound the increase in penalised cost obtained for the other three cases. The only difference is that $(s, e, \mathbf{J})$ is first split up to twice in order to remove the short overlap with the true anomalous region at the beginning and/or the end. For the sake of brevity, we limit ourselves to write out the proof for the third case, for which the result is tightest. If, $\mathbf{J}$ is sparse, we have that

$$\begin{aligned}
&\mathcal{C}\left(x_{(s+1):e}, \mathbf{J}\right) \geq \mathcal{C}\left(x_{(s+1):e_{k'}}, \mathbf{J}\right) + \mathcal{C}\left(x_{(e_{k'}+1):s_{k''}}, \mathbf{J}\right) + \mathcal{C}\left(x_{(s_{k''}+1):e}, \mathbf{J}\right) - 2C\left(\psi + |\mathbf{J}| \log(p)\right) \\
&> \mathcal{C}\left(x_{(s+1):e_{k'}}, \mathbf{J}\right) + \sum_{k \in \mathcal{D}_{s,e}} \left(\mathcal{C}\left(\mathbf{x}_{(s_k+1):e_k}, \mathbf{J}_k\right)\right) + \mathcal{C}\left(x_{(s_{k''}+1):e}, \mathbf{J}\right) - \frac{22}{20} C\left(\psi + |\mathbf{J}| \log(p)\right),
\end{aligned}$$

where the inequality follows from Lemma 36. Similarly, if, $\mathbf{J} = \mathbf{1}$ is dense, we have that

$$\begin{aligned}
&\mathcal{C}\left(x_{(s+1):e}, \mathbf{1}\right) \geq \mathcal{C}\left(x_{(s+1):e_{k'}}, \mathbf{1}\right) + \mathcal{C}\left(x_{(e_{k'}+1):s_{k''}}, \mathbf{1}\right) + \mathcal{C}\left(x_{(s_{k''}+1):e}, \mathbf{1}\right) - 2(C+1)\left(\psi + \sqrt{p\psi}\right) \\
&> \mathcal{C}\left(x_{(s+1):e_{k'}}, \mathbf{1}\right) + \sum_{k \in \mathcal{D}_{s,e}} \left(\mathcal{C}\left(\mathbf{x}_{(s_k+1):e_k}, \mathbf{J}_k\right)\right) + \mathcal{C}\left(x_{(s_{k''}+1):e}, \mathbf{1}\right) - \frac{22}{20} C\left(\psi + \sqrt{p\psi}\right),
\end{aligned}$$

where the inequalities follow from Lemma 36, E_9, and C exceeding a global constant. This finishes the proof.

Proof of Proposition 20

Propositions 23, 21, 24, and 24 give the result.

B.3 Proofs for Lemmata

B.3.1 Proof of Lemma 14

The MGF of $Z = (X - c)^+$ is given by

$$\begin{aligned}\mathbb{E}\left(e^{\lambda Z}\right) &= \mathbb{P}\left(\chi_v^2 < c\right) + \int_c^\infty e^{\lambda(x-c)} \frac{1}{\Gamma\left(\frac{v}{2}\right) 2^{\frac{v}{2}}} x^{\frac{v}{2}-1} e^{-\frac{1}{2}x} dx \\ &= \mathbb{P}\left(\chi_v^2 < c\right) + \frac{e^{-\lambda c}}{\Gamma\left(\frac{v}{2}\right) 2^{\frac{v}{2}}} \int_c^\infty x^{\frac{v}{2}-1} e^{-(1-2\lambda)x} dx\end{aligned}$$

using the substitution $y = (1 - 2\lambda)x$ the above can be shown to be equal to

$$\mathbb{P}\left(\chi_v^2 < c\right) + \frac{e^{-\lambda a}}{\Gamma\left(\frac{v}{2}\right) 2^{\frac{v}{2}} (1-2\lambda)^{\frac{v}{2}}} \int_c^\infty y^{\frac{v}{2}-1} e^{-y} dy = \mathbb{P}\left(\chi_v^2 < c\right) + \frac{e^{-\lambda c}}{(1-2\lambda)^{\frac{v}{2}}} \mathbb{P}\left(\chi_v^2 > c(1-2\lambda)\right).$$

B.3.2 Proof of Lemma 15

As shown by Lemma 14, the MGF of $Z = (x - c)^+$ is given by

$$\mathbb{P}\left(\chi_v^2 < c\right) + \frac{e^{-\lambda c}}{(1-2\lambda)^{v/2}} \mathbb{P}\left(\chi_v^2 > c(1-2\lambda)\right),$$

for $0 \leq \lambda \leq 1/2$. Consequently,

$$\frac{d}{d\lambda}\left(\mathbb{E}\left(e^{\lambda Z}\right)\right) = \frac{2cf(c)}{1-2\lambda} + \left(\frac{v}{1-2\lambda} - c\right) \frac{e^{-\lambda c}}{(1-2\lambda)^{v/2}} \mathbb{P}\left(\chi_v^2 > c(1-2\lambda)\right).$$

Evaluating the above at $\lambda = 0$ shows that the mean of Z is indeed $\mu = 2cf(c) + (v - c)\mathbb{P}\left(\chi_v^2 > c\right)$. We therefore have

$$\begin{aligned}
&\frac{d}{d\lambda}\left(\log\left(\mathbb{E}\left(e^{\lambda Z}\right)\right)\right) - \mu \\
&= \frac{\frac{d}{d\lambda}\left(\mathbb{E}\left(e^{\lambda Z}\right)\right)}{\mathbb{E}\left(e^{\lambda Z}\right)} - \mu = \frac{\frac{2cf(c)}{1-2\lambda} + \left(\frac{v}{1-2\lambda} - c\right)\frac{e^{-\lambda c}}{(1-2\lambda)^{v/2}}\mathbb{P}\left(\chi_v^2 > c(1-2\lambda)\right)}{\mathbb{P}\left(\chi_v^2 < c\right) + \frac{e^{-\lambda c}}{(1-2\lambda)^{v/2}}\mathbb{P}\left(\chi_v^2 > c(1-2\lambda)\right)} - \mu \\
&= \frac{1}{1-2\lambda}\left[\frac{2cf(c) + (v - (1-2\lambda)c)\frac{e^{-\lambda c}}{(1-2\lambda)^{v/2}}\mathbb{P}\left(\chi_v^2 > c(1-2\lambda)\right)}{\mathbb{P}\left(\chi_v^2 < c\right) + \frac{e^{-\lambda c}}{(1-2\lambda)^{v/2}}\mathbb{P}\left(\chi_v^2 > c(1-2\lambda)\right)} - (1-2\lambda)\mu\right] \\
&= \frac{1}{1-2\lambda}\left[\frac{2cf(c) - (v-c)\mathbb{P}\left(\chi_v^2 < c\right) - 2\lambda c\mathbb{P}\left(\chi_v^2 < c\right)}{\mathbb{P}\left(\chi_v^2 < c\right) + \frac{e^{-\lambda c}}{(1-2\lambda)^{v/2}}\mathbb{P}\left(\chi_v^2 > c(1-2\lambda)\right)} + (v - (1-2\lambda)c) - (1-2\lambda)\mu\right] \\
&= \frac{1}{1-2\lambda}\left[\frac{\mu - (v-c) - 2\lambda c\mathbb{P}\left(\chi_v^2 < c\right)}{\mathbb{P}\left(\chi_v^2 < c\right) + \frac{e^{-\lambda c}}{(1-2\lambda)^{v/2}}\mathbb{P}\left(\chi_v^2 > c(1-2\lambda)\right)} + (v - c - \mu) + 2(\mu + c)\lambda\right].
\end{aligned}$$

Next note that

$$\begin{aligned}
\mathbb{P}\left(\chi_v^2 > c(1-2\lambda)\right) &= \int_{c(1-2\lambda)}^{\infty} \frac{1}{2^{\frac{v}{2}}\Gamma\left(\frac{v}{2}\right)} x^{\frac{v}{2}-1} e^{-x/2} dx \\
&= \frac{1}{2^{\frac{v}{2}}\Gamma\left(\frac{v}{2}\right)} e^{\lambda c} \int_c^{\infty} \left(\frac{y}{y - 2\lambda c}\right)^{1-\frac{v}{2}} y^{\frac{v}{2}-1} e^{-y/2} dx.
\end{aligned}$$

When $v \leq 2$, this shows that:

$$\mathbb{P}\left(\chi_v^2 > c\right) < e^{-\lambda c}\mathbb{P}\left(\chi_v^2 > c(1-2\lambda)\right) < \frac{\mathbb{P}\left(\chi_v^2 > c\right)}{(1-2\lambda)^{1-\frac{v}{2}}}. \tag{B.3.1}$$

We can now use this result to further bound the MGF of the truncated χ_1^2. We consider two cases separately:

Case 1: $\mu - (v-c) - 2\lambda c\mathbb{P}\left(\chi_v^2 < c\right) \geq 0$. The lower bound in B.3.1 shows that $\frac{d}{d\lambda}\left(\log\left(\mathbb{E}\left(e^{\lambda Z}\right)\right)\right) - \mu$ is bounded by

$$\begin{aligned}
&\frac{1}{1-2\lambda}\left[\frac{\mu - (v-c) - 2\lambda c\mathbb{P}\left(\chi_v^2 < c\right)}{\mathbb{P}\left(\chi_v^2 < c\right) + \frac{1}{(1-2\lambda)^{\frac{v}{2}}}\mathbb{P}\left(\chi_v^2 > c\right)} + (v - c - \mu) + 2(\mu + c)\lambda\right] \\
&\leq \frac{1}{1-2\lambda}\left[\mu - (v-c) - 2\lambda c\mathbb{P}\left(\chi_v^2 < c\right) + (v - c - \mu) + 2(\mu + c)\lambda\right] = \frac{1}{1-2\lambda}\left[2(\mu + c\mathbb{P}\left(\chi_v^2 > c\right))\lambda\right] \\
&\leq \frac{2\lambda(1-\lambda)}{(1-2\lambda)^2}(\mu + c\mathbb{P}\left(\chi_v^2 > c\right)) = \frac{2\lambda(1-\lambda)}{(1-2\lambda)^2}(2cf(c) + v\mathbb{P}\left(\chi_v^2 > c\right))
\end{aligned}$$

Case 2: $\mu-(v-c)-2\lambda c\mathbb{P}\left(\chi_v^2<c\right)<0$. The upper bound in B.3.1 shows that $\frac{d}{d\lambda}\left(\log\left(\mathbb{E}\left(e^{\lambda Z}\right)\right)\right)-\mu$ is bounded by

$$\begin{aligned}
&\frac{1}{1-2\lambda}\left[\frac{\mu+(v-c)-2\lambda c\mathbb{P}\left(\chi_v^2<c\right)}{\mathbb{P}\left(\chi_v^2<c\right)+\frac{1}{1-2\lambda}\mathbb{P}\left(\chi_v^2>c\right)}+(v-c-\mu)+2(\mu+c)\lambda\right]\\
&=\frac{1}{1-2\lambda}\left[(v-c-\mu)\left(1-\frac{1}{\mathbb{P}\left(\chi_v^2<c\right)+\frac{\mathbb{P}(\chi_v^2>c)}{1-2\lambda}}\right)+2\lambda c\left(1-\frac{\mathbb{P}\left(\chi_v^2<c\right)}{\mathbb{P}\left(\chi_v^2<c\right)+\frac{\mathbb{P}(\chi_v^2>c)}{1-2\lambda}}\right)+2\lambda\mu\right]\\
&=\frac{1}{1-2\lambda}\left[(v-c-\mu)\frac{\frac{1}{1-2\lambda}\mathbb{P}\left(\chi_v^2>c\right)-\mathbb{P}\left(\chi_v^2>c\right)}{\mathbb{P}\left(\chi_v^2<c\right)+\frac{1}{1-2\lambda}\mathbb{P}\left(\chi_v^2>c\right)}+2\lambda c\left(\frac{\frac{1}{1-2\lambda}\mathbb{P}\left(\chi_v^2>v\right)}{\mathbb{P}\left(\chi_v^2<c\right)+\frac{1}{1-2\lambda}\mathbb{P}\left(\chi_v^2>c\right)}\right)+2\lambda\mu\right]\\
&=\frac{1}{1-2\lambda}\left[(v-c-\mu)\frac{2\lambda\mathbb{P}\left(\chi_v^2>c\right)}{1-2\lambda\mathbb{P}\left(\chi_v^2<c\right)}+2\lambda c\frac{\mathbb{P}\left(\chi_v^2>c\right)}{1-2\lambda\mathbb{P}\left(\chi_v^2<c\right)}+2\lambda\mu\right]\\
&=\frac{2\lambda}{1-2\lambda}\left[\mu+(v-\mu)\frac{\mathbb{P}\left(\chi_v^2>c\right)}{1-2\lambda\mathbb{P}\left(\chi_v^2<c\right)}\right]\\
&=\frac{2\lambda}{(1-2\lambda)^2}\left[\mu\left(1-2\lambda\right)+(v-\mu)\mathbb{P}\left(\chi_v^2>c\right)\frac{1-2\lambda}{1-2\lambda\mathbb{P}\left(\chi_v^2<c\right)}\right]\\
&=\frac{2\lambda}{(1-2\lambda)^2}\left[\mu\left(1-2\lambda\right)+(v-\mu)\mathbb{P}\left(\chi_v^2>c\right)-(v-\mu)\mathbb{P}\left(\chi_v^2>c\right)\frac{2\lambda\mathbb{P}\left(\chi_v^2>c\right)}{1-2\lambda\mathbb{P}\left(\chi_v^2<c\right)}\right]
\end{aligned}$$

Using the fact that $\mu+(v-c)-2\lambda c\mathbb{P}\left(\chi_v^2<c\right)<0$ and that $v-\mu\geq 0$, we can bound this by

$$\begin{aligned}
&=\frac{2\lambda}{(1-2\lambda)^2}\left[\mu\left(1-\lambda\right)+(v-\mu)\mathbb{P}\left(\chi_v^2>c\right)-2\lambda c\mathbb{P}\left(\chi_v^2>c\right)^2\right]\\
&=\frac{2\lambda}{(1-2\lambda)^2}\left[\left(\mu+c\mathbb{P}\left(\chi_v^2>c\right)\right)\left(1-\lambda\right)+(v-\mu-c)\mathbb{P}\left(\chi_v^2>c\right)-2\lambda c\mathbb{P}\left(\chi_v^2>c\right)^2+c\lambda\mathbb{P}\left(\chi_v^2>c\right)\right]
\end{aligned}$$

Since $\lambda<\frac{1}{2}$ and $v-c-\mu\leq 0$, we have that

$$\begin{aligned}
&(v-\mu-c)\mathbb{P}\left(\chi_v^2>c\right)-2\lambda c\mathbb{P}\left(\chi_v^2>c\right)^2+c\lambda\mathbb{P}\left(\chi_v^2>c\right)\leq\lambda\mathbb{P}\left(\chi_v^2>c\right)\left(2\left(v-c-\mu\right)+c-2c\mathbb{P}\left(\chi_v^2>c\right)\right)\\
&=\lambda\mathbb{P}\left(\chi_v^2>c\right)\left(2\left(\mathbb{E}\left(\chi_v^2|\chi_v^2<c\right)\mathbb{P}\left(\chi_v^2<c\right)-c\mathbb{P}\left(\chi_v^2<c\right)\right)+c-2c\mathbb{P}\left(\chi_v^2>c\right)\right)\\
&=\lambda\mathbb{P}\left(\chi_v^2>c\right)\left(2\mathbb{E}\left(\chi_v^2|\chi_v^2<c\right)\mathbb{P}\left(\chi_v^2<c\right)-c\right)\leq 0
\end{aligned}$$

where the last inequality follows from the fact that $\mathbb{E}\left(\chi_v^2|\chi_v^2<c\right)\leq c/2$, which is due to the fact that the pdf of the χ_v^2-distribution is decreasing.

Consequently,

$$\frac{d}{d\lambda}\left(\log\left(\mathbb{E}\left(e^{\lambda Z}\right)\right)-\lambda\mu\right)\leq\frac{2\lambda(1-\lambda)}{(1-2\lambda)^2}(2cf(c)+v\mathbb{P}\left(\chi_v^2>c\right))=\frac{d}{d\lambda}\left(\frac{2(2cf(c)+v\mathbb{P}\left(\chi_v^2>c\right))\lambda^2}{2(1-2\lambda)}\right).$$

This shows that

$$\log\left(\mathbb{E}\left(e^{\lambda(Z-\mu)}\right)\right) \leq \frac{2(2cf(c)+v\mathbb{P}\left(\chi_v^2 > c\right))\lambda^2}{2(1-2\lambda)},$$

which finishes the proof.

B.3.3 Proof of Lemma 16

It is sufficient to show that

$$\mathbb{P}\left(Y_i \geq a+x | Y_i \geq a, v_i = 1\right) \geq \mathbb{P}\left(Z > a+x | Z \geq a\right).$$

We have that

$$\mathbb{P}\left(Y_i \geq a+x | Y_i \geq a, v_i = 1\right) = \frac{\mathbb{P}\left(\epsilon_i > \sqrt{a+x} - \mu\right)}{\mathbb{P}\left(\epsilon_i > \sqrt{a} - \mu\right)}$$

The derivative of left hand side with respect to μ is

$$\frac{\mathbb{P}\left(\epsilon_i > \sqrt{a+x} - \mu\right)}{\mathbb{P}\left(\epsilon_i > \sqrt{a} - \mu\right)} \left(\frac{\phi(\sqrt{a+x} - \mu)}{\mathbb{P}\left(\epsilon_i > \sqrt{a+x} - \mu\right)} - \frac{\phi(\sqrt{a} - \mu)}{\mathbb{P}\left(\epsilon_i > \sqrt{a} - \mu\right)} \right)$$

This is greater than 0, since the hazard rate of the Gaussian is increasing. Hence,

$$\begin{aligned}
\mathbb{P}\left(Y_i \geq a+x | Y_i \geq a, v_i = 1\right) &= \frac{\mathbb{P}\left(\epsilon_i > \sqrt{a+x} - \mu\right)}{\mathbb{P}\left(\epsilon_i > \sqrt{a} - \mu\right)} \\
&\geq \frac{\mathbb{P}\left(\epsilon_i > \sqrt{a+x}\right)}{\mathbb{P}\left(\epsilon_i > \sqrt{a}\right)} \\
&= \mathbb{P}\left(Z > a+x | Z \geq a\right).
\end{aligned}$$

B.3.4 Proof of Lemma 17

Let $Z \sim \chi_1^2$ and write $\mu = \mathbb{E}\left((Z-a)^+\right)$. The MGF $G(\lambda)$ of the random variable

$$W = (a-Z) | (Z > a) + \frac{\mu}{\mathbb{P}\left(\chi_1^2 > a\right)}$$

is then

$$G(\lambda) = \exp\left(\frac{\lambda\mu}{\mathbb{P}\left(\chi_1^2 > a\right)}\right)\frac{1}{\mathbb{P}\left(\chi_1^2 > a\right)}\int_0^\infty \frac{1}{\sqrt{2\pi x}}e^{\lambda a - \lambda z x - \frac{1}{2}x}dx$$
$$= \exp\left(\frac{\lambda\mu}{\mathbb{P}\left(\chi_1^2 > a\right)} + \lambda a\right)\frac{\mathbb{P}\left(\chi_1^2 > a(1+2\lambda)\right)}{\mathbb{P}\left(\chi_1^2 > a\right)\sqrt{1+2\lambda}}.$$

Consequently, $\frac{dG(\lambda)}{d\lambda}$ is equal to

$$\frac{1}{\mathbb{P}\left(\chi_1^2 > a\right)}\left[-\frac{2af(a)}{1+2\lambda}e^{-\lambda a} + \left(\frac{\mu}{\mathbb{P}\left(\chi_1^2 > a\right)} + a - \frac{1}{1+2\lambda}\right)\frac{\mathbb{P}\left(\chi_1^2 > a(1+2\lambda)\right)}{\sqrt{1+2\lambda}}\right]\exp\left(\frac{\lambda\mu}{\mathbb{P}\left(\chi_1^2 > a\right)} + \lambda a\right).$$

Therefore,

$$\frac{d\log\left(G(\lambda)\right)}{d\lambda} = \frac{\mu}{\mathbb{P}\left(\chi_1^2 > a\right)} + a - \frac{1}{1+2\lambda} - \frac{2af(a)}{\sqrt{1+2\lambda}}\frac{e^{-\lambda a}}{\mathbb{P}\left(\chi_1^2 > a(1+2\lambda)\right)}$$

Since,

$$\mathbb{P}\left(\chi_1^2 > a(1+2\lambda)\right) = \int_{a(1+2\lambda)}^\infty \frac{1}{\sqrt{2\pi x}}e^{-\frac{x}{2}}dx$$
$$= e^{-\lambda a}\int_a^\infty \frac{1}{\sqrt{2\pi y}}\sqrt{\frac{1}{1+2\lambda\frac{a}{y}}}e^{-\frac{y}{2}}dy \leq e^{-\lambda a}\mathbb{P}\left(\chi_1^2 > a\right),$$

we must also have

$$\frac{d\log\left(G(\lambda)\right)}{d\lambda} < \frac{2\lambda}{1+2\lambda}\left(1 + \frac{2af(a)}{\mathbb{P}\left(\chi_1^2 > a\right)}\right) \leq 2\lambda\left(1 + \frac{2af(a)}{\mathbb{P}\left(\chi_1^2 > a\right)}\right)$$

and therefore

$$G(\lambda) \leq \frac{\lambda^2 2\left(1 + \frac{2af(a)}{\mathbb{P}\left(\chi_1^2 > a\right)}\right)}{2}.$$

This proves that W is sub-Gaussian. Standard tail bounds for sub-Gaussian random variables then imply that independent random variables $W_1, ..., W_k$ obeying the same law as W satisfy

$$\mathbb{P}\left(\sum_{i=1}^k > 2\sqrt{\left(1 + \frac{2af(a)}{\mathbb{P}\left(\chi_1^2 > a\right)}\right)kt}\right) < e^{-t},$$

for positive integers k and all $t \in \mathbb{R}$. This finishes the proof.

B.3.5 Proof of Lemma 18

The equality follows from Lemma 15. To prove the inequality, write $G(\tau) = \tau + 2af(a)$, where $0 \leq \tau \leq 1$ and a is defined by the equation $\mathbb{P}(\chi_1^2 > a) = \tau$. Note that $G(0) = 0$ and

$$\frac{dG}{d\tau} = 1 + \frac{da}{d\tau}(f(a) - af(a)) = 1 - \frac{1}{f(a)}(f(a) - af(a)) = a > 0$$

Hence, $m + 2paf(a) = pG(\frac{m}{p})$ is increasing in m. Moreover the following bounds hold on a:

$$2\tau = 2\mathbb{P}\left(\chi_1^2 > a\right) < \mathbb{P}\left(\chi_2^2 > 2a\right) = \exp(-a).$$

Therefore, we have that

$$G(\tau) \leq \int_0^\tau -2\log(x)dx = -2\tau\log(\tau) + 2\tau = 2\tau\log\left(\frac{1}{\tau}\right) + 2\tau.$$

Noting that $m + 2paf(a) = pG(\frac{m}{p})$, finishes the proof.

B.3.6 Proof of Lemma 19

We know from Lemma 15, that

$$\mathbb{E}\left((\chi_1^2 - b)^+ | \chi_1^2 > b\right) = 1 - b + 2bf(b)\mathbb{P}\left(\chi_1^2 > b\right)^{-1}.$$

Next note that

$$\mathbb{P}\left(\chi_1^2 > b\right) = \int_b^\infty \sqrt{\frac{2}{\pi x}} e^{-x/2} dx \leq \int_b^\infty \sqrt{\frac{2}{\pi b}} e^{-x/2} dx = 2f(b)$$

Hence,

$$\mathbb{E}\left((\chi_1^2 - b)^+ | \chi_1^2 > b\right) = 1 - b + 2bf(b)\mathbb{P}\left(\chi_1^2 > b\right)^{-1} \geq 1 - b + b = 1.$$

This finishes the proof.

B.3.7 Proof of Lemma 20

Let $\eta_1, ..., \eta_{s+w} \overset{i.i.d.}{\sim} N(0,1)$ for some positive integer s. Define

$$Z_s := \max_{0\le a\le w} (s+a)\left(\bar{\eta}_{1:(s+a)}\right)^2.$$

Write $T_a = \sum_{t=1}^{a} \eta_t$ and note that $e^{\lambda T_a}$ is a super-martingale for all $\lambda > 0$. The following holds:

$$\begin{aligned}
&\mathbb{P}(Z_s > u) \le \sum_{i=\lfloor \log_b(s)\rfloor}^{\lceil \log_b(s+w)\rceil} \mathbb{P}\left(\max_{b^i\le s+a\le b^{i+1}} (s+a)\left(\bar{\eta}_{1:(s+a)}\right)^2 > u\right)\\
&\le \sum_{i=\lfloor \log_b(s)\rfloor}^{\lceil \log_b(s+w)\rceil} \mathbb{P}\left(\max_{b^i\le a'\le b^{i+1}} (T_{a'})^2 > b^i u\right)\\
&\le 2\sum_{i=\lfloor \log_b(s)\rfloor}^{\lceil \log_b(s+w)\rceil} \mathbb{P}\left(\max_{b^i\le a'\le b^{i+1}} T_{a'} > \sqrt{b^i u}\right) = 2\sum_{i=\lfloor \log_b(s)\rfloor}^{\lceil \log_b(s+w)\rceil} \min_{\lambda}\left[\mathbb{P}\left(\max_{b^i\le a'\le b^{i+1}} e^{\lambda T_{a'}} > e^{\sqrt{b^i u}\lambda}\right)\right]\\
&\le 2\sum_{i=\lfloor \log_b(s)\rfloor}^{\lceil \log_b(s+w)\rceil} \min_{\lambda}\left[\mathbb{E}\left(e^{\lambda T_{\lfloor b^{i+1}\rfloor}}\right) e^{-\sqrt{b^i u}\lambda}\right] = 2\sum_{i=\lfloor \log_b(s)\rfloor}^{\lceil \log_b(s+w)\rceil} \min_{\lambda}\left[e^{\frac{\lfloor b^{i+1}\rfloor}{2}\lambda^2 - \sqrt{b^i u}\lambda}\right]\\
&\le 2\sum_{i=\lfloor \log_b(s)\rfloor}^{\lceil \log_b(s+w)\rceil} \min_{\lambda}\left[e^{\frac{b^{i+1}}{2}\lambda^2 - \sqrt{b^i u}\lambda}\right] = 2\sum_{i=\lfloor \log_b(s)\rfloor}^{\lceil \log_b(s+w)\rceil} e^{-\frac{u}{2b}} = 2(1+\lceil \log_b(s+w)\rceil - \lfloor \log_b(s)\rfloor)e^{-\frac{u}{2b}}\\
&\le 2(3+\log_b(s+w)-\log_b(s))e^{-\frac{u}{2b}} = 2(3+\log_b(1+w/s))e^{-\frac{u}{2b}} \le 2(3+\log_b(w+1))e^{-\frac{u}{2b}}\\
&\le 6\frac{1+\log(w+1)}{\log(b)}e^{-\frac{u}{2b}}.
\end{aligned}$$

Here the fifth inequality follows from Doob's martingale inequality.

Next note that

$$\begin{aligned}
&\mathbb{P}\left(\max_{0\le f,d\le w: j-f-d-i\ge 0}\left((j-f-d-i+1)\left(\bar{\boldsymbol{\eta}}^{(c)}_{(i+d):(j-f)}\right)^2\right) > u\right)\\
&\le \sum_{d=0}^{w}\mathbb{P}\left(\max_{0\le f\le \min(w, j-i-d)}\left((j-f-d-i+1)\left(\bar{\boldsymbol{\eta}}^{(c)}_{(i+d):(j-f)}\right)^2\right) > u\right)\\
&\le \sum_{d=0}^{w}\mathbb{P}\left(Z_{\max(1,j-i-2w)} > u\right)\\
&\le 6(w+1)\frac{1+\log(w+1)}{\log(b)}e^{-\frac{u}{2b}}.
\end{aligned}$$

B.3.8 Proof of Lemma 21

In this section, we define the event that E_a holds for a given set tuple (i, j) to be $E_a^{(i,j)}$, for $a = 1, ..., 11$. We know from the proof of Propositions 4 that

$$\mathbb{P}\left(E_1^{(i,j)}\right) > 1 - A_1 e^{-\psi}$$

holds. A Bonferroni correction over all possible tuples (i, j) then gives $\mathbb{P}(E_1) > 1 - A_1 n^2 e^{-\psi}$. Furthermore, we have that

$$\mathbb{P}\left(E_2^{(i,j)}\right) = \mathbb{P}\left(\sum_{c=1}^{p}(j-i+1)\left(\bar{\boldsymbol{\eta}}_{i:j}^{(c)}\right)^2 < p + 2\psi + 2\sqrt{p\psi}\right) = \mathbb{P}\left(\chi_p^2 < p + 2\psi + 2\sqrt{p\psi}\right)$$
$$\geq 1 - e^{-\psi},$$

with the inequality following from the tail bounds proven in Laurent and Massart (2000). A Bonferroni correction then gives $\mathbb{P}(E_2) > 1 - n^2 e^{-\psi}$. Next note that for any fixed fixed i, j, and c

$$\mathbb{P}\left((j-i+1)\left(\bar{\boldsymbol{\eta}}_{i:j}^{(c)} + \bar{\boldsymbol{\mu}}_{i:j}^{(c)}\right)^2 > s\right) \geq \mathbb{P}\left((j-i+1)\left(\bar{\boldsymbol{\eta}}_{i:j}^{(c)}\right)^2 > s\right)$$

holds for all $s \geq 0$. Therefore,

$$\mathbb{P}\left(\sum_{c=1}^{p}(j-i+1)\left(\bar{\boldsymbol{\eta}}_{i:j}^{(c)} + \bar{\boldsymbol{\mu}}_{i:j}^{(c)}\right)^2 > p - 2\sqrt{p\psi}\right) \geq \mathbb{P}\left(\sum_{c=1}^{p}(j-i+1)\left(\bar{\boldsymbol{\eta}}_{i:j}^{(c)}\right)^2 > p - 2\sqrt{p\psi}\right) \geq 1 - e^{-\psi},$$

with the last inequality again flowing from Laurent and Massart (2000). A Bonferroni correction then gives $\mathbb{P}(E_3) > 1 - n^2 e^{-\psi}$. Next note that

$$\frac{1}{\sqrt{|S|}}\sum_{c\in S}\sqrt{j-i+1}\bar{\boldsymbol{\eta}}_{i:j} \sim N(0,1)$$

We can then use the well known tail bounds on the Normal distribution to show that

$$\mathbb{P}\left(\left|\frac{1}{\sqrt{|S|}}\sum_{c\in S}\sqrt{j-i+1}\bar{\boldsymbol{\eta}}_{i:j}\right| < \sqrt{2\psi + 2|S|\log(p)}\right) \geq 1 - A_4 p^{-|S|} e^{-\psi},$$

for a constant A_4. A Bonferroni correction over all possible sets S then shows that

$$\begin{aligned}
&\mathbb{P}\left(\left|\frac{1}{\sqrt{|S|}}\sum_{c\in S}\sqrt{j-i+1}\bar{\boldsymbol{\eta}}_{i:j}\right| < \sqrt{2\psi+2|S|\log(p)} \quad \forall S\subset\{1,...,p\}\right)\\
&\geq 1-\sum_{m=1}^{p}|\{S|S\subset\{1,...,p\},|S|=m\}|A_4p^{-|m|}e^{-\psi} \geq 1-\sum_{m=1}^{p}\frac{p!}{(p-m)!m!}A_4p^{-|m|}e^{-\psi}\\
&\geq 1-\sum_{m=1}^{p}\frac{1}{m!}p^mA_4p^{-|m|}e^{-\psi} \geq 1-(A_4e)\,e^{-\psi}.
\end{aligned}$$

A Bonferroni correction over the indices i and j then proves that $\mathbb{P}(E_4) > 1-(A_4e)\,n^2e^{-\psi}$. Next, for fixed i and j,

$$\begin{aligned}
&\mathbb{P}\left(\sum_{c\notin S}(j-i+1)\left(\bar{\boldsymbol{\eta}}_{i:j}^{(c)}+\bar{\boldsymbol{\mu}}_{i:j}^{(c)}\right)^2 > p-2\sqrt{p\psi}-2\psi-2|S|\log(p)\right)\\
&\geq \mathbb{P}\left(\sum_{c\notin S}(j-i+1)\left(\bar{\boldsymbol{\eta}}_{i:j}^{(c)}\right)^2 > p-2\sqrt{p\psi}-2\psi-2|S|\log(p)\right)\\
&\geq 1-\mathbb{P}\left(\sum_{i=1}^{p}(j-i+1)\left(\bar{\boldsymbol{\eta}}_{i:j}^{(c)}\right)^2 \leq p-2\sqrt{p\psi}\right)-\mathbb{P}\left(\sum_{c\in S}(j-i+1)\left(\bar{\boldsymbol{\eta}}_{i:j}^{(c)}\right)^2 > 2\psi-2|S|\log(p)\right)\\
&\geq 1-(1+A_1)e^{-\psi}.
\end{aligned}$$

A Bonferroni correction over all indices i and j then gives $\mathbb{P}(E_5) > 1-(1+A_1)n^2e^{-\psi}$.

Next we note that

$$\left(\sum_{c\in S}\left(\sum_{t=i}^{j}\left(\boldsymbol{\mu}_t^{(c)}-\bar{\boldsymbol{\mu}}_{i:j}^{(c)}\right)\boldsymbol{\eta}_t^{(c)}\right)\right)\left(\sqrt{\sum_{c\in S}\sum_{t=i}^{j}\left(\boldsymbol{\mu}_t^{(c)}-\bar{\boldsymbol{\mu}}_{i:j}^{(c)}\right)^2}\right)^{-1} \sim N(0,1).$$

Consequently,

$$\begin{aligned}
&\mathbb{P}\left(\left|\left(\sum_{c\in S}\left(\sum_{t=i}^{j}\left(\boldsymbol{\mu}_t^{(c)}-\bar{\boldsymbol{\mu}}_{i:j}^{(c)}\right)\boldsymbol{\eta}_t^{(c)}\right)\right)\left(\sqrt{\sum_{c\in S}\sum_{t=i}^{j}\left(\boldsymbol{\mu}_t^{(c)}-\bar{\boldsymbol{\mu}}_{i:j}^{(c)}\right)^2}\right)^{-1}\right| > \sqrt{2\psi+2\,|S\cap W_{i,j}|\log(p)}\right)\\
&\qquad\qquad \leq A_4p^{-|S\cap W_{i,j}|}e^{-\psi},
\end{aligned}$$

for some constant A_4. A Bonferroni correction over the sets S then gives that

$$\begin{aligned}
&\mathbb{P}\left(\frac{\left|\left(\sum_{c\in S}\left(\sum_{t=i}^{j}\left(\boldsymbol{\mu}_t^{(c)}-\bar{\boldsymbol{\mu}}_{i:j}^{(c)}\right)\boldsymbol{\eta}_t^{(c)}\right)\right)\right|}{\sqrt{\sum_{c\in S}\sum_{t=i}^{j}\left(\boldsymbol{\mu}_t^{(c)}-\bar{\boldsymbol{\mu}}_{i:j}^{(c)}\right)^2}}\leq\sqrt{2\psi+2\left|S\cap W_{i,j}\right|\log(p)}\quad\forall S\subset\{1,...,p\}\right)\\
&=\mathbb{P}\left(\frac{\left|\left(\sum_{c\in S\cap W_{i,j}}\left(\sum_{t=i}^{j}\left(\boldsymbol{\mu}_t^{(c)}-\bar{\boldsymbol{\mu}}_{i:j}^{(c)}\right)\boldsymbol{\eta}_t^{(c)}\right)\right)\right|}{\sqrt{\sum_{c\in S\cap W_{i,j}}\sum_{t=i}^{j}\left(\boldsymbol{\mu}_t^{(c)}-\bar{\boldsymbol{\mu}}_{i:j}^{(c)}\right)^2}}\leq\sqrt{2\psi+2\left|S\cap W_{i,j}\right|\log(p)}\quad\forall S\subset\{1,...,p\}\right)\\
&=\mathbb{P}\left(\frac{\left|\left(\sum_{c\in W}\left(\sum_{t=i}^{j}\left(\boldsymbol{\mu}_t^{(c)}-\bar{\boldsymbol{\mu}}_{i:j}^{(c)}\right)\boldsymbol{\eta}_t^{(c)}\right)\right)\right|}{\sqrt{\sum_{c\in W}\sum_{t=i}^{j}\left(\boldsymbol{\mu}_t^{(c)}-\bar{\boldsymbol{\mu}}_{i:j}^{(c)}\right)^2}}\leq\sqrt{2\psi+2\left|W\right|\log(p)}\quad\forall W\subset W_{i,j}\right)\\
&\leq 1-\sum_{W\subset W_{i,j}}\mathbb{P}\left(\frac{\left|\left(\sum_{c\in W}\left(\sum_{t=i}^{j}\left(\boldsymbol{\mu}_t^{(c)}-\bar{\boldsymbol{\mu}}_{i:j}^{(c)}\right)\boldsymbol{\eta}_t^{(c)}\right)\right)\right|}{\sqrt{\sum_{c\in W}\sum_{t=i}^{j}\left(\boldsymbol{\mu}_t^{(c)}-\bar{\boldsymbol{\mu}}_{i:j}^{(c)}\right)^2}}>\sqrt{2\psi+2\left|W\right|\log(p)}\quad\forall W\subset W_{i,j}\right)\\
&\leq 1-\sum_{|W|=1}^{|W_{i:j}|}\frac{p!}{(p-|W|)!(|W|)!}A_4p^{-|W|}e^{-\psi}\leq 1-(A_4e)e^{-\psi}
\end{aligned}$$

We note that

$$\frac{(j-j')(j'-i+1)}{j-i+1}\left(\bar{\boldsymbol{\eta}}_{i:j'}^{(c)}-\bar{\boldsymbol{\eta}}_{(j'+1):j}^{(c)}\right)^2\sim\chi_1^2,$$

and

$$\mathbb{P}\left(\frac{(j-j')(j'-i+1)}{j-i+1}\left(\bar{\mathbf{x}}_{i:j'}^{(c)}-\bar{\mathbf{x}}_{(j'+1):j}^{(c)}\right)^2>t\right)\geq\mathbb{P}\left(\chi_1^2>t\right)\quad\forall t>0.$$

Therefore, proving that constants A_7, A_8, and A_9 exist such that $\mathbb{P}\left(E_7^{(i,j',j)}\right)>1-A_7e^{-\psi}$, $\mathbb{P}\left(E_8^{(i,j',j)}\right)>1-A_8e^{-\psi}$, and $\mathbb{P}\left(E_9^{(i,j',j)}\right)>1-A_9e^{-\psi}$ hold for fixed i, j', and j is equivalent to proving the existence of constants A_1, A_2, and A_3 such that $\mathbb{P}\left(E_1^{(i,j)}\right)>1-A_1e^{-\psi}$, $\mathbb{P}\left(E_2^{(i,j)}\right)>1-A_2e^{-\psi}$, and $\mathbb{P}\left(E_3^{(i,j)}\right)>1-A_3e^{-\psi}$ hold. This was already done earlier in the proof. A Bonferroni correction over all possible i, j', and j then yields $\mathbb{P}\left(E_7\right)>1-A_7n^3e^{-\psi}$, $\mathbb{P}\left(E_8\right)>1-A_8n^3e^{-\psi}$, and $\mathbb{P}\left(E_9\right)>1-A_9n^3e^{-\psi}$.

The fact that

$$\left(\sum_{c\in\mathbf{J}_k}\sqrt{j-i+1}\bar{\boldsymbol{\eta}}^{(c)}_{i:j}\right)\left(\sqrt{2|\mathbf{J}_k|\psi}\right)^{-1}\sim N(0,1)$$

shows that $\mathbb{P}\left(E_{10}^{(i,j)}\right)>1-A_{10}e^{-\psi}$ for some constant A_{10}. The cardinality of the set of allowed tuples (i,j) is strictly less than n^2. Consequently $\mathbb{P}\left(E_{10}\right)>1-A_{10}n^2e^{-\psi}$. The same argument can be used to show that $\mathbb{P}\left(E_{11}^{(i,e_k,j)}\right)>1-A_{11}e^{-\psi}$. A Bonferroni correction over all triplets (i,e_k,j) then proves that $\mathbb{P}\left(E_{11}\right)>1-A_{11}n^3e^{-\psi}$

B.3.9 Proof of Lemma 22

This Lemma can be proven using straightforward algebra.

$$\begin{aligned}\mathcal{S}\left(\mathbf{x}_{i:j},\mathbf{J}\right)&=\sum_{c\in\mathbf{J}}(j-i+1)\left(\bar{\mathbf{x}}^{(c)}_{i:j}\right)^2=\sum_{c\in\mathbf{J}}(j-i+1)^{-1}\left(\bar{\mathbf{x}}^{(c)}_{i:j}\right)^2\\&=\sum_{c\in\mathbf{J}}\left[\frac{\left((j'+1-i)\bar{\mathbf{x}}^{(c)}_{i:j'}+(j-j')\bar{\mathbf{x}}^{(c)}_{(j'+1):j}\right)^2}{j-i+1}\right].\end{aligned}$$

Next we note that the following holds for all a, b, y, and z:

$$\begin{aligned}\frac{(ay+bz)^2}{a+b}&=\frac{a^2y^2+2abyz+b^2z^2}{a+b}=ay^2+bz^2+\frac{-aby^2+2abyz-baz^2}{a+b}\\&=ay^2+bz^2-\frac{ab}{a+b}(y-z)^2.\end{aligned}$$

Thus,

$$\mathcal{S}\left(\mathbf{x}_{i:j},\mathbf{J}\right)=\sum_{c\in S}(j'+1-i)\left(\bar{\mathbf{x}}^{(c)}_{i:j'}\right)^2+\sum_{c\in S}(j-j')\left(\bar{\mathbf{x}}^{(c)}_{(j'+1):j}\right)^2-\sum_{c\in S}\frac{(j-j')(j'-i+1)}{j-i+1}\left(\bar{\mathbf{x}}^{(c)}_{i:j'}-\bar{\mathbf{x}}^{(c)}_{(j'+1):j}\right)^2,$$

which finishes the proof.

B.3.10 Proof of Lemma 23

This result deals with the p term of the penalty incurred for splitting a sparse fitted segment into two and follows directly from E_9 and Lemma 22. Indeed, by Lemma 22

implies that

$$\mathcal{C}\left(\mathbf{x}_{i:j'},\mathbf{1}\right)+\mathcal{C}\left(\mathbf{x}_{(j'+1):j},\mathbf{1}\right)-\mathcal{C}\left(\mathbf{x}_{i:j},\mathbf{1}\right)=p+C\psi+C\sqrt{p\psi}-\sum_{c=1}^{p}\frac{(j-j')(j'-i+1)}{j-i+1}\left(\bar{\mathbf{x}}_{i:j'}^{(c)}-\bar{\mathbf{x}}_{(j'+1):j}^{(c)}\right)^2.$$

Given E_9, the above is bounded by

$$p+C\psi+C\sqrt{p\psi}-p+2\sqrt{p\psi}=C\psi+C\sqrt{p\psi}+2\sqrt{p\psi}.$$

This finishes the proof.

B.3.11 Proof of Lemma 24

This lemma shows that merging two neighbouring fitted segments reduces the penalised cost by $O(C)$ and follows almost immediately from Lemma 22. We consider the cases $|\mathbf{J}_k|\leq k^*$ and $|\mathbf{J}_k|>k^*$ separately. Let $|\mathbf{J}_k|\leq k^*$. Then

$$\begin{aligned}
&\mathcal{C}\left(x_{i:j'},\mathbf{J}_k\right)+\mathcal{C}\left(x_{(j'+1):j},\mathbf{J}_k\right)-\mathcal{C}\left(x_{i:j},\mathbf{J}_k\right)\\
&=C\psi+C|\mathbf{J}_k|\log(p)-\sum_{c\in\mathbf{J}_k}\frac{(j-j')(j'-i+1)}{j-i+1}\left(\bar{\boldsymbol{\eta}}_{i:j'}^{(c)}-\bar{\boldsymbol{\eta}}_{(j'+1):j}^{(c)}\right)^2\\
&\geq C\psi+C|\mathbf{J}_k|\log(p)-2\psi-2|\mathbf{J}_k|\log(p)\geq\frac{79}{80}C\left(\psi+|\mathbf{J}_k|\log(p)\right),
\end{aligned}$$

where the first inequality follows from E_7 and the second one holds if C exceeds some global constant. Now let $|\mathbf{J}_k|\geq k^*$

$$\begin{aligned}
&\mathcal{C}\left(x_{i:j'},\mathbf{1}\right)+\mathcal{C}\left(x_{(j'+1):j},\mathbf{1}\right)-\mathcal{C}\left(x_{i:j},\mathbf{1}\right)\\
&=p+C\psi+C\sqrt{p\psi}-\sum_{c=1}^{p}\frac{(j-j')(j'-i+1)}{j-i+1}\left(\bar{\boldsymbol{\eta}}_{i:j'}^{(c)}-\bar{\boldsymbol{\eta}}_{(j'+1):j}^{(c)}\right)^2\\
&\geq p+C\psi+C\sqrt{p\psi}-2\psi-2\sqrt{p\psi}-p\geq\frac{79}{80}C\left(\psi+\sqrt{p\psi}\right),
\end{aligned}$$

where the first inequality follows from E_8 and the second one holds if C exceeds some global constant.

B.3.12 Proof of Lemma 25

This Lemma proves MVCAPA has power at detecting anomalous regions. We begin by considering the case in which J_k is dense. We have:

$$\mathcal{C}\left(\mathbf{x}_{i:j}, \mathbf{1}\right) = p + C\psi + C\sqrt{p\psi} - \sum_{c=1}^{p} (j-i+1)\left(\mathbf{x}_{i:j}^{(c)}\right)^2$$

$$= p + C\psi + C\sqrt{p\psi} - \sum_{c=1}^{p}(j-i+1)\left(\bar{\boldsymbol{\eta}}_{i:j}^{(c)}\right)^2 - |\mathbf{J}_k|\boldsymbol{\mu}_k^2(j-i+1) - 2\boldsymbol{\mu}_k\sqrt{j-i+1}\sum_{c\in\mathbf{J}_k}\left(\sqrt{j-i+1}\bar{\boldsymbol{\eta}}_{i:j}\right)$$

$$\leq C\psi + (C+2)\sqrt{p\psi} - |\mathbf{J}_k|\boldsymbol{\mu}_k^2(j-i+1) + 2\sqrt{(j-i+1)\boldsymbol{\mu}_k^2}\sqrt{2|\mathbf{J}_k|\psi}$$

$$\leq C\psi + (C+2)\sqrt{p\psi} - \frac{1}{2}|\mathbf{J}_k|\boldsymbol{\mu}_k^2(j-i+1) + 4\psi \leq C\psi + (C+2)\sqrt{p\psi} - \frac{1}{2}|\mathbf{J}_k|\boldsymbol{\mu}_k^2\frac{4C}{\triangle_k^2} + 4\psi$$

$$= (C+4)\psi + (C+2)\sqrt{p\psi} - 2C(\psi + \sqrt{p\psi}) \leq 0$$

with the first inequality following form E_{10} and E_3, the second from the AM-GM inequality, the third from the condition on $j-i+1$, and the last one holds if C exceeds a global constant.

The proof for when J_k is sparse is almost identical. We have that:

$$\mathcal{C}\left(\mathbf{x}_{i:j}, \mathbf{J}_k\right) = C\psi + C|\mathbf{J}_k|\log(p) - \sum_{c\in\mathbf{J}_k}(j-i+1)\left(\boldsymbol{\mu}_k + \bar{\boldsymbol{\eta}}_{i:j}^{(c)}\right)^2$$

$$= C\psi + C|\mathbf{J}_k|\log(p) - \sum_{c\in\mathbf{J}_k}(j-i+1)\left(\bar{\boldsymbol{\eta}}_{i:j}^{(c)}\right)^2 - |\mathbf{J}_k|\boldsymbol{\mu}_k^2(j-i+1) - 2\boldsymbol{\mu}_k\sqrt{j-i+1}\sum_{c\in\mathbf{J}_k}\left(\sqrt{j-i+1}\bar{\boldsymbol{\eta}}_{i:j}\right)$$

$$\leq C\psi + C|\mathbf{J}_k|\log(p) - |\mathbf{J}_k|\boldsymbol{\mu}_k^2(j-i+1) + 2\sqrt{(j-i+1)\boldsymbol{\mu}_k^2}\sqrt{2|\mathbf{J}_k|\psi + 2|\mathbf{J}_k|^2\log(p)}$$

$$\leq C\psi + C|\mathbf{J}_k|\log(p) - \frac{1}{2}|\mathbf{J}_k|\boldsymbol{\mu}_k^2(j-i+1) + 4\psi + 4|\mathbf{J}_k|\log(p)$$

$$s \leq (C+4)\psi + (C+4)|\mathbf{J}_k|\log(p) - \frac{1}{2}|\mathbf{J}_k|\boldsymbol{\mu}_k^2\frac{4C}{\triangle_k^2}$$

$$= (C+4)\psi + (C+4)|\mathbf{J}_k|\log(p) - 2C(\psi + |\mathbf{J}_k|\log(p)) \leq 0,$$

where the first inequality follows from E_4, the second from the AM-GM inequality, the third from the condition on $j-i+1$, and the last one holds if C exceeds a global constant.

B.3.13 Proof of Lemma 26

This Lemma prevents fitted changes from containing too many observations belonging to the typical distribution. We limit ourselves to proving the result for the first case, since the proof of the second case is symmetrical. We begin by proving the result for the case in which $|\mathbf{J}_k| \leq k^*$. Writing, $e' = e_k + \lceil 10\frac{C}{\triangle_k^2} \rceil$ and $s' = e_k - \lceil 10\frac{C}{\triangle_k^2} \rceil$

$$\begin{aligned}\mathcal{C}\left(x_{i:j}, \mathbf{J}_k\right) &\geq \mathcal{C}\left(x_{i:s'}, \mathbf{J}_k\right) + \mathcal{C}\left(x_{(s'+1):e'}, \mathbf{J}_k\right) + \mathcal{C}\left(x_{(e'+1):j}, \mathbf{J}_k\right) - 2C\psi - 2C|\mathbf{J}_k|\log(p) \\ &\geq \mathcal{C}\left(x_{i:s'}, \mathbf{J}_k\right) + \mathcal{C}\left(x_{(s'+1):e'}, \mathbf{J}_k\right) - (C+2)\left(\psi + |\mathbf{J}_k|\log(p)\right)\end{aligned}$$

Next, note that Lemma 22 implies that $\mathcal{C}\left(x_{(s'+1):e'}, \mathbf{J}_k\right)$ is equal to

$$\mathcal{C}\left(x_{(s'+1):e_k}, \mathbf{J}_k\right) + \mathcal{C}\left(x_{(e_k+1):e'}, \mathbf{J}_k\right) - C\left(\psi + |\mathbf{J}_k|\log(p)\right) + \sum_{c \in \mathbf{J}_k} \frac{e'-s'}{2}\left(\boldsymbol{\mu}_k + \bar{\boldsymbol{\eta}}_{(s'+1):e_k} - \bar{\boldsymbol{\eta}}_{(e_k+1):e'}\right)^2$$

Moreover, we have that

$$\begin{aligned}&\sum_{c \in \mathbf{J}_k} \frac{e'-s'}{2}\left(\boldsymbol{\mu}_k + \bar{\boldsymbol{\eta}}_{(s'+1):e_k} - \bar{\boldsymbol{\eta}}_{(e_k+1):e'}\right)^2 \\ &= \frac{e'-s'}{2}|\mathbf{J}_k|\boldsymbol{\mu}_k^2 - \boldsymbol{\mu}_k(e'-s')\sum_{c \in \mathbf{J}_k}\left(\bar{\boldsymbol{\eta}}_{(s'+1):e_k} - \bar{\boldsymbol{\eta}}_{(e_k+1):e'}\right) + \frac{e'-s'}{2}\sum_{c \in \mathbf{J}_k}\left(\bar{\boldsymbol{\eta}}_{(s'+1):e_k} - \bar{\boldsymbol{\eta}}_{(e_k+1):e'}\right)^2 \\ &\geq \frac{e'-s'}{2}|\mathbf{J}_k|\boldsymbol{\mu}_k^2 - 2\sqrt{(e'-s')\boldsymbol{\mu}_k^2}\sqrt{2|\mathbf{J}_k|\psi} \geq \frac{e'-s'}{3}|\mathbf{J}_k|\boldsymbol{\mu}_k^2 - 12\psi = \frac{20}{3}C\left(\psi + |\mathbf{J}_k|\log(p)\right) - 12\psi,\end{aligned}$$

where the first inequality follows from E_{11} and the second inequality from the AM-GM inequality. Combining all the above, we obtain that

$$\begin{aligned}&\mathcal{C}\left(x_{i:j}, \mathbf{J}_k\right) \\ &\geq \mathcal{C}\left(x_{i:s'}, \mathbf{J}_k\right) + \mathcal{C}\left(x_{(s'+1):e_k}, \mathbf{J}_k\right) + \mathcal{C}\left(x_{(e_k+1):e'}, \mathbf{J}_k\right) + \left(\frac{14}{3}C - 14\right)\psi + \left(\frac{14}{3}C - 2\right)|\mathbf{J}_k|\log(p) \\ &\geq \mathcal{C}\left(x_{i:e_k}, \mathbf{J}_k\right) + \frac{19}{20}C\left(\psi + |\mathbf{J}_k|\log(p)\right) + (C-2)\left(\psi + |\mathbf{J}_k|\log(p)\right) + \left(\frac{14}{3}C - 14\right)\psi \\ &\quad + \left(\frac{14}{3}C - 2\right)|\mathbf{J}_k|\log(p) \\ &\geq 6C\left(\psi + |\mathbf{J}_k|\log(p)\right),\end{aligned}$$

where the second inequality follows from Lemma 24 and E_2. The proof for the case in which $|\mathbf{J}_k| > k^*$ is very similar. We the have that

$$\begin{aligned}\mathcal{C}\left(x_{i:j}, \mathbf{1}\right) &\geq \mathcal{C}\left(x_{i:s'}, \mathbf{1}\right) + \mathcal{C}\left(x_{(s'+1):e'}, \mathbf{1}\right) + \mathcal{C}\left(x_{(e'+1):j}, \mathbf{1}\right) - 2C\psi - 2(C+2)\sqrt{p\psi} \\ &\geq \mathcal{C}\left(x_{i:s'}, \mathbf{1}\right) + \mathcal{C}\left(x_{(s'+1):e'}, \mathbf{1}\right) - (C+6)\left(\psi + \sqrt{p\psi}\right),\end{aligned}$$

with the first inequality following from Lemma 22 and the event E_3 the second being due to E_2. The remainder of the proof of the Lemma is very similar to the sparse case and has therefore been omitted.

B.3.14 Proof of Lemma 27

This Lemma shows that the optimal partition can not contain fitted segments containing more than $10\frac{C}{\triangle_k^2}$ observations from both the typical distribution and a dense anomalous region. If $\mathbf{J} = \mathbf{1}$ the result follows a fortiori from Lemma 26. Assume now that $|\mathbf{J}| \leq k^*$. As in the proof of Lemma 26, we limit ourselves to proving the first case, the proof of the other one being symmetrical. The following holds:

$$\begin{aligned}\mathcal{C}\left(x_{i:j}, \mathbf{J}\right) &= \mathcal{C}\left(x_{i:j}, \mathbf{1}\right) - (p + C\psi + C\sqrt{p\psi}) + \sum_{c \notin \mathbf{J}} (j - i + 1)\left(\bar{\mathbf{x}}_{i:j}\right)^2 + C\psi + C|\mathbf{J}|\log(p) \\ &\geq \mathcal{C}\left(x_{i:j}, \mathbf{1}\right) - C(\psi + \sqrt{p\psi}) - 2\sqrt{p\psi} - 2\psi - 2|\mathbf{J}|\log(p) + C\psi + C|\mathbf{J}|\log(p) \\ &\geq \mathcal{C}\left(x_{i:e_k}, \mathbf{1}\right) + 4C(\psi + \sqrt{p\psi}),\end{aligned}$$

where the first inequality follows from the event E_5 and the second one from Lemma 26 and a choice of C exceeding some global constant.

B.3.15 Proof of Lemma 28

This lemma shows that merging two neighbouring fitted segments reduces penalised cost by $O(C)$ – even when they are separated by a gap. The proof is very similar to that of Lemma 24. In fact, Lemma 28 follows a fortiori from Lemma 24 when $j' = j''$. When $j' \neq j''$ we consider the $|\mathbf{J}_k| \leq k^*$ and $|\mathbf{J}_k| > k^*$ separately. Let $|\mathbf{J}_k| \leq k^*$. Then,

$$\begin{aligned}
&\mathcal{C}\left(x_{i:j'}, \mathbf{J}_k\right) + \mathcal{C}\left(x_{(j''+1):j}, \mathbf{J}_k\right) - \mathcal{C}\left(x_{i:j}, \mathbf{J}_k\right) \\
&\geq \mathcal{C}\left(x_{i:j'}, \mathbf{J}_k\right) + \left[\mathcal{C}\left(x_{(j'+1):j''}, \mathbf{J}_k\right) - C\psi - C|\mathbf{J}_k|\log(p)\right] + \mathcal{C}\left(x_{(j''+1):j}, \mathbf{J}_k\right) - \mathcal{C}\left(x_{i:j}, \mathbf{J}_k\right) \\
&\geq -C\psi - C|\mathbf{J}_k|\log(p) + \frac{79}{80}C\left(\psi + |\mathbf{J}_k|\log(p)\right) + \frac{79}{80}C\left(\psi + |\mathbf{J}_k|\log(p)\right) \\
&\geq \frac{19}{20}C\left(\psi + |\mathbf{J}_k|\log(p)\right),
\end{aligned}$$

where the second inequality follows from applying Lemma 24 twice. The proof for the case in which $|\mathbf{J}_k| > k^*$ is very similar. We have that

$$\begin{aligned}
&\mathcal{C}\left(x_{i:j'}, \mathbf{1}\right) + \mathcal{C}\left(x_{(j''+1):j}, \mathbf{1}\right) - \mathcal{C}\left(x_{i:j}, \mathbf{1}\right) \\
&\geq \mathcal{C}\left(x_{i:j'}, \mathbf{1}\right) + \left[\mathcal{C}\left(x_{(j'+1):j''}, \mathbf{1}\right) - C\psi - (C+2)\sqrt{p\psi}\right] + \mathcal{C}\left(x_{(j''+1):j}, \mathbf{1}\right) - \mathcal{C}\left(x_{i:j}, \mathbf{1}\right) \\
&\geq -C\psi - (C+2)\sqrt{p\psi} + \frac{79}{80}C\left(\psi + \sqrt{p\psi}\right) + \frac{79}{80}C\left(\psi + \sqrt{p\psi}\right) \geq \frac{19}{20}C\left(\psi + \sqrt{p\psi}\right),
\end{aligned}$$

where the first inequality follows from E_3, the third inequality follows from applying Lemma 24 twice, and the third holds if C exceeds a global constant.

B.3.16 Proof of Lemma 29

This Lemma shows that if a fitted segment contains observations belonging to the typical distribution it can be trimmed to containing only anomalous observations

without increasing the penalised cost by more than $O(1)$. We begin by proving the sparse case

$$\begin{aligned}
\mathcal{C}\left(\mathbf{x}_{i:j}, \mathbf{J}\right) &\geq \mathcal{C}\left(\mathbf{x}_{i:j'}, \mathbf{J}\right) + \left(\mathcal{C}\left(\mathbf{x}_{(j'+1):j}, \mathbf{J}\right) - C\psi - C|\mathbf{J}|\log(p)\right) \\
&\geq \mathcal{C}\left(\mathbf{x}_{i:j'}, \mathbf{J}\right) - 2\psi - 2|\mathbf{J}|\log(p) \\
&\geq \left(\mathcal{C}\left(\mathbf{x}_{i:(i'-1)}, \mathbf{J}\right) - C\psi - C|\mathbf{J}|\log(p)\right) + \mathcal{C}\left(\mathbf{x}_{i':j'}, \mathbf{J}\right) - 2\psi - 2|\mathbf{J}|\log(p) \\
&\geq \mathcal{C}\left(\mathbf{x}_{i':j'}, \mathbf{J}\right) - 4\psi - 4|\mathbf{J}|\log(p),
\end{aligned}$$

where the first and third inequality follows from the fact that introducing free splits reduces un-penalised cost whilst the second and third inequality follows from E_1. Note that if $j' = j$ and/or $i' = i$ the first and second and/or the third and forth step are not necessary. The result nevertheless holds. A similar proof can be derived for the dense case:

$$\begin{aligned}
\mathcal{C}\left(\mathbf{x}_{i:j}, \mathbf{J}\right) &\geq \mathcal{C}\left(\mathbf{x}_{i:j'}, \mathbf{J}\right) + \left(\mathcal{C}\left(\mathbf{x}_{(j'+1):j}, \mathbf{J}\right) - C\psi - C\sqrt{p\psi}\right) - 2\sqrt{p\psi} \\
&\geq \mathcal{C}\left(\mathbf{x}_{i:j'}, \mathbf{J}\right) - 2\psi - 4\sqrt{p\psi} \\
&\geq \left(\mathcal{C}\left(\mathbf{x}_{i:(i'-1)}, \mathbf{J}\right) - C\psi - C\sqrt{p\psi}\right) + \mathcal{C}\left(\mathbf{x}_{i':j'}, \mathbf{J}\right) - 2\psi - 6\sqrt{p\psi} \\
&\geq \mathcal{C}\left(\mathbf{x}_{i':j'}, \mathbf{J}\right) - 4\psi - 8\sqrt{p\psi},
\end{aligned}$$

with the first and third inequalities following form Lemma 23, and the second and fourth from E_2.

B.3.17 Proof of Lemma 30

This Lemma links the savings of a fitted segment to the signal strength of the corresponding segment. We have that

$$\alpha\left(C\psi + C|\mathbf{J}|\log(p)\right) \leq \sum_{c\in\mathbf{J}}\left(\boldsymbol{\mu}_k + \bar{\boldsymbol{\eta}}_{i:j}^{(c)}\right)^2 (j-i+1)$$

$$= |\mathbf{J}\cap\mathbf{J}_k|(j-i+1)\boldsymbol{\mu}_k^2 + 2\sqrt{j-i+1}\boldsymbol{\mu}_k \sum_{c\in\mathbf{J}\cap\mathbf{J}_k}\sqrt{j-i+1}\bar{\boldsymbol{\eta}}_{i:j}^{(c)} + \sum_{c\in\mathbf{J}}\left(\sqrt{j-i+1}\bar{\boldsymbol{\eta}}_{i:j}^{(c)}\right)^2$$

$$\leq |\mathbf{J}\cap\mathbf{J}_k|(j-i+1)\boldsymbol{\mu}_k^2 + 2\sqrt{j-i+1}|\boldsymbol{\mu}_k|\sqrt{2\psi|\mathbf{J}\cap\mathbf{J}_k| + 2|\mathbf{J}\cap\mathbf{J}_k|^2\log(p)} + 2\psi + 2|\mathbf{J}|\log(p)$$

$$\leq |\mathbf{J}|(j-i+1)\boldsymbol{\mu}_k^2 + 2\sqrt{j-i+1}|\boldsymbol{\mu}_k|\sqrt{2\psi|\mathbf{J}| + 2|\mathbf{J}|^2\log(p)} + 2\psi + 2|\mathbf{J}|\log(p)$$

$$= \left(\sqrt{|\mathbf{J}|(j-i+1)\boldsymbol{\mu}_k^2} + \sqrt{2\psi + 2|\mathbf{J}|\log(p)}\right)^2,$$

with the first inequality following from E_1 and E_4 and th second from the fact that $|\mathbf{J}\cap\mathbf{J}_k| \leq |\mathbf{J}|$.This therefore implies that

$$\sqrt{|\mathbf{J}|(j-i+1)\boldsymbol{\mu}_k^2} \geq \left(\sqrt{\alpha C} - \sqrt{2}\right)\sqrt{\psi + |\mathbf{J}|\log(p)}$$

B.3.18 Proof of Lemma 31

This Lemma shows that if removing a fitted sparse segment does not result in a reduction in penalised cost of $O(\frac{1}{20}C)$, the increase in penalised cost incurred for replacing it with the sparse ground truth is $O(\frac{1}{20}C)$. We will use a very similar strategy to the one we used to prove Lemma 32. We begin by noting that

$$\mathcal{C}\left(\mathbf{x}_{i:j}, \mathbf{J}_k\right) - \mathcal{C}\left(\mathbf{x}_{i:j}, \mathbf{J}\right) = C\left(|\mathbf{J}_k| - |\mathbf{J}|\right)\log(p) - \sum_{c\in\mathbf{J}_k\setminus\mathbf{J}}(j-i+1)\left(\mu + \bar{\boldsymbol{\eta}}_{i:j}^{(c)}\right)^2 + \sum_{c\in\mathbf{J}\setminus\mathbf{J}_k}(j-i+1)\left(\bar{\boldsymbol{\eta}}_{i:j}^{(c)}\right)^2 \tag{B.3.2}$$

If $|\mathbf{J}| > \frac{19}{20}|\mathbf{J}_k|$, E_1 bounds (B.3.2) by

$$C\left(|\mathbf{J}_k| - |\mathbf{J}|\right)\log(p) + 2\psi + 2|\mathbf{J}|\log(p) \leq \frac{1}{10}C|\mathbf{J}_k|\log(p) + 2\psi + \left(2 - \frac{1}{20}C\right)|\mathbf{J}|\log(p)$$

$$\leq \frac{1}{10}C|\mathbf{J}_k|\log(p) + 2\psi,$$

with the last inequality holding if C exceeds some global constant. If $|\mathbf{J}| \leq \frac{19}{20}|\mathbf{J}_k|$ we write $\mathbf{A} = \mathbf{J}_k \setminus \mathbf{J}$ and bound (B.3.2) by

$$C\left(|\mathbf{J}_k| - |\mathbf{J}|\right)\log(p) + 2\psi + 2|\mathbf{J}|\log(p) - |\mathbf{A}|\boldsymbol{\mu}_k^2(j-i+1) + 2\sqrt{\boldsymbol{\mu}_k^2(j-i+1)}\sqrt{|\mathbf{A}|\psi + |\mathbf{A}|^2\log(p)}$$

using E_1 and E_4. Lemma 30 implies that

$$\sqrt{(j-i+1)\boldsymbol{\mu}_k^2} \geq \frac{1}{\sqrt{|\mathbf{J}|}}\left(\sqrt{\frac{19}{20}C} - 2\right)\sqrt{\psi + |\mathbf{J}|\log(p)}.$$

Consequently, copying parts of the proof of Lemma 32 , we have that

$$|\mathbf{A}|\boldsymbol{\mu}_k^2(j-i+1) - 2\sqrt{\boldsymbol{\mu}_k^2(j-i+1)}\sqrt{|\mathbf{A}|\psi + |\mathbf{A}|^2\log(p)} > \frac{37}{40}C|\mathbf{A}|\log(p),$$

which shows that (B.3.2) is bounded by

$$\begin{aligned} C\left(|\mathbf{J}_k| - |\mathbf{J}|\right)\log(p) + 2\psi + 2|\mathbf{J}|\log(p) - \frac{37}{40}C|\mathbf{A}|\log(p) &\leq \frac{1}{10}C|\mathbf{J}_k|\log(p) + 2\psi + (2 - \frac{1}{40}C)|\mathbf{J}|\log(p) \\ &\leq \frac{1}{10}C|\mathbf{J}_k|\log(p) + 2\psi, \end{aligned}$$

where the first inequality follows from the fact that $|\mathbf{J}_k| < |\mathbf{J}| + |\mathbf{A}|$ and the second one holds if C exceeds a global constant. This finishes the proof.

B.3.19 Proof of Lemma 32

This Lemma shows that if removing a fitted sparse segment does not result in a reduction in penalised cost of $O(\frac{1}{20}C)$, the increase in penalised cost incurred for replacing it with the dense ground truth is $O(\frac{1}{20}C)$. We have that

$$\mathcal{C}\left(\mathbf{x}_{i:j}, \mathbf{1}\right) - \mathcal{C}\left(\mathbf{x}_{i:j}, \mathbf{J}\right) = p + C\sqrt{p\psi} - C|\mathbf{J}|\log(p) - \sum_{c \notin \mathbf{J}}(j-i+1)\left(\bar{\mathbf{x}}_{i:j}^{(c)}\right)^2 \quad \text{(B.3.3)}$$

We consider 2 cases separately. If $|\mathbf{J}| > \frac{19}{20}k^*$, the event E_5 implies that the above can be bounded by

$$p + C\sqrt{p\psi} - C|\mathbf{J}|\log(p) - \left(p - 2\sqrt{p\psi} - 2\psi - 2|\mathbf{J}|\log(p)\right) \geq 2\psi + (C+2)\sqrt{p\psi} - (C-2)\frac{19}{20}\sqrt{p\psi}$$
$$\geq \frac{1}{10}C\sqrt{p\psi} + 2\psi,$$

provided C exceeds some global constant. If $|\mathbf{J}| \leq \frac{19}{20}k^*$, we introduce the set $\mathbf{A} = \mathbf{J}_k \setminus \mathbf{J}$. The quantity in (B.3.3) is then equal to

$$p + C\sqrt{p\psi} - C|\mathbf{J}|\log(p) - |\mathbf{A}|(j-i+1)\boldsymbol{\mu}_k^2 + 2\sqrt{(j-i+1)}\boldsymbol{\mu}_k \sum_{c\in\mathbf{A}} \sqrt{(j-i+1)}\bar{\boldsymbol{\eta}}_{i:j}^{(c)}$$
$$- \sum_{c\notin\mathbf{J}} (j-i+1)\left(\bar{\boldsymbol{\eta}}_{i:j}^{(c)}\right)^2$$
$$\leq (C+2)\sqrt{p\psi} - (C-2)|\mathbf{J}|\log(p) + 2\psi - |\mathbf{A}|(j-i+1)\boldsymbol{\mu}_k^2$$
$$+ 2\sqrt{(j-i+1)\boldsymbol{\mu}_k^2}\sqrt{2|\mathbf{A}|\psi + 2|\mathbf{A}|^2\log(p)},$$

where the inequality flows from E_1, E_4, and E_5. If C exceeds a fixed constant, the above is less than

$$\frac{41}{40}C\sqrt{p\psi} - \frac{37}{40}C|\mathbf{J}|\log(p) + 2\psi - |\mathbf{A}|(j-i+1)\boldsymbol{\mu}_k^2 + 2\sqrt{(j-i+1)\boldsymbol{\mu}_k^2}\sqrt{2|\mathbf{A}|\psi + 2|\mathbf{A}|^2\log(p)} \tag{B.3.4}$$

Lemma 30 now implies that

$$\sqrt{(j-i+1)\boldsymbol{\mu}_k^2} \geq \frac{1}{\sqrt{|\mathbf{J}|}}\left(\sqrt{\frac{19}{20}C} - 2\right)\sqrt{\psi + |\mathbf{J}|\log(p)}.$$

Therefore

$$|\mathbf{A}|\sqrt{(j-i+1)\boldsymbol{\mu}_k^2} - 2\sqrt{2|\mathbf{A}|\psi + 2|\mathbf{A}|^2\log(p)}$$
$$\geq \left(\sqrt{\frac{19}{20}C} - 2\right)\sqrt{\frac{|\mathbf{A}|}{|\mathbf{J}|}|\mathbf{A}|\psi + |\mathbf{A}|^2\log(p)} - 2\sqrt{2|\mathbf{A}|\psi + 2|\mathbf{A}|^2\log(p)}$$
$$\geq \left(\sqrt{\frac{19}{20}C} - 2\right)\sqrt{\frac{1}{20}|\mathbf{A}|\psi + |\mathbf{A}|^2\log(p)} - 2\sqrt{2|\mathbf{A}|\psi + 2|\mathbf{A}|^2\log(p)},$$

which exceeds 0 if C exceeds a global constant. Therefore

$$
\begin{aligned}
&|\mathbf{A}|^2(j-i+1)\boldsymbol{\mu}_k^2 - 2|\mathbf{A}|\sqrt{(j-i+1)\boldsymbol{\mu}_k^2}\sqrt{2|\mathbf{A}|\psi + 2|\mathbf{A}|^2\log(p)} \\
&\geq \frac{|\mathbf{A}|}{|\mathbf{J}|}\left(\sqrt{\frac{19}{20}C} - 2\right)^2(\psi + |\mathbf{J}|\log(p)) - 2\frac{\sqrt{\frac{19}{20}C} - 2}{|\mathbf{J}|}\sqrt{2|\mathbf{A}|\psi + 2|\mathbf{A}|^2\log(p)}\sqrt{\psi + |\mathbf{J}|\log(p)} \\
&\geq \left(\sqrt{\frac{19}{20}C} - 2\right)^2\left(\frac{|\mathbf{A}|}{|\mathbf{J}|}\psi + |\mathbf{A}|\log(p)\right) - 2\sqrt{\frac{19}{20}C}\sqrt{2\psi + 2|\mathbf{A}|\log(p)}\sqrt{\frac{|\mathbf{A}|}{|\mathbf{J}|}\psi + |\mathbf{A}|\log(p)} \\
&\geq \left(\sqrt{\frac{19}{20}C} - 2\right)^2\left(\frac{|\mathbf{A}|}{|\mathbf{J}|}\psi + |\mathbf{A}|\log(p)\right) - \sqrt{\frac{19}{20}C}\left(\left(2 + \frac{|\mathbf{A}|}{|\mathbf{J}|}\right)\psi + 3|\mathbf{A}|\log(p)\right) \\
&= \left(\left(\sqrt{\frac{19}{20}C} - 2\right)^2 - 3\sqrt{\frac{19}{20}C}\right)|\mathbf{A}|\log(p) + \left(\left(\left(\sqrt{\frac{19}{20}C} - 2\right)^2 - \sqrt{\frac{19}{20}C}\right)\frac{|\mathbf{A}|}{|\mathbf{J}|} - 2\right)\psi,
\end{aligned}
$$

where the third inequality follows from the AM-GM-inequality. If C exceeds a fixed constant this will exceed

$$\frac{37}{40}C|\mathbf{A}|\log(p),$$

Hence the quantity in (B.3.4) is bounded by

$$\frac{41}{40}C\sqrt{p\psi} - \frac{37}{40}C\left(|\mathbf{J}| + |\mathbf{A}|\right)\log(p) + 2\psi \leq \frac{41}{40}C\sqrt{p\psi} - \frac{37}{40}Ck^*\log(p) + 2\psi = \frac{1}{10}C\sqrt{p\psi} + 2\psi.$$

This finishes the proof.

B.3.20 Proof of Lemma 33

This Lemma bounds the increase in penalised cost incurred when transitioning from a fitted dense segment to the sparse ground truth. We have that

$$\begin{aligned}
&\mathcal{C}\left(\mathbf{x}_{i:j}, \mathbf{J}_k\right) - \mathcal{C}\left(\mathbf{x}_{i:j}, \mathbf{1}\right) = C|\mathbf{J}_k|\log(p) - \left(C\sqrt{p\psi} + p\right) + \sum_{c \notin \mathbf{J}_k}(j-i+1)\left(\bar{\boldsymbol{\eta}}^c_{i:j}\right)^2 \\
&\leq C|\mathbf{J}_k|\log(p) - \left(C\sqrt{p\psi} + p\right) + \left(p + 2\psi + 2\sqrt{p\psi}\right) \\
&= C|\mathbf{J}_k|\log(p) - C\sqrt{p\psi} + 2\sqrt{p\psi} + 2\psi \\
&\leq \frac{13}{20}C|\mathbf{J}_k|\log(p) - \frac{6}{10}C\sqrt{p\psi} + 2\psi \leq \frac{1}{10}C|\mathbf{J}_k|\log(p) - \frac{1}{20}C\sqrt{p\psi} + 2\psi,
\end{aligned}$$

for large enough C. Here the first inequality follows from E_2 and the second inequality holds because $|\mathbf{J}_k| \leq k^*$.

B.3.21 Proof of Lemma 34

The proof is very similar to that of Lemma 32 and has therefore been omitted.

B.3.22 Proof of Lemma 35

The proof is very similar to that of Lemma 31 and has therefore been omitted.

B.3.23 Proof of Lemma 36

This Lemma shows that splitting up long fitted changes containing multiple sparse anomalous regions along the ground truth reduces the penalised cost by $O(C)$ We

begin by considering

$$
\begin{aligned}
&\mathcal{C}\left(\mathbf{x}_{s,e}, \mathbf{1}\right) - \sum_{k \in \mathcal{D}_{s,e}} \left(\mathcal{C}\left(\mathbf{x}_{(s_k+1):e_k}, \mathbf{J}_k\right)\right) \\
&= p + C\psi + C\sqrt{p\psi} + \sum_{c=1}^{p}\left(\sum_{t=s}^{e}\left(\mathbf{x}_t^{(c)} - \bar{\mathbf{x}}_{s:e}^{(c)}\right)^2\right) - \sum_{c=1}^{p}\sum_{t:\nexists k:\ c\in\mathbf{J}_k \wedge t\in[s_k+1,e_k]}\left(\boldsymbol{\eta}_t^{(c)}\right)^2 \\
&- \sum_{k\in\mathcal{D}_{s,e}}\left(\sum_{c\in\mathbf{J}_k}\left(\sum_{t=s_k+1}^{e_k}\left(\mathbf{x}_t^{(c)} - \bar{\mathbf{x}}_{(s_k+1):e_k}^{(c)}\right)^2\right) + C\psi + C|\mathbf{J}_k|\log(p)\right) \\
&\geq p + C\psi + C\sqrt{p\psi} + \sum_{c=1}^{p}\left(\sum_{t=s}^{e}\left(\boldsymbol{\mu}_t^{(c)} - \bar{\boldsymbol{\mu}}_{s:e}^{(c)} + \boldsymbol{\eta}_t^{(c)} - \bar{\boldsymbol{\eta}}_{s:e}^{(c)}\right)^2\right) \\
&- \sum_{k\in\mathcal{D}_{s,e}}\left(\sum_{c\in\mathbf{J}_k}\left(\sum_{t=s_k+1}^{e_k}\left(\boldsymbol{\eta}_t^{(c)}\right)^2\right) + C\psi + C|\mathbf{J}_k|\log(p)\right) - \sum_{c=1}^{p}\sum_{t:\nexists k:\ c\in\mathbf{J}_k \wedge t\in[s_k+1,e_k]}\left(\boldsymbol{\eta}_t^{(c)}\right)^2 \\
&= p + C\psi + C\sqrt{p\psi} + \sum_{c=1}^{p}\left(\sum_{t=s}^{e}\left(\boldsymbol{\eta}_t^{(c)} - \bar{\boldsymbol{\eta}}_{s:e}^{(c)}\right)^2\right) + \sum_{c=1}^{p}\left(\sum_{t=s}^{e}\left(\boldsymbol{\mu}_t^{(c)} - \bar{\boldsymbol{\mu}}_{s:e}^{(c)}\right)^2\right) \\
&+ 2\sum_{c=1}^{p}\left(\sum_{t=s}^{e}\left(\boldsymbol{\mu}_t^{(c)} - \bar{\boldsymbol{\mu}}_{s:e}^{(c)}\right)\left(\boldsymbol{\eta}_t^{(c)} - \bar{\boldsymbol{\eta}}_{s:e}^{(c)}\right)\right) - \sum_{c=1}^{p}\left(\sum_{t=s}^{e}\left(\boldsymbol{\eta}_t^{(c)}\right)^2\right) - \sum_{k\in\mathcal{D}_{s,e}}\left(C\psi + C|\mathbf{J}_k|\log(p)\right) \\
&= p + C\psi + C\sqrt{p\psi} - \sum_{c=1}^{p}\left((e-s+1)\left(\bar{\boldsymbol{\eta}}_{s:e}^{(c)}\right)^2\right) + \sum_{c=1}^{p}\left(\sum_{t=s}^{e}\left(\boldsymbol{\mu}_t^{(c)} - \bar{\boldsymbol{\mu}}_{s:e}^{(c)}\right)^2\right) \\
&- \sum_{k\in\mathcal{D}_{s,e}}\left(C\psi + C|\mathbf{J}_k|\log(p)\right) + 2\sum_{c=1}^{p}\left(\sum_{t=s}^{e}\left(\boldsymbol{\mu}_t^{(c)} - \bar{\boldsymbol{\mu}}_{s:e}^{(c)}\right)\left(\boldsymbol{\eta}_t^{(c)}\right)\right) \\
&\geq \frac{19}{20}C\left(\psi + \sqrt{p\psi}\right) + \sum_{c=1}^{p}\left(\sum_{t=s}^{e}\left(\boldsymbol{\mu}_t^{(c)} - \bar{\boldsymbol{\mu}}_{s:e}^{(c)}\right)^2\right) - \sum_{k\in\mathcal{D}_{s,e}}\left(C\psi + C|\mathbf{J}_k|\log(p)\right) \\
&- 2\sqrt{\sum_{c=1}^{p}\sum_{t=s}^{e}\left(\boldsymbol{\mu}_t^{(c)} - \bar{\boldsymbol{\mu}}_{s:e}^{(c)}\right)^2}\sqrt{2\psi + 2\left|W_{s,e}\right|\log(p)} \\
&\geq \frac{19}{20}C\left(\psi + \sqrt{p\psi}\right) + \frac{1}{2}\sum_{c=1}^{p}\left(\sum_{t=s}^{e}\left(\boldsymbol{\mu}_t^{(c)} - \bar{\boldsymbol{\mu}}_{s:e}^{(c)}\right)^2\right) - \sum_{k\in\mathcal{D}_{s,e}}\left(C\psi + C|\mathbf{J}_k|\log(p)\right) - 8\psi - 8\left|W_{s,e}\right|\log(p),
\end{aligned}
$$

where the first inequality follows from the fact that the residual sum of squares is minimised at the mean, the second inequality follows from E_2 and E_6, and the last inequality follows from the AM-GM inequality.

Next note that

$$\sum_{t=s}^{e}\left(\boldsymbol{\mu}_t^{(c)} - \bar{\boldsymbol{\mu}}_{s:e}^{(c)}\right)^2$$

corresponds to the residual sum of squares obtained by fitting $\boldsymbol{\mu}_e^{(c)}, ..., \boldsymbol{\mu}_s^{(c)}$ as a single segment. Consequently, breaking it up into smaller segments does not increase un-

penalised cost. More precisely, for any partition $\tau_{s:e} = \{s, \tau_1, ..., \tau_m, e\}$ of the segment $(s+1, e)$,

$$\sum_{t=s+1}^{e} \left(\boldsymbol{\mu}_t^{(c)} - \bar{\boldsymbol{\mu}}_{(s+1):e}^{(c)}\right)^2 \geq \sum_{k=0}^{m} \left(\sum_{t=\tau_m+1}^{\tau_{m+1}} \left(\boldsymbol{\mu}_t^{(c)} - \bar{\boldsymbol{\mu}}_{(\tau_m+1):\tau_{m+1}}^{(c)}\right)^2 \right)$$

holds. In particular, we therefore have that

$$\begin{aligned}
&\frac{1}{2}\sum_{c=1}^{p}\left(\sum_{t=s}^{e}\left(\boldsymbol{\mu}_t^{(c)} - \bar{\boldsymbol{\mu}}_{s:e}^{(c)}\right)^2\right) \\
&\geq \sum_{k:e_k\in[s,e]} \frac{1}{2}\left(\sum_{c\in\mathbf{J}_k}\left(\sum_{e_k-\lceil\frac{10C}{\triangle_k^2}\rceil}^{e_k+\lfloor\frac{10C}{\triangle_k^2}\rfloor}\left(\boldsymbol{\mu}_t^{(c)} - \bar{\boldsymbol{\mu}}^{(c)}_{\left(e_k-\lceil\frac{10C}{\triangle_k^2}\rceil\right):\left(e_k+\lfloor\frac{10C}{\triangle_k^2}\rfloor\right)}\right)^2\right)\right) \\
&+ \sum_{k:s_k\in[s,e]} \frac{1}{2}\left(\sum_{c\in\mathbf{J}_k}\left(\sum_{s_k-\lfloor\frac{10C}{\triangle_k^2}\rfloor}^{s_k+\lceil\frac{10C}{\triangle_k^2}\rceil}\left(\boldsymbol{\mu}_t^{(c)} - \bar{\boldsymbol{\mu}}^{(c)}_{\left(s_k-\lfloor\frac{10C}{\triangle_k^2}\rfloor\right):\left(s_k+\lceil\frac{10C}{\triangle_k^2}\rceil\right)}\right)^2\right)\right) \\
&= \frac{1}{2}\sum_{k:e_k\in[s,e]}\left(|\mathbf{J}_k|2\left\lceil\frac{10C}{\triangle_k^2}\right\rceil\frac{\boldsymbol{\mu}_k^2}{4}\right) + \frac{1}{2}\sum_{k:s_k\in[s,e]}\left(|\mathbf{J}_k|\frac{20C}{\triangle_k^2}\frac{\boldsymbol{\mu}_k^2}{4}\right) \geq \frac{1}{2}\sum_{k\in\mathcal{D}_{s,e}}\left(|\mathbf{J}_k|\frac{20C}{\triangle_k^2}\frac{\boldsymbol{\mu}_k^2}{4}\right) \\
&= \sum_{k\in\mathcal{D}_{s,e}}\frac{5}{2}C\left(\psi + |\mathbf{J}_k|\log(p)\right),
\end{aligned}$$

where the first inequality follows from using a partition which cuts $\frac{10C}{\triangle_k^2}$ either side of the starting points and end points of true anomalous regions contained in $[s, e]$ and the second inequality follows from the definition of $\mathcal{D}_{s,e}$. Consequently, we have that

$$\begin{aligned}
&\mathcal{C}\left(\mathbf{x}_{s,e}, \mathbf{1}\right) - \sum_{k\in\mathcal{D}_{s,e}}\left(\mathcal{C}\left(\mathbf{x}_{(s_k+1):e_k}, \mathbf{J}_k\right)\right) \\
&\geq \frac{19}{20}C\left(\psi + \sqrt{p\psi}\right) + \sum_{k\in\mathcal{D}_{s,e}}\frac{5}{2}C\left(\psi + |\mathbf{J}_k|\log(p)\right) - \sum_{k\in\mathcal{D}_{s,e}}\left(C\psi + C|\mathbf{J}_k|\log(p)\right) - 8\psi - 8\left|W_{s,e}\right|\log(p) \\
&\geq \frac{19}{20}C\left(\psi + \sqrt{p\psi}\right),
\end{aligned}$$

where the first inequality follows from assembling the previous two results, and the

second one holds if C exceeds a global constant. We also have that:

$$
\begin{aligned}
&\mathcal{C}\left(\mathbf{x}_{s,e}, \mathbf{J}\right) - \sum_{k \in \mathcal{D}_{s,e}} \left(\mathcal{C}\left(\mathbf{x}_{(s_k+1):e_k}, \mathbf{J}_k\right)\right) = C\psi + C|\boldsymbol{J}|\log(p) + \sum_{c \in \mathbf{J}} \left(\sum_{t=s}^{e} \left(\mathbf{x}_t^{(c)} - \bar{\mathbf{x}}_{s:e}^{(c)} \right)^2 \right) \\
&- \sum_{k \in \mathcal{D}_{s,e}} \left(\sum_{c \in \mathbf{J}_k \cap \mathbf{J}} \left(\sum_{t=s_k+1}^{e_k} \left(\mathbf{x}_t^{(c)} - \bar{\mathbf{x}}_{(s_k+1):e_k}^{(c)} \right)^2 \right) + C\psi + C|\mathbf{J}_k|\log(p) \right) \\
&- \sum_{c \in \mathbf{J}} \sum_{t : \nexists k:\, c \in \mathbf{J}_k \wedge t \in [s_k+1, e_k]} \left(\boldsymbol{\eta}_t^{(c)} \right)^2 + \sum_{k \in \mathcal{D}_{s,e}} \left(\sum_{c \in \mathbf{J}_k \setminus \mathbf{J}} \left((e_k - s_k) \left(\bar{\mathbf{x}}_{(s_k+1):e_k}^{(c)} \right)^2 \right) \right) \\
&\geq C\psi + C|\boldsymbol{J}|\log(p) + \sum_{c \in \mathbf{J}} \left(\sum_{t=s}^{e} \left(\boldsymbol{\mu}_t^{(c)} - \bar{\boldsymbol{\mu}}_{s:e}^{(c)} \right)^2 \right) + \sum_{c \in \mathbf{J}} \left(\sum_{t=s}^{e} \left(\boldsymbol{\eta}_t^{(c)} - \bar{\boldsymbol{\eta}}_{s:e}^{(c)} \right)^2 \right) \\
&+ 2 \sum_{c \in \mathbf{J}} \left(\sum_{t=s}^{e} \left(\boldsymbol{\eta}_t^{(c)} - \bar{\boldsymbol{\eta}}_{s:e}^{(c)} \right) \left(\boldsymbol{\mu}_t^{(c)} - \bar{\boldsymbol{\mu}}_{s:e}^{(c)} \right) \right) - \sum_{c \in \mathbf{J}} \left(\sum_{t=s}^{e} \left(\boldsymbol{\eta}_t^{(c)} \right)^2 \right) - \sum_{k \in \mathcal{D}_{s,e}} \left(C\psi + C|\mathbf{J}_k|\log(p) \right) \\
&+ \sum_{k \in \mathcal{D}_{s,e}} \left(\sum_{c \in \mathbf{J}_k \setminus \mathbf{J}} (e_k - s_k)\boldsymbol{\mu}_k^2 + 2(e_k - s_k)\boldsymbol{\mu}_k \sum_{c \in \mathbf{J}_k \setminus \mathbf{J}} \left(\bar{\boldsymbol{\eta}}_{(s_k+1):e_k}^{(c)} \right) \right) \\
&\geq (C-2)(\psi + |\boldsymbol{J}|\log(p)) - \sum_{k \in \mathcal{D}_{s,e}} \left(C\psi + C|\mathbf{J}_k|\log(p) \right) + \sum_{c \in \mathbf{J}} \left(\sum_{t=s}^{e} \left(\boldsymbol{\mu}_t^{(c)} - \bar{\boldsymbol{\mu}}_{s:e}^{(c)} \right)^2 \right) \\
&+ \sum_{k \in \mathcal{D}_{s,e}} \left((e_k - s_k)|\mathbf{J}_k \setminus \mathbf{J}|\boldsymbol{\mu}_k^2 - 2\sqrt{(e_k - s_k)\boldsymbol{\mu}_k^2 |\mathbf{J}_k \setminus \mathbf{J}| (2\psi + 2|\mathbf{J}_k \setminus \mathbf{J}|\log(p))} \right) \\
&+ 2 \sum_{c \in \mathbf{J}} \left(\sum_{t=s}^{e} \left(\boldsymbol{\mu}_t^{(c)} - \bar{\boldsymbol{\mu}}_{s:e}^{(c)} \right) \boldsymbol{\eta}_t^{(c)} \right).
\end{aligned}
$$

This can further be bounded below by

$$
\begin{aligned}
&(C-2)(\psi + |\boldsymbol{J}|\log(p)) - \sum_{k \in \mathcal{D}_{s,e}} \left(C\psi + C|\mathbf{J}_k|\log(p) \right) + \sum_{c \in \mathbf{J}} \left(\sum_{t=s}^{e} \left(\boldsymbol{\mu}_t^{(c)} - \bar{\boldsymbol{\mu}}_{s:e}^{(c)} \right)^2 \right) \\
&- 2\sqrt{\sum_{c \in \mathbf{J}} \left(\sum_{t=s}^{e} \left(\boldsymbol{\mu}_t^{(c)} - \bar{\boldsymbol{\mu}}_{s:e}^{(c)} \right)^2 \right)} \sqrt{2\psi + 2|W_{s:e}|\log(p)} \\
&+ \sum_{k \in \mathcal{D}_{s,e}} \left(\frac{1}{2}(e_k - s_k)|\mathbf{J}_k \setminus \mathbf{J}|\boldsymbol{\mu}_k^2 - 8(\psi + |\mathbf{J}_k \setminus \mathbf{J}|\log(p)) \right) \\
&\geq (C-2)(\psi + |\boldsymbol{J}|\log(p)) - \sum_{k \in \mathcal{D}_{s,e}} (C+8) \left(\psi + |\mathbf{J}_k|\log(p) \right) + \sum_{c \in \mathbf{J}} \left(\sum_{t=s}^{e} \left(\boldsymbol{\mu}_t^{(c)} - \bar{\boldsymbol{\mu}}_{s:e}^{(c)} \right)^2 \right) \\
&+ \frac{1}{2} \sum_{k \in \mathcal{D}_{s,e}} \left((e_k - s_k)|\mathbf{J}_k \setminus \mathbf{J}|\boldsymbol{\mu}_k^2 \right) - 8(\psi + |W_{s:e}|\log(p)) - \frac{1}{2} \sum_{c \in \mathbf{J}} \left(\sum_{t=s}^{e} \left(\boldsymbol{\mu}_t^{(c)} - \bar{\boldsymbol{\mu}}_{s:e}^{(c)} \right)^2 \right) \\
&\geq \frac{19}{20} C(\psi + |\mathbf{J}|\log(p)) - \sum_{k \in \mathcal{D}_{s,e}} 2C \left(\psi + |\mathbf{J}_k|\log(p) \right) \\
&+ \frac{1}{2} \left(\sum_{c \in \mathbf{J}} \left(\sum_{t=s}^{e} \left(\boldsymbol{\mu}_t^{(c)} - \bar{\boldsymbol{\mu}}_{s:e}^{(c)} \right)^2 \right) + \sum_{k \in \mathcal{D}_{s,e}} \left((e_k - s_k)|\mathbf{J}_k \setminus \mathbf{J}|\boldsymbol{\mu}_k^2 \right) \right).
\end{aligned}
$$

A very similar argument as the one used for the dense case can be used to show that

$$\frac{1}{2}\sum_{c\in\mathbf{J}}\left(\sum_{t=s}^{e}\left(\boldsymbol{\mu}_t^{(c)}-\bar{\boldsymbol{\mu}}_{s:e}^{(c)}\right)^2\right)\geq\sum_{k\in\mathcal{D}_{s,e}}\frac{5}{2}C|\mathbf{J}_k\cap\mathbf{J}|\frac{\boldsymbol{\mu}_k^2}{\triangle_k^2}.$$

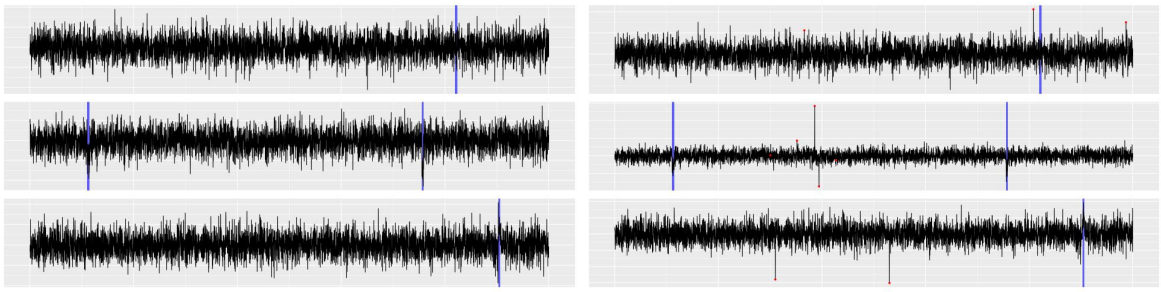
Consequently,

$$\begin{aligned}
&\mathcal{C}\left(\mathbf{x}_{s,e},\mathbf{J}\right)-\sum_{k\in\mathcal{D}_{s,e}}\left(\mathcal{C}\left(\mathbf{x}_{(s_k+1):e_k},\mathbf{J}_k\right)\right)\\
&\geq\frac{19}{20}C(\psi+|\mathbf{J}|\log(p))-\sum_{k\in\mathcal{D}_{s,e}}2C\left(\psi+|\mathbf{J}_k|\log(p)\right)+\sum_{k\in\mathcal{D}_{s,e}}\left[\frac{5}{2}C|\mathbf{J}_k\cap\mathbf{J}|\frac{\boldsymbol{\mu}_k^2}{\triangle_k^2}+\frac{5C\boldsymbol{\mu}_k^2}{2\triangle_k^2}|\mathbf{J}_k\setminus\mathbf{J}|\right]\\
&=\frac{19}{20}C(\psi+|\mathbf{J}|\log(p))-\sum_{k\in\mathcal{D}_{s,e}}2C\left(\psi+|\mathbf{J}_k|\log(p)\right)+\sum_{k\in\mathcal{D}_{s,e}}\frac{5}{2}C\left(\psi+|\mathbf{J}_k|\log(p)\right)\\
&\geq\frac{19}{20}C(\psi+|\mathbf{J}|\log(p)),
\end{aligned}$$

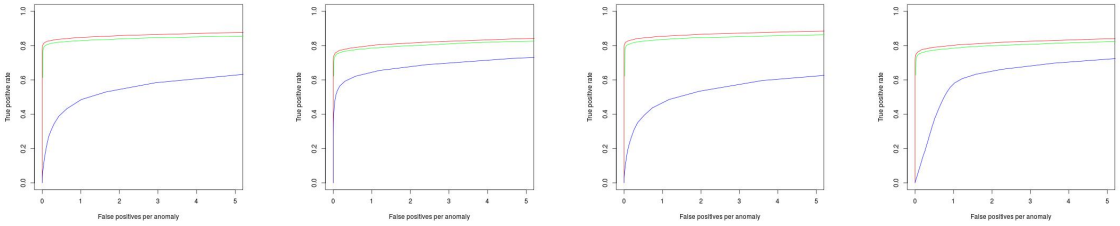
where the first inequality follows from the condition on the segment length e_k-s_k.

B.4 Further Simulations And Tables

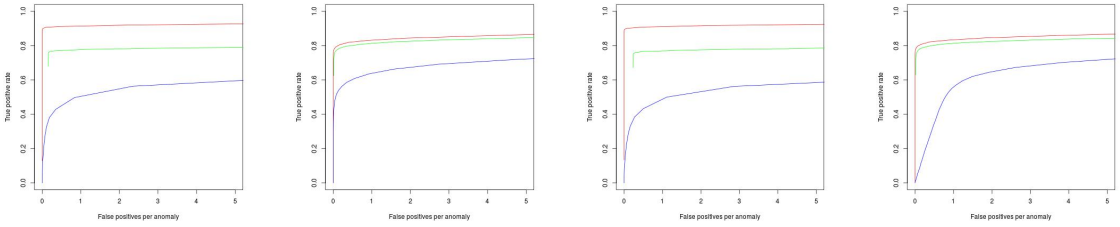
In this section, we present additional results from the simulation study and application section. Figures B.4.1 to B.4.4 display the full comparison between MVCAPA, PASS, and Inspect over the four settings and data generating processes described in Section 4.7. We repeated setting 1 and 3 from the main paper with joint changes in mean and variance. The number, location, rate of occurrence, and strength of the change in mean is as in the mean paper. The only difference is that within each anomaly the variance changes away from the typical variance, to a new, $\Gamma^{-1}(5, 5)$-distributed variance. The results for settings 1 and 3 are displayed in Figures B.4.5 and Figures B.4.6 respectively. Table B.4.7 gives the results of PASS and MVCAPA at detecting known CNVs from data from chromosome 6.



(a) Example (b) Example with pt. anomalies

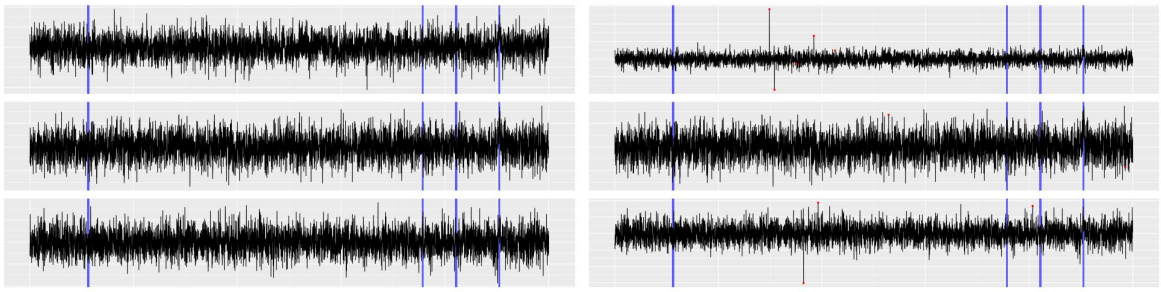


(c) p=10 (d) p=10, AR (e) p=10, PAs (f) p=10, AR, PAs

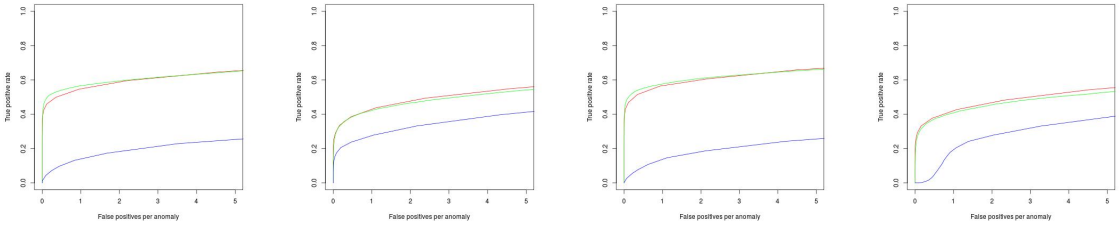


(g) p=100 (h) p=10, T (i) p=100, PAs (j) p=10, T, PAs

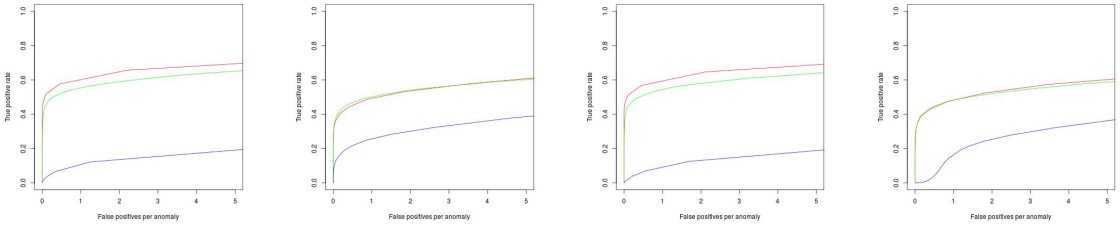
Figure B.4.1: Example series and ROC curves for setting 1. MVCAPA is in red, PASS in green, and Inspect in blue. The x-axis denotes the number of false discoveries normalised by the total number of real anomalies present in the data.



(a) Example (b) Example with pt. anomalies

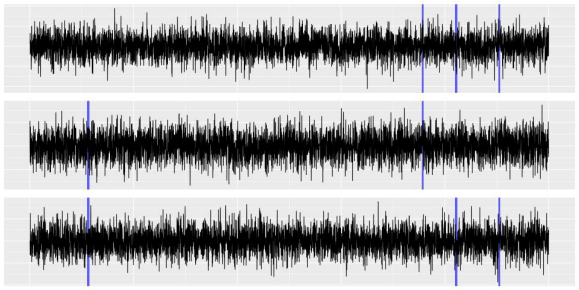


(c) p=10 (d) p=10, AR (e) p=10, PAs (f) p=10, AR, PAs

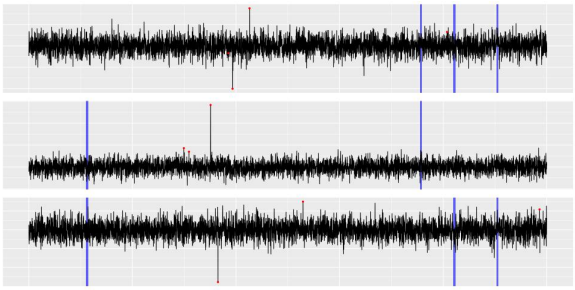


(g) p=100 (h) p=10, T (i) p=100, PAs (j) p=10, T, PAs

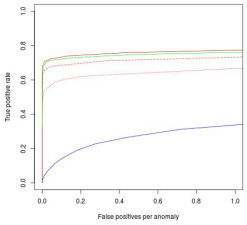
Figure B.4.2: Example series and ROC curves for setting 2. MVCAPA is in red, PASS in green, and Inspect in blue. The x-axis denotes the number of false discoveries normalised by the total number of real anomalies present in the data.



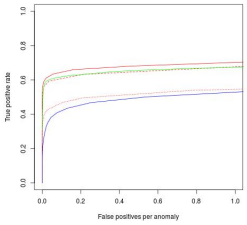
(a) Example



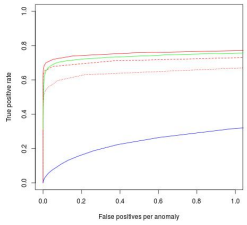
(b) Example with pt. anomalies



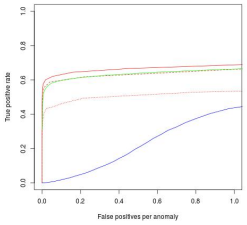
(c) p=10



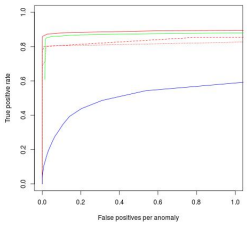
(d) p=10, AR



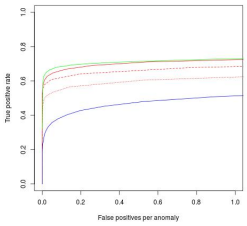
(e) p=10, PAs



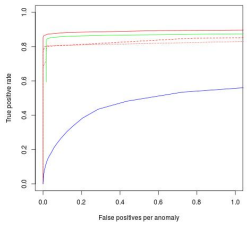
(f) p=10, AR, PAs



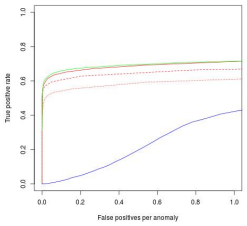
(g) p=100



(h) p=10, T

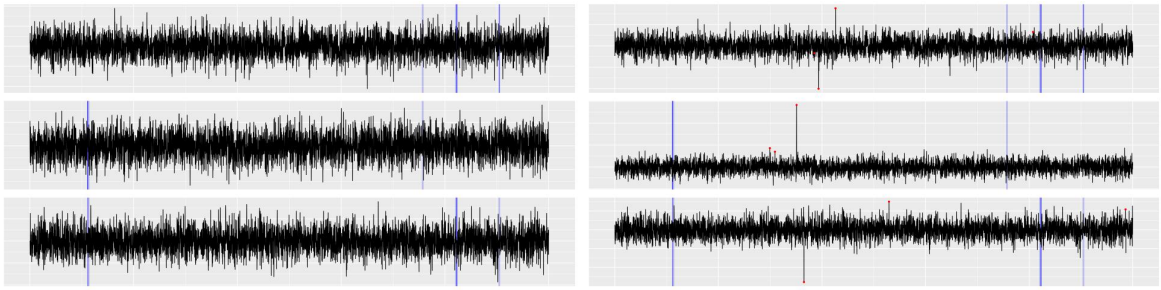


(i) p=100, PAs

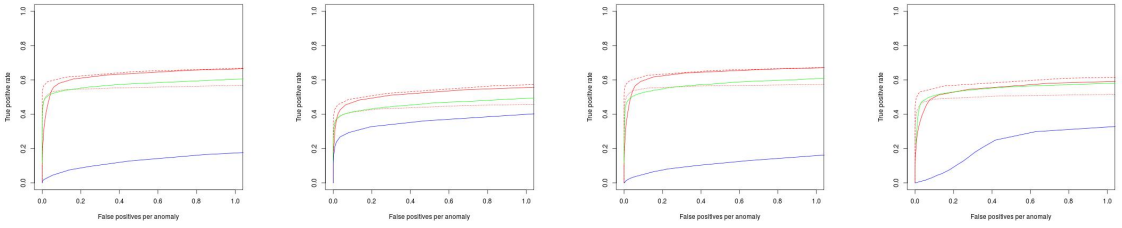


(j) p=10, T, PAs

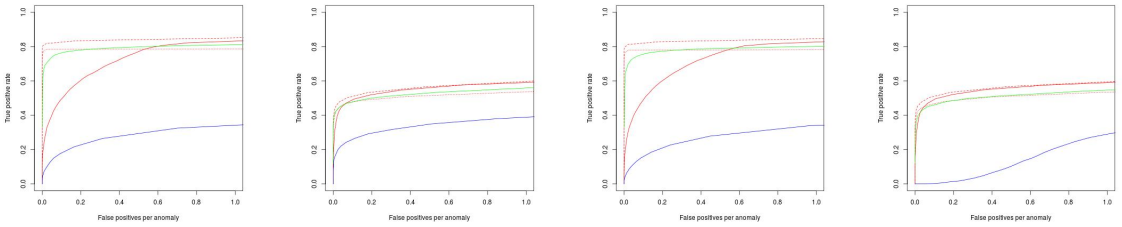
Figure B.4.3: Example series and ROC curves for setting 3. MVCAPA is in red, PASS in green, and Inspect in blue. The solid red line corresponds to $w = 0$, the dashed one to $w = 10$ and the dotted one to $w = 20$. The x-axis denotes the number of false discoveries normalised by the total number of real anomalies present in the data.



(a) Example (b) Example with pt. anomalies

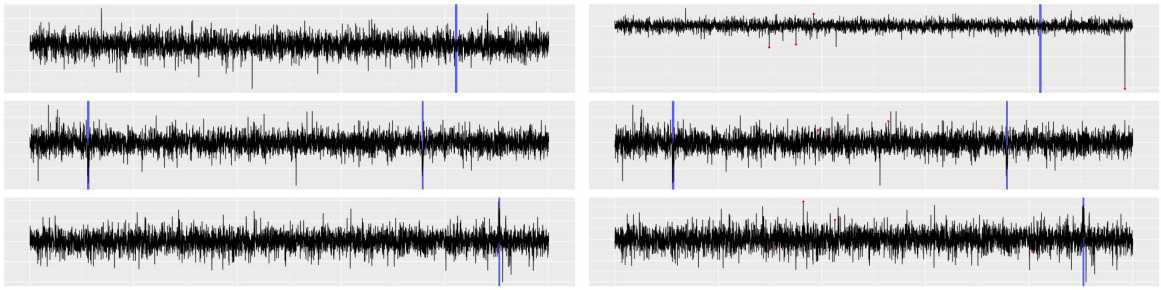


(c) p=10 (d) p=10, AR (e) p=10, PAs (f) p=10, AR, PAs



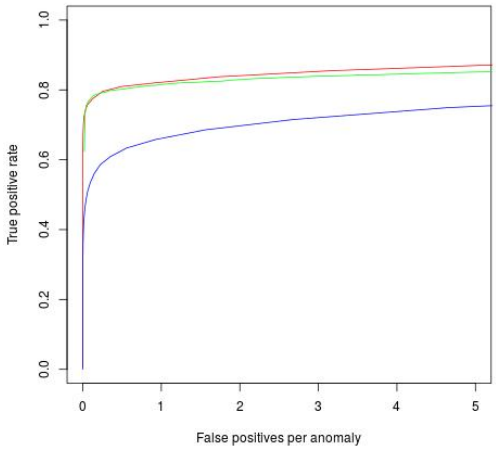
(g) p=100 (h) p=10, T (i) p=100, PAs (j) p=10, T, PAs

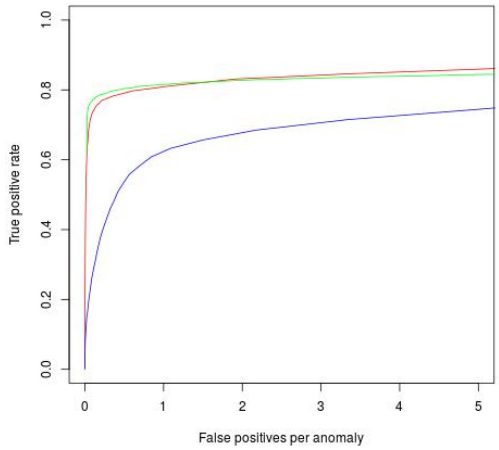
Figure B.4.4: Example series and ROC curves for setting 4. MVCAPA is in red, PASS in green, and Inspect in blue. The solid red line corresponds to $w = 0$, the dashed one to $w = 10$ and the dotted one to $w = 20$. The x-axis denotes the number of false discoveries normalised by the total number of real anomalies present in the data.



(a) Example

(b) Example, with pt. anomalies

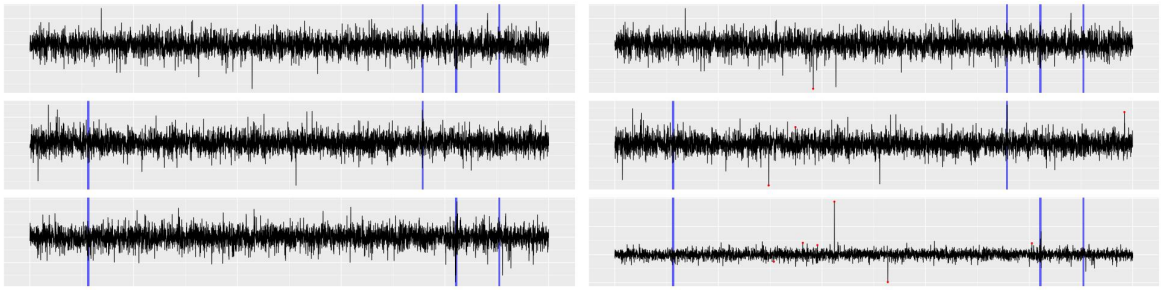




(c) p=10

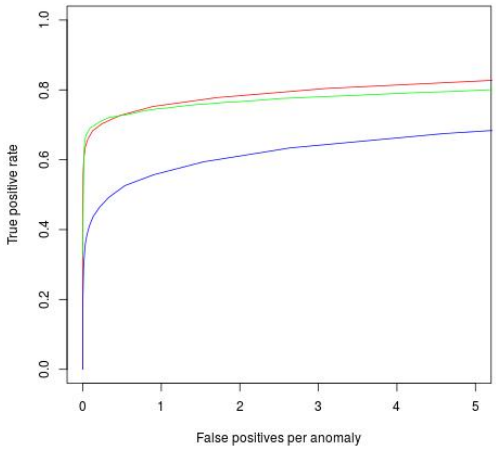
(d) p=10, with pt. anomalies

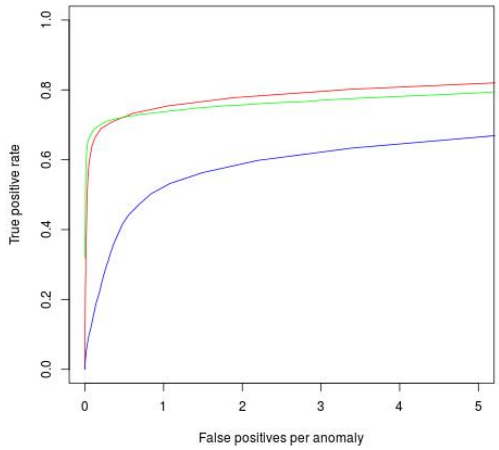
Figure B.4.5: Example series and ROC curves for setting 1. MVCAPA is in red, PASS in green, and Inspect in blue. The x-axis denotes the number of false discoveries normalised by the total number of real anomalies present in the data.



(a) Example

(b) Example, with pt. anomalies





(c) p=10

(d) p=10, with pt. anomalies

Figure B.4.6: Example series and ROC curves for setting 3. MVCAPA is in red, PASS in green, and Inspect in blue. The x-axis denotes the number of false discoveries normalised by the total number of real anomalies present in the data.

Truth	PASS			MVCAPA ($w = 40$)			MVCAPA ($w = 0$)		
Start	Rep 1	Rep 2	Rep 3	Rep 1	Rep 2	Rep 3	Rep 1	Rep 2	Rep 3
202314	✓	✓	✓	✓	✓	✓	✓	✓	✓
243582	✓	✓	✓	✓	✓	✓	✓	✓	✓
29945146	✓			✓	✓		✓	✓	
30569918					✓				
31388628				✓	✓		✓	✓	
31388628				✓	✓			✓	
32562531					✓	✓		✓	✓
32605305		✓		✓	✓	✓	✓	✓	✓
32717397	✓			✓	✓	✓	✓	✓	✓
74648424				✓	✓		✓	✓	
77073620									
77155147				✓	✓		✓	✓	
77496587	✓								
78936685		✓	✓	✓	✓	✓	✓	✓	✓
103844990	✓	✓	✓	✓	✓	✓	✓	✓	✓
126226035			✓	✓			✓		✓
139645437									
165647651	✓		✓						✓

...

Figure B.4.7: Analysis of Chromosome 6 as detailed in the caption of Figure 4.8.1. Note that the chromosome contains two different CNVs (of different lengths) beginning at 31388628.

B.5 Pseudocode

Algorithm 5 Update

Input: A vector of past lagged savings $\mathbf{S}_T^{(j)}$.

A new saving S.

A maximum lag $w \geq 0$.

1: **for** $k \in \{w, ..., 1\}$ **do**

2: $\quad \mathbf{S}_T^{(j)}(k) \leftarrow \mathbf{S}_T^{(j)}(k-1)$

3: **end for**

4: $\mathbf{S}_T^{(j)}(0) \leftarrow S$

5: $\hat{\mathbf{E}}_T^{(j)} \leftarrow \arg\max_{0 \leq k \leq w} \left(\mathbf{S}_T^{(j)}\right)(k)$

6: $\hat{\mathbf{C}}_T^{(j)} \leftarrow \max_{0 \leq k \leq w} \left(\mathbf{S}_T^{(j)}\right)(k)$

Output An updated by-end-lag savings vector $\mathbf{S}_T^{(j)}$, and optimal end-lag $\hat{\mathbf{E}}_T^{(j)}$ and the corresponding saving $\hat{\mathbf{C}}_T^{(j)}$.

Algorithm 6 ComputeSaving

Input: A vector of savings $\mathbf{C}_T^{(1:p)}$.

Penalty constants $\beta_{1:p}$ for the components of a collective anomalies.

1: $\sigma_1, ..., \sigma_p \leftarrow order(\mathbf{C}_T^{(1)}, ..., \mathbf{C}_T^{(p)})$ ▷ In decreasing order

2: $\mathbf{C}_T \leftarrow \max_{1 \leq k \leq p} \left(\sum_{i=1}^{k} \mathbf{C}_T^{(\sigma_i)} - \beta_i\right)$

3: $\hat{k} \leftarrow \arg\max_{1 \leq k \leq p} \left(\sum_{i=1}^{k} \mathbf{C}_T^{(\sigma_i)} - \beta_i\right)$

4: $\mathbf{CP}(T) \leftarrow \{\sigma_1, ..., \sigma_{\hat{k}}\}$

Output The optimal set of components $\mathbf{CP}(T)$, as well as the corresponding penalised saving $\mathbf{C}_T$.

Algorithm 7 ComputePtSaving

Input: A vector of observations $\mathbf{x}_t^{(1:p)}$.

Penalty constants β' for a point anomaly.

1: $\mathbf{C}'_t \leftarrow \sum_{i=1}^{p} \left(\left(\mathbf{x}_t^{(i)} \right)^2 - \beta' \right)^+$

2: $\mathbf{CP}'_t \leftarrow \left\{ i | i \in \{1, ..., p\} : \left(\mathbf{x}_t^{(i)} \right)^2 > \beta' \right\}$

Output The optimal set of components $\mathbf{CP}'_t$, as well as the corresponding penalised saving $\mathbf{C}'_t$.

Algorithm 8 MVCAPA Algorithm (No Pruning)

Input: A set of multivariate observations of the form, $(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)$ where $\mathbf{x}_i \in \mathbb{R}^p$.

Penalty constants $\beta_{1:p}$ and β' for the components of a collective anomaly and for point anomalies.

A minimum segment length $l \geq 2$, a maximum segment length $m \geq l$, a maximum lag $w \geq 0$.

Initialise: Set $C(0) = 0$, $Anom(0) = NULL$, $Comp(0) = NULL$, $Lags(0) = NULL$

for $j \in \{1, ..., p\}$ **do**

  $\hat{\mu}^{(j)} \leftarrow MEDIAN(\mathbf{x}_1^{(j)}, \mathbf{x}_2^{(j)}, \ldots, \mathbf{x}_n^{(j)})$ ▷ Obtain robust estimates of the mean and variance

  $\hat{\sigma}^{(j)} \leftarrow IQR(\mathbf{x}_1^{(j)}, \mathbf{x}_2^{(j)}, \ldots, \mathbf{x}_n^{(j)})$

  for $i \in \{1, ..., n\}$ **do**

    $\mathbf{x}_i^{(j)} \leftarrow \frac{\mathbf{x}_i^{(j)} - \hat{\mu}^{(j)}}{\hat{\sigma}^{(j)}}$ ▷ Centralise the data

    for $k \in \{0, ..., w\}$ **do**

      $\mathbf{S}_i^{(j)}(k) \leftarrow 0$ ▷ Initialise saving per end-lag

    end for

  end for

end for

for $t \in \{1, ..., n\}$ **do**

  for $T \in \{1, ..., t\} \cap \{t - m, ..., t - l + 1\}$ **do**

    for $j \in \{1, ..., p\}$ **do**

      $S \leftarrow (t + 1 - T) \left(\frac{1}{t+1-T} \sum_{i=T}^{t} \mathbf{x}_i^{(j)} \right)^2$ ▷ Calculate saving without any lag

      $\mathbf{S}_T^{(j)}, \tilde{\mathbf{E}}_T^{(j)}, \tilde{\mathbf{C}}_T^{(j)} \leftarrow Update(\mathbf{S}_T^{(j)}, S, w)$ ▷ Update saving per end-lag, and associated saving

    end for

  end for

continues on next page

for $T \in \{1, ..., t\} \cap \{t - m, ..., t - l + 1\}$ **do**

  for $j \in \{1, ..., p\}$ **do**

    $\mathbf{C}_T^{(j)} \leftarrow \max_{0 \leq t' \leq w} \left(\tilde{\mathbf{C}}_{T+t'}^{(j)} \right)$ ▷ Find the lowest starting cost

    $\mathbf{L}_T^{(j)} \leftarrow \arg\max_{0 \leq t' \leq w} \left(\tilde{\mathbf{C}}_{T+t'}^{(j)} \right)$ ▷ Find the best start lag

    $\mathbf{E}_T^{(j)} \leftarrow \tilde{\mathbf{E}}_{T+\mathbf{L}_T^{(j)}}^{(j)}$ ▷ And deduce the best end lag

  end for

end for

for $T \in \{1, ..., t\} \cap \{t - m, ..., t - l + 1\}$ **do**

  $\mathbf{C}_T, \mathbf{Cp}(T) \leftarrow ComputeSaving(\mathbf{C}_T^{(1:p)}, \beta_{1:p})$

end for

$\mathbf{C}'_t, \mathbf{Cp}' \leftarrow ComputePtSaving(\mathbf{x}_t^{(1:p)}, \beta')$ ▷ Cost and components of point anomaly

$C_1(t) \leftarrow \max_{t-m+1 \leq T \leq t-l+1} \left[C(k) + \mathbf{C}_T \right]$ ▷ Collective Anom.

$s \leftarrow C(t-1) + \arg\max_{t-m+1 \leq T \leq t-l+1} \left[C(k) + \mathbf{C}_T \right]$

$C_2(t) \leftarrow C(t-1)$ ▷ No Anomaly

$C_3(t) \leftarrow C(t-1) + C'_t$ ▷ Point Anomaly

$C(m) \leftarrow \max \left[C_1(m), C_2(m), C_3(m) \right]$

continues on next page

34: **switch** $\arg\max\left[C_1(m), C_2(m), C_3(m)\right]$ **do** ▷ Select type of anomaly giving the lowest cost

35: **case** 1 :

36: $Anom(m) \leftarrow [Anom(s), (s+1, m)]$

37: $Comp(m) \leftarrow [Comp(s), \mathbf{Cp}(s)]$

38: $Lags(m) \leftarrow [Lags(s), (\mathbf{L}_s^{(1:p)}, \mathbf{E}_s^{(1:p)})]$

39: **case** 2 :

40: $Anom(m) \leftarrow Anom(m-1)$

41: **case** 3 :

42: $Anom(m) \leftarrow [Anom(m-1), (m)]$

43: $Comp(m) \leftarrow [Comp(m-1), \mathbf{Cp}']$

44: **end for**

Output The points and segments recorded in $Anom(n)$, the sets of components in $Comp(n)$ and the sets of start and end lags in $Lags(n)$.

Appendix C

CE-BASS

C.1 Theorems and Derivations

C.1.1 Theorem 5

Theorem 5. *Let the prior for the hidden state* $\boldsymbol{X}_t$ *be* $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ *and an observation* $\boldsymbol{Y}_{t+1} := \boldsymbol{Y}$ *be available. Then the samples for* $\tilde{\boldsymbol{V}}_{t+1}^{(i,i)}$ *from*

$$\tilde{\sigma}_i \Gamma\left(a_i + \frac{1}{2}, a_i + \frac{\tilde{\sigma}_i}{2\boldsymbol{\Sigma}_A^{(i,i)}}\left(\frac{\left(\hat{\boldsymbol{\Sigma}}^{-1}\right)^{(i,:)}(\boldsymbol{Y} - \boldsymbol{CA\mu})}{\left(\hat{\boldsymbol{\Sigma}}^{-1}\right)^{(i,i)}}\right)^2\right)$$

have associated weight

$$\frac{1}{M} r_i \frac{\Gamma(a_i + \frac{1}{2})}{\Gamma(a_i)} \sqrt{\tilde{\sigma}_i} \frac{a_i^{a_i}}{\left(a_i + \frac{\tilde{\sigma}_i}{2\boldsymbol{\Sigma}_A^{(i,i)}}\left(\frac{\left(\hat{\boldsymbol{\Sigma}}^{-1}\right)^{(i,:)}(\boldsymbol{Y} - \boldsymbol{CA\mu})}{\left(\hat{\boldsymbol{\Sigma}}^{-1}\right)^{(i,i)}}\right)^2\right)^{a_i + \frac{1}{2}}} \frac{\exp\left(-\frac{1}{2}(\boldsymbol{Y} - \boldsymbol{CA\mu})^T \hat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{Y} - \boldsymbol{CA\mu})\right)}{\sqrt{|\hat{\boldsymbol{\Sigma}}|}\sqrt{\left(\tilde{\boldsymbol{V}}^{(i,i)} + \boldsymbol{\Sigma}_A^{(i,i)}\left(\hat{\boldsymbol{\Sigma}}^{-1}\right)^{(i,i)}\right)}}$$
$$\exp\left(\frac{1}{2}\left(1 + \left(\frac{\tilde{\boldsymbol{V}}_{t+1}^{(i,i)}}{\boldsymbol{\Sigma}_A^{(i,i)}\left(\hat{\boldsymbol{\Sigma}}^{-1}\right)^{(i,i)}}\right)^2 \frac{\boldsymbol{\Sigma}_A^{(i,i)}\left(\hat{\boldsymbol{\Sigma}}^{-1}\right)^{(i,i)}}{\boldsymbol{\Sigma}_A^{(i,i)}\left(\hat{\boldsymbol{\Sigma}}^{-1}\right)^{(i,i)} + \tilde{\boldsymbol{V}}_{t+1}^{(i,i)}}\right)\left(\frac{\left(\hat{\boldsymbol{\Sigma}}^{-1}\right)^{(i,:)}(\boldsymbol{Y} - \boldsymbol{CA\mu})}{\sqrt{\left(\hat{\boldsymbol{\Sigma}}^{-1}\right)^{(i,i)}}}\right)^2\right).$$

Proof: We wish to sample from the posterior distribution of $\tilde{\mathbf{V}}_{t+1}^{(i,i)}$ which is pro-

portional to

$$r_i f_i\left(\tilde{\mathbf{V}}_{t+1}^{(i,i)}\right) \frac{\exp\left(-\frac{1}{2}\left(\mathbf{Y}-\mathbf{CA}\boldsymbol{\mu}\right)^T\left(\hat{\boldsymbol{\Sigma}}+\frac{\boldsymbol{\Sigma}_A^{(i,i)}}{\tilde{\mathbf{V}}_{t+1}^{(i,i)}}\mathbf{I}^{(i)}\right)^{-1}\left(\mathbf{Y}-\mathbf{CA}\boldsymbol{\mu}\right)\right)}{\sqrt{\left|\hat{\boldsymbol{\Sigma}}+\frac{\boldsymbol{\Sigma}_A^{(i,i)}}{\tilde{\mathbf{V}}_{t+1}^{(i,i)}}\mathbf{I}^{(i)}\right|}}, \tag{C.1.1}$$

where $f_i()$ denotes the PDF of a $\tilde{\sigma}_i\Gamma(a_i,a_i)$-distribution. The intractable part in the above consists of

$$\left(\hat{\boldsymbol{\Sigma}}+\frac{\boldsymbol{\Sigma}_A^{(i,i)}}{\tilde{\mathbf{V}}_{t+1}^{(i,i)}}\mathbf{I}^{(i)}\right)^{-1},$$

where $\mathbf{I}^{(i)}=\mathbf{e}_i\mathbf{e}_i^T$ is a matrix which is 0 everywhere with the exception of the ith entry of the ith row, which is 1. Note that $\mathbf{I}^{(i)}$ has rank 1 and therefore, by the Sherman Morrison formula,

$$\begin{aligned}\left(\hat{\boldsymbol{\Sigma}}+\frac{\boldsymbol{\Sigma}_A^{(i,i)}}{\tilde{\mathbf{V}}_{t+1}^{(i,i)}}\mathbf{I}^{(i)}\right)^{-1} &= \hat{\boldsymbol{\Sigma}}^{-1}-\frac{\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{I}^{(i)}\hat{\boldsymbol{\Sigma}}^{-1}}{1+tr(\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{I}^{(i)})\frac{\boldsymbol{\Sigma}_A^{(i,i)}}{\tilde{\mathbf{V}}_{t+1}^{(i,i)}}}\frac{\boldsymbol{\Sigma}_A^{(i,i)}}{\tilde{\mathbf{V}}_{t+1}^{(i,i)}}\\ &= \hat{\boldsymbol{\Sigma}}^{-1}-\frac{1}{tr(\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{I}^{(i)})}\frac{\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{I}^{(i)}\hat{\boldsymbol{\Sigma}}^{-1}}{1+\frac{1}{tr(\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{I}^{(i)})\boldsymbol{\Sigma}_A^{(i,i)}}\tilde{\mathbf{V}}_{t+1}^{(i,i)}}.\end{aligned}$$

Furthermore, given $tr(\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{I}^{(i)})=\left(\hat{\boldsymbol{\Sigma}}^{-1}\right)^{(i,i)}$, the above is equal to

$$\hat{\boldsymbol{\Sigma}}^{-1}-\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{I}^{(i)}\hat{\boldsymbol{\Sigma}}^{-1}\left[\frac{1}{\left(\hat{\boldsymbol{\Sigma}}^{-1}\right)^{(i,i)}}-\left(\frac{1}{\left(\hat{\boldsymbol{\Sigma}}^{-1}\right)^{(i,i)}}\right)^2\frac{\tilde{\mathbf{V}}_{t+1}^{(i,i)}}{\boldsymbol{\Sigma}_A^{(i,i)}}+\left(\frac{\tilde{\mathbf{V}}_{t+1}^{(i,i)}}{\boldsymbol{\Sigma}_A^{(i,i)}\left(\hat{\boldsymbol{\Sigma}}^{-1}\right)^{(i,i)}}\right)^2\frac{1}{\left(\hat{\boldsymbol{\Sigma}}^{-1}\right)^{(i,i)}+\frac{\tilde{\mathbf{V}}_{t+1}^{(i,i)}}{\boldsymbol{\Sigma}_A^{(i,i)}}}\right].$$

Crucially, the first term is constant in $\tilde{\mathbf{V}}_{t+1}^{(i,i)}$, while the second is linear in $\tilde{\mathbf{V}}_{t+1}^{(i,i)}$ and therefore conjugate to the prior of $\tilde{\mathbf{V}}_{t+1}^{(i,i)}$. The last term is quadratic in $\tilde{\mathbf{V}}_{t+1}^{(i,i)}$ and therefore vanishing much faster than the other two terms as $\tilde{\mathbf{V}}_{t+1}^{(i,i)}$ goes to 0, i.e. as the anomaly becomes stronger.

A very similar result for rank 1 updates of determinants, the matrix determinant Lemma, can be used to show that

$$\left|\hat{\boldsymbol{\Sigma}}+\frac{\boldsymbol{\Sigma}_A^{(i,i)}}{\tilde{\mathbf{V}}_{t+1}^{(i,i)}}\mathbf{I}^{(i)}\right| = \left|\hat{\boldsymbol{\Sigma}}\right|\left(1+\frac{\boldsymbol{\Sigma}_A^{(i,i)}}{\tilde{\mathbf{V}}_{t+1}^{(i,i)}}\left(\hat{\boldsymbol{\Sigma}}^{-1}\right)^{(i,i)}\right).$$

Furthermore, given that

$$-\frac{1}{2}\left(\mathbf{Y}-\mathbf{CA}\boldsymbol{\mu}\right)^{T}\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{I}^{(j)}\hat{\boldsymbol{\Sigma}}^{-1}\left(\mathbf{Y}-\mathbf{CA}\boldsymbol{\mu}\right)$$

is equal to

$$-\frac{1}{2}\left(\left(\hat{\boldsymbol{\Sigma}}^{-1}\right)^{(i,:)}\left(\mathbf{Y}-\mathbf{CA}\boldsymbol{\mu}\right)\right)^{2},$$

we can rewrite the posterior of $\tilde{\mathbf{V}}_{t+1}^{(i,i)}$ in Equation (C.1.1) as

$$r_i f(\mathbf{V}_{t+1}^{(i,i)})\sqrt{|\tilde{\mathbf{V}}_{t+1}^{(i,i)}|}\exp\left(-\frac{\tilde{\mathbf{V}}_{t+1}^{(i,i)}}{2\boldsymbol{\Sigma}_A^{(i,i)}}\left(\frac{\left(\hat{\boldsymbol{\Sigma}}^{-1}\right)^{(i,:)}\left(\mathbf{Y}-\mathbf{CA}\boldsymbol{\mu}\right)}{\left(\hat{\boldsymbol{\Sigma}}^{-1}\right)^{(i,i)}}\right)^{2}\right)\frac{\exp\left(-\frac{1}{2}\left(\mathbf{Y}-\mathbf{CA}\boldsymbol{\mu}\right)^{T}\hat{\boldsymbol{\Sigma}}^{-1}\left(\mathbf{Y}-\mathbf{CA}\boldsymbol{\mu}\right)\right)}{\sqrt{|\hat{\boldsymbol{\Sigma}}|}\sqrt{\left(\tilde{\mathbf{V}}^{(i,i)}+\boldsymbol{\Sigma}_A^{(i,i)}\left(\hat{\boldsymbol{\Sigma}}^{-1}\right)^{(i,i)}\right)}}$$

$$\exp\left(\frac{1}{2}\left(1+\left(\frac{\tilde{\mathbf{V}}_{t+1}^{(i,i)}}{\boldsymbol{\Sigma}_A^{(i,i)}\left(\hat{\boldsymbol{\Sigma}}^{-1}\right)^{(i,i)}}\right)^{2}\frac{\boldsymbol{\Sigma}_A^{(i,i)}\left(\hat{\boldsymbol{\Sigma}}^{-1}\right)^{(i,i)}}{\boldsymbol{\Sigma}_A^{(i,i)}\left(\hat{\boldsymbol{\Sigma}}^{-1}\right)^{(i,i)}+\tilde{\mathbf{V}}_{t+1}^{(i,i)}}\right)\left(\frac{\left(\hat{\boldsymbol{\Sigma}}^{-1}\right)^{(i,:)}\left(\mathbf{Y}-\mathbf{CA}\boldsymbol{\mu}\right)}{\sqrt{\left(\hat{\boldsymbol{\Sigma}}^{-1}\right)^{(i,i)}}}\right)^{2}\right)$$

Using conjugacy, we can therefore sample M particles for $\tilde{\mathbf{V}}^{(i,i)}$ from

$$\tilde{\sigma}_i\Gamma\left(a_i+\frac{1}{2},a_i+\frac{\tilde{\sigma}_i}{2\boldsymbol{\Sigma}_A^{(i,i)}}\left(\frac{\left(\hat{\boldsymbol{\Sigma}}^{-1}\right)^{(i,:)}\left(\mathbf{Y}-\mathbf{CA}\boldsymbol{\mu}\right)}{\left(\hat{\boldsymbol{\Sigma}}^{-1}\right)^{(i,i)}}\right)^{2}\right)$$

and give each particle an importance weight proportional to

$$\frac{1}{M}r_i\frac{\Gamma(a_i+\frac{1}{2})}{\Gamma(a_i)}\sqrt{\tilde{\sigma}_i}\frac{a_i^{a_i}}{\left(a_i+\frac{\tilde{\sigma}_i}{2\boldsymbol{\Sigma}_A^{(i,i)}}\left(\frac{\left(\hat{\boldsymbol{\Sigma}}^{-1}\right)^{(i,:)}\left(\mathbf{Y}-\mathbf{CA}\boldsymbol{\mu}\right)}{\left(\hat{\boldsymbol{\Sigma}}^{-1}\right)^{(i,i)}}\right)^{2}\right)^{a_i+\frac{1}{2}}}\frac{\exp\left(-\frac{1}{2}\left(\mathbf{Y}-\mathbf{CA}\boldsymbol{\mu}\right)^{T}\hat{\boldsymbol{\Sigma}}^{-1}\left(\mathbf{Y}-\mathbf{CA}\boldsymbol{\mu}\right)\right)}{\sqrt{|\hat{\boldsymbol{\Sigma}}|}\sqrt{\left(\tilde{\mathbf{V}}^{(i,i)}+\boldsymbol{\Sigma}_A^{(i,i)}\left(\hat{\boldsymbol{\Sigma}}^{-1}\right)^{(i,i)}\right)}}$$

$$\exp\left(\frac{1}{2}\left(1+\left(\frac{\tilde{\mathbf{V}}_{t+1}^{(i,i)}}{\boldsymbol{\Sigma}_A^{(i,i)}\left(\hat{\boldsymbol{\Sigma}}^{-1}\right)^{(i,i)}}\right)^{2}\frac{\boldsymbol{\Sigma}_A^{(i,i)}\left(\hat{\boldsymbol{\Sigma}}^{-1}\right)^{(i,i)}}{\boldsymbol{\Sigma}_A^{(i,i)}\left(\hat{\boldsymbol{\Sigma}}^{-1}\right)^{(i,i)}+\tilde{\mathbf{V}}_{t+1}^{(i,i)}}\right)\left(\frac{\left(\hat{\boldsymbol{\Sigma}}^{-1}\right)^{(i,:)}\left(\mathbf{Y}-\mathbf{CA}\boldsymbol{\mu}\right)}{\sqrt{\left(\hat{\boldsymbol{\Sigma}}^{-1}\right)^{(i,i)}}}\right)^{2}\right).$$

C.1.2 Theorem 6

Theorem 6. *Let the prior for the hidden state $\boldsymbol{X}_t$ be $N(\boldsymbol{\mu},\boldsymbol{\Sigma})$ and an observation $\boldsymbol{Y}_{t+1}:=\boldsymbol{Y}$ be available. Then the samples for $\tilde{\boldsymbol{W}}^{(j,j)}$ from*

$$\hat{\sigma}_i\Gamma\left(b_j+\frac{1}{2},b_j+\frac{\hat{\sigma}_j}{2\boldsymbol{\Sigma}_I^{(j,j)}}\left(\frac{\left(\boldsymbol{C}^{T}\right)^{(j,:)}\hat{\boldsymbol{\Sigma}}^{-1}\left(\boldsymbol{Y}-\boldsymbol{CA}\boldsymbol{\mu}\right)}{\left(\boldsymbol{C}^{T}\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{C}\right)^{(j,j)}}\right)^{2}\right)$$

have associated weight

$$\frac{1}{M}s_j\frac{\Gamma(b_i+\frac{1}{2})}{\Gamma(b_j)}\sqrt{\hat{\sigma}_j}\frac{b_j^{b_j}}{\left(b_j+\frac{\hat{\sigma}_i}{2\boldsymbol{\Sigma}_I^{(j,j)}}\left(\frac{(\boldsymbol{C}^T)^{(j,:)}\hat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{Y}-\boldsymbol{CA\mu})}{(\boldsymbol{C}^T\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{C})^{(j,j)}}\right)^2\right)^{b_i+\frac{1}{2}}}\frac{\exp\left(-\frac{1}{2}(\boldsymbol{Y}-\boldsymbol{CA\mu})^T\hat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{Y}-\boldsymbol{CA\mu})\right)}{\sqrt{|\hat{\boldsymbol{\Sigma}}|}\sqrt{\left(\tilde{\boldsymbol{W}}^{(j,j)}+\boldsymbol{\Sigma}_I^{(j,j)}\left(\boldsymbol{C}^T\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{C}\right)^{(j,j)}\right)}}$$

$$\exp\left(\frac{1}{2}\left(1+\left(\frac{\tilde{\boldsymbol{W}}^{(j,j)}}{\boldsymbol{\Sigma}_I^{(j,j)}\left(\boldsymbol{C}^T\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{C}\right)^{(j,j)}}\right)^2\frac{\boldsymbol{\Sigma}_I^{(j,j)}\left(\boldsymbol{C}^T\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{C}\right)^{(j,j)}}{\boldsymbol{\Sigma}_I^{(j,j)}\left(\boldsymbol{C}^T\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{C}\right)^{(j,j)}+\tilde{\boldsymbol{W}}_{t+1}^{(j,j)}}\right)\left(\frac{(\boldsymbol{C}^T)^{(j,:)}\hat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{Y}-\boldsymbol{CA\mu})}{\sqrt{\left(\boldsymbol{C}^T\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{C}\right)^{(j,j)}}}\right)^2\right)$$

The proof is almost identical to that of Theorem 5 and has been omitted.

C.1.3 Theorem 7

Theorem 7. *Let the prior for the hidden state* $\boldsymbol{X}_t$ *be* $N(\boldsymbol{\mu},\boldsymbol{\Sigma})$ *and an observation* $\boldsymbol{Y}_{t+1}:=\boldsymbol{Y}$ *be available. Then the proposal particle* $(\boldsymbol{I}_p,\boldsymbol{I}_q)$ *for* $(\boldsymbol{V}_t,\boldsymbol{W}_t)$ *has weight proportional to*

$$(1-\sum_{i=1}^{p}r_i-\sum_{j=1}^{q}s_j)\frac{\exp\left(-\frac{1}{2}(\boldsymbol{Y}-\boldsymbol{CA\mu})^T\hat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{Y}-\boldsymbol{CA\mu})\right)}{\sqrt{|\hat{\boldsymbol{\Sigma}}|}}.$$

This is immediate from the Gaussian likelihood and the Bernoulli priors for $\lambda_t^{(i)}$ and $\gamma_t^{(j)}$.

C.1.4 Proof of Theorem 4

Removing the likelihood term common to all particles the importance weights can be summarised as being

$$\frac{1}{M}r_i\frac{\Gamma(a_i+\frac{1}{2})}{\Gamma(a_i)}\sqrt{\tilde{\sigma}_i}\frac{a_i^{a_i}}{\left(a_i+\frac{\tilde{\sigma}_i}{2\boldsymbol{\Sigma}_A^{(i,i)}}\left(\frac{(\hat{\boldsymbol{\Sigma}}^{-1})^{(i,:)}(\mathbf{Y}-\mathbf{CA}\boldsymbol{\mu})}{(\hat{\boldsymbol{\Sigma}}^{-1})^{(i,i)}}\right)^2\right)^{a_i+\frac{1}{2}}}\frac{1}{\sqrt{\left(\tilde{\mathbf{V}}^{(i,i)}+\boldsymbol{\Sigma}_A^{(i,i)}\left(\hat{\boldsymbol{\Sigma}}^{-1}\right)^{(i,i)}\right)}}$$

$$\exp\left(\frac{1}{2}\left(1+\left(\frac{\tilde{\mathbf{V}}_{t+1}^{(i,i)}}{\boldsymbol{\Sigma}_A^{(i,i)}\left(\hat{\boldsymbol{\Sigma}}^{-1}\right)^{(i,i)}}\right)^2\frac{\boldsymbol{\Sigma}_A^{(i,i)}\left(\hat{\boldsymbol{\Sigma}}^{-1}\right)^{(i,i)}}{\boldsymbol{\Sigma}_A^{(i,i)}\left(\hat{\boldsymbol{\Sigma}}^{-1}\right)^{(i,i)}+\tilde{\mathbf{V}}_{t+1}^{(i,i)}}\right)\left(\frac{\left(\hat{\boldsymbol{\Sigma}}^{-1}\right)^{(i,:)}(\mathbf{Y}-\mathbf{CA}\boldsymbol{\mu})}{\sqrt{\left(\hat{\boldsymbol{\Sigma}}^{-1}\right)^{(i,i)}}}\right)^2\right).$$

for the particles containing an anomaly in the ith additive component, and

$$\frac{1}{M} s_j \frac{\Gamma(b_i+\frac{1}{2})}{\Gamma(b_j)} \sqrt{\hat{\sigma}_j} \frac{b_j^{b_j}}{\left(b_j + \frac{\hat{\sigma}_j}{2\boldsymbol{\Sigma}_I^{(j,j)}} \left(\frac{(\mathbf{C}^T)^{(j,:)} \hat{\boldsymbol{\Sigma}}^{-1} (\mathbf{Y}-\mathbf{C}\mathbf{A}\boldsymbol{\mu})}{(\mathbf{C}^T\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{C})^{(j,j)}}\right)^2\right)^{b_i+\frac{1}{2}}} \frac{1}{\sqrt{\left(\tilde{\mathbf{W}}^{(j,j)} + \boldsymbol{\Sigma}_I^{(j,j)} \left(\mathbf{C}^T\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{C}\right)^{(j,j)}\right)}}$$

$$\exp\left(\frac{1}{2}\left(1 + \left(\frac{\tilde{\mathbf{W}}^{(j,j)}}{\boldsymbol{\Sigma}_I^{(j,j)}\left(\mathbf{C}^T\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{C}\right)^{(j,j)}}\right)^2 \frac{\boldsymbol{\Sigma}_I^{(j,j)}\left(\mathbf{C}^T\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{C}\right)^{(j,j)}}{\boldsymbol{\Sigma}_I^{(j,j)}\left(\mathbf{C}^T\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{C}\right)^{(j,j)} + \tilde{\mathbf{W}}_{t+1}^{(j,j)}}\right)\left(\frac{(\mathbf{C}^T)^{(j,:)}\hat{\boldsymbol{\Sigma}}^{-1}(\mathbf{Y}-\mathbf{C}\mathbf{A}\boldsymbol{\mu})}{\sqrt{\left(\mathbf{C}^T\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{C}\right)^{(j,j)}}}\right)^2\right)$$

for the particles containing an anomaly in the jth innovative component.

As mentioned in Section II that the mean of the proposal of the ith additive component behaves asymptotically as

$$(2a_i+1)\boldsymbol{\Sigma}_A^{(i,i)} \left(\frac{\left(\hat{\boldsymbol{\Sigma}}^{-1}\right)^{(i,i)}}{\left(\hat{\boldsymbol{\Sigma}}^{-1}\right)^{(i,:)}(\mathbf{Y}-\mathbf{C}\mathbf{A}\boldsymbol{\mu})}\right)^2.$$

Furthermore, the standard deviation is on the same scale. We therefore have that

$$\tilde{\mathbf{V}}_{t+1}^{(i,i)} \sim \frac{1}{\delta^2}$$

as $\delta \to \infty$. The weight of an anomaly in the ith additive component therefore asymptotically behaves as

$$\frac{a_i^{a_i}\frac{1}{M} r_i \frac{\Gamma(a_i+\frac{1}{2})}{\Gamma(a_i)} \exp\left(\frac{1}{2}\delta^2\right)}{\left(\frac{\tilde{\sigma}_i}{2\boldsymbol{\Sigma}_A^{(i,i)}\left(\hat{\boldsymbol{\Sigma}}^{-1}\right)^{(i,i)}}\delta^2\right)^{a_i}}$$

when $\mathbf{Y}-\mathbf{C}\mathbf{A}\boldsymbol{\mu} = \frac{1}{\sqrt{\left(\hat{\boldsymbol{\Sigma}}^{-1}\right)^{(i,i)}}}\delta\mathbf{e}_i$ as $\delta \to \infty$. A very similar reasoning can be used to show that the weight of an anomaly in the jth innovative component converges to

$$\frac{b_j^{b_j}\frac{1}{M} s_j \frac{\Gamma(b_j+\frac{1}{2})}{\Gamma(b_j)} \exp\left(\frac{1}{2}\delta^2\right)}{\left(\frac{\hat{\sigma}_j}{2\boldsymbol{\Sigma}_I^{(j,j)}\left(\mathbf{C}^T\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{C}\right)^{(j,j)}}\delta^2\right)^{b_j}}$$

when $\mathbf{Y}-\mathbf{C}\mathbf{A}\boldsymbol{\mu} = \frac{\mathbf{C}^{(:,ij}}{\sqrt{\left(\mathbf{C}^T\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{C}\right)^{(j,j)}}}\delta$ as $\delta \to \infty$.

The result then follows when all the b_js and the a_is are equal to the same constant c and

$$\tilde{\sigma}_i = \boldsymbol{\Sigma}_A^{(i,i)}\left(\hat{\boldsymbol{\Sigma}}^{-1}\right)^{(i,i)} \quad \text{and} \quad \hat{\sigma}_j = \boldsymbol{\Sigma}_I^{(j,j)}\left(\mathbf{C}^T\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{C}\right)^{(j,j)}.$$

C.1.5 Theorem 8

Theorem 8. *Let the prior for the hidden state* $\boldsymbol{X}_{t-k}$ *be* $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$*. Then the samples for* $\tilde{\boldsymbol{W}}_{t-k+1}^{(j,j)}$ *from*

$$\hat{\sigma_j}\Gamma\left(b_j+\frac{1}{2}, b_j+\frac{\hat{\sigma_j}}{2\boldsymbol{\Sigma}_I^{(j,j)}}\left(\frac{\left(\left(\tilde{\boldsymbol{C}}^{(k)}\right)^T\right)^{(j,:)}\left(\hat{\boldsymbol{\Sigma}}^{(k)}\right)^{-1}\tilde{\boldsymbol{z}}_{t+1-k}^{(k)}}{\left(\left(\tilde{\boldsymbol{C}}^{(k)}\right)^T\left(\hat{\boldsymbol{\Sigma}}^{(k)}\right)^{-1}\tilde{\boldsymbol{C}}^{(k)}\right)^{(j,j)}}\right)^2\right),$$

where $\tilde{\boldsymbol{z}}_{t+1-k}^{(k)} = \tilde{\boldsymbol{Y}}_{t+1-k}^{(k)} - \tilde{\boldsymbol{C}}^{(k)}\boldsymbol{A}\boldsymbol{\mu}$ *have associated weight*

$$\frac{\frac{\frac{1}{M}s_i\left(1-\sum_{i'=1}^p r_{i'}-\sum_{j'=1}^q s_{j'}\right)^k\frac{\Gamma(b_j+\frac{1}{2})}{\Gamma(b_j)}\sqrt{\hat{\sigma}_j}b_j^{b_j}}{\left(b_i+\frac{\hat{\sigma}_j}{2\boldsymbol{\Sigma}_I^{(j,j)}}\left(\frac{\left(\left(\tilde{\boldsymbol{C}}^{(k)}\right)^T\right)^{(j,:)}(\hat{\boldsymbol{\Sigma}}^{(k)})^{-1}\left(\tilde{\boldsymbol{z}}_{t+1-k}^{(k)}\right)}{\left(\left(\tilde{\boldsymbol{C}}^{(k)}\right)^T(\hat{\boldsymbol{\Sigma}}^{(k)})^{-1}\tilde{\boldsymbol{C}}^{(k)}\right)^{(j,j)}}\right)^2\right)^{b_j+\frac{1}{2}}}\exp\left(-\frac{1}{2}\left(\tilde{\boldsymbol{z}}_{t+1-k}^{(k)}\right)^T\left(\hat{\boldsymbol{\Sigma}}^{(k)}\right)^{-1}\left(\tilde{\boldsymbol{z}}_{t+1-k}^{(k)}\right)\right)}{\sqrt{\left|\hat{\boldsymbol{\Sigma}}^{(k)}\right|}\sqrt{\left(\boldsymbol{W}^{(j,j)}+\boldsymbol{\Sigma}_I^{(j,j)}\left(\left(\tilde{\boldsymbol{C}}^{(k)}\right)^T\left(\hat{\boldsymbol{\Sigma}}^{(k)}\right)^{-1}\left(\tilde{\boldsymbol{C}}^{(k)}\right)\right)^{(j,j)}\right)}}$$

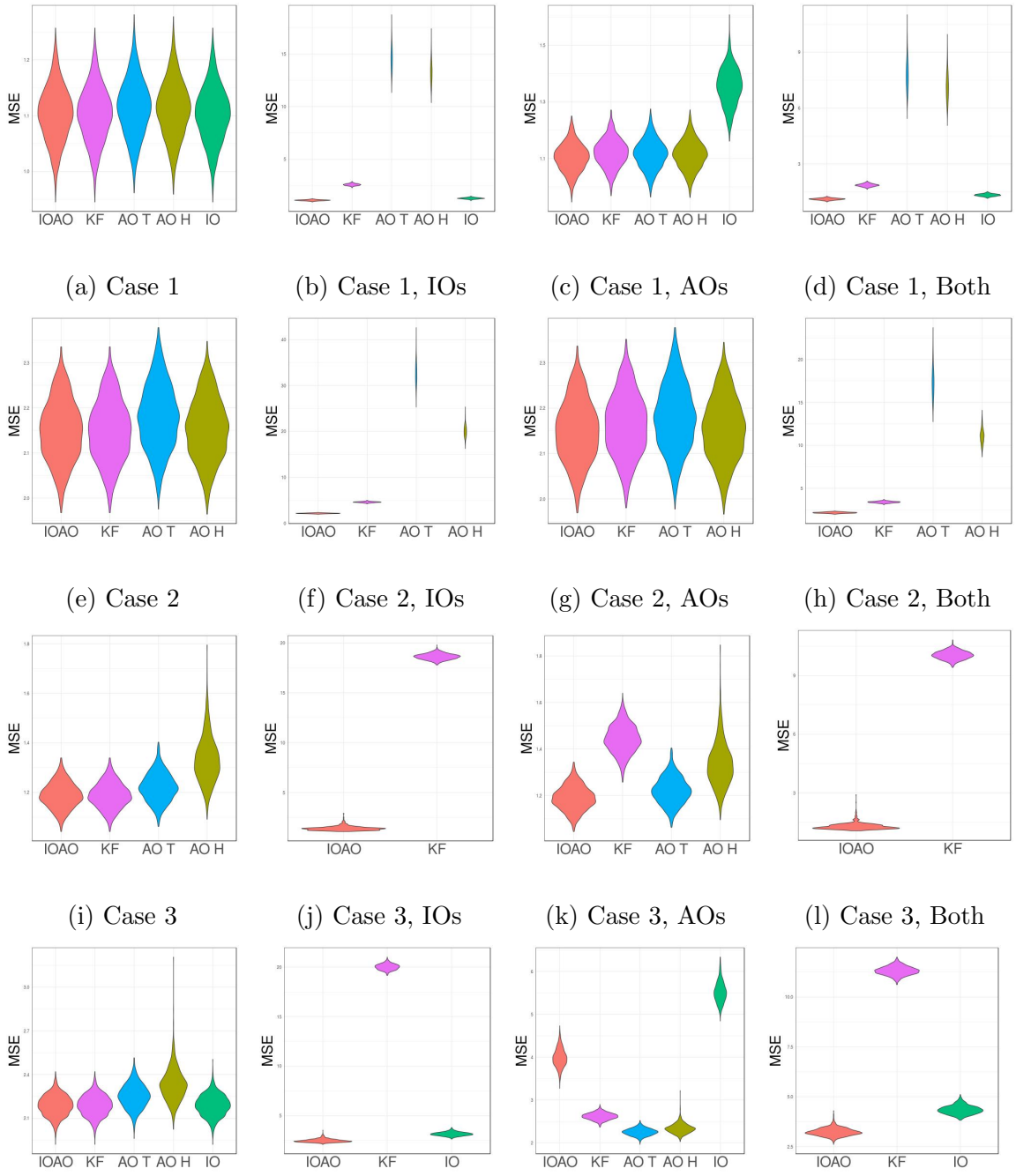
$$\exp\left(\frac{1}{2}\left(1+\left(\frac{\boldsymbol{W}_{t+1}^{(j,j)}}{\boldsymbol{\Sigma}_I^{(j,j)}\left(\left(\tilde{\boldsymbol{C}}^{(k)}\right)^T\left(\hat{\boldsymbol{\Sigma}}^{(k)}\right)^{-1}\left(\tilde{\boldsymbol{C}}^{(k)}\right)\right)^{(j,j)}}\right)^2\frac{\boldsymbol{\Sigma}_I^{(j,j)}\left(\left(\tilde{\boldsymbol{C}}^{(k)}\right)^T\left(\hat{\boldsymbol{\Sigma}}^{(k)}\right)^{-1}\left(\tilde{\boldsymbol{C}}^{(k)}\right)\right)^{(j,j)}}{\boldsymbol{\Sigma}_I^{(j,j)}\left(\left(\tilde{\boldsymbol{C}}^{(k)}\right)^T\left(\hat{\boldsymbol{\Sigma}}^{(k)}\right)^{-1}\left(\tilde{\boldsymbol{C}}^{(k)}\right)\right)^{(j,j)}+\boldsymbol{W}_{t+1}^{(j,j)}}\right.\right.$$
$$\left.\left.\left(\frac{\left(\left(\tilde{\boldsymbol{C}}^{(k)}\right)^T\right)^{(j,:)}\left(\hat{\boldsymbol{\Sigma}}^{(k)}\right)^{-1}\left(\tilde{\boldsymbol{Y}}_{t+1-k}^{(k)}-\left(\tilde{\boldsymbol{C}}^{(k)}\right)\boldsymbol{A}\boldsymbol{\mu}_{t-k}\right)}{\sqrt{\left(\left(\tilde{\boldsymbol{C}}^{(k)}\right)^T\left(\hat{\boldsymbol{\Sigma}}^{(k)}\right)^{-1}\left(\tilde{\boldsymbol{C}}^{(k)}\right)\right)^{(j,j)}}}\right)^2\right)$$

Proof: Identical (up to variable names) to that of Theorem 6.

C.2 Additional Simulations

Violin plots for the predictive mean squared error are displayed in Figure C.2.1

C.3 Complete pseudocode



(a) Case 1 (b) Case 1, IOs (c) Case 1, AOs (d) Case 1, Both

(e) Case 2 (f) Case 2, IOs (g) Case 2, AOs (h) Case 2, Both

(i) Case 3 (j) Case 3, IOs (k) Case 3, AOs (l) Case 3, Both

(m) Case 4 (n) Case 4, IOs (o) Case 4, AOs (p) Case 4, Both

Figure C.2.1: Average predictive mean squared error of the five filters over the four different scenarios under a range of models. Lower values correspond to better performance. Methods are omitted if they can not be applied to the setting or if their performance is too poor.

Algorithm 9 KF_Upd($\mathbf{Y}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{C}, \mathbf{A}, \boldsymbol{\Sigma}_A, \boldsymbol{\Sigma}_I$)

$\boldsymbol{\mu}_p \leftarrow \mathbf{A}\boldsymbol{\mu}$

$\boldsymbol{\Sigma}_p \leftarrow \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T + \boldsymbol{\Sigma}_I$

$\mathbf{z} = \mathbf{Y} - \boldsymbol{\mu}_p$

$\hat{\boldsymbol{\Sigma}} \leftarrow \mathbf{C}\boldsymbol{\Sigma}_p\mathbf{C}^T + \boldsymbol{\Sigma}_A$

$\mathbf{K} \leftarrow \boldsymbol{\Sigma}_p\mathbf{C}^T\hat{\boldsymbol{\Sigma}}^{-1}$

$\boldsymbol{\mu}_{new} \leftarrow \boldsymbol{\mu}_p + \mathbf{K}\mathbf{z}$

$\boldsymbol{\Sigma}_{new} \leftarrow (\mathbf{I} - \mathbf{K}\mathbf{C})\,\boldsymbol{\Sigma}_p$

Output: $(\boldsymbol{\mu}_{new}, \boldsymbol{\Sigma}_{new})$

Algorithm 10 Sample_typical($\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{Y}, \mathbf{A}, \mathbf{C}, \boldsymbol{\Sigma}_A, \boldsymbol{\Sigma}_I$)

$\mathbf{V} \leftarrow \mathbf{I}_p$

$\mathbf{W} \leftarrow \mathbf{I}_q$

$\hat{\boldsymbol{\Sigma}} \leftarrow \mathbf{C}\left(\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T + \boldsymbol{\Sigma}_I\right)\mathbf{C}^T + \boldsymbol{\Sigma}_A$

$\mathbf{z} \leftarrow \mathbf{Y} - \mathbf{C}\mathbf{A}\boldsymbol{\mu}$

$prob \leftarrow \left(1 - \sum_{i=1}^{p} r_i - \sum_{j=1}^{q} s_j\right) \exp\left(-\frac{1}{2}\mathbf{z}^T\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{z}\right) / \sqrt{\left|\hat{\boldsymbol{\Sigma}}\right|}$

Output: $(\mathbf{V}, \mathbf{W}, prob)$

Algorithm 11 Sample_add_comp$(i, \mathbf{z}, \hat{\mathbf{\Sigma}}, \mathbf{\Sigma}_A, M)$

1: $\mathbf{V} \leftarrow \mathbf{I}_p$

2: $\mathbf{V} \leftarrow \mathbf{I}_q$

3: $\mathbf{V}^{(i,i)} \leftarrow \tilde{\sigma}_i \Gamma \left(a_i + \frac{1}{2}, a_i + \frac{\tilde{\sigma}_i}{2\mathbf{\Sigma}_A^{(i,i)}} \left(\frac{\left(\hat{\mathbf{\Sigma}}^{-1}\right)^{(i,:)} \mathbf{z}}{\left(\hat{\mathbf{\Sigma}}^{-1}\right)^{(i,i)}} \right)^2 \right)$

4:

$$prob \leftarrow \frac{1}{M} r_i \frac{\Gamma(a_i + \frac{1}{2})}{\Gamma(a_i)} \frac{a_i^{a_i}}{\left(a_i + \frac{\tilde{\sigma}_i}{2\mathbf{\Sigma}_A^{(i,i)}} \left(\frac{\left(\hat{\mathbf{\Sigma}}^{-1}\right)^{(i,:)} \mathbf{z}}{\left(\hat{\mathbf{\Sigma}}^{-1}\right)^{(i,i)}} \right)^2 \right)^{a_i + \frac{1}{2}}} \frac{\sqrt{\tilde{\sigma}_i} \exp\left(-\frac{1}{2} \mathbf{z}^T \hat{\mathbf{\Sigma}}^{-1} \mathbf{z} \right)}{\sqrt{|\hat{\mathbf{\Sigma}}|} \sqrt{\left(\tilde{\mathbf{V}}^{(i,i)} + \mathbf{\Sigma}_A^{(i,i)} \left(\hat{\mathbf{\Sigma}}^{-1} \right)^{(i,i)} \right)}}$$
$$\exp\left(\frac{1}{2} \left(1 + \left(\frac{\tilde{\mathbf{V}}_{t+1}^{(i,i)}}{\mathbf{\Sigma}_A^{(i,i)} \left(\hat{\mathbf{\Sigma}}^{-1} \right)^{(i,i)}} \right)^2 \frac{\mathbf{\Sigma}_A^{(i,i)} \left(\hat{\mathbf{\Sigma}}^{-1} \right)^{(i,i)}}{\mathbf{\Sigma}_A^{(i,i)} \left(\hat{\mathbf{\Sigma}}^{-1} \right)^{(i,i)} + \tilde{\mathbf{V}}_{t+1}^{(i,i)}} \right) \left(\frac{\left(\hat{\mathbf{\Sigma}}^{-1} \right)^{(i,:)} \mathbf{z}}{\sqrt{\left(\hat{\mathbf{\Sigma}}^{-1} \right)^{(i,i)}}} \right)^2 \right).$$

Output: $(\mathbf{V}, \mathbf{W}, prob)$

Algorithm 12 Sample_add$(\boldsymbol{\mu}, \mathbf{\Sigma}, \mathbf{Y}, \mathbf{A}, \mathbf{C}, \mathbf{\Sigma}_A, \mathbf{\Sigma}_I, M)$

1: $\hat{\mathbf{\Sigma}} \leftarrow \mathbf{C} \left(\mathbf{A} \mathbf{\Sigma} \mathbf{A}^T + \mathbf{\Sigma}_I \right) \mathbf{C}^T + \mathbf{\Sigma}_A$

2: $\mathbf{z} \leftarrow \mathbf{Y} - \mathbf{C}\mathbf{A}\boldsymbol{\mu}$

3: $Add\_Pt \leftarrow \{\}$ ▷ Additive Anom. Particles

4: **for** $i \in \{1, ..., p\}$ **do**

5: $\quad Add\_Pt \leftarrow Add\_Pt \cup \{\text{Sample\_add\_comp}(i, \mathbf{z}, \hat{\mathbf{\Sigma}}, \mathbf{\Sigma}_A, M)\}$

6: **end for**

Output: Add_Pt

Algorithm 13 Sample_inn_comp$(j, \mathbf{z}, \hat{\mathbf{\Sigma}}, \mathbf{\Sigma}_I, M)$

$\mathbf{V} \leftarrow \mathbf{I}_p$

$\mathbf{V} \leftarrow \mathbf{I}_q$

$\mathbf{W}^{(i,i)} \leftarrow \hat{\sigma}_i \Gamma\left(b_i + \frac{1}{2}, b_i + \frac{\hat{\sigma}_i}{2\mathbf{\Sigma}_I^{(i,i)}}\left(\frac{\left(\mathbf{C}^T\right)^{(i,:)}\hat{\mathbf{\Sigma}}^{-1}\mathbf{z}}{\left(\mathbf{C}^T\hat{\mathbf{\Sigma}}^{-1}\mathbf{C}\right)^{(i,i)}}\right)^2\right)$

$$prob \leftarrow \frac{1}{M} s_j \frac{\Gamma(b_i + \frac{1}{2})}{\Gamma(b_j)} \frac{b_j^{b_j}}{\left(b_j + \frac{\hat{\sigma}_i}{2\mathbf{\Sigma}_I^{(j,j)}}\left(\frac{\left(\mathbf{C}^T\right)^{(j,:)}\hat{\mathbf{\Sigma}}^{-1}\mathbf{z}}{\left(\mathbf{C}^T\hat{\mathbf{\Sigma}}^{-1}\mathbf{C}\right)^{(j,j)}}\right)^2\right)^{b_i+\frac{1}{2}}} \frac{\sqrt{\hat{\sigma}_j}\exp\left(-\frac{1}{2}\mathbf{z}^T\hat{\mathbf{\Sigma}}^{-1}\mathbf{z}\right)}{\sqrt{|\hat{\mathbf{\Sigma}}|}\sqrt{\left(\tilde{\mathbf{W}}^{(j,j)} + \mathbf{\Sigma}_I^{(j,j)}\left(\mathbf{C}^T\hat{\mathbf{\Sigma}}^{-1}\mathbf{C}\right)^{(j,j)}\right)}}$$
$$\exp\left(\frac{1}{2}\left(1 + \left(\frac{\tilde{\mathbf{W}}^{(j,j)}}{\mathbf{\Sigma}_I^{(j,j)}\left(\mathbf{C}^T\hat{\mathbf{\Sigma}}^{-1}\mathbf{C}\right)^{(j,j)}}\right)^2 \frac{\mathbf{\Sigma}_I^{(j,j)}\left(\mathbf{C}^T\hat{\mathbf{\Sigma}}^{-1}\mathbf{C}\right)^{(j,j)}}{\mathbf{\Sigma}_I^{(j,j)}\left(\mathbf{C}^T\hat{\mathbf{\Sigma}}^{-1}\mathbf{C}\right)^{(j,j)} + \tilde{\mathbf{W}}_{t+1}^{(j,j)}}\right)\left(\frac{\left(\mathbf{C}^T\right)^{(j,:)}\hat{\mathbf{\Sigma}}^{-1}\mathbf{z}}{\sqrt{\left(\mathbf{C}^T\hat{\mathbf{\Sigma}}^{-1}\mathbf{C}\right)^{(j,j)}}}\right)^2\right)$$

Output: $(\mathbf{V}, \mathbf{W}, prob)$

Algorithm 14 Sample_inn$(\boldsymbol{\mu}, \mathbf{\Sigma}, \mathbf{Y}, \mathbf{A}, \mathbf{C}, \mathbf{\Sigma}_A, \mathbf{\Sigma}_I, M)$

$\hat{\mathbf{\Sigma}} \leftarrow \mathbf{C}\left(\mathbf{A}\mathbf{\Sigma}\mathbf{A}^T + \mathbf{\Sigma}_I\right)\mathbf{C}^T + \mathbf{\Sigma}_A$

$\mathbf{z} \leftarrow \mathbf{Y} - \mathbf{C}\mathbf{A}\boldsymbol{\mu}$

$Inn\_Pt \leftarrow \{\}$ ▷ Innovative Anom. Particles

for $i \in \{1, ..., q\}$ **do**

    $Inn\_Pt \leftarrow Inn\_Pt \cup \{\text{Sample\_inn\_comp}(i, \mathbf{z}, \hat{\mathbf{\Sigma}}, \mathbf{\Sigma}_I, M)\}$

end for

Output: Inn_Pt

Algorithm 15 Sample_Particles($M, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{Y}, \mathbf{A}, \mathbf{C}, \boldsymbol{\Sigma}_A, \boldsymbol{\Sigma}_I$)

$Desc \leftarrow \{\}$ ▷ To store Descendants

$Desc \leftarrow Desc \cup \text{Sample\_typical}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{Y}, \mathbf{A}, \mathbf{C}, \boldsymbol{\Sigma}_A, \boldsymbol{\Sigma}_I)$

for $i \in 1, ..., M$ **do**

    $Desc \leftarrow Desc \cup \text{Sample\_add}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{Y}, \mathbf{A}, \mathbf{C}, \boldsymbol{\Sigma}_A, \boldsymbol{\Sigma}_I, M)$

end for

for $i \in 1, ..., M$ **do**

    $Desc \leftarrow Desc \cup \text{Sample\_inn}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{Y}, \mathbf{A}, \mathbf{C}, \boldsymbol{\Sigma}_A, \boldsymbol{\Sigma}_I, M)$

end for

Output: *Desc*

Algorithm 16 BS_inn ($\boldsymbol{\mu}, \boldsymbol{\Sigma}, \tilde{\mathbf{Y}}, \mathbf{A}, \mathbf{C}, \boldsymbol{\Sigma}_A, \boldsymbol{\Sigma}_I, M, horizon$)

$\tilde{\mathbf{C}} \leftarrow \mathbf{C}\left[\left(\mathbf{A}^0\right)^T, ..., \left(\mathbf{A}^{horizon}\right)^T\right]^T$

$\tilde{\mathbf{z}} \leftarrow \tilde{\mathbf{Y}} - \tilde{\mathbf{C}}\mathbf{A}\boldsymbol{\mu}$

$\tilde{\boldsymbol{\Sigma}} \leftarrow \tilde{\mathbf{C}}\left(\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T + \mathbf{I}_{horizon} \otimes \boldsymbol{\Sigma}_I\right)\tilde{\mathbf{C}}^T + \mathbf{I}_{horizon} \otimes \boldsymbol{\Sigma}_A$

$Cd \leftarrow \{\}$ ▷ To store Candidates.

for $i \in \{1, .., q\}$ **do**

    if $horizon \in \mathcal{B}_i$ **then**

        for $j \in \{1, ..., M\}$ **do**

            $Cd \leftarrow Cd \cup \{\text{Sample\_inn\_comp}(i, \tilde{\mathbf{z}}, \tilde{\boldsymbol{\Sigma}}, \mathbf{A}, \tilde{\mathbf{C}}, \boldsymbol{\Sigma}_I, M \cdot |\mathcal{B}_i|)\}$

        end for

    end if

end for

Output: *Cand*

Algorithm 17 Basic Particle Filter (No Back-sampling)

Input: An initial state estimate $(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$

A number of descendants, $M' = M(p+q)+1$

A number of particles to be maintained, N.

A stream of observations $\mathbf{Y}_1, \mathbf{Y}_2, ...$

Initialise: Set $Particles(0) = \{(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)\}$

for $t \in \mathbb{N}^+$ **do**

  $Candidates \leftarrow \{\}$

  for $(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \in Particles(t-1)$ **do**

    $(\mathbf{V}, \mathbf{W}, prob) \leftarrow$ Sample_Particles$(M, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{Y}_t, \mathbf{A}, \mathbf{C}, \boldsymbol{\Sigma}_A, \boldsymbol{\Sigma}_I)$

    $Candidates \leftarrow Candidates \cup \{(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{V}, \mathbf{W}, prob)\}$

  end for

  $Descendants \leftarrow$ Subsample$(N, Candidates)$

  $Particles(t) \leftarrow \{\}$

  for $(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{V}, \mathbf{W}, prob) \in Descendants$ **do**

    $(\boldsymbol{\mu}_{new}, \boldsymbol{\Sigma}_{new}) \leftarrow$ KF_Upd$(\mathbf{Y}_t, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{C}, \mathbf{A}, \mathbf{V}^{1/2}\boldsymbol{\Sigma}_A, \mathbf{W}^{1/2}\boldsymbol{\Sigma}_I)$

    $Particles(t) \leftarrow Particles(t) \cup \{(\boldsymbol{\mu}_{new}, \boldsymbol{\Sigma}_{new})\}$

  end for

end for

Algorithm 18 Particle Filter (With Back Sampling) – CE-BASS

Input: An initial state estimate $(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$.

A number of descendants, $M' = M(p+q)+1$.

A number of particles to be maintained, N.

A stream of observations $\mathbf{Y}_1, \mathbf{Y}_2, ...$

Initialise: Set $Particles(0) = \{(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0, 1)\}$

Set $max\_horizon = \max\left(\cup_{i=1}^{q} \mathcal{B}_i\right)$

for $t \in \mathbb{N}^+$ **do**

  $Cand \leftarrow \{\}$ ▷ To Store Candidates

  for $(\boldsymbol{\mu}, \boldsymbol{\Sigma}, prob_{prev}) \in Particles(t-1)$ **do**

    $(\mathbf{V}, \mathbf{W}, prob) \leftarrow$ Sample_typical$(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{Y}_t, \mathbf{A}, \mathbf{C}, \boldsymbol{\Sigma}_A, \boldsymbol{\Sigma}_I)$

    $Cand \leftarrow Cand \cup \{(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{V}, \mathbf{W}, prob \cdot prob_{prev}, 1)\}$

    $Add\_Des \leftarrow$ Sample_add$(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{Y}_t, \mathbf{A}, \mathbf{C}, \boldsymbol{\Sigma}_A, \boldsymbol{\Sigma}_I, M)$

    for $(\mathbf{V}, \mathbf{W}, prob) \in Add\_Des$ **do**

      $Cand \leftarrow Cand \cup \{(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{V}, \mathbf{W}, prob \cdot prob_{prev}, 1)\}$

    end for

  end for

  for $hor \in \{1, ..., max\_horizon\}$ **do**

    for $(\boldsymbol{\mu}, \boldsymbol{\Sigma}, prob_{prev}) \in Particles(t-hor)$ **do**

      $\tilde{\mathbf{Y}} \leftarrow \left[\mathbf{Y}_{t-hor+1}^T, ..., \mathbf{Y}_t^T\right]^T$

      $Inn\_Des \leftarrow$ BS_inn$(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \tilde{\mathbf{Y}}, \mathbf{A}, \mathbf{C}, \boldsymbol{\Sigma}_A, \boldsymbol{\Sigma}_I, M, hor)$

      for $(\mathbf{V}, \mathbf{W}, prob) \in Inn\_Des$ **do**

        $Cand \leftarrow Cand \cup \{(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{V}, \mathbf{W}, prob \cdot prob_{prev}, hor)\}$

      end for

    end for

  end for

  $Desc \leftarrow$ Subsample$(N, Cand)$ ▷ Sampling proportional to *prob*

continues on next page

21: $Particles(t) \leftarrow \{\}$

22: **for** $(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{V}, \mathbf{W}, prob, hor) \in Desc$ **do**

23: $(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \leftarrow \text{KF\_Upd}(\mathbf{Y}_{t+1-hor}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{C}, \mathbf{A}, \mathbf{V}^{1/2}\boldsymbol{\Sigma}_A, \mathbf{W}^{1/2}\boldsymbol{\Sigma}_I)$

24: **if** $hor > 1$ **then**

25: **for** $i \in \{2, ..., hor\}$ **do**

26: $(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \leftarrow \text{KF\_Upd}(\mathbf{Y}_{t+i-hor}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{C}, \mathbf{A}, \boldsymbol{\Sigma}_A, \boldsymbol{\Sigma}_I)$

27: **end for**

28: **end if**

29: $Particles(t) \leftarrow Particles(t) \cup \{(\boldsymbol{\mu}, \boldsymbol{\Sigma}, prob \cdot \frac{|Cand|}{|Desc|})\}$

30: **end for**

31: **end for**

Bibliography

Gabriel Agamennoni, Juan I Nieto, and Eduardo M Nebot. An outlier-robust Kalman filter. In *2011 IEEE International Conference on Robotics and Automation*, pages 1551–1558. IEEE, 2011.

Gabriel Agamennoni, Juan I Nieto, and Eduardo M Nebot. Approximate inference in state-space models with heavy-tailed noise. *IEEE Transactions on Signal Processing*, 60(10):5024–5037, 2012.

Pratik Agarwal, Gian Diego Tipaldi, Luciano Spinello, Cyrill Stachniss, and Wolfram Burgard. Robust map optimization using dynamic covariance scaling. In *2013 IEEE International Conference on Robotics and Automation*, pages 62–69. Ieee, 2013.

Subutai Ahmad, Alexander Lavin, Scott Purdy, and Zuha Agha. Unsupervised real-time anomaly detection for streaming data. *Neurocomputing*, 262:134–147, 2017.

Mohiuddin Ahmed, Abdun Naser Mahmood, and Md Rafiqul Islam. A survey of anomaly detection techniques in financial domain. *Future Generation Computer Systems*, 55:278–288, 2016.

Emil Artin. *The gamma function*. Courier Dover Publications, 2015.

M Sanjeev Arulampalam, Simon Maskell, Neil Gordon, and Tim Clapp. A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking. *IEEE Transactions on signal processing*, 50(2):174–188, 2002.

John AD Aston and Claudia Kirch. Evaluating stationarity via change-point alternatives with applications to fMRI data. *The Annals of Applied Statistics*, 6(4): 1906–1948, 2012.

John AD Aston, Claudia Kirch, et al. Evaluating stationarity via change-point alternatives with applications to fmri data. *The Annals of Applied Statistics*, 6(4): 1906–1948, 2012.

Lawrence Bardwell and Paul Fearnhead. Bayesian detection of abnormal segments in multiple time series. *Bayesian Analysis*, 12(1):193–218, 2017.

Stéphane Boucheron and Maud Thomas. Concentration inequalities for order statistics. *Electronic Communications in Probability*, 17, 2012.

Kris Boudt, Peter J Rousseeuw, Steven Vanduffel, and Tim Verdonck. The minimum regularized covariance determinant estimator. *Statistics and Computing*, 30(1): 113–128, 2020.

Markus Breunig, Hans-Peter Kriegel, Raymond T. Ng, and Jörg Sander. LOF: Identifying density-based local outliers. *Proceedings of the 2000 ACM Sigmod International Conference on Management of Data*, pages 93–104, 2000.

T Cai, X Jessie Jeng, and Jiashun Jin. Optimal detection of heterogeneous and het-

eroscedastic mixtures. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(5):629–662, 2011a.

Tony Cai, Jessie Jeng, and Jiashun Jin. Optimal detection of heterogeneous and heteroscedastic mixtures. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(5):629–662, 2011b.

Tony T Cai, Jessie X Jeng, and Hongzhe Li. Robust detection and identification of sparse segments in ultrahigh dimensional data analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(5):773–797, 2012.

Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3):15, 2009.

Guobin Chang. Robust Kalman filtering based on Mahalanobis distance as outlier judging criterion. *Journal of Geodesy*, 88(4):391–401, 2014.

Lubin Chang, Baiqing Hu, Guobin Chang, and An Li. Robust derivative-free kalman filter based on huber's m-estimation methodology. *Journal of Process Control*, 23 (10):1555–1561, 2013.

Rong Chen and Jun S Liu. Mixture Kalman filters. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(3):493–508, 2000.

Yudong Chen, Tengyao Wang, and Richard J Samworth. High-dimensional, multiscale online changepoint detection. *arXiv preprint arXiv:2003.03668*, 2020.

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*, pages 475–507, 2015.

Robert B Cleveland, William S Cleveland, and Irma Terpenning. STL: A seasonal-trend decomposition procedure based on loess. *Journal of Official Statistics*, 6(1): 3, 1990.

International HapMap Consortium. The international hapmap project. *Nature*, 426 (6968):789, 2003.

David Donoho and Jiashun Jin. Higher criticism for detecting sparse heterogeneous mixtures. *The Annals of Statistics*, 32(3):962–994, 2004.

Farida Enikeeva and Zaid Harchaoui. High-dimensional change-point detection under sparse alternatives. *The Annals of Statistics*, 47(4):2051–2079, 2019.

Paul Fearnhead and Peter Clifford. On-line inference for hidden Markov models via particle filters. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(4):887–899, 2003.

Paul Fearnhead and Hans R. Künsch. Particle filters and data assimilation. *Annual Review of Statistics and Its Application*, 5(1):421–449, 2018.

Paul Fearnhead and Guillem Rigaill. Changepoint detection in the presence of outliers. *Journal of the American Statistical Association*, 114(525):169–183, 2019a.

Paul Fearnhead and Guillem Rigaill. Changepoint detection in the presence of outliers. *Journal of the American Statistical Association*, 114(525):169–183, 2019b.

A. T. M. Fisch, D. J. Grose, I. A. Eckley, and P. Fearnhead. *anomaly: An R package for detecting anomalies in data.*, 2020. URL `https://CRAN.R-project.org/package=anomaly`. R package version 4.0.0.

Alexander Fisch, Idris A Eckley, and Paul Fearnhead. A linear time method for the detection of point and collective anomalies. *arXiv preprint arXiv:1806.01947*, 2018a.

Alexander T M Fisch, Idris A Eckley, and Paul Fearnhead. A linear time method for the detection of point and collective anomalies. *arXiv preprint arXiv:1806.01947*, 2018b.

Piotr Fryzlewicz. Wild binary segmentation for multiple change-point detection. *The Annals of Statistics*, 42(6):2243–2281, 2014.

Mital A Gandhi and Lamine Mili. Robust Kalman filter based on a generalized maximum-likelihood-type estimator. *IEEE Transactions on Signal Processing*, 58 (5):2509–2520, 2009.

Jonathan Goh, Sridhar Adepu, Marcus Tan, and Zi Shan Lee. Anomaly detection in cyber physical systems using recurrent neural networks. In *2017 IEEE 18th International Symposium on High Assurance Systems Engineering (HASE)*, pages 140–145. IEEE, 2017.

N. J. Gordon, D. J. Salmond, and A. F. M. Smith. Novel approach to nonlinear/non-gaussian bayesian state estimation. *IEE Proceedings F - Radar and Signal Processing*, 140(2):107–113, 1993.

Sudipto Guha, Nina Mishra, Gourav Roy, and Okke Schrijvers. Robust random cut forest based anomaly detection on streams. In *International Conference on Machine Learning*, pages 2712–2721, 2016.

Frank R. Hampel, Elvezio M. Ronchetti, Peter J. Rousseeuw, and Werner A. Stahel. *Robust Statistics - The Approach Based on Influence Functions*. Wiley, 1986.

Frank Rudolf Hampel. *Contributions to the theory of robust estimation*. PhD thesis, University of California Berkeley, 1968.

P Jeffrey Harrison and Colin F Stevens. Bayesian forecasting. *Journal of the Royal Statistical Society: Series B (Methodological)*, 38(3):205–228, 1976.

Yulong Huang, Yonggang Zhang, Ning Li, Zhemin Wu, and Jonathon A Chambers. A novel robust student's t-based Kalman filter. *IEEE Transactions on Aerospace and Electronic Systems*, 53(3):1545–1554, 2017.

Yulong Huang, Yonggang Zhang, Yuxin Zhao, and Jonathon A Chambers. A novel robust gaussian-student's t mixture distribution based Kalman filter. *IEEE Transactions on Signal Processing*, 2019.

Peter J Huber. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):73–101, 1964a.

Peter J Huber. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):73–101, 1964b.

Brad Jackson, Jeffrey D Scargle, David Barnes, Sundararajan Arabhi, Alina Alt, Peter Gioumousis, Elyus Gwin, Paungkaew Sangtrakulcharoen, Linda Tan, and Tun Tao Tsai. An algorithm for optimal partitioning of data on an interval. *IEEE Signal Processing Letters*, 12(2):105–108, 2005.

Nicholas A James, Arun Kejariwal, and David S Matteson. Leveraging cloud data to mitigate user experience from 'breaking bad'. In *Big Data (Big Data), 2016 IEEE International Conference on*, pages 3499–3508. IEEE, 2016.

X Jessie Jeng, T Tony Cai, and Hongzhe Li. Optimal sparse segment identification with application in copy number variation analysis. *Journal of the American Statistical Association*, 105(491):1156–1166, 2010.

X Jessie Jeng, T Tony Cai, and Hongzhe Li. Simultaneous discovery of rare and common segment variants. *Biometrika*, 100(1):157–172, 2012.

Jana Jurečková and Jan Picek. *Robust statistical methods with R*. CRC Press, 2005.

Rudolph Emil Kalman. A new approach to linear filtering and prediction problems. *Transactions of the ASME–Journal of Basic Engineering*, 82(Series D):35–45, 1960.

Rebecca Killick and Idris Eckley. changepoint: An R package for changepoint analysis. *Journal of statistical software*, 58(3):1–19, 2014.

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.

Rebecca Killick, Kaylea Haynes, and Idris A. Eckley. *changepoint: An R package for changepoint analysis*, 2018. URL `https://CRAN.R-project.org/package=changepoint`. R package version 2.3.

Genshiro Kitagawa. Non-gaussian state—space modeling of nonstationary time series. *Journal of the American statistical association*, 82(400):1032–1041, 1987.

Beatrice Laurent and Pascal Massart. Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, pages 1302–1338, 2000.

Marc Lavielle and Eric Moulines. Least-squares estimation of an unknown number of shifts in a time series. *Journal of time series analysis*, 21(1):33–59, 2000.

Alexander Lavin and Subutai Ahmad. Evaluating real-time anomaly detection algorithms–the numenta anomaly benchmark. In *2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA)*, pages 38–44. IEEE, 2015.

Bruce Levin and Jennie Kline. The cusum test of homogeneity with an application in spontaneous abortion epidemiology. *Statistics in Medicine*, 4(4):469–488, 1985.

Housen Li, Axel Munk, Hannes Sieling, et al. Fdr-control in multiscale change-point segmentation. *Electronic Journal of Statistics*, 10(1):918–959, 2016.

Chiao-Feng Lin, Adam C Naj, and Li-San Wang. Analyzing copy number variation using snp array data: protocols for calling cnv and association tests. *Current protocols in human genetics*, 79(1):1–27, 2013.

Haoyang Liu, Chao Gao, and Richard J Samworth. Minimax rates in sparse, high-dimensional changepoint detection. *arXiv preprint arXiv:1907.10012*, 2019.

Ting Fung Ma and Chun Yip Yau. A pairwise likelihood-based approach for change-point detection in multivariate time series models. *Biometrika*, 103(2):409–421, 2016.

Hyeyoung Maeng and Piotr Fryzlewicz. Detecting linear trend changes and point anomalies in data sequences. *arXiv preprint arXiv:1906.01939*, 2019.

Ritesh Maheshwari, Yang Yang, Ruixuan Hou, Baolei Li, and Liang Zhang. luminol: Anomaly detection and correlation library, 2014. URL `"https://github.com/linkedin/luminol"`.

Robert Maidstone, Toby Hocking, Guillem Rigaill, and Paul Fearnhead. On optimal multiple changepoint algorithms for large data. *Statistics and Computing*, 27(2): 519–533, 2017.

Timothy D Morton, Stephen T Bryson, Jeffrey L Coughlin, Jason F Rowe, Ganesh Ravichandran, Erik A Petigura, Michael R Haas, and Natalie M Batalha. False positive probabilities for all kepler objects of interest: 1284 newly validated planets and 428 likely false positives. *The Astrophysical Journal*, 822(2):86, 2016.

Nora Muler and Víctor J Yohai. Robust estimates for ARCH processes. *Journal of Time Series Analysis*, 23(3):341–375, 2002.

Nora Muler and Victor J Yohai. Robust estimates for GARCH models. *Journal of Statistical Planning and Inference*, 138(10):2918–2940, 2008.

Nora Muler, Daniel Pena, and Víctor J Yohai. Robust estimation for ARMA models. *The Annals of Statistics*, 37(2):816–840, 2009.

Susan E Mullally. Data validation time series file: Description of file format and content. 2016.

Adam B Olshen, E S Venkatraman, Robert Lucito, and Michael Wigler. Circular binary segmentation for the analysis of array-based dna copy number data. *Biostatistics*, 5(4):557–572, 2004.

Ewan S Page. Continuous inspection schemes. *Biometrika*, 41(1/2):100–115, 1954.

Marco AF Pimentel, David A Clifton, Lei Clifton, and Lionel Tarassenko. A review of novelty detection. *Signal Processing*, 99:215–249, 2014.

Michael K Pitt and Neil Shephard. Filtering via simulation: Auxiliary particle filters. *Journal of the American statistical association*, 94(446):590–599, 1999.

Rolf-Dieter Reiss. *Approximate distributions of order statistics: with applications to nonparametric statistics*. Springer science & business media, 2012.

Lucy F Robinson, Tor D Wager, and Martin A Lindquist. Change point estimation in multi-subject fMRI studies. *Neuroimage*, 49(2):1581–1592, 2010.

Peter Rousseeuw and Víctor J Yohai. Robust regression by means of S-estimators. In *Robust and nonlinear time series analysis*, pages 256–272. Springer, 1984.

Peter J Rousseeuw. Least median of squares regression. *Journal of the American statistical association*, 79(388):871–880, 1984.

Peter J Rousseeuw. Multivariate estimation with high breakdown point. *Mathematical statistics and applications*, 8:283–297, 1985.

Peter Ruckdeschel, Bernhard Spangl, and Daria Pupashenko. Robust Kalman tracking and smoothing with propagating and non-propagating outliers. *Statistical Papers*, 55(1):93–123, 2014.

P Sartoretti and J Schneider. On the detection of satellites of extrasolar planets with the method of transits. *Astronomy and Astrophysics Supplement Series*, 134(3): 553–560, 1999.

Armin Schwartzman, Yulia Gavrilov, and Robert J Adler. Multiple testing of local maxima for detection of peaks in 1d. *Annals of statistics*, 39(6):3290, 2011.

Alastair J Scott and M Knott. A cluster analysis method for grouping means in the analysis of variance. *Biometrics*, pages 507–512, 1974.

Galen R Shorack and Jon A Wellner. *Empirical processes with applications to statistics*, volume 59. Siam, 2009.

David Siegmund, Benjamin Yakir, and Nancy R Zhang. Detecting simultaneous vari-

ant intervals in aligned sequences. *The Annals of Applied Statistics*, pages 645–668, 2011.

Padhraic Smyth. Markov monitoring with unknown states. *IEEE Journal on Selected Areas in Communications*, 12(9):1600–1612, 1994.

Chi Song, Xiaoyi Min, and Heping Zhang. The screening and ranking algorithm for change-points detection in multiple samples. *The annals of applied statistics*, 10(4): 2102, 2016.

Otto Struve. Proposal for a project of high-precision stellar radial velocity work. *The Observatory*, 72:199–200, 1952.

Priyanga Dilini Talagala, Rob J Hyndman, Kate Smith-Miles, Sevvandi Kandanaarachchi, and Mario A Muñoz. Anomaly detection in streaming nonstationary temporal data. *Journal of Computational and Graphical Statistics*, pages 1–21, 2019.

Andreas Theissler. Detecting known and unknown faults in automotive systems using ensemble-based anomaly detection. *Knowledge-Based Systems*, 123:163–173, 2017.

SO Tickle, IA Eckley, P Fearnhead, and K Haynes. Parallelisation of a common changepoint detection method. *arXiv preprint arXiv:1810.03591*, 2018.

Jo-Anne Ting, Evangelos Theodorou, and Stefan Schaal. Learning an outlier-robust Kalman filter. In *European Conference on Machine Learning*, pages 748–756. Springer, 2007.

John W Tukey. A survey of sampling from contaminated distributions. *Contributions to probability and statistics*, 2:448–485, 1960.

ES Venkatraman and Adam B Olshen. A faster circular binary segmentation algorithm for the analysis of array cgh data. *Bioinformatics*, 23(6):657–663, 2007.

Tengyao Wang and Richard J Samworth. Inspectchangepoint: high-dimensional changepoint estimation via sparse projection. *R Package Version*, 1, 2016.

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.

Dingjie Xu, Chen Shen, and Feng Shen. A robust particle filtering algorithm with non-gaussian measurement noise using student-t distribution. *IEEE Signal Processing Letters*, 21(1):30–34, 2013.

Qiwei Yao. Tests for change-points with epidemic alternatives. *Biometrika*, 80(1): 179–191, 1993.

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

Nancy R Zhang, David O Siegmund, Hanlee Ji, and Jun Z Li. Detecting simultaneous changepoints in multiple sequences. *Biometrika*, 97(3):631–645, 2010.

Hongshan Zhao, Huihai Liu, Wenjing Hu, and Xihui Yan. Anomaly detection and fault

analysis of wind turbine components based on deep learning network. *Renewable energy*, 127:825–834, 2018.

Zifeng Zhao and Chun Yip Yau. Alternating dynamic programming for multiple epidemic change-point estimation. *arXiv preprint arXiv:1907.06810*, 2019.