Subset Multivariate Collective And Point Anomaly Detection

Alexander T. M. Fisch, *Idris A. Eckley and Paul Fearnhead
Department of Mathematics and Statistics, Lancaster University

September 29, 2021

Abstract

In recent years, there has been a growing interest in identifying anomalous structure within multivariate data sequences. We consider the problem of detecting collective anomalies, corresponding to intervals where one, or more, of the data sequences behaves anomalously. We first develop a test for a single collective anomaly that has power to simultaneously detect anomalies that are either rare, that is affecting few data sequences, or common. We then show how to detect multiple anomalies in a way that is computationally efficient but avoids the approximations inherent in binary segmentation-like approaches. This approach is shown to consistently estimate the number and location of the collective anomalies – a property that has not previously been shown for competing methods. Our approach can be made robust to point anomalies and can allow for the anomalies to be imperfectly aligned. We show the practical usefulness of allowing for imperfect alignments through a resulting increase in power to detect regions of copy number variation.

Keywords: Copy Number Variations, Dynamic Programming, Epidemic Changepoints, Outliers, Robust Statistics.

*This work was supported by EPSRC grant numbers EP/N031938/1 (StatScale) and EP/L015692/1 (STOR-i). The authors also acknowledge British Telecommunications plc (BT) for financial support, David Yearling and Kjeld Jensen in BT Research & Innovation for discussions. Thanks also to Lawrence Bardwell for providing us with data and code reproducing results from his paper.

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. Interest is either in removing anomalies, so that robust inferences can be made by analysing non-anomalous data (Rousseeuw & Bossche 2018), or detecting anomalies because they are indicative of features of interest. Examples of the latter include anomalous sensor measurements indicating failure of equipment (Cui et al. 2018), or unusual patterns in computer network data indicating a cyber attack (Metelli & Heard 2019). Comprehensive reviews of the area can be found in Chandola et al. (2009) and Pimentel et al. (2014). One of the main challenges in anomaly detection is that anomalies can come in different guises. Chandola et al. (2009) categorises anomalies into one of three categories: global, contextual, or collective. The first two of these categories are point anomalies, i.e. single observations which are anomalous with respect to the global, or local, data context respectively. Conversely, a collective anomaly is defined as a sequence of observations which together form an anomalous pattern.

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 observations at n discrete time-points, each 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

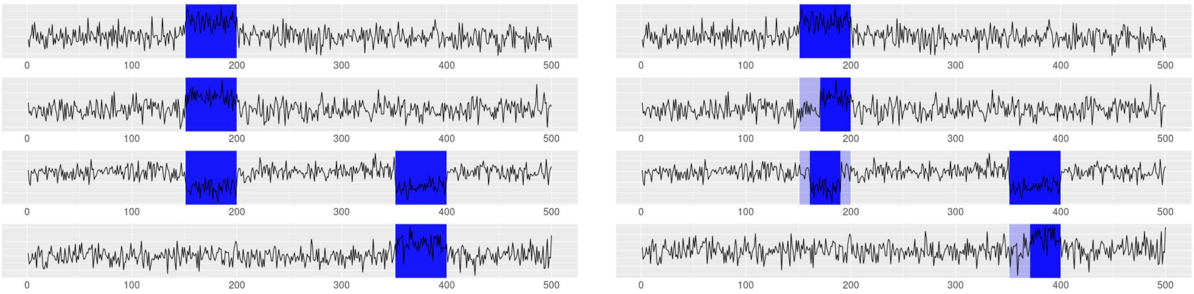


Figure 1: An example of time series with $K = 2$ collective anomalies, highlighted in blue, where they are perfectly aligned (left) and are mis-aligned (right). Using notation from Section 2 these collective anomalies occur from times $s_1 = 150$ to $e_1 = 200$ and $s_2 = 350$ to $e_2 = 400$; the affected components are $\mathbf{J}_1 = \{1, 2, 3\}$ and $\mathbf{J}_2 = \{3, 4\}$. For the mis-aligned anomalies, the time series have anomalous behaviour in dark-blue regions, and behave normally in light-blue regions.

contemporaneously across all affected sequences, in practice one might expect some time delays as illustrated by Figure 1. We will consider two different scenarios for the alignment of related collective anomalies across different components. First, 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 assumes that concurrent collective anomalies start and end at similar but not necessarily identical time points.

There has been much less work on detecting collective anomalies than point anomalies. It is possible to use point anomaly methods to detect a collective anomaly, by applying them to data from suitably chosen periods of time. For example, Talagala et al. (2021) use this approach to detect anomalous daily behaviour of pedestrians in Melbourne, while the online method of Talagala et al. (2020) produces a bivariate summary of data within a moving window and detects a collective anomaly based on a measure of how unusual this summary is. One problem with these approaches is the need to specify an appropriate period of time or window length.

Other approaches aimed at detecting collective anomalies include state space models 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 Smyth (1994) and Bardwell & Fearnhead (2017). 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 difficult to specify.

The epidemic changepoint model (Levin & Kline 1985) provides an alternative detection framework, built on an assumption that there is a normal, non-anomalous, 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 distribution that describes the non-anomalous behaviour. Epidemic changepoints can be inferred by using classical changepoint methods (Killick et al. 2012, Fryzlewicz 2014, Wang & Samworth 2018), but this leads to sub-optimal power, as it does not exploit the fact that the parameter associated with each non-anomalous segment is the same.

Instead, many epidemic changepoint methods are fit using the 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 (Jeng et al. 2012). A range of different epidemic changepoint methods have been proposed that use circular binary segmentation: the methods of Zhang et al. (2010) for dense changes, LRS (Jeng et al. 2010) for sparse changes, and higher criticism based methods like PASS (Jeng et al. 2012) for both types of changes.

The approach in this paper 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 2. Whilst penalised cost methods have proved popular for changepoint detection (e.g. Davis et al. 2006, Killick et al. 2012, amongst many others), they have been less used for detecting collective anomalies or epidemic changepoints. Such an 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. Whilst this paper focuses mainly on detecting changes in mean in Gaussian-like data, as is common for penalised cost based methods, the framework can easily extend to detecting changes in count or categorical data (Hocking et al. 2015), or to detecting changes in other features such as variance or correlation (Davis et al. 2006) or distribution (Haynes, Fearnhead & Eckley 2017).

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 optimise 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. 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.

As well as demonstrating the performance of MVCAPA empirically on simulated and real data, we also present a number of theoretical results that both guide the implementation of the method and show its strong statistical properties. One of the challenges with implementing a penalised cost 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 focusing 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.

We give finite sample consistency results for MVCAPA for the problem of detecting anomalies that change the mean in Gaussian data in Section 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 mis-alignment of the collective anomalies. We are unaware of similar consistency results for detecting collective anomalies in multivariate data. Whilst there are similar consistency results for the related problem of detecting changes in multivariate data (Wang & Samworth 2018, Cho & Fryzlewicz 2015), these are for 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, and this greatly increases the technical challenge – as one of the main issues is showing that a single true collective anomaly is not fit using multiple collective anomalies, something that can be easily excluded if our inference procedure assumes a diverging minimum segment length.

2 Model and Inference for a Single Collective Anomaly

2.1 Penalised Cost Approach

We begin with 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

$$\boldsymbol{\theta}^{(i)}(t) = \boldsymbol{\theta}_k^{(i)} \text{ if } s_k < t \leq e_k \text{ and } i \in \mathbf{J}_k \tag{1}$$

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 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 using robust statistical methods, see for example Fisch et al. (2018), or our approach in Section 7.

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

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

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, $\mathcal{C}_i\left(\mathbf{x}_{s+1:e}^{(i)}, \boldsymbol{\theta}\right) = \sum_{i=s+1}^{e}((\mathbf{x}_t^{(i)} - \boldsymbol{\mu})/\boldsymbol{\sigma}_0^{(i)})^2$, where $\boldsymbol{\mu}$ is the segment mean, and $\boldsymbol{\sigma}_0^{(i)}$ is the common standard deviation of the ith component. This leads to a saving

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

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

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

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 $\sum_{i \in \mathbf{J}} \mathcal{S}_i\left(s,e\right)$. As the start or the end of the window, or the set of components affected, are unknown, we 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

$$\max_{\mathbf{J}, s \leq e-l} \left[\sum_{i \in \mathbf{J}} \mathcal{S}_i\left(s,e\right) - P(|\mathbf{J}|) \right], \tag{2}$$

where $P(\cdot)$ 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 (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 (2).

To efficiently maximise (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 $P(\cdot)$. Let the order statistics of $\mathcal{S}_1\left(s,e\right), \ldots, \mathcal{S}_p\left(s,e\right)$ be $\mathcal{S}_{(1)}\left(s,e\right) \geq \ldots \geq \mathcal{S}_{(p)}\left(s,e\right)$, and define the penalised saving statistic of the segment $\mathbf{x}_{(s+1):e}$,

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

Then (2) is obtained by maximising $\mathcal{S}\left(s,e\right)$ 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

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

2.2 Choosing Appropriate Penalties

The choice of penalties will impact both the false error rate, the probability of detecting an anomaly if there are none, 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 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 $a\chi_v^2$ 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_v^2 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; or when data from each component follow an independent AR(1)-models with bounded positive auto-correlation parameter (Lavielle & Moulines 2000).

Let $\hat{K}$ be the estimated number of collective anomalies. Our bounds on the false positive rate will be based on showing

$$\mathbb{P}\left(\hat{K} = 0\right) \geq 1 - Ae^{-\psi(p,n)}, \tag{3}$$

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 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 2), 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 (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

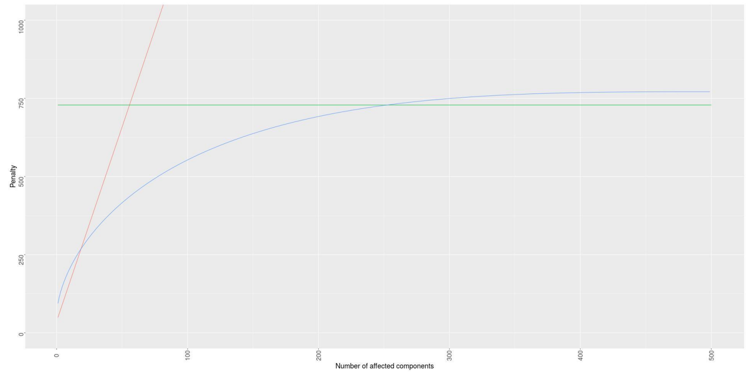


Figure 2: A comparison of the 3 penalty regimes for a χ^2_1-distributed saving when $p = 500$ and $\psi = 2\log(10000)$. Regime 1 is in green, regime 2 in red and regime 3 in blue.

$1 \leq j \leq p$ and $\alpha = a\left(pv + 2\sqrt{pv\psi} + 2\psi\right)$.

Under this penalty, we would infer that any detected anomaly region affects all components. This 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 is 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

$\mathbb{P}\left(\chi^2_v > c_j\right) = j/p$.

As can be seen from Figure 2 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 1 *Assume that there are no collective anomalies. Assume also that the savings $S_i(s,e)$ are independent across components and each 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 $\mathbb{P}\{\hat{K} = 0\} \geq 1 - An^2e^{-\psi}$.*

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 1, that this composite penalty regime achieves $\mathbb{P}\{\hat{K} = 0\} \geq 1 - 3An^2e^{-\psi}$ for the same global constant A as in Proposition 1, when $S_i(s,e)$ is stochastically bounded by $a\chi^2_v$.

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

$$\mathbf{x}_t^{(i)} = v^{(i)}\mu + \boldsymbol{\eta}_t^{(i)}, \quad v^{(i)} \sim \begin{cases} 0 & \text{with prob. } 1 - p^{-\xi}, \\ 1 & \text{with prob. } p^{-\xi}, \end{cases} \quad \text{and} \quad \boldsymbol{\eta}_t^{(i)} \overset{i.i.d.}{\sim} N(0,1), \quad \text{for} \quad s < t \leq e, \tag{4}$$

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 and $\xi > \frac{1}{2}$ corresponding to sparse changes (Jeng et al. 2012, Enikeeva & 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. (2011) 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 2 *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). 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 MV-CAPA 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*

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

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 $\frac{1}{2} < \xi \leq \frac{3}{4}$.

3 Inference for Multiple Anomalies

A natural way of extending the methodology introduced in Section 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

$$\sum_{k=1}^{\hat{K}} \mathcal{S}\left(\hat{s}_k, \hat{e}_k\right), \tag{5}$$

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 & Rigaill 2019). 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 we define $\mathcal{S}'\left(\mathbf{x}_t\right)$ to be the reduction in cost obtained by truncating the cost of a subset of the components of the observation $\mathbf{x}_t^{(i)}$. For example, if our anomalies correspond to changes in mean and we are using square error loss, or equivalently the Gaussian log-likelihood based cost, then

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

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

$$\sum_{k=1}^{\hat{K}} \mathcal{S}\left(\hat{s}_k, \hat{e}_k\right) + \sum_{t \in O} \mathcal{S}'\left(\mathbf{x}_t\right), \tag{6}$$

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.

The 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 of no point anomalies. When Tukey's biweight loss is used, the following proposition holds for any penalty β':

Proposition 3 *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 2.2.

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 (6), though a similar approach applies to optimising (5). 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 the dynamic programme is $O(Mn)$.

If no maximum segment length is specified, the computational cost 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. (2018) to reduced this computational cost. This is discussed in Section 1.1 of the Supplementary Material. 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; see Section 7.3.

5 Accuracy of Detecting Multiple 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. 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,

$$\boldsymbol{\mu}^{(i)}(t) = \boldsymbol{\mu}_k \quad \text{if} \quad s_k < t \leq e_k \quad \text{and} \quad i \in \mathbf{J}_k \tag{7}$$

and 0 otherwise.

Consider also the following choice of penalty.

$$\sum_{i=1}^{j} \boldsymbol{\beta}_i = \begin{cases} C\psi + Cj\log(p) & \text{if} \quad j \leq k^*, \\ p + C\psi + C\sqrt{p\psi} & \text{if} \quad j > k^*. \end{cases} \tag{8}$$

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 1 *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 $\tau = \{(\hat{s}_1, \hat{e}_1, \hat{\boldsymbol{J}}_1), ..., (\hat{s}_{\hat{K}}, \hat{e}_{\hat{K}}, \hat{\boldsymbol{J}}_{\hat{K}})\}$ obtained by applying MVCAPA using the penalty regime specified in (8) and no minimum segment length, on data $\boldsymbol{x}$ which follows the distribution specified in (7) satisfies*

$$\mathbb{P}\left(\hat{K} = K, \;\; |\hat{s}_k - s_k| < \frac{10C}{\triangle_k^2}, \;\; |\hat{e}_k - e_k| < \frac{10C}{\triangle_k^2}\right) > 1 - An^3 e^{-\psi}, \tag{9}$$

provided that, for $k = 1, ..., K$,

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

The result is proved in the Supplementary Material using combinatorial arguments. This finite sample result holds for a fixed C, which is independent of n, p, K, and/or $\triangle_k$.

Theorem 1 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 (9) 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

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

6 Incorporating Lags

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 (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

$$\boldsymbol{\theta}^{(i)}(t) = \boldsymbol{\theta}_k^{(i)} \quad \text{if} \quad s_k + \mathbf{d}_k^{(i)} < t \leq e_k - \mathbf{f}_k^{(i)} \quad \text{and} \quad i \in \mathbf{J}_k, \tag{10}$$

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 2 can easily be extended to incorporate lags. The only modification this requires is to re-define the saving $\mathcal{S}_i\left(s, e\right)$ to be

$$\max_{\substack{0 \leq \mathbf{d}^{(i)}, \mathbf{f}^{(i)} \leq w \\ e-s-\mathbf{d}^{(i)}-\mathbf{f}^{(i)} \geq l}} \left[\mathcal{C}_i \left(\mathbf{x}^{(i)}_{(s+1+\mathbf{d}^{(i)}):(e-\mathbf{f}^{(i)})}, \boldsymbol{\theta}_0^{(i)} \right) - \min_{\boldsymbol{\theta}} \left(\mathcal{C}_i \left(\mathbf{x}^{(i)}_{(s+1+\mathbf{d}^{(i)}):(e-\mathbf{f}^{(i)})}, \boldsymbol{\theta} \right) \right) \right], \tag{11}$$

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

$$\sum_{k=1}^{\hat{K}} \mathcal{S}\left(\hat{s}_k, \hat{e}_k\right) + \sum_{t \in O} \mathcal{S}'\left(\mathbf{x}_t\right), \tag{12}$$

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 1.2 of the Supplementary Material that if we use the penalty regimes from Section 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 show in Section 2.6 of the Supplementary Material that we can improve on this and use the following weaker penalty and still control the false positive probability.

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

An important question is how to choose the maximum lag length, w. In practice this needs to be guided by the application, and any knowledge about the degree to which common anomalies may be misaligned. In Section 2.6 in the Supplementary Material we provide theory which shows the gain in power that can be achieved by incorporating lags when collective anomalies are misaligned. However, too large a value of w may lead to a loss of power due to the inflation of the penalties that are needed, or, in extreme cases, may mean that separate anomalies are fit as a single mis-aligned anomaly. We investigate the choice of w empirically in Section 7.

6.2 Computational Considerations

The dynamic programming approach described in Section 4 can also be used to minimise the penalised negative saving in Equation (12). Solving the dynamic programme requires the computation of $\mathcal{S}_i\left(t, m\right)$ 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

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

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 $\mathcal{S}_i\left(t, m\right)$ 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 Supplementary Material.

7 Simulation Study

We now compare the performance of MVCAPA to that of other popular methods. In particular, we compare detection accuracy, precision, as well as the runtime with PASS (Jeng et al. 2012) and Inspect (Wang & 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 & 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 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 $k = p$.

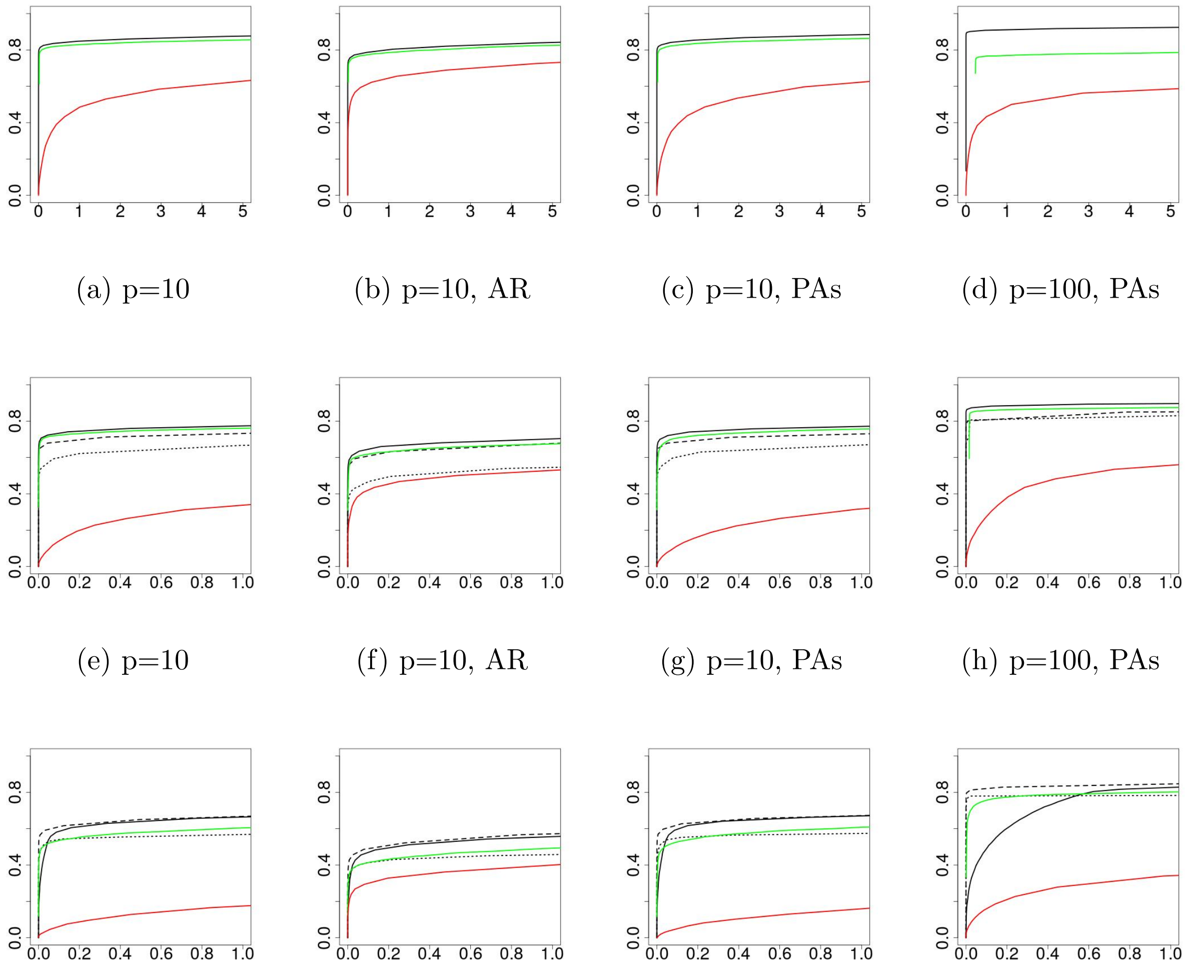
3. A regime close to the boundary between sparse and dense changes, i.e. $k = 2$ when $p = 10$ and $k = 6$ when $p = 100$ with $\sigma = \log(p)$ and $w = 0$.

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.

7.1 Detection Accuracy

All methods are sensitive to the choice of a threshold parameter that defines how much evidence of an anomaly or change there needs to be before one is flagged. To make our results robust to this choice, we investigate how each method performs as we vary its threshold parameter, and plot the proportion of anomalies detected against the number of false positives. 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, for MVCAPA, we used the composite penalty regime for $w = 0$ and penalty regime 2' for $w > 0$, and varied its threshold by re-scaling the penalty by a constant that is varied. 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



(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 3: Proportion of anomalies detected against the ratio of false positives to true positives for setting 1, 3, and 4 (top row to bottom row). MVCAPA is in black, PASS in green, and Inspect in red. The solid black 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.

to be a true 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. Qualitatively similar results were obtained as we vary the definition of a true positive: see Figure 8 in the Supplementary Material.

A subset of the results for three of the settings considered can be found in Figure 3. The full results for the four settings can be found in Figures 2 to 5 of the Supplementary Material. 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 3e and 3i, 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 3i and 3k shows that the performance of MVCAPA is hardly affected by the presence of point anomalies.

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

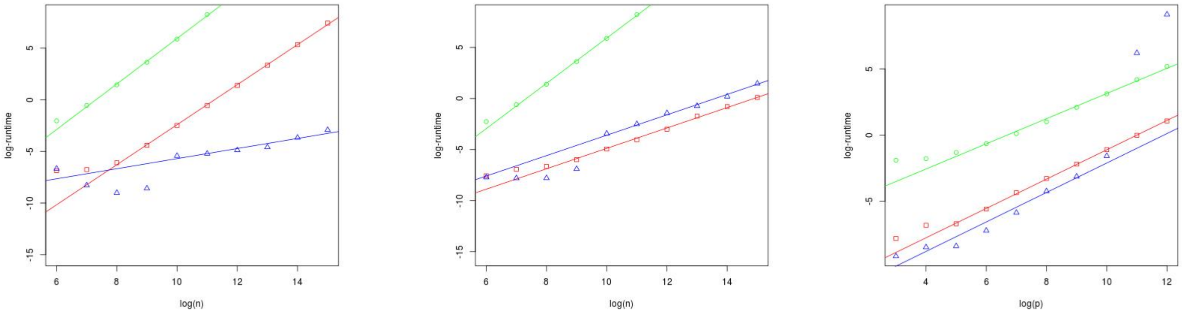
We have also considered the case in which collective anomalies are characterised by joint changes in mean and variance. Detection accuracy plots for that setting can be found in Figures 6 and 7 in the Supplementary Material.

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 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.

7.3 Runtime

We compared the scaling of the runtime of MVCAPA, PASS, and Inspect in both the number of observations n, as well as the number of components p. To evaluate the scaling in n we set $p = 10$ and varied n on data without any anomalies. We repeated this analysis with collective anomalies appearing (on average) every 100 observations. The results of these two analyses can be found in Figures 4a and 4b respectively. We note that the slope of MVCAPA is very close to 2, in the anomaly-free setting and very close to 1 in the setting in which the number of anomalies increases linearly with the number of



(a) No anomalies

(b) Regular anomalies

(c) Large p

Figure 4: A comparison of run-times for MVCAPA (red squared), PASS (green circles), and Inspect (blue triangles). Robust lines of best fit have been added. All logarithms are in base 2. Left-hand and middle plots are for $p = 10$ and varying n for data with, respectively, no collective anomalies and a collective anomaly every 100 observations on average. Right-hand plot is for $n = 100$ and varying p.

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 2: 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.

observations, suggesting quadratic and linear behaviour respectively, whilst the slopes of PASS and Inspect are close to 2 and 1 respectively in both cases.

Turning to the scaling of the three methods in p, we set $n = 100$ and varied p. The results of this analysis can be found in Figure 4c. We note that the slopes of all methods are close to 1 suggesting linear behaviour. However, Inspect becomes very slow once p exceeds a certain threshold.

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 (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 & Fearnhead (2017) to compare MVCAPA with PASS. We 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 methods by how accurately they detect known CNVs for a given test size. We used known CNVs from the HapMap project (International HapMap Consortium 2003) as true positives and tuned the penalties and thresholds so that 4% of the genome was flagged as anomalous for all methods.

The results of this analysis on Chromosome 16 can be found in Table 2 while the results for Chromosome 6 can be found in Table 10 in the Supplementary Material. These tables show that MVCAPA shows much more consistency across folds than PASS. We also see that allowing for lags generally led to a better performance, 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 took 17 minutes.

9 Discussion

Although we have focused on anomalies that relate to changes in mean, by choosing a cost based on an appropriate model the framework can be applied to detecting a variety of types of change. Moreover one can use different likelihood-based costs for different components and thus detect collective anomalies for data with a mix of data types. It is also possible to use a similar penalised cost to detect changepoints in multivariate data, and this extension has been considered in Tickle et al. (2020).

We have presented theory that indicates how to choose appropriate penalties for the method. However this theory relies on various assumptions, for example that the data is independent across series and across time. Results for other penalised cost procedures (Lavielle & Moulines 2000) suggest that one way of accounting for these assumptions not holding is to increase the penalties. We suggest doing this by taking the penalties suggested by theory and scaling them all by a constant. The choice of constant can be made in a data-driven way through analysis of performance on test data (Hocking et al. 2013), or by comparing changes the fit as we vary the constant (Haynes, Eckley & Fearnhead 2017). Alternatively we can model the dependencies, see Tveten et al. (2020) for an extension of MVCAPA to allow for correlation in the data across components.

MVCAPA does not quantify uncertainty about estimates of individual collective anomalies. This is similar to other epidemic changepoint and changepoint methods that output a single estimate of their locations. Recent work (Hyun et al. 2021, Jewell et al. 2019) has shown how to obtain valid p-values that quantify the uncertainty that each estimated change is close to an actual change; the output of these methods can then be used to give, e.g., control of the false discovery rate for the estimated changes. These ideas should be able to be extended to cover the collective anomaly problem, and hence quantify uncertainty of each collective anomaly detected by MVCAPA.

References

Bardwell, L. & Fearnhead, P. (2017), 'Bayesian detection of abnormal segments in multiple time series', *Bayesian Analysis* **12**(1), 193–218.

Boucheron, S. & Thomas, M. (2012), 'Concentration inequalities for order statistics', *Electronic Communications in Probability* **17**(51), 1–12.

Cai, T., Jeng, J. & Jin, J. (2011), 'Optimal detection of heterogeneous and heteroscedastic mixtures', *Journal of the Royal Statistical Society: Series B* **73**(5), 629–662.

Chandola, V., Banerjee, A. & Kumar, V. (2009), 'Anomaly detection: A survey', *ACM computing surveys (CSUR)* **41**(3), 15.

Cho, H. & Fryzlewicz, P. (2015), 'Multiple-change-point detection for high dimensional time series via sparsified binary segmentation', *Journal of the Royal Statistical Society: Series B* **77**(2), 475–507.

Cui, Y., Bangalore, P. & Tjernberg, L. B. (2018), An anomaly detection approach based on machine learning and scada data for condition monitoring of wind turbines, *in* '2018 IEEE International Conference on Probabilistic Methods Applied to Power Systems (PMAPS)', IEEE, pp. 1–6.

Davis, R. A., Lee, T. C. M. & Rodriguez-Yam, G. A. (2006), 'Structural break estimation for nonstationary time series models', *Journal of the American Statistical Association* **101**(473), 223–239.

Enikeeva, F. & Harchaoui, Z. (2019), 'High-dimensional change-point detection under sparse alternatives', *The Annals of Statistics* **47**(4), 2051–2079.

Fearnhead, P. & Rigaill, G. (2019), 'Changepoint detection in the presence of outliers', *Journal of the American Statistical Association* **114**(525), 169–183.

Fisch, A., Eckley, I. A. & Fearnhead, P. (2018), 'A linear time method for the detection of point and collective anomalies'. arXiv:1806.01947.

Fryzlewicz, P. (2014), 'Wild binary segmentation for multiple change-point detection', *The Annals of Statistics* **42**(6), 2243–2281.

Haynes, K., Eckley, I. A. & Fearnhead, P. (2017), 'Computationally efficient changepoint detection for a range of penalties', *Journal of Computational and Graphical Statistics* **26**(1), 134–143.

Haynes, K., Fearnhead, P. & Eckley, I. A. (2017), 'A computationally efficient nonparametric approach for changepoint detection', *Statistics and Computing* **27**(5), 1293–1305.

Hocking, T., Rigaill, G. & Bourque, G. (2015), Peakseg: constrained optimal segmentation and supervised penalty learning for peak detection in count data, *in* 'International Conference on Machine Learning', pp. 324–332.

Hocking, T., Rigaill, G., Vert, J.-P. & Bach, F. (2013), Learning sparse penalties for change-point detection using max margin interval regression, *in* 'International Conference on Machine Learning', pp. 172–180.

Hyun, S., Lin, K. Z., G'Sell, M. & Tibshirani, R. J. (2021), 'Post-selection inference for changepoint detection algorithms with application to copy number variation data', *Biometrics* .

International HapMap Consortium (2003), 'The international HapMap project', *Nature* **426**(6968), 789.

Jeng, X. J., Cai, T. T. & Li, H. (2010), 'Optimal sparse segment identification with application in copy number variation analysis', *Journal of the American Statistical Association* **105**(491), 1156–1166.

Jeng, X. J., Cai, T. T. & Li, H. (2012), 'Simultaneous discovery of rare and common segment variants', *Biometrika* **100**(1), 157–172.

Jewell, S., Fearnhead, P. & Witten, D. (2019), 'Testing for a change in mean after change-point detection'. arxiv.1910.04291.

Killick, R., Fearnhead, P. & Eckley, I. A. (2012), 'Optimal detection of changepoints with a linear computational cost', *Journal of the American Statistical Association* **107**(500), 1590–1598.

Laurent, B. & Massart, P. (2000), ‘Adaptive estimation of a quadratic functional by model selection’, *Annals of Statistics* **28**(5), 1302–1338.

Lavielle, M. & Moulines, E. (2000), ‘Least-squares estimation of an unknown number of shifts in a time series’, *Journal of Time Series Analysis* **21**(1), 33–59.

Levin, B. & Kline, J. (1985), ‘The cusum test of homogeneity with an application in spontaneous abortion epidemiology’, *Statistics in Medicine* **4**(4), 469–488.

Lin, C.-F., Naj, A. C. & Wang, L.-S. (2013), ‘Analyzing copy number variation using SNP array data: protocols for calling CNV and association tests’, *Current Protocols in Human Genetics* **79**(1), 1–27.

Metelli, S. & Heard, N. (2019), ‘On Bayesian new edge prediction and anomaly detection in computer networks’, *Annals of Applied Statistics* **13**(4), 2586–2610.

Olshen, A. B., Venkatraman, E. S., Lucito, R. & Wigler, M. (2004), ‘Circular binary segmentation for the analysis of array-based DNA copy number data’, *Biostatistics* **5**(4), 557–572.

Pimentel, M. A. F., Clifton, D. A., Clifton, L. & Tarassenko, L. (2014), ‘A review of novelty detection’, *Signal Processing* **99**, 215–249.

Rousseeuw, P. J. & Bossche, W. V. D. (2018), ‘Detecting deviating data cells’, *Technometrics* **60**(2), 135–145.

Smyth, P. (1994), ‘Markov monitoring with unknown states’, *IEEE Journal on Selected Areas in Communications* **12**(9), 1600–1612.

Talagala, P. D., Hyndman, R. J. & Smith-Miles, K. (2021), 'Anomaly detection in high-dimensional data', *Journal of Computational and Graphical Statistics* **30**(2), 360–375.

Talagala, P. D., Hyndman, R. J., Smith-Miles, K., Kandanaarachchi, S. & Muñoz, M. A. (2020), 'Anomaly detection in streaming nonstationary temporal data', *Journal of Computational and Graphical Statistics* **29**(1), 13–27.

Tickle, S., Eckley, I. & Fearnhead, P. (2020), 'A computationally efficient, high-dimensional multiple changepoint procedure with application to global terrorism incidence'. arXiv:2011.03599.

Tveten, M., Eckley, I. A. & Fearnhead, P. (2020), 'Scalable changepoint and anomaly detection in cross-correlated data with an application to condition monitoring'. arXiv:2010.06937.

Wang, T. & Samworth, R. J. (2016), 'Inspectchangepoint: high-dimensional changepoint estimation via sparse projection', *R Package Version* **1**.

Wang, T. & Samworth, R. J. (2018), 'High dimensional change point estimation via sparse projection', *Journal of the Royal Statistical Society: Series B* **80**(1), 57–83.

Zhang, N. R., Siegmund, D. O., Ji, H. & Li, J. Z. (2010), 'Detecting simultaneous changepoints in multiple sequences', *Biometrika* **97**(3), 631–645.

Subset Multivariate Collective And Point Anomaly Detection: Supplementary Material

Alexander T. M. Fisch Idris A. Eckley and Paul Fearnhead

Department of Mathematics and Statistics, Lancaster University

1 Additional Theoretical Results

1.1 Pruning Without Lags

The following proposition holds:

Proposition 4 *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

$$S(m) - \alpha - \sum_{1}^{p} \boldsymbol{\beta}_i > S(t) + \mathcal{S}\left(t, m\right)$$

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.

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 5 *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 $\mathcal{S}_i\left(s,e\right)$ as defined in (11) *satisfies*

$$\mathbb{P}\left(\mathcal{S}_i\left(s,e\right) > x\right) \leq (w+1)^2 \mathbb{P}\left(a\chi^2_v > x\right).$$

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.

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. (2018). Indeed, the following generalisation of Proposition 4 holds:

Proposition 6 *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.

2 Proofs for Theorems and Propositions

2.1 Proof of Proposition 1

We will prove the existence of such a constant for each the three penalty regimes. The result follows from this.

2.1.1 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

$$\begin{aligned}
&\mathbb{P}\left(\sum_{c \in J_m} \mathcal{S}_c\left(s,e\right) < a\left(pv + 2\sqrt{pv\psi} + 2\psi\right), \quad \forall S_m \subset \{1,...,p\} : |J_m| = m, \;\; 1 \leq m \leq p\right) \\
&= \mathbb{P}\left(\sum_{c=1}^{p} \mathcal{S}_c\left(s,e\right) < a\left(pv + 2\sqrt{pv\psi} + 2\psi\right)\right) \\
&\geq \mathbb{P}\left(\chi^2_{pv} < pv + 2\psi + 2\sqrt{pv\psi}\right) \geq 1 - e^{-\psi},
\end{aligned}$$

where the first inequality follows from the stochastic bound on $\mathcal{S}_c\left(s,e\right)$ and the second inequality follows from the bounds on the chi-squared distribution in Laurent & Massart (2000). A Bonferroni correction over all possible pairs s, e then finishes the proof.

2.1.2 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

$$\begin{aligned}
&\mathbb{P}\left(\sum_{c \in S_m} Y_c < 2(1+\epsilon)a\psi + 2ma(1+\epsilon)\log(p), \quad \forall S_m \subset \{1,...,p\} : |S_m| = m, \;\; 1 \leq m \leq p\right) \\
&\geq \mathbb{P}\left(\sum_{i=1}^{p}\left(\frac{Y_i - 2(1+\epsilon)a\log(p)}{a}\right)^{+} < 2(1+\epsilon)\psi\right) \\
&\geq \mathbb{P}\left(\sum_{i=1}^{p}\left(Z_i - 2(1+\epsilon)\log(p)\right)^{+} < 2(1+\epsilon)\psi\right) \\
&\geq 1 - \mathbb{E}\left(e^{\lambda(Z_1 - 2(1+\epsilon)\log(p))^{+}}\right)^{p} e^{-2\lambda(1+\epsilon)\psi},
\end{aligned}$$

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 1 *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,

$$\begin{aligned}
&\mathbb{E}\left(e^{\lambda(Z_1 - 2(1+\epsilon)\log(p))^+}\right)^p e^{-\psi} \\
&= \left[\mathbb{P}\left(\chi^2_v < 2(1+\epsilon)\log(p)\right) + \frac{e^{-2(1+\epsilon)\lambda\log(p)}}{(1-2\lambda)^{\frac{v}{2}}}\mathbb{P}\left(\chi^2_v > 2(1+\epsilon)(1-2\lambda)\log(p)\right)\right]^p e^{-\psi} \\
&\leq \left[1 + \frac{e^{-2(1+\epsilon)\lambda\log(p)}}{(1-2\lambda)^{\frac{v}{2}}}\right]^p e^{-\psi} = \left[1 + \frac{1}{p}\left(\frac{1+\epsilon}{\epsilon}\right)^{\frac{v}{2}}\right]^p e^{-\psi} \leq \exp\left(\left(\frac{1+\epsilon}{\epsilon}\right)^{\frac{v}{2}}\right)e^{-\psi}.
\end{aligned}$$

A Bonferroni correction over all possible pairs s, e then finishes the proof.

2.1.3 Regime 3

Let $1 \leq s \leq e \leq n$. 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 $Z_{(1)} \geq ... \geq Z_{(p)}$ The probability that the segment (s,e) is not flagged up as anomalous is given by

$$\begin{aligned}
&\mathbb{P}\left(\sum_{i=1}^{m} Y_{(i)} < a\left(2(\psi + \log(p)) + mv + 2pc_m f(c_m) + 2\sqrt{(mv + 2pc_m f(c_m))(\psi + \log(p))}\right), \quad 1 \leq m \leq p\right) \\
&\geq 1 - \sum_{m=1}^{p}\mathbb{P}\left(\sum_{i=1}^{m}\left(\frac{Y_{(i)} - ac_m}{a}\right) > 2(\psi + \log(p)) + m(v - c_m) + 2pc_m f(c_m) + 2\sqrt{(mv + 2pc_m f(c_m))(\psi + \log(p))}\right) \\
&\geq 1 - \sum_{m=1}^{p}\mathbb{P}\left(\sum_{i=1}^{m}\left(\frac{Y_{(i)} - ac_m}{a}\right)^+ > 2(\psi + \log(p)) + m(v - c_m) + 2pc_m f(c_m) + 2\sqrt{(vm + 2pc_m f(c_m))(\psi + \log(p))}\right) \\
&\geq 1 - \sum_{m=1}^{p}\mathbb{P}\left(\sum_{i=1}^{p}\left(\frac{Y_i - ac_m}{a}\right)^+ < 2(\psi + \log(p)) + m(v - c_m) + 2pc_m f(c_m) + 2\sqrt{(vm + 2pc_m f(c_m))(\psi + \log(p))}\right). \\
&\geq 1 - \sum_{m=1}^{p}\mathbb{P}\left(\sum_{i=1}^{p}(Z_i - c_m)^+ < 2(\psi + \log(p)) + m(v - c_m) + 2pc_m f(c_m) + 2\sqrt{(mv + 2pc_m f(c_m))(\psi + \log(p))}\right).
\end{aligned}$$

We will now use the following lemma, which shows that $(Z_c - c_m)^+$ is sub-gamma.

Lemma 2 *Let $Z \sim \chi^2_v$ for $v \leq 2$. Then $(Z-c)^+ - [2cf(c) + (v-c)\mathbb{P}\left(\chi^2_v > c\right)]$ is sub-gamma with scale parameter 2 and variance $V = 4cf(c) + 2v\mathbb{P}\left(\chi^2_v > c\right)$.*

Using Lemma 2 and the bounds on sub-gamma random-variables in Boucheron & Thomas (2012), we have that

$$
\begin{aligned}
&\sum_{m=1}^{p} \mathbb{P}\left(\sum_{i=1}^{p} (Z_i - c_m)^+ < 2(\psi + \log(p)) + m(v - c_m) + 2pc_m f(c_m) + 2\sqrt{(mv + 2pc_m f(c_m))(\psi + \log(p))}\right) \\
&= \sum_{m=1}^{p} \mathbb{P}\left(\sum_{i=1}^{p} \Big((Z_i - c_m)^+ - (v - c_m)\mathbb{P}\left(\chi^2_v > c_m\right) + 2c_m f(c_m)\Big) < 2(\psi + \log(p)) + 2\sqrt{pv\left(\mathbb{P}\left(\chi^2_v > c_m\right) + 2c_m f(c_m)\right)\left(\psi + \log(p)\right)}\right) \\
&\leq \sum_{m=1}^{p} p^{-1} e^{-\psi} = e^{-\psi}.
\end{aligned}
$$

A Bonferroni correction over all possible pairs s, e then finishes the proof.

2.2 Proof of Proposition 2

Proof of Proposition 2: 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

$$\mathbb{P}\left(\exists m : \sum_{i=1}^{m} X_{(i)} \geq \alpha + \sum_{i=1}^{m} \boldsymbol{\beta}_i\right) \geq \mathbb{P}\left(\sum_{i=1}^{p} X_i > p + 2\sqrt{p\psi} + 2\psi\right) = \mathbb{P}\left(\sum_{i=1}^{p} \epsilon_i^2 + \sum_{i=1}^{p} v_i \left(2\mu\epsilon_i + \mu^2\right) > p + 2\sqrt{p\psi} + 2\psi\right)$$

$$\geq 1 - \mathbb{P}\left(\sum_{i=1}^{p} \epsilon_i^2 < p - 2\sqrt{p\psi}\right) - \mathbb{P}\left(\sum_{i=1}^{p} v_i \left(2\mu\epsilon_i + \mu^2\right) < 4\sqrt{p\psi} + 2\psi\right)$$

$$\geq 1 - e^{-\psi} - \mathbb{P}\left(N\left(\mu^2\left(\sum_{i=1}^{p} v_i\right), 4\mu^2\left(\sum_{i=1}^{p} v_i\right)\right) < 4\sqrt{p\psi} + 2\psi\right)$$

Furthermore,

$$\mathbb{P}\left(N\left(\mu^2\left(\sum_{i=1}^{p} v_i\right), 4\mu^2\left(\sum_{i=1}^{p} v_i\right)\right) > 4\sqrt{p\psi} + 2\psi\right) > \mathbb{P}\left(N\left(k\mu^2, 4k\mu^2\right) > 4\sqrt{p\psi} + 2\psi\right)\mathbb{P}\left(\sum_{i=1}^{p} v_i > k\right),$$

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

$$\mathbb{P}\left(\sum_{i=1}^{p} v_i > \frac{1}{2}p^{1-\xi}\right) \to 1 \quad as \quad p \to \infty.$$

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

$$1 - (1 - \mathbb{P}(X_1 > 2\psi + 2\log(p), v_1 = 1))^p = 1 - \left(1 - p^{-\xi}\mathbb{P}\left(\left(\epsilon_1 + \sqrt{2r_p\log(p)}\right)^2 > 2\psi + 2\log(p)\right)\right)^p$$
$$> 1 - \left(1 - p^{-\xi}\mathbb{P}\left(N(0,1) > \sqrt{2\psi + 2\log(p)} - \sqrt{2r_p\log(p)}\right)\right)^p$$
$$\geq 1 - \exp\left(-p^{1-\xi}\mathbb{P}\left(N(0,1) > \sqrt{2\psi + 2\log(p)} - \sqrt{2r_p\log(p)}\right)\right),$$

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 $\mathbb{P}(N(0,1) > x) > \frac{1}{\sqrt{2\pi}}\frac{x}{x^2+1}\exp\left(-\frac{x^2}{2}\right)$, 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:

$$Y_i = \begin{cases} [(\mu + \epsilon_i)^+]^2 & \text{if} \quad v_i = 1 \\ \epsilon_i^2 & \text{if} \quad v_i = 0, \end{cases}$$

where $(x)^+$ denotes the positive part of x. Note that $Y_i \leq X_i$. We also introduce the following four technical lemmata

Lemma 3 *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 4 *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 5 *Let a_k be defined implicitly as $\mathbb{P}\left(\chi_1^2 > a_k\right) = \frac{k}{p}$ and let $f(\cdot)$ denote the probability density function of the χ_1^2 distribution. Then*

$$p\mathbb{E}\left((\chi_1^2 - a_k)^+\right) + ka_k = k + 2pa_kf(a_k) \leq 2k + 2k\log(p/k)$$

Lemma 6 *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 $p\mathbb{P}\left(\chi_1^2 > b\right) \leq \tilde{k} \leq p$ 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 $\tilde{\mu} = \mathbb{E}\left((\chi_1^2 - a_{\tilde{k}})^+\right)$. The following holds:

$$\mathbb{P}\left(\exists m : \sum_{i=1}^{m} X_{(i)} \geq \alpha + \sum_{i=1}^{m}\boldsymbol{\beta}_i\right) \geq \mathbb{P}\left(\sum_{i=1}^{\tilde{k}} Y_{(i)} \geq 2\psi + 2\log(p) + \tilde{k}a_{\tilde{k}} + p\tilde{\mu} + 2\sqrt{(\tilde{k}a_{\tilde{k}} + p\tilde{\mu})(\psi + 2\log(p))}\right)$$

$$\geq \mathbb{P}\left(\sum_{i=1}^{\tilde{k}} Y_{(i)} \geq 2\psi + 2\log(p) + \tilde{k}a_{\tilde{k}} + p\tilde{\mu} + 2\sqrt{(\tilde{k}a_{\tilde{k}} + p\tilde{\mu})(\psi + 2\log(p))}\middle| \sum_{i=1}^{p}\mathbb{I}\left(Y_i > b\right) \geq \tilde{k}\right)\mathbb{P}\left(\sum_{i=1}^{p}\mathbb{I}\left(Y_i > b\right) \geq \tilde{k}\right),$$

where the first inequality follows from substituting the third penalty regime (using the equality from Lemma 5) and the second inequality follows from conditioning on the number of Y_i exceeding b. Next note that

$$\mathbb{P}\left(\sum_{i=1}^{\tilde{k}} Y_{(i)} \geq 2\psi + 2\log(p) + \tilde{k}a_{\tilde{k}} + p\tilde{\mu} + 2\sqrt{(\tilde{k}a_{\tilde{k}} + p\tilde{\mu})(\psi + \log(p))}\,\middle|\, \sum_{i=1}^{p} \mathbb{I}\left(Y_i > b\right) \geq \tilde{k}\right)$$

$$\geq \mathbb{P}\left(\sum_{i=1}^{\tilde{k}} Y_{(i)} \geq 2\psi + 2\log(p) + \tilde{k}a_{\tilde{k}} + p\tilde{\mu} + 2\sqrt{(\tilde{k}a_{\tilde{k}} + p\tilde{\mu})(\psi + \log(p))}\,\middle|\, \sum_{i=1}^{p} \mathbb{I}\left(Y_i > b\right) = \tilde{k}\right)$$

$$= \mathbb{P}\left(\sum_{i=1}^{\tilde{k}} (Y_i - b)^+ \geq 2\psi + 2\log(p) - \tilde{k}\left(b - a_{\tilde{k}}\right) + p\tilde{\mu} + 2\sqrt{(\tilde{k}a_{\tilde{k}} + p\tilde{\mu})(\psi + \log(p))}\,\middle|\, Y_1, ..., Y_{\tilde{k}} > b\right)$$

Let $Z_1, ..., Z_{\tilde{k}}$ be i.i.d. χ_1^2 distributed. Lemma 3 then implies that the above exceeds

$$\mathbb{P}\left(\sum_{i=1}^{\tilde{k}} (Z_i - b)^+ \geq 2\psi + 2\log(p) - \tilde{k}\left(b - a_{\tilde{k}}\right) + p\tilde{\mu} + 2\sqrt{(\tilde{k}a_{\tilde{k}} + p\tilde{\mu})(\psi + \log(p))}\,\middle|\, Z_1, ..., Z_{\tilde{k}} > b\right).$$

Using the inequality in Lemma 5 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

$$\mathbb{P}\left(\sum_{i=1}^{\tilde{k}} W_i \geq p\tilde{\mu} - \tilde{k}\left(b - a_{\tilde{k}}\right) + 8\sqrt{\tilde{k}\log(p)^2}\right), \tag{13}$$

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 (13) 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 4. Given Lemma 6 and the fact that Lemma 5 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.

2.3 Proof of Proposition 3

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

2.4 Proof of Propositions 4 and 6

We give the proof of Proposition 6, as Proposition 4 corresponds to as special case. We write

$$\mathcal{S}\left(s,e,\mathbf{d},\mathbf{f},\mathbf{J}\right) = \sum_{i\in\mathbf{J}}\left(\mathcal{C}_i\left(\mathbf{x}^{(i)}_{(s+1+\mathbf{d}^{(i)}):(e-\mathbf{f}^{(i)})}, \boldsymbol{\theta}^{(i)}_0\right) - \min_{\boldsymbol{\theta}}\left[\mathcal{C}_i\left(\mathbf{x}^{(i)}_{(s+1+\mathbf{d}^{(i)}):(e-\mathbf{f}^{(i)})}, \boldsymbol{\theta}\right)\right]\right) - \alpha - \sum_{i=1}^{|\mathbf{J}|}\boldsymbol{\beta}_i$$

and note that

$$\mathcal{S}\left(t,m\right) = \max_{\mathbf{d},\mathbf{f},\mathbf{J}:\ m-t-d-f\geq l}\left[\mathcal{S}\left(t,m,\mathbf{d},\mathbf{f},\mathbf{J}\right)\right]$$

The proof of Proposition 6 is then a corollary of the observation that for all $\mathbf{d}, \mathbf{f} < w$

$$
\begin{aligned}
\mathcal{S}\left(t, m', \mathbf{d}, \mathbf{f}, \mathbf{J}\right) &= \sum_{i \in \mathbf{J}} \left(\mathcal{C}_i \left(\mathbf{x}^{(i)}_{(t+1+\mathbf{d}^{(i)}):(m'-\mathbf{f}^{(i)})}, \boldsymbol{\theta}^{(i)}_0 \right) - \min_{\boldsymbol{\theta}} \left[\mathcal{C}_i \left(\mathbf{x}^{(i)}_{(t+1+\mathbf{d}^{(i)}):(m'-\mathbf{f}^{(i)})}, \boldsymbol{\theta} \right) \right] \right) - \alpha - \sum_{i=1}^{|\mathbf{J}|} \boldsymbol{\beta}_i \\
&\leq \sum_{i \in \mathbf{J}} \left(\mathcal{C}_i \left(\mathbf{x}^{(i)}_{(t+1+\mathbf{d}^{(i)}):(m'-\mathbf{f}^{(i)})}, \boldsymbol{\theta}^{(i)}_0 \right) - \min_{\boldsymbol{\theta}} \left[\mathcal{C}_i \left(\mathbf{x}^{(i)}_{(t+1+\mathbf{d}^{(i)}):m}, \boldsymbol{\theta} \right) \right] - \min_{\boldsymbol{\theta}} \left[\mathcal{C}_i \left(\mathbf{x}^{(i)}_{(m+1):(m'-\mathbf{f}^{(i)})}, \boldsymbol{\theta} \right) \right] \right) \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad - \alpha - 2 \sum_{i=1}^{|\mathbf{J}|} \boldsymbol{\beta}_i + \sum_{i=1}^{p} \boldsymbol{\beta}_i \\
&= \mathcal{S}\left(t, m, \mathbf{d}, 0, \mathbf{J}\right) + \mathcal{S}\left(m, m', 0, \mathbf{f}, \mathbf{J}\right) + \sum_{i=1}^{p} \boldsymbol{\beta}_i + \alpha,
\end{aligned}
$$

since $m - t - w \geq l$ and $m' - m - w \geq l$. Indeed, the above inequality shows that

$$
\begin{aligned}
&S(t) + \max_{\mathbf{d} \leq w, \mathbf{f} \leq w, \mathbf{J}:\ m-t-d-f \geq l} \left[\mathcal{S}\left(t, m', \mathbf{d}, \mathbf{f}, \mathbf{J}\right) \right] \\
&\leq S(t) + \max_{\mathbf{d} \leq w, \mathbf{f} \leq w, \mathbf{J}:\ m-t-d-f \geq l} \left[\mathcal{S}\left(t, m, \mathbf{d}, \mathbf{0}, \mathbf{J}\right) + \mathcal{S}\left(m, m', \mathbf{0}, \mathbf{f}, \mathbf{J}\right) \right] + \alpha + \sum_{i=1}^{p} \boldsymbol{\beta}_i \\
&\leq S(t) + \max_{\mathbf{d} \leq w, \mathbf{J}:\ m-t-d \geq l} \left[\mathcal{S}\left(t, m, \mathbf{d}, \mathbf{0}, \mathbf{J}\right) \right] + \max_{\mathbf{f} \leq w, \mathbf{J}:\ m-t-f \geq l} \left[\mathcal{S}\left(m, m', \mathbf{0}, \mathbf{f}, \mathbf{J}\right) \right] + \alpha + \sum_{i=1}^{p} \boldsymbol{\beta}_i \\
&\leq S(t) + \max_{\mathbf{d} \leq w, \mathbf{f} \leq w, \mathbf{J}:\ m-t-d-f \geq l} \left[\mathcal{S}\left(t, m, \mathbf{d}, \mathbf{f}, \mathbf{J}\right) \right] + \max_{\mathbf{d} \leq w, \mathbf{f} \leq w, \mathbf{J}:\ m-t-d-f \geq l} \left[\mathcal{S}\left(m, m', \mathbf{d}, \mathbf{f}, \mathbf{J}\right) \right] + \alpha + \sum_{i=1}^{p} \boldsymbol{\beta}_i \\
&< S(m) + \max_{\mathbf{d}, \mathbf{f}, \mathbf{J}:\ m-t-d-f \geq l} \left[\mathcal{S}\left(m, m', \mathbf{d}, \mathbf{f}, \mathbf{J}\right) \right] \leq S(m').
\end{aligned}
$$

This finishes the proof.

2.5 Proof of Proposition 5

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:

$$\begin{aligned}\mathbb{P}\left(\mathcal{S}_i\left(s,e\right)>y\right) &\leq \sum_{d=0}^{w}\sum_{f=0}^{w}\mathbb{P}\left(\mathcal{C}_i\left(\mathbf{x}^{(i)}_{(s+1+d):(e-f}, \boldsymbol{\theta}_0^{(i)}\right)-\min_{\boldsymbol{\theta}}\left(\mathcal{C}_i\left(\mathbf{x}^{(i)}_{(s+1+d):(e-f)}, \boldsymbol{\theta}\right)\right)>y\right)\\ &\leq \sum_{d=0}^{w}\sum_{f=0}^{w}\mathbb{P}\left(a\chi^2_v+b>y\right)=(w+1)^2\mathbb{P}\left(a\chi^2_v+b>y\right)\end{aligned}$$

2.6 Results when using Lags: Penalty Regime and Power

For anomalies which correspond to changes in mean, the following result shows that Penalty Regime 2' will control the false positive rate.

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 $\epsilon > 0$, $\mathcal{S}_i\left(s,e\right)$, as defined in (11) is stochastically bounded by*

$$(1+\epsilon)\chi_1^2+2(1+\epsilon)\left(\log\left(w+1\right)+\log(6)-\log(\log(1+\epsilon))+\log(1+\log(w+1))\right), \quad (14)$$

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

Proof: The proof follows almost directly from the following Lemma:

Lemma 7 *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

$$\max_{0\leq j,d\leq w: j-f-d-i\geq 0}\left((j-f-d-i+1)\left(\bar{\boldsymbol{\eta}}^{(c)}_{(i+d):(j-f)}\right)^2\right) \quad (15)$$

is stochastically bounded by

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

□

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 $e-s=w$.

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 2.3. Let

$$\mathbf{x}_t^{(i)} = v^{(i)}\mathbb{I}\left(s+d_i < t \leq e - f_i\right)\mu + \boldsymbol{\eta}_t^{(i)}, \quad v^{(i)} \sim \begin{cases} 0 & \text{with prob. } 1-p^{-\xi}, \\ 1 & \text{with prob. } p^{-\xi}, \end{cases} \tag{16}$$

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* (16) *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* $\log(n) = o(\log(p))$.

Proof: 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 2 since

$$\sqrt{e-s-w}\bar{\mathbf{x}}^{(i)}_{(s+1+d_i):(e-f_i)} \overset{i.i.d.}{\sim} N(\sqrt{2r_p\log(p(w+1))},1).$$

$\square$

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. (2011).

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 $p \to \infty$.

2.7 Proof of Theorem 1

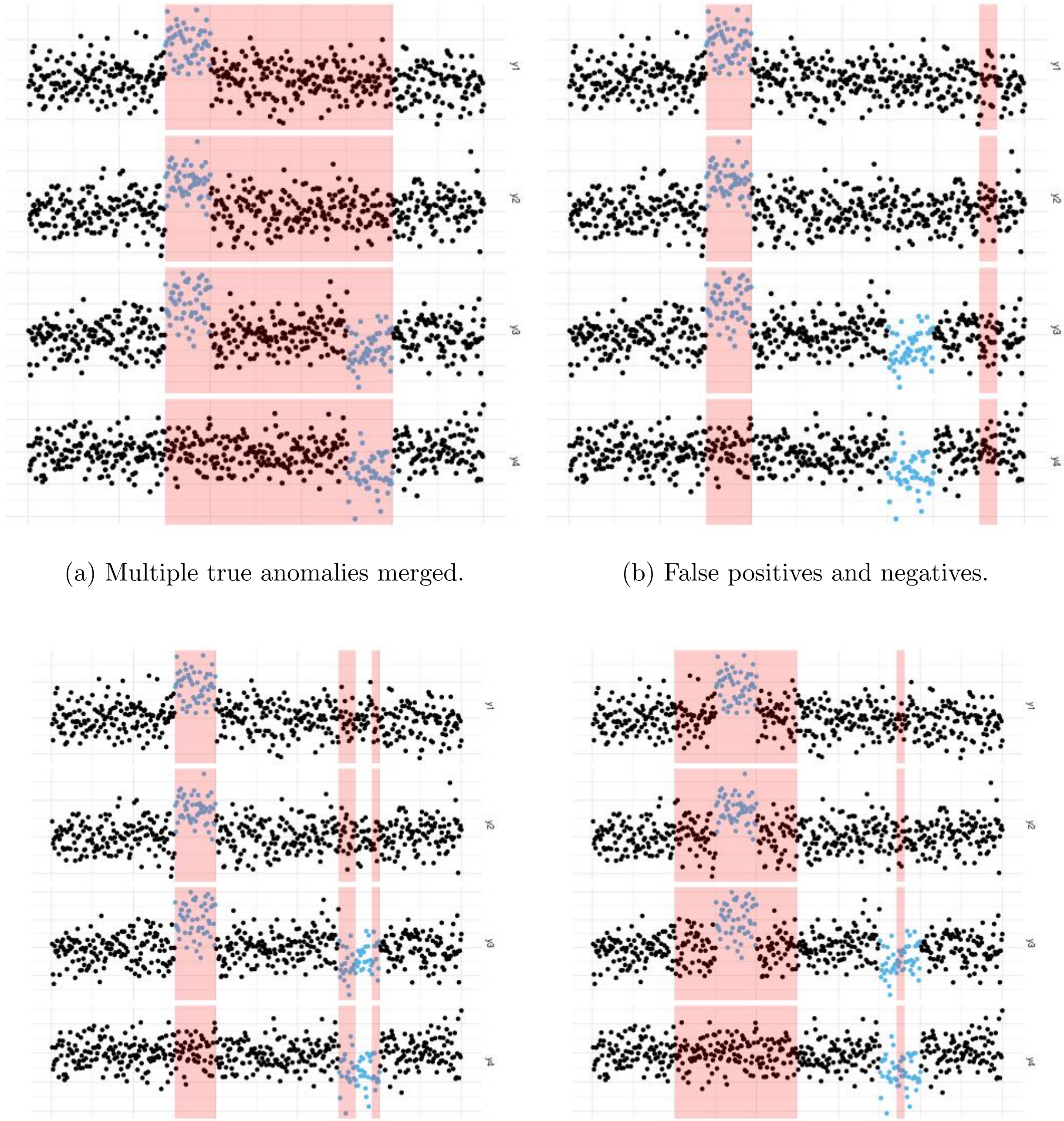
We define the penalised cost of a segment $\mathbf{x}_{i:j}$ under a partition $\tau = \{\hat{\tau}_1, ..., \hat{\tau}_{\hat{K}}\}$, where $\hat{\tau}_k = (\hat{s}_k, \hat{e}_k, \hat{\mathbf{J}}_k)$ to be

$$\mathcal{C}\left(\mathbf{x}_{i:j}, \hat{\tau}\right) = \sum_{k=1}^{\hat{K}} \left[\mathcal{C}(\mathbf{x}_{(\hat{s}_k+1):\hat{e}_k}, \hat{\mathbf{J}}_k)\right].$$

Here the penalised cost of introducing the kth anomalous window is

$$\mathcal{C}\left(\mathbf{x}_{(s+1):e}, \{(s,e,\mathbf{J})\}\right) = \mathcal{C}(\mathbf{x}_{(s+1):e}, \mathbf{J}) := -\mathcal{S}(\mathbf{x}_{(s+1):e}, \mathbf{J}) + \sum_{i=1}^{|\mathbf{J}|} \boldsymbol{\beta_i}. := -(e-s)\sum_{i\in\mathbf{J}} \mathcal{C}\left(\bar{\mathbf{x}}^{(i)}_{(s+1):e}\right)^2 + \sum_{i=1}^{|\mathbf{J}|} \boldsymbol{\beta_i},$$

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 $\bar{\mathbf{x}}^{(i)}_{(s+1):e} := (e-s)^{-1}\sum_{t=s+1}^{e} \mathbf{x}_t^i$ 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 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 (17). True anomalies are indicated in blue.

We also define the following event sets over all pairs i, j such that $1 \leq i \leq j \leq n$

$$\begin{aligned}
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\{ \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| \leq \sqrt{\sum_{c \in S} \sum_{t=i}^{j} \left(\boldsymbol{\mu}_t^{(c)} - \bar{\boldsymbol{\mu}}_{i:j}^{(c)} \right)^2} \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\{ \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} \quad \forall i,j : \exists k \in \{1,...,K\} : e_{k-1} < i, \ \ j \leq s_{k+1} \right\},
\end{aligned}$$

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 8 *There exists a constant A such that*

$$\mathbb{P}\left(E\right) > 1 - An^3 e^{-\psi}$$

We now define the set of good partitions $\mathcal{B}_C$ to be

$$\mathcal{B}_C = \left\{ \tau : |\tau| = K, \;\; |\hat{s}_k - s_k| \leq \frac{10C}{\triangle_k^2} \;\; |\hat{e}_k - e_k| \leq \frac{10C}{\triangle_k^2} \right\}. \tag{17}$$

It is sufficient to prove the following proposition in order to prove Theorem 1

Proposition 10 *Let the assumptions of Theorem 1 hold. Given E holds and C exceeds a global constant, the partition τ_0 minimising the penalised cost $\mathcal{C}\left(\boldsymbol{x}_{1:n}, \tau\right)$ satisfies $\tau_0 \in \mathcal{B}_C$*

The main ideas of the proof of Proposition 10 are that given E:

I Each fitted anomalous segment overlaps with at most one true anomalous segment; this excludes the situation depicted in Figure 1a.

II Each fitted anomalous segment overlaps with at least one true anomalous region; this excludes the situation depicted in Figure 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 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 1d.

We will prove these properties which exclude the various types of poor partitions in Figure 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 10. In the subsequent proofs we will use a certain number of technical Lemmata, all proved in Section 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 9 *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 10 *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 11 *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 12 *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 13 *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 14 *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 15 *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 16 *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 17 *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 18 *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 $|\boldsymbol{J}_k| \leq k^*$; 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 19 *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*

$$\mathcal{S}\left(\boldsymbol{x}_{i:j}, \boldsymbol{J}\right) \geq \frac{19}{20} C \left(|\boldsymbol{J}| \log(p) + \psi\right),$$

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 20 *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 kth 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 21 *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 22 *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.

2.7.1 Property III

We can prove that a fitted segment must overlap with at least one true anomalous segments:

Proposition 11 *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 11: 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.

2.7.2 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,

$$\mathcal{T}_1 = \{\tau : |\tau| = K \ \wedge (\forall (s,e,\mathbf{J}) \in \tau \ \ \exists k : s_{k+1} \geq e \quad \wedge \quad e_{k-1} \leq s \quad \wedge \quad [s+1,e] \cap [s_k+1, e_k] \neq \emptyset) \\ \wedge (\forall k \ \ \exists (s,e,\mathbf{J}) \in \tau : [s+1,e] \cap [s_k+1,e_k] \neq \emptyset)\}.$$

The following proposition then holds:

Proposition 12 *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 12: 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 12, 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 13 and 14 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 14 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 13, 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 $|\hat{\mathbf{J}}| \leq k^*$,

$$\begin{aligned}
&\mathcal{C}\left(\mathbf{x}_{(\hat{s}_k+1):\hat{e}_k},\hat{\mathbf{J}}\right) \geq \mathcal{C}\left(\mathbf{x}_{(\hat{s}_k+1):\hat{e}_k},\mathbf{J}_k\right) - \sum_{c\in\hat{\mathbf{J}}\setminus\mathbf{J}_k}(\hat{e}_k-\hat{s}_k)\left(\bar{\boldsymbol{\eta}}_{(\hat{s}_k+1):\hat{e}_k}\right)^2 + C(|\hat{\mathbf{J}}|-|\mathbf{J}_k|)\log(p)\\
&\geq \mathcal{C}\left(\mathbf{x}_{(\hat{s}_k+1):\left(e_k-\lceil\frac{10C}{\triangle_k^2}\rceil\right)},\mathbf{J}_k\right)+\mathcal{C}\left(\mathbf{x}_{\left(e_k-\lfloor\frac{10C}{\triangle_k^2}\rfloor\right):\hat{e}_k},\mathbf{J}_k\right) - 2C|\mathbf{J}_k|\log(p) - C\psi - 2\psi - 2|\hat{\mathbf{J}}\setminus\mathbf{J}_k|\log(p) + C|\hat{\mathbf{J}}|\log(p)\\
&\geq \mathcal{C}\left(\mathbf{x}_{(\hat{s}_k+1):\left(e_k-\lceil\frac{10C}{\triangle_k^2}\rceil\right)},\mathbf{J}_k\right)+\mathcal{C}\left(\mathbf{x}_{\left(e_k-\lfloor\frac{10C}{\triangle_k^2}\rfloor\right):e_k},\mathbf{J}_k\right) - 2C\psi - 2C|\mathbf{J}_k|\log(p),
\end{aligned}$$

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 13 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.

2.7.3 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

$$\mathcal{T}_2 = \{\tau : \forall (s, e, \mathbf{J}) \in \tau \;\; \exists k : s_{k+1} \geq e \quad \wedge \quad e_{k-1} \leq s \quad \wedge \quad [s+1, e] \cap [s_k + 1, e_k] \neq \emptyset\}.$$

and note that $\mathcal{T}_1 \subset \mathcal{T}_2$. The following proposition holds:

Proposition 13 *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 13: 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 12, 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 15, 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 16. If $|\mathbf{J}_1| \leq k^*$, we first note that Lemma 16 implies that

$$\mathcal{C}(x_{a':b}, \mathbf{J}_1) \leq \mathcal{C}(\mathbf{x}_{a:b}, \mathbf{J}_1) + 8\psi + 8|\mathbf{J}_1| \log(p) \leq \mathcal{C}(\mathbf{x}_{a:b}, \mathbf{J}_1) + 8\psi + 8\sqrt{p\psi} \tag{18}$$

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 19 shows that

$$\mathcal{C}(x_{a':b}, \mathbf{1}) \leq \mathcal{C}(x_{a':b}, \mathbf{J}_1) + \frac{1}{10} C \left(\psi + \sqrt{p\psi}\right). \tag{19}$$

Combining (18) and (19) 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 15, 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 16 and 20, 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 18 has to be used instead of Lemma 19.

2.7.4 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 14 *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 13. 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 15 *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*

$$\mathcal{C}\left(\boldsymbol{x}_{1:n}, \tau''\right) \leq \mathcal{C}\left(\boldsymbol{x}_{1:n}, \tau'\right) - \frac{6}{10} \left(\sum_{(s,e,\boldsymbol{J}) \in \tau' \cap \mathcal{A}_1} \left(C\psi + C|\boldsymbol{J}| \log(p)\right) + \sum_{(s,e,\boldsymbol{1}) \in \tau' \cap \mathcal{A}_2} \left(C\psi + C\sqrt{p\psi} \right) \right),$$

if C exceeds a global constant.

Proposition 16 *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*

$$\mathcal{C}\left(\boldsymbol{x}_{1:n}, \tau'\right) \leq \mathcal{C}\left(\boldsymbol{x}_{1:n}, \tau\right) + \frac{11}{20} \left(\sum_{(s,e,\boldsymbol{J}) \in \tau \cap \mathcal{A}_1} \left(C\psi + C|\boldsymbol{J}| \log(p)\right) + \sum_{(s,e,\boldsymbol{1}) \in \tau \cap \mathcal{A}_2} \left(C\psi + C\sqrt{p\psi} \right) \right),$$

with equality if and only if $\tau \in \mathcal{T}_2$, if C exceeds a global constant.

Note that Proposition 15 does not assume that τ' is optimal. Using these two propositions it is easy to derive the following:

Proof of Proposition 14: Assume that the optimal partition τ is such that $\tau \notin \mathcal{T}_2$. Then, by Proposition 16 there exists a partition $\tau' \in \mathcal{T}_2$ such that

$$\mathcal{C}\left(\mathbf{x}_{1:n}, \tau\right) > \mathcal{C}\left(\mathbf{x}_{1:n}, \tau'\right) - \frac{11}{20} \left(\sum_{(s,e,\mathbf{J}) \in \tau \cap \mathcal{A}_1} \left(C\psi + C|\mathbf{J}| \log(p)\right) + \sum_{(s,e,\mathbf{1}) \in \tau \cap \mathcal{A}_2} \left(C\psi + C\sqrt{p\psi} \right) \right),$$

Moreover, Proposition 15 implies that there exists another partition $\tau'' \in \mathcal{T}_2$ such that

$$\mathcal{C}\left(\mathbf{x}_{1:n}, \tau'\right) \geq \mathcal{C}\left(\mathbf{x}_{1:n}, \tau''\right) + \frac{6}{10}\left(\sum_{(s,e,\mathbf{J}) \in \tau' \cap \mathcal{A}_1} \left(C\psi + C|\mathbf{J}|\log(p)\right) + \sum_{(s,e,\mathbf{1}) \in \tau' \cap \mathcal{A}_2} \left(C\psi + C\sqrt{p\psi}\right)\right),$$

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 15: Proposition 13 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 13. 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

$$\tau'_k = \left\{(s, e, \mathbf{J}) \in \tau' : [s+1, e] \cap [s_k + 1, e_k] \neq \emptyset\right\}.$$

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 22 and 21 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 20 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 15 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 12 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 15 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 13. This finishes the proof.

Proof of Proposition 16: 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 $\frac{10C}{\triangle'^2_k}$ 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 14 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 23 *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 23. 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 23, E_9, and C exceeding a global constant. This finishes the proof.

2.7.5 Proof of Proposition 10

Proof of Proposition 10: Propositions 13, 11, 14, and 14 give the result.

3 Proofs for Lemmata

3.1 Proof of Lemma 1

The MGF of $Z = (X - c)^+$ is given by

$$\mathbb{E}\left(e^{\lambda Z}\right) = \mathbb{P}\left(\chi_v^2 < c\right) + \int_c^\infty e^{\lambda(x-c)}\frac{1}{\Gamma\left(\frac{v}{2}\right)2^{\frac{v}{2}}}x^{\frac{v}{2}-1}e^{-\frac{1}{2}x}dx = \mathbb{P}\left(\chi_v^2 < c\right) + \frac{e^{-\lambda c}}{\Gamma\left(\frac{v}{2}\right)2^{\frac{v}{2}}}\int_c^\infty x^{\frac{v}{2}-1}e^{-(1-2\lambda)x}dx$$

using the substitution $y = (1-2\lambda)x$ the above can be shown to be equal to

$$\mathbb{P}\left(\chi_v^2 < c\right) + \frac{e^{-\lambda a}}{\Gamma\left(\frac{v}{2}\right)2^{\frac{v}{2}}\left(1-2\lambda\right)^{\frac{v}{2}}}\int_c^\infty y^{\frac{v}{2}-1}e^{-y}dy = \mathbb{P}\left(\chi_v^2 < c\right) + \frac{e^{-\lambda c}}{(1-2\lambda)^{\frac{v}{2}}}\mathbb{P}\left(\chi_v^2 > c(1-2\lambda)\right).$$

3.2 Proof of Lemma 2

As shown by Lemma 1, 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

$$\frac{d}{d\lambda}\left(\log\left(\mathbb{E}\left(e^{\lambda Z}\right)\right)\right) - \mu = \frac{\frac{d}{d\lambda}\left(\mathbb{E}\left(e^{\lambda Z}\right)\right)}{\mathbb{E}\left(e^{\lambda Z}\right)} - \mu = \frac{\frac{2cf(c)}{1-2\lambda} + \left(\frac{v}{1-2\lambda} - c\right)\frac{e^{-\lambda c}}{(1-2\lambda)^{v/2}}\mathbb{P}\left(\chi_v^2 > c(1-2\lambda)\right)}{\mathbb{P}\left(\chi_v^2 < c\right) + \frac{e^{-\lambda c}}{(1-2\lambda)^{v/2}}\mathbb{P}\left(\chi_v^2 > c(1-2\lambda)\right)} - \mu$$

$$= \frac{1}{1-2\lambda}\left[\frac{2cf(c) + \left(v - (1-2\lambda)\,c\right)\frac{e^{-\lambda c}}{(1-2\lambda)^{v/2}}\mathbb{P}\left(\chi_v^2 > c(1-2\lambda)\right)}{\mathbb{P}\left(\chi_v^2 < c\right) + \frac{e^{-\lambda c}}{(1-2\lambda)^{v/2}}\mathbb{P}\left(\chi_v^2 > c(1-2\lambda)\right)} - (1-2\lambda)\,\mu\right]$$

$$= \frac{1}{1-2\lambda}\left[\frac{2cf(c) - (v-c)\,\mathbb{P}\left(\chi_v^2 < c\right) - 2\lambda c\mathbb{P}\left(\chi_v^2 < c\right)}{\mathbb{P}\left(\chi_v^2 < c\right) + \frac{e^{-\lambda c}}{(1-2\lambda)^{v/2}}\mathbb{P}\left(\chi_v^2 > c(1-2\lambda)\right)} + \left(v - (1-2\lambda)\,c\right) - (1-2\lambda)\,\mu\right]$$

$$= \frac{1}{1-2\lambda}\left[\frac{\mu - (v-c) - 2\lambda c\mathbb{P}\left(\chi_v^2 < c\right)}{\mathbb{P}\left(\chi_v^2 < c\right) + \frac{e^{-\lambda c}}{(1-2\lambda)^{v/2}}\mathbb{P}\left(\chi_v^2 > c(1-2\lambda)\right)} + (v - c - \mu) + 2(\mu + c)\lambda\right].$$

Next note that

$$\mathbb{P}\left(\chi_v^2 > c(1-2\lambda)\right) = \int_{c(1-2\lambda)}^{\infty}\frac{1}{2^{\frac{v}{2}}\Gamma\left(\frac{v}{2}\right)}x^{\frac{v}{2}-1}e^{-x/2}dx = \frac{1}{2^{\frac{v}{2}}\Gamma\left(\frac{v}{2}\right)}e^{\lambda c}\int_{c}^{\infty}\left(\frac{y}{y-2\lambda c}\right)^{1-\frac{v}{2}}y^{\frac{v}{2}-1}e^{-y/2}dx.$$

When $v \leq 2$, this shows that:

$$\mathbb{P}\left(\chi_v^2 > c\right) < e^{-\lambda c}\mathbb{P}\left(\chi_v^2 > c(1-2\lambda)\right) < \frac{\mathbb{P}\left(\chi_v^2 > c\right)}{(1-2\lambda)^{1-\frac{v}{2}}}. \tag{20}$$

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 20 shows that $\frac{d}{d\lambda}\left(\log\left(\mathbb{E}\left(e^{\lambda Z}\right)\right)\right) - \mu$ is bounded by

$$\frac{1}{1-2\lambda}\left[\frac{\mu-(v-c)-2\lambda c\mathbb{P}\left(\chi_v^2<c\right)}{\mathbb{P}\left(\chi_v^2<c\right)+\frac{1}{(1-2\lambda)^{\frac{v}{2}}}\mathbb{P}\left(\chi_v^2>c\right)}+(v-c-\mu)+2(\mu+c)\lambda\right]$$
$$\leq \frac{1}{1-2\lambda}\left[\mu-(v-c)-2\lambda c\mathbb{P}\left(\chi_v^2<c\right)+(v-c-\mu)+2(\mu+c)\lambda\right] = \frac{1}{1-2\lambda}\left[2(\mu+c\mathbb{P}\left(\chi_v^2>c\right))\lambda\right]$$
$$\leq \frac{2\lambda(1-\lambda)}{(1-2\lambda)^2}(\mu+c\mathbb{P}\left(\chi_v^2>c\right)) = \frac{2\lambda(1-\lambda)}{(1-2\lambda)^2}(2cf(c)+v\mathbb{P}\left(\chi_v^2>c\right))$$

Case 2: $\mu - (v-c) - 2\lambda c \mathbb{P}\left(\chi_v^2 < c\right) < 0$. The upper bound in 20 shows that $\frac{d}{d\lambda}\left(\log\left(\mathbb{E}\left(e^{\lambda Z}\right)\right)\right) - \mu$ is bounded by

$$\frac{1}{1-2\lambda}\left[\frac{\mu+(v-c)-2\lambda c\mathbb{P}\left(\chi_v^2<c\right)}{\mathbb{P}\left(\chi_v^2<c\right)+\frac{1}{1-2\lambda}\mathbb{P}\left(\chi_v^2>c\right)}+(v-c-\mu)+2(\mu+c)\lambda\right]$$
$$=\frac{1}{1-2\lambda}\left[(v-c-\mu)\left(1-\frac{1}{\mathbb{P}\left(\chi_v^2<c\right)+\frac{1}{1-2\lambda}\mathbb{P}\left(\chi_v^2>c\right)}\right)+2\lambda c\left(1-\frac{\mathbb{P}\left(\chi_v^2<c\right)}{\mathbb{P}\left(\chi_v^2<c\right)+\frac{1}{1-2\lambda}\mathbb{P}\left(\chi_v^2>c\right)}\right)+2\lambda\mu\right]$$
$$=\frac{1}{1-2\lambda}\left[(v-c-\mu)\frac{\frac{1}{1-2\lambda}\mathbb{P}\left(\chi_v^2>c\right)-\mathbb{P}\left(\chi_v^2>c\right)}{\mathbb{P}\left(\chi_v^2<c\right)+\frac{1}{1-2\lambda}\mathbb{P}\left(\chi_v^2>c\right)}+2\lambda c\left(\frac{\frac{1}{1-2\lambda}\mathbb{P}\left(\chi_v^2>v\right)}{\mathbb{P}\left(\chi_v^2<c\right)+\frac{1}{1-2\lambda}\mathbb{P}\left(\chi_v^2>c\right)}\right)+2\lambda\mu\right]$$
$$=\frac{1}{1-2\lambda}\left[(v-c-\mu)\frac{2\lambda\mathbb{P}\left(\chi_v^2>c\right)}{1-2\lambda\mathbb{P}\left(\chi_v^2<c\right)}+2\lambda c\frac{\mathbb{P}\left(\chi_v^2>c\right)}{1-2\lambda\mathbb{P}\left(\chi_v^2<c\right)}+2\lambda\mu\right]$$
$$=\frac{2\lambda}{1-2\lambda}\left[\mu+(v-\mu)\frac{\mathbb{P}\left(\chi_v^2>c\right)}{1-2\lambda\mathbb{P}\left(\chi_v^2<c\right)}\right]$$
$$=\frac{2\lambda}{(1-2\lambda)^2}\left[\mu\left(1-2\lambda\right)+(v-\mu)\mathbb{P}\left(\chi_v^2>c\right)\frac{1-2\lambda}{1-2\lambda\mathbb{P}\left(\chi_v^2<c\right)}\right]$$
$$=\frac{2\lambda}{(1-2\lambda)^2}\left[\mu\left(1-2\lambda\right)+(v-\mu)\mathbb{P}\left(\chi_v^2>c\right)-(v-\mu)\mathbb{P}\left(\chi_v^2>c\right)\frac{2\lambda\mathbb{P}\left(\chi_v^2>c\right)}{1-2\lambda\mathbb{P}\left(\chi_v^2<c\right)}\right]$$

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

$$=\frac{2\lambda}{(1-2\lambda)^2}\left[\mu\left(1-\lambda\right)+(v-\mu)\mathbb{P}\left(\chi_v^2>c\right)-2\lambda c\mathbb{P}\left(\chi_v^2>c\right)^2\right]$$
$$=\frac{2\lambda}{(1-2\lambda)^2}\left[\left(\mu+c\mathbb{P}\left(\chi_v^2>c\right)\right)\left(1-\lambda\right)+(v-\mu-c)\mathbb{P}\left(\chi_v^2>c\right)-2\lambda c\mathbb{P}\left(\chi_v^2>c\right)^2+c\lambda\mathbb{P}\left(\chi_v^2>c\right)\right]$$

Since $\lambda < \frac{1}{2}$ and $v - c - \mu \leq 0$, we have that

$$(v-\mu-c)\mathbb{P}\left(\chi_v^2 > c\right) - 2\lambda c\mathbb{P}\left(\chi_v^2 > c\right)^2 + c\lambda\mathbb{P}\left(\chi_v^2 > c\right) \leq \lambda\mathbb{P}\left(\chi_v^2 > c\right)\left(2\left(v-c-\mu\right) + c - 2c\mathbb{P}\left(\chi_v^2 > c\right)\right)$$
$$= \lambda\mathbb{P}\left(\chi_v^2 > c\right)\left(2\left(\mathbb{E}\left(\chi_v^2|\chi_v^2 < c\right)\mathbb{P}\left(\chi_v^2 < c\right) - c\mathbb{P}\left(\chi_v^2 < c\right)\right) + c - 2c\mathbb{P}\left(\chi_v^2 > c\right)\right)$$
$$= \lambda\mathbb{P}\left(\chi_v^2 > c\right)\left(2\mathbb{E}\left(\chi_v^2|\chi_v^2 < c\right)\mathbb{P}\left(\chi_v^2 < c\right) - c\right) \leq 0$$

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.

3.3 Proof of Lemma 3

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,

$$\mathbb{P}\left(Y_i \geq a + x|Y_i \geq a, v_i = 1\right) = \frac{\mathbb{P}\left(\epsilon_i > \sqrt{a+x} - \mu\right)}{\mathbb{P}\left(\epsilon_i > \sqrt{a} - \mu\right)} \geq \frac{\mathbb{P}\left(\epsilon_i > \sqrt{a+x}\right)}{\mathbb{P}\left(\epsilon_i > \sqrt{a}\right)} = \mathbb{P}\left(Z > a + x|Z \geq a\right).$$

3.4 Proof of Lemma 4

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

$$G(\lambda) = \exp\left(\frac{\lambda\mu}{\mathbb{P}\left(\chi_1^2 > a\right)}\right)\frac{1}{\mathbb{P}\left(\chi_1^2 > a\right)}\int_0^\infty \frac{1}{\sqrt{2\pi x}}e^{\lambda a-\lambda zx-\frac{1}{2}x}dx = \exp\left(\frac{\lambda\mu}{\mathbb{P}\left(\chi_1^2 > a\right)}+\lambda a\right)\frac{\mathbb{P}\left(\chi_1^2 > a(1+2\lambda)\right)}{\mathbb{P}\left(\chi_1^2 > a\right)\sqrt{1+2\lambda}}.$$

Consequently,

$$\frac{dG(\lambda)}{d\lambda} = \frac{1}{\mathbb{P}\left(\chi_1^2 > a\right)}\left[-\frac{2af(a)}{1+2\lambda}e^{-\lambda a} + \left(\frac{\mu}{\mathbb{P}\left(\chi_1^2 > a\right)}+a-\frac{1}{1+2\lambda}\right)\frac{\mathbb{P}\left(\chi_1^2 > a(1+2\lambda)\right)}{\sqrt{1+2\lambda}}\right]\exp\left(\frac{\lambda\mu}{\mathbb{P}\left(\chi_1^2 > a\right)}+\lambda a\right)$$

and 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.

3.5 Proof of Lemma 5

The equality follows from Lemma 2. To prove the inequality, write $G(\tau) = \tau + 2af(a)$, where $0 \leq \tau \leq 1$ and a is defined by the equation $\mathbb{P}\left(\chi_1^2 > a\right) = \tau$. Note that $G(0) = 0$ and

$$\frac{dG}{d\tau} = 1 + \frac{da}{d\tau}\left(f(a) - af(a)\right) = 1 - \frac{1}{f(a)}\left(f(a) - af(a)\right) = a > 0$$

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.

3.6 Proof of Lemma 6

We know from Lemma 2, 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.

3.7 Proof of Lemma 7

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:

$$\begin{aligned}
&\mathbb{P}\left(Z_s > u\right) \le \sum_{i=\lfloor \log_b(s) \rfloor}^{\lceil \log_b(s+w) \rceil} \mathbb{P}\left(\max_{b^i \le s+a \le b^{i+1}} (s+a)\left(\bar{\eta}_{1:(s+a)}\right)^2 > u\right) \le \sum_{i=\lfloor \log_b(s) \rfloor}^{\lceil \log_b(s+w) \rceil} \mathbb{P}\left(\max_{b^i \le a' \le b^{i+1}} \left(T_{a'}\right)^2 > b^i u\right) \\
&\le 2 \sum_{i=\lfloor \log_b(s) \rfloor}^{\lceil \log_b(s+w) \rceil} \mathbb{P}\left(\max_{b^i \le a' \le b^{i+1}} T_{a'} > \sqrt{b^i u}\right) = 2 \sum_{i=\lfloor \log_b(s) \rfloor}^{\lceil \log_b(s+w) \rceil} \min_{\lambda} \left[\mathbb{P}\left(\max_{b^i \le a' \le b^{i+1}} e^{\lambda T_{a'}} > e^{\sqrt{b^i u}\lambda}\right)\right] \\
&\le 2 \sum_{i=\lfloor \log_b(s) \rfloor}^{\lceil \log_b(s+w) \rceil} \min_{\lambda} \left[\mathbb{E}\left(e^{\lambda T_{\lfloor b^{i+1} \rfloor}}\right) e^{-\sqrt{b^i u}\lambda}\right] = 2 \sum_{i=\lfloor \log_b(s) \rfloor}^{\lceil \log_b(s+w) \rceil} \min_{\lambda} \left[e^{\frac{\lfloor b^{i+1} \rfloor}{2}\lambda^2 - \sqrt{b^i u}\lambda}\right] \\
&\le 2 \sum_{i=\lfloor \log_b(s) \rfloor}^{\lceil \log_b(s+w) \rceil} \min_{\lambda} \left[e^{\frac{b^{i+1}}{2}\lambda^2 - \sqrt{b^i u}\lambda}\right] = 2 \sum_{i=\lfloor \log_b(s) \rfloor}^{\lceil \log_b(s+w) \rceil} e^{-\frac{u}{2b}} = 2(1 + \lceil \log_b(s+w) \rceil - \lfloor \log_b(s) \rfloor) e^{-\frac{u}{2b}} \\
&\le 2(3 + \log_b(s+w) - \log_b(s)) e^{-\frac{u}{2b}} = 2(3 + \log_b(1 + w/s)) e^{-\frac{u}{2b}} \le 2(3 + \log_b(w+1)) e^{-\frac{u}{2b}} \\
&\le 6 \frac{1 + \log(w+1)}{\log(b)} e^{-\frac{u}{2b}}.
\end{aligned}$$

Here the fifth inequality follows from Doob's martingale inequality.

Next note that

$$\begin{aligned}
&\mathbb{P}\left(\max_{0 \le f, d \le w : j-f-d-i \ge 0} \left((j-f-d-i+1)\left(\bar{\boldsymbol{\eta}}^{(c)}_{(i+d):(j-f)}\right)^2\right) > u\right) \\
&\le \sum_{d=0}^{w} \mathbb{P}\left(\max_{0 \le f \le \min(w, j-i-d)} \left((j-f-d-i+1)\left(\bar{\boldsymbol{\eta}}^{(c)}_{(i+d):(j-f)}\right)^2\right) > u\right) \le \sum_{d=0}^{w} \mathbb{P}\left(Z_{\max(1, j-i-2w)} > u\right) \\
&\le 6(w+1) \frac{1 + \log(w+1)}{\log(b)} e^{-\frac{u}{2b}}
\end{aligned}$$

3.8 Proof of Lemma 8

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 1 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 & 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 & 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,

$$\mathbb{P}\left(\left|\left(\sum_{c\in S}\left(\sum_{t=i}^{j}\left(\boldsymbol{\mu}_t^{(c)}-\bar{\boldsymbol{\mu}}_{i:j}^{(c)}\right)\boldsymbol{\eta}_t^{(c)}\right)\right)\left(\sqrt{\sum_{c\in S}\sum_{t=i}^{j}\left(\boldsymbol{\mu}_t^{(c)}-\bar{\boldsymbol{\mu}}_{i:j}^{(c)}\right)^2}\right)^{-1}\right| > \sqrt{2\psi+2\,|S\cap W_{i,j}|\log(p)}\right)\leq A_4p^{-|S\cap W_{i,j}|}e^{-\psi},$$

for some constant A_4. A Bonferroni correction over the sets S then gives that

$$\begin{aligned}
&\mathbb{P}\left(\left|\left(\sum_{c\in S}\left(\sum_{t=i}^{j}\left(\boldsymbol{\mu}_t^{(c)}-\bar{\boldsymbol{\mu}}_{i:j}^{(c)}\right)\boldsymbol{\eta}_t^{(c)}\right)\right)\right| \le \sqrt{\sum_{c\in S}\sum_{t=i}^{j}\left(\boldsymbol{\mu}_t^{(c)}-\bar{\boldsymbol{\mu}}_{i:j}^{(c)}\right)^2}\sqrt{2\psi+2\left|S\cap W_{i,j}\right|\log(p)}\quad \forall S\subset\{1,...,p\}\right)\\
&=\mathbb{P}\left(\left|\left(\sum_{c\in S\cap W_{i,j}}\left(\sum_{t=i}^{j}\left(\boldsymbol{\mu}_t^{(c)}-\bar{\boldsymbol{\mu}}_{i:j}^{(c)}\right)\boldsymbol{\eta}_t^{(c)}\right)\right)\right| \le \sqrt{\sum_{c\in S\cap W_{i,j}}\sum_{t=i}^{j}\left(\boldsymbol{\mu}_t^{(c)}-\bar{\boldsymbol{\mu}}_{i:j}^{(c)}\right)^2}\sqrt{2\psi+2\left|S\cap W_{i,j}\right|\log(p)}\quad \forall S\subset\{1,...,p\}\right)\\
&=\mathbb{P}\left(\left|\left(\sum_{c\in W}\left(\sum_{t=i}^{j}\left(\boldsymbol{\mu}_t^{(c)}-\bar{\boldsymbol{\mu}}_{i:j}^{(c)}\right)\boldsymbol{\eta}_t^{(c)}\right)\right)\right| \le \sqrt{\sum_{c\in W}\sum_{t=i}^{j}\left(\boldsymbol{\mu}_t^{(c)}-\bar{\boldsymbol{\mu}}_{i:j}^{(c)}\right)^2}\sqrt{2\psi+2\left|W\right|\log(p)}\quad \forall W\subset W_{i,j}\right)\\
&\le 1-\sum_{W\subset W_{i,j}}\mathbb{P}\left(\left|\left(\sum_{c\in W}\left(\sum_{t=i}^{j}\left(\boldsymbol{\mu}_t^{(c)}-\bar{\boldsymbol{\mu}}_{i:j}^{(c)}\right)\boldsymbol{\eta}_t^{(c)}\right)\right)\right| > \sqrt{\sum_{c\in W}\sum_{t=i}^{j}\left(\boldsymbol{\mu}_t^{(c)}-\bar{\boldsymbol{\mu}}_{i:j}^{(c)}\right)^2}\sqrt{2\psi+2\left|W\right|\log(p)}\quad \forall W\subset W_{i,j}\right)\\
&\le 1-\sum_{|W|=1}^{|W_{i:j}|}\frac{p!}{(p-|W|)!(|W|)!}A_4p^{-|W|}e^{-\psi}\le 1-(A_4e)e^{-\psi}
\end{aligned}$$

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

$$\mathbb{P}\left(\frac{(j-j')(j'-i+1)}{j-i+1}\left(\bar{\mathbf{x}}_{i:j'}^{(c)}-\bar{\mathbf{x}}_{(j'+1):j}^{(c)}\right)^2>t\right)\ge\mathbb{P}\left(\chi_1^2>t\right)\quad\forall t>0.$$

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

$$\left(\sum_{c\in\mathbf{J}_k}\sqrt{j-i+1}\,\bar{\boldsymbol{\eta}}_{i:j}^{(c)}\right)\left(\sqrt{2|\mathbf{J}_k|\psi}\right)^{-1}\sim N(0,1)$$

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 $\mathbb{P}\left(E_{11}\right) > 1 - A_{11}n^3e^{-\psi}$

3.9 Proof of Lemma 9

This Lemma can be proven using straightforward algebra.

$$\begin{aligned}\mathcal{S}\left(\mathbf{x}_{i:j}, \mathbf{J}\right) &= \sum_{c\in\mathbf{J}}(j-i+1)\left(\bar{\mathbf{x}}_{i:j}^{(c)}\right)^2 = \sum_{c\in\mathbf{J}}(j-i+1)^{-1}\left(\bar{\mathbf{x}}_{i:j}^{(c)}\right)^2 \\ &= \sum_{c\in\mathbf{J}}\left[\frac{\left((j'+1-i)\bar{\mathbf{x}}_{i:j'}^{(c)} + (j-j')\bar{\mathbf{x}}_{(j'+1):j}^{(c)}\right)^2}{j-i+1}\right].\end{aligned}$$

Next we note that the following holds for all a, b, y, and z:

$$\frac{(ay+bz)^2}{a+b} = \frac{a^2y^2+2abyz+b^2z^2}{a+b} = ay^2+bz^2+\frac{-aby^2+2abyz-baz^2}{a+b} = ay^2+bz^2-\frac{ab}{a+b}(y-z)^2.$$

Thus,

$$\mathcal{S}\left(\mathbf{x}_{i:j}, \mathbf{J}\right) = \sum_{c\in S}(j'+1-i)\left(\bar{\mathbf{x}}_{i:j'}^{(c)}\right)^2 + \sum_{c\in S}(j-j')\left(\bar{\mathbf{x}}_{(j'+1):j}^{(c)}\right)^2 - \sum_{c\in S}\frac{(j-j')(j'-i+1)}{j-i+1}\left(\bar{\mathbf{x}}_{i:j'}^{(c)} - \bar{\mathbf{x}}_{(j'+1):j}^{(c)}\right)^2,$$

which finishes the proof.

3.10 Proof of Lemma 10

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 9. Indeed, by Lemma 9 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.

3.11 Proof of Lemma 11

This lemma shows that merging two neighbouring fitted segments reduces the penalised cost by $O(C)$ and follows almost immediately from Lemma 9. We consider the cases $|\mathbf{J}_k| \leq k^*$ and $|\mathbf{J}_k| > k^*$ separately. Let $|\mathbf{J}_k| \leq k^*$. Then

$$\begin{aligned}
&\mathcal{C}\left(x_{i:j'}, \mathbf{J}_k\right) + \mathcal{C}\left(x_{(j'+1):j}, \mathbf{J}_k\right) - \mathcal{C}\left(x_{i:j}, \mathbf{J}_k\right) \\
&= C\psi + C|\mathbf{J}_k|\log(p) - \sum_{c \in \mathbf{J}_k} \frac{(j-j')(j'-i+1)}{j-i+1}\left(\bar{\boldsymbol{\eta}}^{(c)}_{i:j'} - \bar{\boldsymbol{\eta}}^{(c)}_{(j'+1):j}\right)^2 \\
&\geq C\psi + C|\mathbf{J}_k|\log(p) - 2\psi - 2|\mathbf{J}_k|\log(p) \geq \frac{79}{80}C\left(\psi + |\mathbf{J}_k|\log(p)\right),
\end{aligned}$$

where the first inequality follows from E_7 and the second one holds if C exceeds some global constant. Now let $|\mathbf{J}_k| \geq k^*$

$$\begin{aligned}
&\mathcal{C}\left(x_{i:j'}, \mathbf{1}\right) + \mathcal{C}\left(x_{(j'+1):j}, \mathbf{1}\right) - \mathcal{C}\left(x_{i:j}, \mathbf{1}\right) = p + C\psi + C\sqrt{p\psi} - \sum_{c=1}^{p} \frac{(j-j')(j'-i+1)}{j-i+1}\left(\bar{\boldsymbol{\eta}}^{(c)}_{i:j'} - \bar{\boldsymbol{\eta}}^{(c)}_{(j'+1):j}\right)^2 \\
&\geq p + C\psi + C\sqrt{p\psi} - 2\psi - 2\sqrt{p\psi} - p \geq \frac{79}{80}C\left(\psi + \sqrt{p\psi}\right),
\end{aligned}$$

where the first inequality follows from E_8 and the second one holds if C exceeds some global constant.

3.12 Proof of Lemma 12

This Lemma proves MVCAPA has power at detecting anomalous regions. We begin by considering the case in which J_k is dense. We have:

$$
\begin{aligned}
&\mathcal{C}\left(\mathbf{x}_{i:j}, \mathbf{1}\right) = p + C\psi + C\sqrt{p\psi} - \sum_{c=1}^{p} (j-i+1)\left(\mathbf{x}_{i:j}^{(c)}\right)^2 \\
&= p + C\psi + C\sqrt{p\psi} - \sum_{c=1}^{p}(j-i+1)\left(\bar{\boldsymbol{\eta}}_{i:j}^{(c)}\right)^2 - |\mathbf{J}_k|\boldsymbol{\mu}_k^2(j-i+1) - 2\boldsymbol{\mu}_k\sqrt{j-i+1}\sum_{c\in\mathbf{J}_k}\left(\sqrt{j-i+1}\bar{\boldsymbol{\eta}}_{i:j}\right) \\
&\leq C\psi + (C+2)\sqrt{p\psi} - |\mathbf{J}_k|\boldsymbol{\mu}_k^2(j-i+1) + 2\sqrt{(j-i+1)\boldsymbol{\mu}_k^2}\sqrt{2|\mathbf{J}_k|\psi} \\
&\leq C\psi + (C+2)\sqrt{p\psi} - \frac{1}{2}|\mathbf{J}_k|\boldsymbol{\mu}_k^2(j-i+1) + 4\psi \leq C\psi + (C+2)\sqrt{p\psi} - \frac{1}{2}|\mathbf{J}_k|\boldsymbol{\mu}_k^2\frac{4C}{\triangle_k^2} + 4\psi \\
&= (C+4)\psi + (C+2)\sqrt{p\psi} - 2C(\psi + \sqrt{p\psi}) \leq 0
\end{aligned}
$$

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:

$$
\begin{aligned}
&\mathcal{C}\left(\mathbf{x}_{i:j}, \mathbf{J}_k\right) = C\psi + C|\mathbf{J}_k|\log(p) - \sum_{c\in\mathbf{J}_k}(j-i+1)\left(\boldsymbol{\mu}_k + \bar{\boldsymbol{\eta}}_{i:j}^{(c)}\right)^2 \\
&= C\psi + C|\mathbf{J}_k|\log(p) - \sum_{c\in\mathbf{J}_k}(j-i+1)\left(\bar{\boldsymbol{\eta}}_{i:j}^{(c)}\right)^2 - |\mathbf{J}_k|\boldsymbol{\mu}_k^2(j-i+1) - 2\boldsymbol{\mu}_k\sqrt{j-i+1}\sum_{c\in\mathbf{J}_k}\left(\sqrt{j-i+1}\bar{\boldsymbol{\eta}}_{i:j}\right) \\
&\leq C\psi + C|\mathbf{J}_k|\log(p) - |\mathbf{J}_k|\boldsymbol{\mu}_k^2(j-i+1) + 2\sqrt{(j-i+1)\boldsymbol{\mu}_k^2}\sqrt{2|\mathbf{J}_k|\psi + 2|\mathbf{J}_k|^2\log(p)} \\
&\leq C\psi + C|\mathbf{J}_k|\log(p) - \frac{1}{2}|\mathbf{J}_k|\boldsymbol{\mu}_k^2(j-i+1) + 4\psi + 4|\mathbf{J}_k|\log(p) \leq (C+4)\psi + (C+4)|\mathbf{J}_k|\log(p) - \frac{1}{2}|\mathbf{J}_k|\boldsymbol{\mu}_k^2\frac{4C}{\triangle_k^2} \\
&= (C+4)\psi + (C+4)|\mathbf{J}_k|\log(p) - 2C(\psi + |\mathbf{J}_k|\log(p)) \leq 0,
\end{aligned}
$$

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.

3.13 Proof of Lemma 13

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 9 implies that

$$\mathcal{C}\left(x_{(s'+1):e'}, \mathbf{J}_k\right) = \mathcal{C}\left(x_{(s'+1):e_k}, \mathbf{J}_k\right) + \mathcal{C}\left(x_{(e_k+1):e'}, \mathbf{J}_k\right) - C\left(\psi + |\mathbf{J}_k|\log(p)\right) + \sum_{c\in\mathbf{J}_k} \frac{e'-s'}{2}\left(\boldsymbol{\mu}_k + \bar{\boldsymbol{\eta}}_{(s'+1):e_k} - \bar{\boldsymbol{\eta}}_{(e_k+1):e'}\right)^2$$

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

$$\begin{aligned}&\mathcal{C}\left(x_{i:j}, \mathbf{J}_k\right) \\ &\geq \mathcal{C}\left(x_{i:s'}, \mathbf{J}_k\right) + \mathcal{C}\left(x_{(s'+1):e_k}, \mathbf{J}_k\right) + \mathcal{C}\left(x_{(e_k+1):e'}, \mathbf{J}_k\right) + \left(\frac{14}{3}C - 14\right)\psi + \left(\frac{14}{3}C - 2\right)|\mathbf{J}_k|\log(p) \\ &\geq \mathcal{C}\left(x_{i:e_k}, \mathbf{J}_k\right) + \frac{19}{20}C\left(\psi + |\mathbf{J}_k|\log(p)\right) + (C-2)\left(\psi + |\mathbf{J}_k|\log(p)\right) + \left(\frac{14}{3}C - 14\right)\psi + \left(\frac{14}{3}C - 2\right)|\mathbf{J}_k|\log(p) \\ &\geq 6C\left(\psi + |\mathbf{J}_k|\log(p)\right),\end{aligned}$$

where the second inequality follows from Lemma 11 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 9 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.

3.14 Proof of Lemma 14

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 13. Assume now that $|\mathbf{J}| \leq k^*$. As in the proof of Lemma 13, 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 13 and a choice of C exceeding some global constant.

3.15 Proof of Lemma 15

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 11. In fact, Lemma 15 follows a fortiori from Lemma 11 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 11 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 11 twice, and the third holds if C exceeds a global constant.

3.16 Proof of Lemma 16

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 10, and the second and fourth from E_2.

3.17 Proof of Lemma 17

This Lemma links the savings of a fitted segment to the signal strength of the corresponding segment. We have that

$$\begin{aligned}
&\alpha\left(C\psi + C|\mathbf{J}|\log(p)\right) \leq \sum_{c\in\mathbf{J}}\left(\boldsymbol{\mu}_k + \bar{\boldsymbol{\eta}}_{i:j}^{(c)}\right)^2 (j-i+1) \\
&= |\mathbf{J}\cap\mathbf{J}_k|(j-i+1)\boldsymbol{\mu}_k^2 + 2\sqrt{j-i+1}\boldsymbol{\mu}_k \sum_{c\in\mathbf{J}\cap\mathbf{J}_k}\sqrt{j-i+1}\bar{\boldsymbol{\eta}}_{i:j}^{(c)} + \sum_{c\in\mathbf{J}}\left(\sqrt{j-i+1}\bar{\boldsymbol{\eta}}_{i:j}^{(c)}\right)^2 \\
&\leq |\mathbf{J}\cap\mathbf{J}_k|(j-i+1)\boldsymbol{\mu}_k^2 + 2\sqrt{j-i+1}|\boldsymbol{\mu}_k|\sqrt{2\psi|\mathbf{J}\cap\mathbf{J}_k| + 2|\mathbf{J}\cap\mathbf{J}_k|^2\log(p)} + 2\psi + 2|\mathbf{J}|\log(p) \\
&\leq |\mathbf{J}|(j-i+1)\boldsymbol{\mu}_k^2 + 2\sqrt{j-i+1}|\boldsymbol{\mu}_k|\sqrt{2\psi|\mathbf{J}| + 2|\mathbf{J}|^2\log(p)} + 2\psi + 2|\mathbf{J}|\log(p) \\
&= \left(\sqrt{|\mathbf{J}|(j-i+1)\boldsymbol{\mu}_k^2} + \sqrt{2\psi + 2|\mathbf{J}|\log(p)}\right)^2,
\end{aligned}$$

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

$$\sqrt{|\mathbf{J}|(j-i+1)\boldsymbol{\mu}_k^2} \geq \left(\sqrt{\alpha C} - \sqrt{2}\right)\sqrt{\psi + |\mathbf{J}|\log(p)}$$

3.18 Proof of Lemma 18

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 19. We begin by noting that

$$\mathcal{C}\left(\mathbf{x}_{i:j}, \mathbf{J}_k\right) - \mathcal{C}\left(\mathbf{x}_{i:j}, \mathbf{J}\right) = C\left(|\mathbf{J}_k| - |\mathbf{J}|\right)\log(p) - \sum_{c \in \mathbf{J}_k \setminus \mathbf{J}} (j-i+1)\left(\mu + \bar{\boldsymbol{\eta}}_{i:j}^{(c)}\right)^2 + \sum_{c \in \mathbf{J} \setminus \mathbf{J}_k} (j-i+1)\left(\bar{\boldsymbol{\eta}}_{i:j}^{(c)}\right)^2 \quad (21)$$

If $|\mathbf{J}| > \frac{19}{20}|\mathbf{J}_k|$, E_1 bounds (21) by

$$\begin{aligned} C\left(|\mathbf{J}_k| - |\mathbf{J}|\right)\log(p) + 2\psi + 2|\mathbf{J}|\log(p) &\leq \frac{1}{10}C|\mathbf{J}_k|\log(p) + 2\psi + \left(2 - \frac{1}{20}C\right)|\mathbf{J}|\log(p) \\ &\leq \frac{1}{10}C|\mathbf{J}_k|\log(p) + 2\psi, \end{aligned}$$

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 (21) 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 17 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 19 , 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 (21) 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.

3.19 Proof of Lemma 19

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

$$\mathcal{C}\left(\mathbf{x}_{i:j}, \mathbf{1}\right) - \mathcal{C}\left(\mathbf{x}_{i:j}, \mathbf{J}\right) = p + C\sqrt{p\psi} - C|\mathbf{J}|\log(p) - \sum_{c \notin \mathbf{J}} (j-i+1)\left(\bar{\mathbf{x}}_{i:j}^{(c)}\right)^2 \tag{22}$$

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 (22) 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

$$\frac{41}{40}C\sqrt{p\psi} - \frac{37}{40}C|\mathbf{J}|\log(p) + 2\psi - |\mathbf{A}|(j-i+1)\boldsymbol{\mu}_k^2 + 2\sqrt{(j-i+1)\boldsymbol{\mu}_k^2}\sqrt{2|\mathbf{A}|\psi + 2|\mathbf{A}|^2\log(p)} \tag{23}$$

Lemma 17 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

$$\begin{aligned}
&|\mathbf{A}|\sqrt{(j-i+1)\boldsymbol{\mu}_k^2} - 2\sqrt{2|\mathbf{A}|\psi + 2|\mathbf{A}|^2\log(p)} \\
&\geq \left(\sqrt{\frac{19}{20}C} - 2\right)\sqrt{\frac{|\mathbf{A}|}{|\mathbf{J}|}|\mathbf{A}|\psi + |\mathbf{A}|^2\log(p)} - 2\sqrt{2|\mathbf{A}|\psi + 2|\mathbf{A}|^2\log(p)} \\
&\geq \left(\sqrt{\frac{19}{20}C} - 2\right)\sqrt{\frac{1}{20}|\mathbf{A}|\psi + |\mathbf{A}|^2\log(p)} - 2\sqrt{2|\mathbf{A}|\psi + 2|\mathbf{A}|^2\log(p)},
\end{aligned}$$

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 (23) 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.

3.20 Proof of Lemma 20

This Lemma bounds the increase in penalised cost incurred when transitioning from a fitted dense segment to the sparse ground truth. We have that

$$\mathcal{C}\left(\mathbf{x}_{i:j}, \mathbf{J}_k\right) - \mathcal{C}\left(\mathbf{x}_{i:j}, \mathbf{1}\right) = C|\mathbf{J}_k|\log(p) - \left(C\sqrt{p\psi} + p\right) + \sum_{c \notin \mathbf{J}_k} (j-i+1)\left(\bar{\boldsymbol{\eta}}_{i:j}^{c}\right)^2$$
$$\leq C|\mathbf{J}_k|\log(p) - \left(C\sqrt{p\psi} + p\right) + \left(p + 2\psi + 2\sqrt{p\psi}\right) = C|\mathbf{J}_k|\log(p) - C\sqrt{p\psi} + 2\sqrt{p\psi} + 2\psi$$
$$\leq \frac{13}{20}C|\mathbf{J}_k|\log(p) - \frac{6}{10}C\sqrt{p\psi} + 2\psi \leq \frac{1}{10}C|\mathbf{J}_k|\log(p) - \frac{1}{20}C\sqrt{p\psi} + 2\psi,$$

for large enough C. Here the first inequality follows from E_2 and the second inequality holds because $|\mathbf{J}_k| \leq k^*$.

3.21 Proof of Lemma 21

The proof is very similar to that of Lemma 19 and has therefore been omitted.

3.22 Proof of Lemma 22

The proof is very similar to that of Lemma 18 and has therefore been omitted.

3.23 Proof of Lemma 23

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_{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)-\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\\
&\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) + 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) \\
&+ \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) \\
&\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) \\
&- 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.

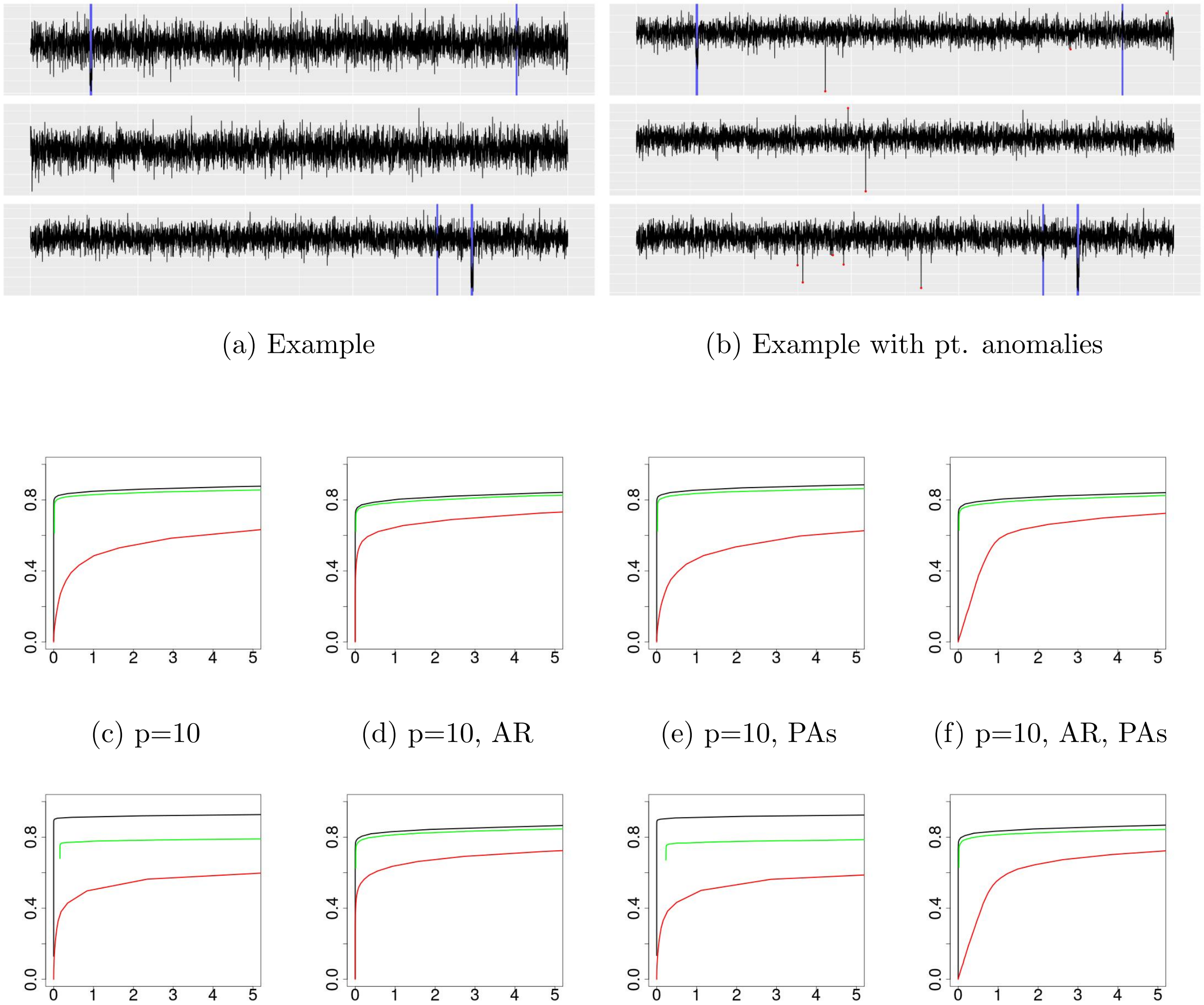
4 Further Simulations And Tables

In this section, we present additional results from the simulation study and application section. Figures 2 to 5 display the full comparison between MVCAPA, PASS, and Inspect over the four settings and data generating processes described in Section 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 6 and Figures 7 respectively.

Figure 8 investigates the robustness of the ROC analysis with regards to different tolerance levels for true positives. Specifically it considers a setting where $\sigma^2 = 2\log(p)$, $k = 1$, $p = 10$, and $n = 5000$ with detected anomalies being counted as true positives if within 10, 20, or 30 observations of a real anomaly. It should be noticed that the ranking remains very similar.

Figure 9 investigates the accuracy of MVCAPA using a range of different lags. It should be noted that this analysis is vulnerable to selection bias – as the max lag is increased weaker anomalies become detectable which pollutes the average.

Table 10 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

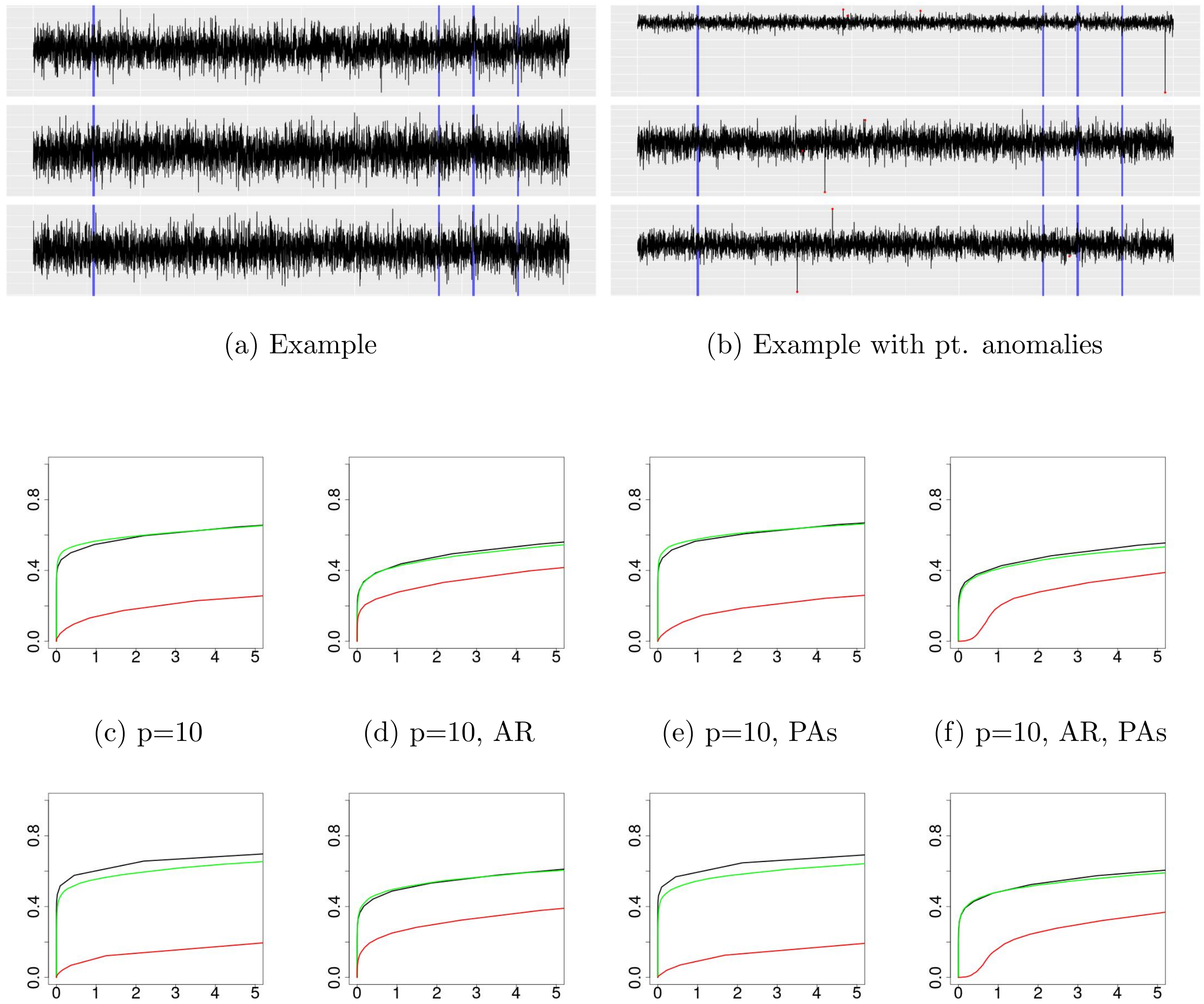
(g) p=100

(h) p=10, T

(i) p=100, PAs

(j) p=10, T, PAs

Figure 2: Example series and detection accuracy plots for setting 1. MVCAPA is in black, PASS in green, and Inspect in red. 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

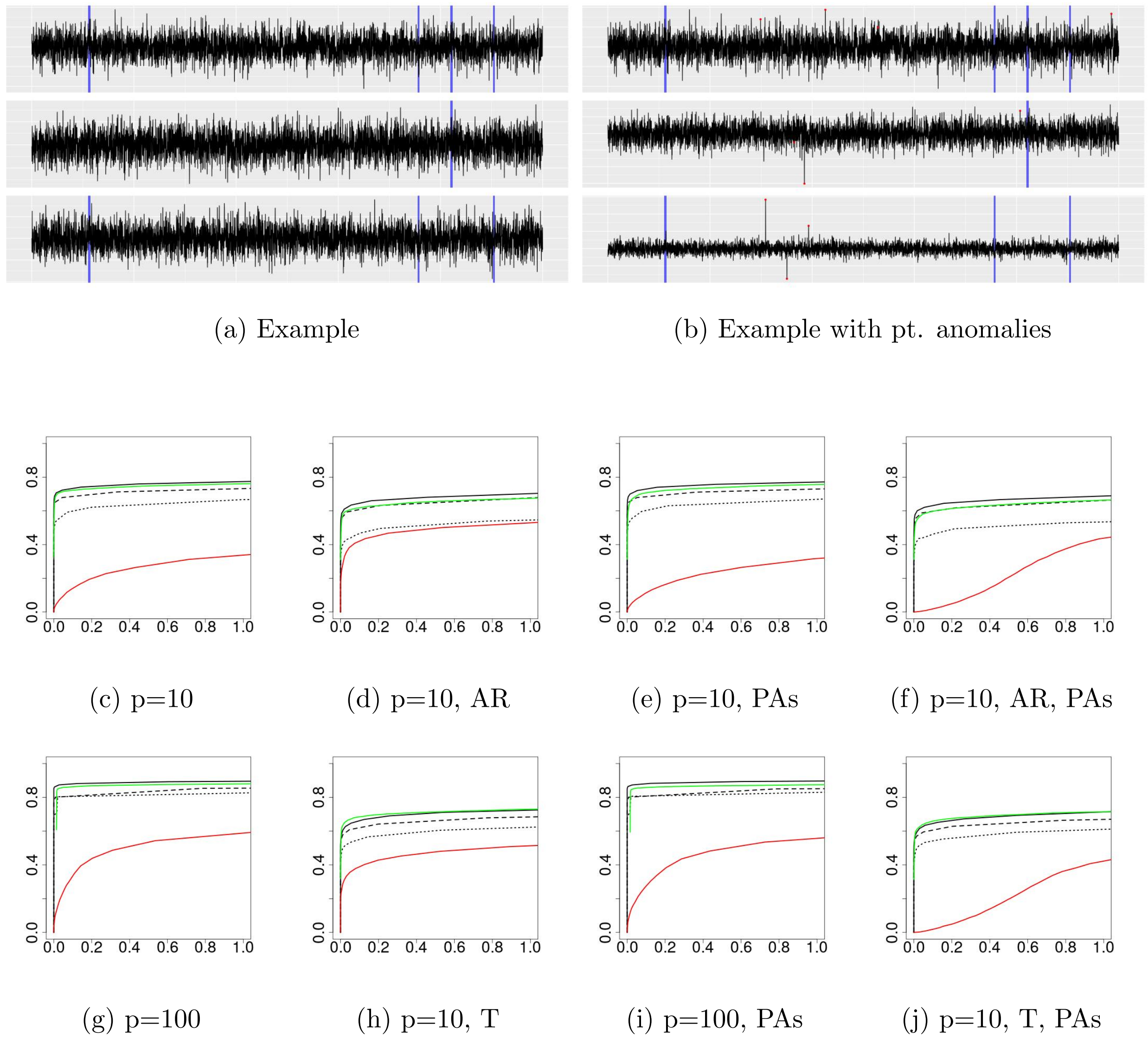
(g) p=100

(h) p=10, T

(i) p=100, PAs

(j) p=10, T, PAs

Figure 3: Example series and detection accuracy plots for setting 2. MVCAPA is in black, PASS in green, and Inspect in red. 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

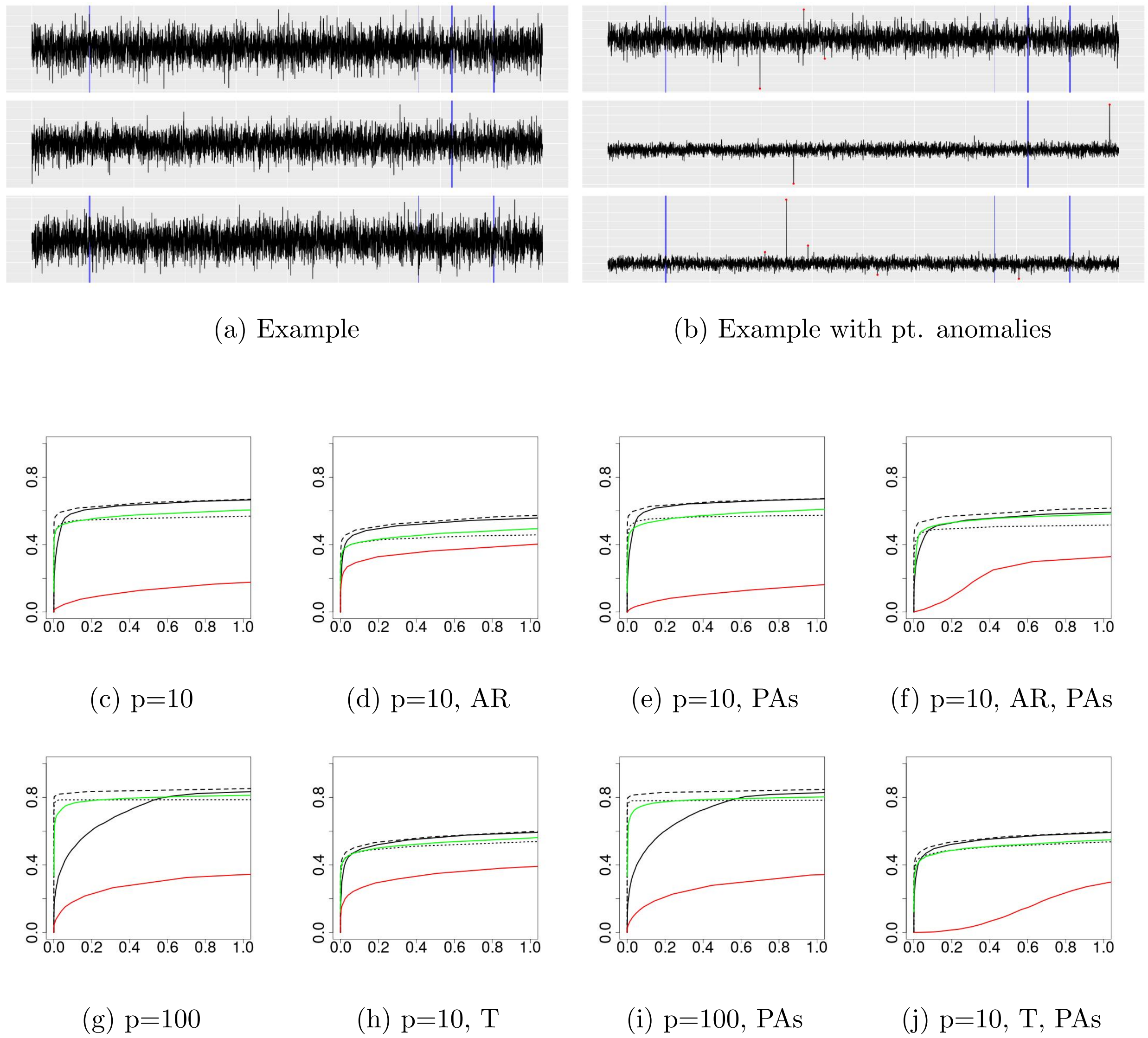
(g) p=100

(h) p=10, T

(i) p=100, PAs

(j) p=10, T, PAs

Figure 4: Example series and detection accuracy plots for setting 3. MVCAPA is in black, PASS in green, and Inspect in red. The solid black 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

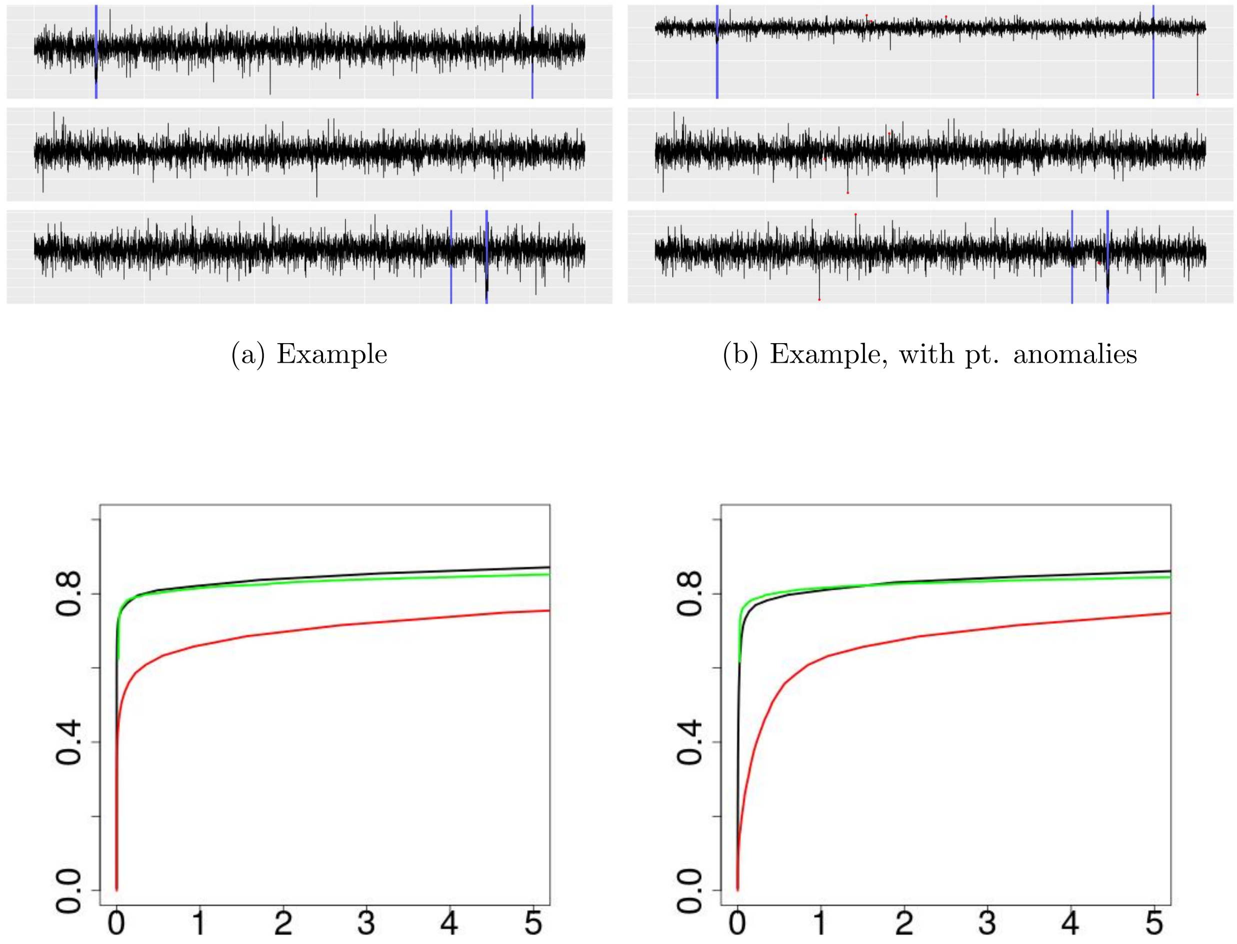
(g) p=100

(h) p=10, T

(i) p=100, PAs

(j) p=10, T, PAs

Figure 5: Example series and detection accuracy plots for setting 4. MVCAPA is in black, PASS in green, and Inspect in red. The solid black 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.



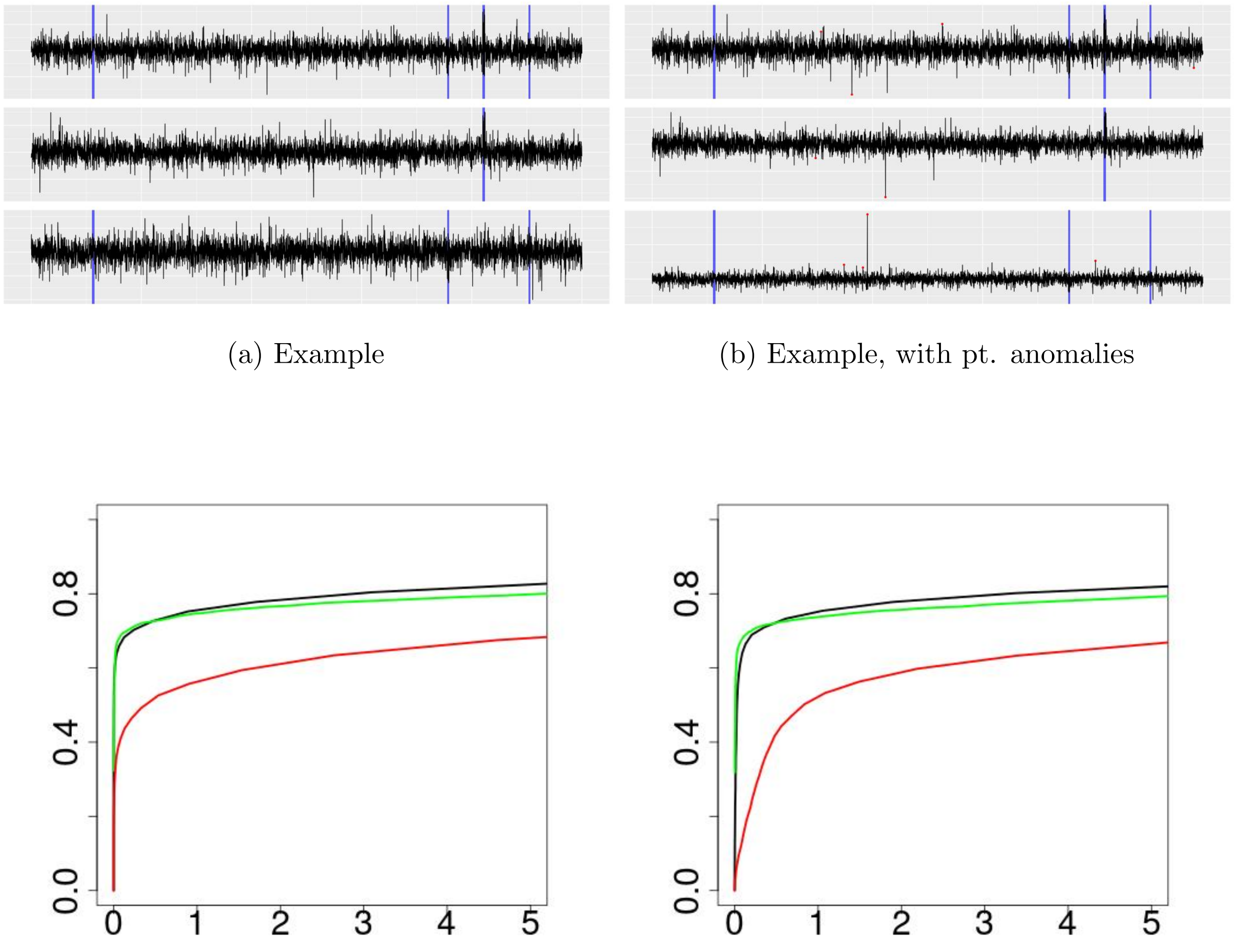
(a) Example

(b) Example, with pt. anomalies

(c) p=10

(d) p=10, with pt. anomalies

Figure 6: Example series and detection accuracy plots for setting 1 with change in both mean and variance. MVCAPA is in black, PASS in green, and Inspect in red. The x-axis denotes the number of false discoveries normalised by the total number of real anomalies present in the data.



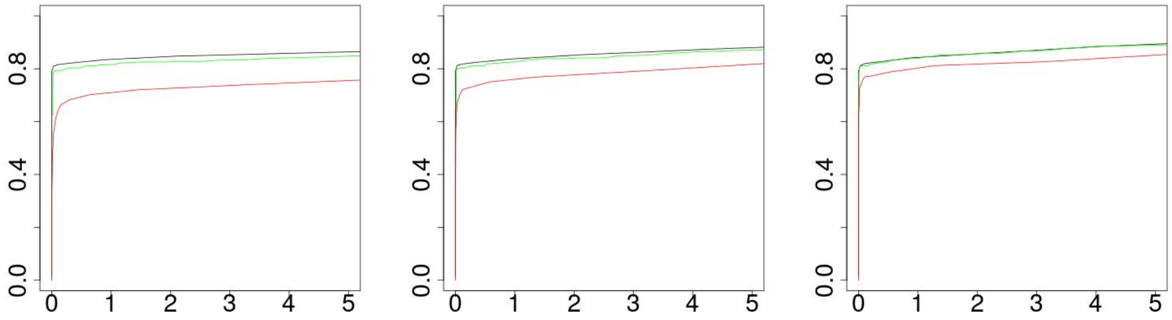
(a) Example

(b) Example, with pt. anomalies

(c) p=10

(d) p=10, with pt. anomalies

Figure 7: Example series and detection accuracy plots for setting 3 with a change in both mean and variance. MVCAPA is in black, PASS in green, and Inspect in red. The x-axis denotes the number of false discoveries normalised by the total number of real anomalies present in the data.

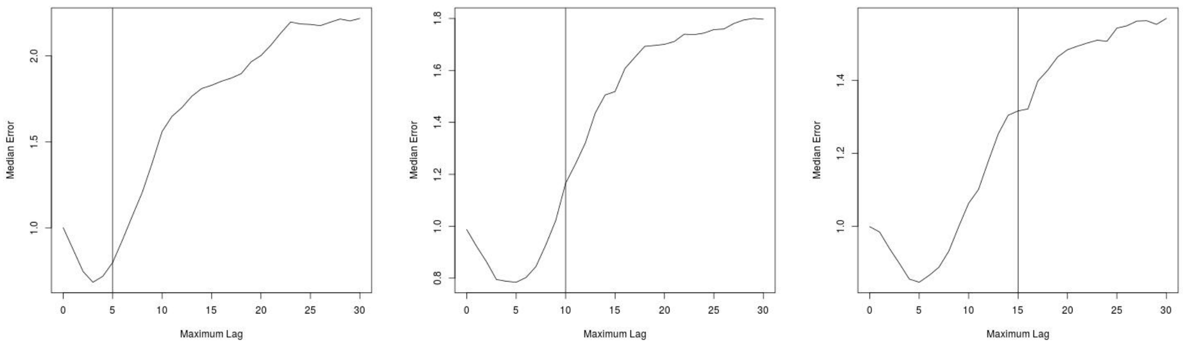


(a) Tolerance 10

(b) Tolerance 20

(c) Tolerance 30

Figure 8: ROC curves for MVCAPA, Pass, and Inspect under different levels of tolerance. For these results, $\sigma^2 = 2\log(p)$, $k = 1$, $p = 10$, and $n = 5000$ with, from left to right, detected anomalies being counted as true positives if within 10, 20, or 30 observations of a real anomaly



(a) Max Lag 5

(b) Max Lag 10

(c) Max Lag 15

Figure 9: Accuracy of MVCAPA for setting 2 (with varying maximum lags) under a misspecified lag. The vertical line represents the actual maximum lag.

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 10: Analysis of Chromosome 6 as detailed in the caption of Figure 2. Note that the chromosome contains two different CNVs (of different lengths) beginning at 31388628.

5 Pseudocode

Algorithm 1 Update

```
Input:  A vector of past lagged savings $\mathbf{S}_T^{(j)}$.
        A new saving $S$.
        A maximum lag $w \geq 0$.
for $k \in \{w, ..., 1\}$ do
    $\mathbf{S}_T^{(j)}(k) \leftarrow \mathbf{S}_T^{(j)}(k-1)$
end for
$\mathbf{S}_T^{(j)}(0) \leftarrow S$
$\hat{\mathbf{E}}_T^{(j)} \leftarrow \arg\max_{0 \leq k \leq w} \left(\mathbf{S}_T^{(j)}\right)(k)$
$\hat{\mathbf{C}}_T^{(j)} \leftarrow \max_{0 \leq k \leq w} \left(\mathbf{S}_T^{(j)}\right)(k)$
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 2 ComputeSaving

```
Input:  A vector of savings $\mathbf{C}_T^{(1:p)}$ .
        Penalty constants $\beta_{1:p}$ for the components of a collective anomalies.
$\sigma_1, ..., \sigma_p \leftarrow order(\mathbf{C}_T^{(1)}, ..., \mathbf{C}_T^{(p)})$        ▷ In decreasing order
$\mathbf{C}_T \leftarrow \max_{1 \leq k \leq p} \left(\sum_{i=1}^{k} \mathbf{C}_T^{(\sigma_i)} - \beta_i\right)$
$\hat{k} \leftarrow \arg\max_{1 \leq k \leq p} \left(\sum_{i=1}^{k} \mathbf{C}_T^{(\sigma_i)} - \beta_i\right)$
$\mathbf{CP}(T) \leftarrow \{\sigma_1, ..., \sigma_{\hat{k}}\}$
Output The optimal set of components $\mathbf{CP}(T)$, as well as the corresponding penalised saving $\mathbf{C}_T$.
```

Algorithm 3 ComputePtSaving

```
Input:  A vector of observations $\mathbf{x}_t^{(1:p)}$.
        Penalty constants $\beta'$ for a point anomaly.
$\mathbf{C}'_t \leftarrow \sum_{i=1}^{p} \left( \left( \mathbf{x}_t^{(i)} \right)^2 - \beta' \right)^+$
$\mathbf{CP}'_t \leftarrow \left\{ i | i \in \{1, ..., p\} : \left( \mathbf{x}_t^{(i)} \right)^2 > \beta' \right\}$
Output The optimal set of components $\mathbf{CP}'_t$, as well as the corresponding penalised saving $\mathbf{C}'_t$.
```

Algorithm 4 MVCAPA Algorithm (No Pruning)

```
Input:      A set of multivariate observations of the form, $(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n)$ where $\mathbf{x}_i \in \mathbb{R}^p$.
            Penalty constants $\beta_{1:p}$ and $\beta'$ 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
    $\hat{\mu}^{(j)} \leftarrow MEDIAN(\mathbf{x}_1^{(j)}, \mathbf{x}_2^{(j)}, \dots, \mathbf{x}_n^{(j)})$        ▷ Obtain robust estimates of the mean and variance
    $\hat{\sigma}^{(j)} \leftarrow IQR(\mathbf{x}_1^{(j)}, \mathbf{x}_2^{(j)}, \dots, \mathbf{x}_n^{(j)})$
    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
    $C(m) \leftarrow \max\left[C_1(m), C_2(m), C_3(m)\right]$
    switch $\arg\max\left[C_1(m), C_2(m), C_3(m)\right]$ do        ▷ Select type of anomaly giving the lowest cost
        case 1 :
            $Anom(m) \leftarrow [Anom(s), (s+1, m)]$
            $Comp(m) \leftarrow [Comp(s), \mathbf{Cp}(s)]$
            $Lags(m) \leftarrow [Lags(s), (\mathbf{L}_s^{(1:p)}, \mathbf{E}_s^{(1:p)})]$
        case 2 :
            $Anom(m) \leftarrow Anom(m-1)$
        case 3 :
            $Anom(m) \leftarrow [Anom(m-1), (m)]$
            $Comp(m) \leftarrow [Comp(m-1), \mathbf{Cp}']$
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 $Lags(n)$.