UiO : University of Oslo

Martin Tveten

Scalable change and anomaly detection in cross-correlated data

Thesis submitted for the degree of Philosophiae Doctor

Department of Mathematics
Faculty of Mathematics and Natural Sciences

2021

Series of dissertations submitted to the
Faculty of Mathematics and Natural Sciences, University of Oslo
No. 2356

ISSN 1501-7710

Cover: Hanne Baadsgaard Utigard.
Print production: Reprosentralen, University of Oslo.

Preface

This doctoral project started in August 2017 as a continuation of my master's thesis. Both projects have been supervised mainly by Ingrid Glad as part of Big Insight, a Centre for Research-Based Innovation funded by the Norwegian Research Council (project 237718). This thesis is the final result, consisting of four articles encapsulated by an introduction to provide context and background.

Being a PhD student has been an invaluable and enjoyable experience, although at times frustrating. Initially, I was supposed to get a head start by turning parts of my master's thesis into a paper during the first semester. Almost two years later, one paper had grown into two (Paper I and Paper II), and I never seemed to be satisfied with them. At this point, I was invited to join an applied Big Insight project in collaboration with the Norwegian Computing Centre and ABB on monitoring the temperature of ship motors (Paper III). Simultaneously, Ingrid got me in contact with Idris Eckley at Lancaster University. He and Paul Fearnhead were kind enough to welcome me to Lancaster as part of the Statscale program during the autumn of 2019 for a highly productive, inspiring and slightly rainy three months, where they initiated and supervised me on Paper IV. In the end, I feel like all the work has come nicely together into the common themes of changes, anomalies, scalable computation and cross-correlation.

The root cause of this thesis is Ingrid, who sparked my interest in statistics and almost five years ago introduced me to changepoint models. She has been exceedingly supportive throughout and is always enthusiastic, positive and solution-oriented. I am grateful to my co-supervisor Nils Lid Hjort's keen eyes for details and for letting me into his office to answer questions whenever needed. Thanks to Idris and Paul for being open to collaboration, encouraging and interested, as well as for the regular supervision on Skype. I also thank Kristoffer Hellton and Morten Stakkeland for letting me play around with the ship sensor data, Solveig Engebretsen for running my code countless times at any time of the day, and the remaining co-authors Ola Haug and Magne Aldrin. It was great fun working with Jonas Moss on the kdensity R package, implementing the 25 year old doctoral work of Ingrid (supervised by Nils). The past three years have been vastly enriched—both socially and academically—by my fellow students and colleagues in the statistics group at the Mathematics Department in Oslo, Big Insight, and the Statscale room in Lancaster. For teaching me basic C++, I owe Daniel Grose a skiing lesson. Finally, I am grateful to my friends for helping me recharge my batteries over weekends, to Trude for enduring me during thesis work and lock-down, and to my family for always supporting what I do.

Martin Tveten
Blindern, January 2021

List of Papers

Paper I

Tveten, M. (2019). Which principal components are most sensitive in the change detection problem? *Stat*, 8(e252).

Paper II

Tveten, M. and Glad, I. K. (2019). Online detection of sparse changes in high-dimensional data streams using tailored projections. *Manuscript.*

Paper III

Hellton, K. H., Tveten, M., Stakkeland, M., Engebretsen, S., Haug, O. and Aldrin, M. (2020). Real-time prediction of propulsion motor overheating using machine learning. *Submitted for publication.*

Paper IV

Tveten, M., Eckley, I. A. and Fearnhead, P. (2020). Scalable changepoint and anomaly detection in cross-correlated data with an application to condition monitoring. *Invited to submit a revision to Annals of Applied Statistics.*

Additional paper

The following paper (considered outside of the thesis' scope) was also written during the doctoral training period:

Moss, J. and Tveten, M. (2019). kdensity: An R package for kernel density estimation with parametric starts and asymmetric kernels. *Journal of Open Source Software*, 4(42), 1566.

Contents

Chapter 1

Introduction

Both in science and industry, the sizes of data sets are growing. But without appropriate tools for turning the data into insight, the value of harvesting more data is severely limited. This has created a surge in demand for statistical methods capable of handling enormous data sets, both in the sense of offering reasonable computing time as well as being methodologically sound. That is, modern statistical methods should not only be consistent, powerful and accurate, but also computationally *scalable*.

Apart from consisting of many measurements, big data sets can be extremely diverse. In long, multivariate time series, the typical assumption of stationarity frequently does not hold in practice. The problem of detecting if and when some distributional properties of the data change over time has therefore found increasing applied interest in recent years. For example, Eckley et al. (2020) use change detection methodology to remotely detect the location of gas emission sources utilising data obtained from sensors mounted on an airplane, Gao et al. (2020) use it for monitoring the surface-temperature of organ transplants, and Lévy-Leduc and Roueff (2009) search for anomalies in large amounts of network traffic data. Other areas where detecting changes has become an integral part include software reliability engineering (Mendiratta et al., 2019), research on telecommunications networks (Bardwell et al., 2019) and econometrics (Hlávka et al., 2017).

The motivating application for this thesis is detection of anomalies in sensor-monitored machinery. In this setting, several sensors are placed on different locations of a machine, for instance a pump or a motor, to measure the temperature, pressure or other quantities of interest over time. The machine is monitored to detect if it is not operating as supposed to, either to optimise performance or to avoid costly or dangerous failures. This applied problem translates well to a statistical change detection problem, as a significant change in the sensor data relative to its normal behaviour often signals that something is off with the machine. For example, if the hourly mean temperature of a motor is higher than it normally is, this may indicate that something is wrong with the cooling system. An idealised example of such temperature monitoring is shown to the left in Figure 1.1. For illustrational purposes, there are only four sensors in this example, but in practice, there may be several hundred sensors making measurements every second. Monitoring the sensor readings by eye is therefore not feasible. In addition, subtler changes can be detected when combining the information across all the sensors in a principled way.

A feature of the sensor data encountered in the present thesis we particularly focus on is cross-correlation—correlation between the sensors at any given time, due to, e.g., the proximity of the sensors (Figure 1.1, right). Handling and

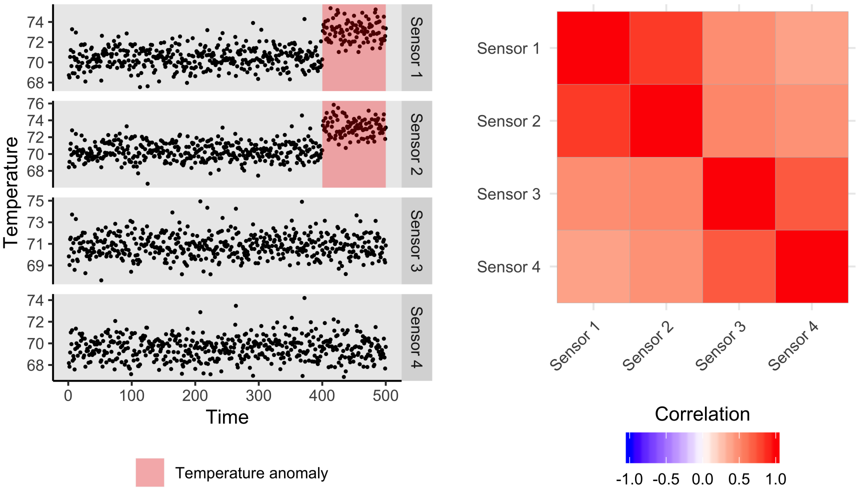

Figure 1.1: A multivariate times series of simulated temperature recordings from four imagined sensors on a ship's motor. Around time-point 400, a part of the cooling system breaks down. As a result, the temperature recordings of sensor 1 and 2 increase to a consistently higher level; the mean temperature has changed. The robustly estimated correlations between the sensors are shown in the matrix to the right. As some sensors are imagined to be relatively close to each other, the correlation between them is strong and positive. See Paper III and Paper IV for a similar type of anomaly detection in real data.

understanding the impact of cross-correlation when combining information from all sensors is important to obtain accurate and trustworthy results. Detecting changes in the correlation structure itself may also be of interest. Moreover, cross-correlation has received relatively little attention in the change detection literature so far, despite its near ubiquitous presence in high-dimensional time series.

There are two different modes of change detection resulting in different but related statistical problems. In *online* change detection, data are collected and analysed in real-time, and the aim is to control the rate of false alarms, but detect true changes as quickly as possible. *Offline* change detection, on the other hand, concerns the retrospective analysis of a historical data set, with the aim of accurately estimating the number and locations of changes. In the sensor-monitoring example, an online method would be used as the real-time monitoring system of the motor, while an offline method could be used to analyse and prepare a training data set for the online method.

We study both online and offline change and anomaly detection for cross-correlated, multivariate time series. Our contributions lie in the intersection of computation and methodology in the form of novel methods that are scalable to scenarios with many sensors or other variables. We also apply change detection methods to new real world problems. Throughout, the focus is mainly on frequentist methods and parametric models. However, alternatives outside this scope will be touched upon along the way in this introduction.

The rest of the thesis is organised as follows: Chapters 2 and 3 provide background material for putting the papers into context. Chapter 2 starts by formally defining the change detection problem in the offline setting. General computational and methodological frameworks are then introduced in the univariate setting as a stepping stone to the more complex multivariate methods. Next, the anomaly detection problem is presented as a special case of the change detection problem. We finish the chapter by pointing to methods and problems surrounding the scope of the thesis. Chapter 3 partly builds on Chapter 2 to introduce the online version of change detection in a similar fashion. Summaries of the four papers then follow in Chapter 4, emphasising their main contributions. In Chapter 5, I discuss parts of my work in more detail and point to important venues of future research. The four papers in full length conclude the thesis.

Before we continue, some general remarks on notation is due. For a compact presentation, we write $x_{s:e} := \{x_s, \dots, x_e\}$, where $s < e$. Bold types are used to indicate that an object is a vector rather than a scalar, for example $\mathbf{x}_{s:e} := \{\mathbf{x}_s, \dots, \mathbf{x}_e\}$, where $\mathbf{x}_t = (x_t^{(1)}, \dots, x_t^{(p)})^\intercal$. We also let $[n] := \{1, \dots, n\}$ for $n \in \mathbb{N}$.

Chapter 2

Offline change and anomaly detection

Offline change detection methods take as input a p-variate time series of fixed length, $\mathbf{x}_t$ for $t = 1, \ldots, n$, and aim to answer one or a mix of the four problems:

(P1) Is the data stationary or does its distribution change over time?

(P2) If there are changes, how many changes are there?

(P3) Given a number of changes, at what times do they occur?

(P4) Given the times of change, how does the distribution change?

To be precise, consider the following general model for $\mathbf{x}_1, \ldots, \mathbf{x}_n$: Let the *changepoints* $1 < \tau_1 < \ldots < \tau_K < n$ denote $K < n$ unknown time-points where the data-generating mechanism for $\mathbf{x}_t$ changes abruptly. As a consequence, the observations are divided into $K + 1$ stationary segments with different distribution functions $F_0(\mathbf{x}), \ldots, F_K(\mathbf{x})$. I.e., the data follow a piecewise stationary distribution given by

$$\mathbf{x}_t \sim F_k \quad \text{for } t = \tau_k + 1, \ldots, \tau_{k+1} \text{ and } k = 0, \ldots, K, \tag{2.1}$$

where we define $\tau_0 := 0$ and $\tau_{K+1} := n$. In this model, (P1) is the *testing* problem of whether $K = 0$ or $K > 0$, while (P2)-(P4) are the problems of *estimating* K, $\tau_1, \ldots, \tau_K$ and $F_0, \ldots, F_K$, respectively, preferably combined with measures of estimation uncertainty. Depending on the problem at hand, the ideal goal is to construct the most powerful test or the most accurate estimator.

In most of this thesis, we will not consider models quite as general as (2.1). Firstly, we will mostly work with real-valued vector observations that can be described by a parametric family of densities $f(\mathbf{x}|\boldsymbol{\theta})$, where f is constant, changes occur in the parameter vector $\boldsymbol{\theta}$. Now the model in (2.1) becomes

$$\mathbf{x}_t \sim f(\mathbf{x}|\boldsymbol{\theta}_k) \quad \text{for } t = \tau_k + 1, \ldots, \tau_{k+1} \text{ and } k = 0, \ldots, K, \tag{2.2}$$

where $\boldsymbol{\theta}_{k-1} \neq \boldsymbol{\theta}_k$ for all k. Secondly, we primarily focus on models where the $\mathbf{x}_t$'s are independent in time. Thirdly, as mentioned in the introduction, our focus lies on frequentist methods. Some Bayesian alternatives are given at the end of this chapter.

The prototypical setup is to let f be a normal density with mean $\boldsymbol{\theta}$, as detecting changes in the mean is arguably the most important problem in practice. Plenty of other setups exist, however, for example changes in variance (Hsu, 1977; Inclán and Tiao, 1994), covariance matrix (Wang et al., 2018), parameters of vector autoregressive models (Wang et al., 2020b), Poisson rates (Henderson and Matthews, 1993), parameters in exponential families (Worsley, 1986). Non-parametric methods for detecting general distributional changes as in (2.1) also exist (Pettitt, 1979; Csörgő and Horváth, 1988).

This chapter gives a brief overview of important methodological developments on the described offline change detection problem. There are a number of more comprehensive reviews in the literature that can be consulted for more details, for instance Truong et al. (2020), Aminikhanghahi and Cook (2017), Niu et al. (2016), Jandhyala et al. (2013), Chen and Gupta (2011), as well as in the theses of e.g. Tickle (2020), Maeng (2019) and Maidstone (2016). We begin by an introduction to some general ideas and frameworks for change detection in the univariate setting.

2.1 General ideas and frameworks—univariate data

Due to the literature on change detection being so vast, there are several ways of categorising all the different change detection methods. Following the review article of Truong et al. (2020) and the work of Killick et al. (2012), I have chosen to structure the exposition based on viewing the offline change detection problem as a problem of optimising a constrained or penalised cost. From this point of view, an offline change detection method consists of three elements: A cost for fitting observations to a specific model, $C(x_{s:e}) \geq 0$, a penalty or constraint for the complexity of the model to avoid overfitting, $P(\tau_{1:K}) \geq 0$, and a search procedure for solving

$$\min_{\tau_{1:K}} \left[\sum_{k=0}^{K} C(x_{(\tau_k+1):\tau_{k+1}}) + P(\tau_{1:K}) \right]. \tag{2.3}$$

The minimising arguments of (2.3) are the changepoint estimates $\hat{\tau}_{1:\hat{K}}$, where $\hat{K}$ is the estimated number of changepoints. In this section we think of the x_t's as univariate observations to fix ideas, but the general framework (2.3) easily carries over to the multivariate setting of Section 2.2.

Note that within this framework, (P1) is answered implicitly through the estimates $\hat{\tau}_{1:\hat{K}}$; the null hypothesis of stationarity is accepted if $\hat{K} = 0$ and rejected otherwise. (P2) and (P3) are solved directly, while (P4) is often answered by construction of the cost function or by a post-processing step given the estimated segmentation.

The cost function is a measure of how well the observations fit the model—the lower the cost, the better the fit—and there is an abundance of costs with different properties available. A prominent example from the changepoint literature is the log-likelihood cost (e.g. Hinkley (1970), Gombay and Horvath (1994), Eckley et al. (2011) and Aue and Horváth (2013)), defined by

$$C(x_{s:e}) = -2 \sup_{\boldsymbol{\theta}} \sum_{t=s}^{e} \log f(x_t | \boldsymbol{\theta}) \tag{2.4}$$

for independent and identically distributed (i.i.d.) observations. Using the log-likelihood cost results in a penalised maximum likelihood approach to change detection. As in many other contexts, the maximum likelihood approach results in estimators with desirable properties, such as consistency of the estimated changepoints under the true model and certain regularity conditions (He and Severini, 2010). The maximum likelihood approach to offline change detection can be traced back to Hinkley (1970), who studied (P3) (estimating the location of a change) in the case of a single change in the mean of Gaussian data with known variance. Other examples of costs include quadratic loss (Chen and Gupta, 2011), absolute loss (Bai, 1995), outlier-robust costs (Huber, 2004; Hušková, 2013; Chakar et al., 2017; Fearnhead and Rigaill, 2019) and nonparametric costs (Zou et al., 2014b). A selection of common costs can be found in Truong et al. (2020).

The penalty function measures the complexity of a given changepoint model. It is essential in obtaining an accurate estimate of the number of changes, K, as it governs how much the cost must be reduced for it to be worth adding an additional changepoint, thereby increasing the model complexity. Excluding a penalty in the change detection problem with an unknown number of changes would result in maximal overfitting as the optimum of (2.3) would be to add a changepoint at every observation, i.e. $\hat{\tau}_{1:\hat{K}} = [n-1]$. The most common penalty function is linear in the number of changepoints; $P(\tau_{1:K}) = \beta K$. This penalty includes standard model selection tools like Akaike's information criterion (Akaike, 1974) when $\beta = 2d$ and the Bayesian or Schwarz' information criterion (Schwarz, 1978) when $\beta = d \log n$, where d is the number of additional parameters in the model per changepoint added. An example of a non-linear penalty that is tailored to the change in mean problem is the modified Bayesian information criterion (Zhang and Siegmund, 2007), given by $P(\tau_{1:K}) = 3K \log n + \sum_{k=0}^{K} \log\big((\tau_{k+1} - \tau_k)/n\big)$. This penalty favours models with evenly spaced changes. More examples of penalties will emerge as we go along in this chapter.

When it comes to search methods, there are particularly two popular classes of algorithms we will treat in more detail. The first approach is based on *model selection* and solves (2.3) *exactly* by a dynamic programming scheme. The second and oldest approach solves (2.3) *approximately* by recursively applying *tests* for the existence of a single changepoint to narrower and narrower windows of the data. After presenting these two classes of algorithms, we go on a quick tour of notable alternatives.

Dynamic programming-based methods Multiple change detection methods based on dynamic programming define recursions for finding the exact optimum of (2.3). The optimal partitioning method of Jackson et al. (2005) is a cornerstone among such algorithms. It can only be used for linear-in-K penalties, but in return, it finds the optimum in $O(n^2)$ time, provided computation of the cost does not depend on n. This is the case for most costs as long as independence between observations in different segments is assumed. The key to optimal partitioning is to define $F(t)$ as the optimal penalised cost for data $x_{1:t}$. It starts by $F(0) = -\beta$, and then proceeds by computing

$$F(t) = \min_{i<t} \big\{ F(i) + C\big(x_{(i+1):t}\big) + \beta \big\}. \tag{2.5}$$

The optimal cost is given by $F(n)$.

Although a reduction from exponential to quadratic in n computing time is remarkable, it is still prohibitive for sufficiently large n. Motivated by this, the pruned exact linear time (PELT) algorithm of Killick et al. (2012) refines optimal partitioning by only considering relevant i's in the minimisation in (2.5) at each step t. This is made possible by the observation that adding a changepoint always reduces the cost. Therefore, if at time $t_2 > t_1$, the inequality

$$F(t_1) + C(x_{(t_1+1):t_2}) + \beta \geq F(t_2) \tag{2.6}$$

holds, then t_1 can never be the most recent changepoint for all $t_3 > t_2$. In other words, t_1 can be "pruned" from the set of candidate changepoints after time t_2. The effect of pruning in practice is roughly to automatically discard times before a true changepoint. Consequently, PELT can scale linearly in n if the expected number of true changepoints also scales linearly with n, but it remains quadratic like optimal partitioning in the worst-case scenario of no changes. Parallelisation can further reduce the computational burden (Tickle et al., 2020), though at the price of sacrificing exactness of the solution. Even without parallelisation, the computational savings achieved by PELT is massive for many practical problems, making it an increasingly popular method. We also derive a PELT type algorithm in Paper IV.

If only changes in a single parameter is of interest, a very fast alternative to the inequality type pruning in PELT is so-called functional pruning in the functional pruning optimal partitioning algorithm of Maidstone et al. (2017). This type of pruning results in a substantial increase in candidate changepoints being pruned, irrespective of the true number of changes present. Functional pruning optimal partitioning can also be used to fit models where parameters are dependent across segments, as opposed to PELT.

As noted, optimal partitioning, PELT and functional pruning optimal partitioning can only be used with a linear penalty. If a non-linear penalty is preferred, the segment neighbourhood algorithm of Auger and Lawrence (1989) is an alternative. Segment neighbourhood passes through the data recursively as optimal partitioning, but also conditions on the number of changepoints in a particular segment. That is, it starts by computing the optimal segmentation for a single change, before recursively updating the optimal segmentation for one added change until a user-input maximum number of changes $\overline{K} < n$ is reached. Consequently, segment neighbourhood requires $O(\overline{K}n^2)$ operations to find the optimum. If K is completely unknown, this means cubic scaling in n, which limits its use to small data sets. As for optimal partitioning, the speed of segment neighbourhood can be improved by pruning techniques (Rigaill, 2010; Maidstone et al., 2017). Using a linear penalty with PELT or functional pruning optimal partitioning, however, remains a vastly more computationally viable option for large data sets.

Binary segmentation-based methods Another large class of multiple change detection algorithms emerges from the following idea: Let $T(\tau, x_{1:n})$ be a test statistic for a changepoint at τ in the series of observations $x_{1:n}$. This could be any of your favourite tests for a difference in distribution between the sample $x_{1:\tau}$ and $x_{(\tau+1):n}$—a t-test or likelihood ratio test for example. A natural test for the presence of a single changepoint is then to compute $\hat{T}(x_{1:n}) = \max_\tau T(\tau, x_{1:n})$ and compare it with a threshold b. If $\hat{T}(x_{1:n})$ is above b, a change is detected and estimated to be located at the maximising changepoint, $\hat{\tau}$. By splitting the sample at $\hat{\tau}$, the same procedure can be applied to each of the two segments $x_{1:\hat{\tau}}$ and $x_{(\hat{\tau}+1):n}$ to identify further changes, and so forth on each segment as long as the test is significant. This is the binary segmentation algorithm and it "is arguably the most established search method used within the changepoint literature" (Killick et al., 2012). It is often attributed to Vostrikova (1981), Scott and Knott (1974) and Edwards and Cavalli-Sforza (1965).

The way binary segmentation approximates the optimisation problem (2.3) becomes more apparent by considering test statistics of the form

$$T(\tau, x_{1:n}) = C(x_{1:n}) - C(x_{1:\tau}) - C(x_{(\tau+1):n}). \tag{2.7}$$

For a log-likelihood cost, (2.7) is the likelihood ratio test. Maximising this test over τ is the same as finding the single changepoint which provides the greatest decrease in cost. The threshold b governs how much the cost must be reduced when adding a changepoint for it to be considered a change, and can thus be viewed as a linear penalty in the number of changepoints.

There are at least three advantages of using binary segmentation. Firstly, it is computationally fast, only requiring $O(n \log n)$ operations. Secondly, it is easy to implement and modular. Thirdly, it is conceptually simple as it essentially reduces the multiple changepoint problem to a single changepoint problem, which can be further reduced to a (multiple) testing problem. Binary segmentation has also been shown to be consistent (Venkatraman, 1993) in scenarios where adjacent changepoints are sufficiently far apart. In total, this makes binary segmentation applicable to a wide range of old and new change detection problems. All that is needed is a test statistic for discriminating between distributional features of interest.

The main disadvantage of binary segmentation is so-called masking, which is due to its particular approximative nature. A typical example is when changes occur frequently and two close-by changes cancel each other out in the test for a single change. Generally, masking refers to change scenarios where at least one change is missed.

As a result, several tweaked versions of binary segmentation have recently been proposed to make it robust to a larger range of changepoint configurations. Circular binary segmentation of Olshen et al. (2004) is an early modification for detecting changes that switch back and forth between two distributional regimes. Later, the wild binary segmentation algorithm of Fryzlewicz (2014) has drawn much attention as it provably provides error-rate-optimal changepoint estimates (both (P2) and (P3)) in a certain sense (Wang et al., 2018, 2020a). Rather than deterministically splitting each segment at the optimal single changepoint, wild binary segmentation draws intervals at random to search for a single change. Achieving the mentioned optimal rates, however, may require a very large amount of intervals, hence losing the computational advantage over the exact search methods, like PELT. Recent further improvements include wild binary segmentation 2 (Fryzlewicz, 2020), the narrowest-over-threshold method (Baranowski et al., 2019) and seeded binary segmentation (Kovács et al., 2020).

It should be mentioned that the most popular test statistic to use within binary segmentation is the cumulative sum (CUSUM) statistic. It can be traced all the way back to the first articles on change detection by Page (1954, 1955), who considered the online version of (P1) (testing for the presence of a change) in the context of industrial quality control. Hinkley (1971) later considered (P3) (estimating the location of a change) for Page's CUSUM in the offline setting with a single change.

In modern offline change detection literature (e.g. Wang and Samworth (2018); Fryzlewicz (2014); Aue and Horváth (2013)), the CUSUM statistic mostly does not refer to Page's CUSUM, but to the statistic

$$T(\tau, x_{1:n}) = \sqrt{\frac{\tau(n-\tau)}{n}} \left(\frac{1}{n-\tau} \sum_{t=\tau+1}^{n} x_t - \frac{1}{\tau} \sum_{t=1}^{\tau} x_t \right). \tag{2.8}$$

This statistic is equivalent to the positive root of the likelihood ratio statistic for a single change at τ in the mean of Gaussian data with known variance, and it serves as a blueprint for many other change detection tests. For example, Inclán and Tiao (1994) derive a test for a change in the variance by using cumulative sums of x_t^2, and Lee et al. (2003) further extend this idea by switching x_t in (2.8) with an appropriate function $g(x_t)$ for detecting a general parameter of interest. The simple form of CUSUM tests is what drives their popularity, as it facilitates both quick computation and theoretical analysis. An important result is that a large class of CUSUMs converge in distribution to a Brownian bridge (e.g. Lee et al. (2003)), which is helpful for tuning the threshold b in certain scenarios.

Not all CUSUMs fit nicely into the story of costs, penalisation and search methods. However, some CUSUMs are related to likelihood ratios (Inclán and Tiao, 1994) and squared error loss. As such, they can be viewed as another layer of approximation in (2.3) in addition to binary segmentation. Despite being approximative in general, the theoretical results on the consistency and optimality of wild or plain binary segmentation mentioned here use CUSUM type test statistics (Venkatraman, 1993; Wang et al., 2018, 2020a).

Other search methods There is a growing number of search methods and approaches apart from those we have seen so far based on dynamic programming and binary segmentation. We now briefly present a selection of these alternatives.

Binary segmentation can be described as a "top-down" search method as it starts with the entire stretch of data, before splitting it into smaller and smaller pieces. A natural alternative is therefore a "bottom-up" search method, where one initially starts with a changepoint at every observation, before merging segments until some criterion is met. Such methods are still new to the change detection field, only recently having been explored by Matteson and James (2014) and Fryzlewicz (2018). These articles, however, suggest that such methods can be competitive with binary segmentation type methods, especially in scenarios with frequent changes.

Another alternative set of methods related to binary segmentation are moving sum methods, proposed for change detection by Preuss et al. (2015) and Eichinger and Kirch (2018), building on similar approaches to testing, e.g. Hušková and Slabý (2001). Moving sum methods, like binary segmentation, are based on testing for a single changepoint, but do so by sliding a window of a certain bandwidth across the time series, testing for a change at the window's midpoint. Given an appropriate bandwidth, moving sum methods can also be shown to be consistent for the number and location of changes, and are quick to compute as well as conceptually simple. Their main drawback is that performance crucially depends on a well-tuned bandwidth parameter.

Other model selection approaches also exist, where the simultaneous multiscale changepoint estimator for detecting changes in the mean proposed by Frick et al. (2014) has recieved much attention. Their take on the change detection problem is to minimise the number of changepoints over all potential piecewise constant mean signals within the acceptance region of a multiscale test. They show that this corresponds to a certain penalised cost, facilitating quick computation, and prove that the family-wise error rate of the number of estimated changes is controlled. Moreover, confidence sets for the locations of the changepoints as well as the piecewise constant mean can also be constructed. Pein et al. (2017) extend the simultaneous multiscale changepoint estimator to heterogeneous data, and Li et al. (2016) propose a related method for controlling the false discovery rate rather than the family-wise error rate, as control of family-wise error rate often leads to underestimating the number of changes. Unfortunately, the framework underpinning these multiscale methods only works for univariate data.

A model selection penalty that is linear in the number of changepoints is connected to an L_0-penalty on the sums of differences of a piecewise constant mean. Harchaoui and Lévy-Leduc (2010) exploit the link between L_0 and L_1 penalisation to create a computationally efficient changepoint estimator, similar to the famous LASSO regression estimator (Tibshirani, 1996). However, the L_1-penalty does not balance type I and type II error optimally for change detection (Cho and Fryzlewicz, 2011).

The final class of change detection methods based on model selection we mention is the data-driven penalty selection methods based on "slope heuristics" of Birgé and Massart (2001, 2007), described in Baudry et al. (2012). These methods aim to automatise tuning of penalties, which is often a delicate problem in practice. Their detection performance is good, but they are restricted to small data sets due to poor computational scaling in the sample size.

2.2 Multivariate methods

Data recordings are increasingly often multivariate and high-dimensional rather than univariate in the current "big data" era. This has led to a massive growth in research on multivariate change detection methods over the past ten years. Before reviewing a selection of the literature, we highlight some of the additional challenges connected to multivariate changepoint analysis compared to the univariate setting.

A naive way of detecting multivariate changes is to apply a univariate method to each time series and put a changepoint at each time-point the ensemble of univariate methods detects a change. However, such an approach would suffer from many false positives due to multiple testing, it does not account for dependence between the variables, and it is not be able to borrow strength across signals to detect changes that are small in each variable, but large when seen as a whole. Moreover, the ensemble of univariate methods might not scale well computationally as the number of variables, p, grows. These are the main reasons for taking what we can call a "fully" multivariate changepoint approach.

Now recall the problem formulation in this chapter's introduction, the changepoint models (2.1) and (2.2) in particular. The space of possible distributions per segment, F_k, is now vastly more complex; imagine the possibility of different marginal distributions per variable and different forms of dependence between them. Even under a family of parametric models $f(\mathbf{x}|\boldsymbol{\theta})$, the number of choices for f and ways in which $\boldsymbol{\theta}$ can change becomes exponentially larger in p. A specific additional question in the multivariate setting that has been addressed in the literature (e.g. Jirak (2015) and Fisch et al. (2019b)), and we pursue in this thesis, is the following:

(P5) Given that there is a change, which of the p variables change?

In the case where $\theta_k^{(i)}$ is the k'th segment mean for variable i, for example, the aim is to estimate the subsets $\mathbf{J}_k \subseteq [p]$ of non-zero elements in $\boldsymbol{\theta}_k - \boldsymbol{\theta}_{k-1}$ for $k = 1, \ldots, K$. Indicating which variables change is important to be able to diagnose what the cause of a change may be.

Complicating things further, there is a big difference between trying to detect changes that occur in more or less than $c\sqrt{p}$ variables, for some non-zero constant c (see e.g. Enikeeva and Harchaoui (2019), Cai et al. (2011) or Jeng et al. (2013)). If more than or exactly $c\sqrt{p}$ variables change, we are in a *dense* regime, and if less than $c\sqrt{p}$ variables change, we are in a *sparse* regime. The intuition behind there being two regimes can be explained as follows: In the dense regime, many variables change such that it is beneficial to aggregate information equally across all variables in the search of a change. If this type of aggregation is used in the sparse regime, on the other hand, the noise from the non-changing variables is more likely to drown out the signal from the few changing variables, making the detection problem harder. The boundary between the two regimes just happens to be at $c\sqrt{p}$ in the limit as $p \to \infty$ for changes in the mean of i.i.d. Gaussian observations with known variance. The consequence is that different methods are optimal for separating the null hypothesis of no change from sparse and dense alternatives, respectively. In addition, it is primarily in the sparse regime it is relevant to ask (P5). It is likely that a boundary between sparse and dense changes also exists for other types of changes and data distributions, but the the exact nature of such a general law is an open problem, to the best of my knowledge.

Changes in the mean As in the univariate setting, changes in the mean vector is the most well-studied problem. Early contributions all consider tests for a single, dense change. As we have seen in Section 2.1, all such tests can be embedded in a binary segmentation type algorithm to detect multiple changes. Srivastava and Worsley (1986) study the likelihood ratio test for a single change in the mean of multivariate Gaussian data when the correlation matrix is unknown but constant. Horváth et al. (1999) later consider a scaled version of the same statistic, but derives its limiting distribution under a more general model with temporally m-dependent noise.

A large portion of modern work concentrates on the problem of testing for a single change, but from a high-dimensional angle. This either means that $p \to \infty$ in theoretical analysis of the method, or that interest lies on methods that are computationally scalable to potentially very large p. Many such tests are based on aggregating information across local test statistics per variable, $T(\tau, x_{1:n}^{(1)}), \ldots, T(\tau, x_{1:n}^{(p)})$, where $T(\cdot, \cdot)$ often is the CUSUM (2.8), but could in principle be any test. Early high-dimensional work focused on models assuming independence between variables $i = 1, \ldots, p$—what we call *cross-independence*—and assumed that the change is dense. For example, Bai (2010), Horváth and Hušková (2012) and Zhang et al. (2010) all propose an L_2-aggregation of their local statistics under these assumptions. The two former allow for temporal dependence and deal with estimation of a change whose presence is known *a priori*, i.e., (P3) assuming that a change has occured somewhere. Zhang et al. (2010) consider the testing problem (P2) and formulate a model where the change is allowed to be sparse, but their test statistic does not deal with the potential sparsity of the change, nor (P5).

Subsequently, the problem of detecting sparse changes in cross-independent models received increasing attention, as in many practical problems it is clear that only a few variables are likely to be affected. A typical example is the detection of DNA copy number variants, where some variants might only be shared across a few samples. Siegmund et al. (2011) incorporated a prior guess p_0 on the fraction of affected variables. Cho and Fryzlewicz (2015) use a hard-thresholded L_1-aggregation of local CUSUM statistics. Jirak (2015) proposes an L_∞-aggregation, i.e., the maximum of the absolute local CUSUM statistics, and is the first to study (P5). Enikeeva and Harchaoui (2019) propose a statistic based on ordered local CUSUM statistic in combination with an L_2-aggregated CUSUM test to obtain optimal rates of convergence for both sparse and dense changes in independent Gaussian data. Cho (2016) suggests to aggregate the ordered local CUSUMs by another coordinate-wise CUSUM transformation. Lastly, Wang and Samworth (2018) derive an optimal projection (i.e., aggregation) of CUSUMs, and offer a consistent estimator of this projection direction by a sparse singular value decomposition on the CUSUM transformed data. Note that Jirak (2015) and Wang and Samworth (2018) also extend their methods to allow for cross-dependence.

There are few penalised cost-based methods for the high-dimensional setting. Two contributions in this direction are Fisch et al. (2019b) and Tickle (2020, Chapter 4), who derive methods for detecting both sparse and dense changes in cross-independent data that are easy to adapt to any parametric model for the marginal distributions.

Most recent high-dimensional literature considering cross-dependent data focus on dense changes (Westerlund, 2019; Bhattacharjee et al., 2019; Li et al., 2019; Wang and Shao, 2020). An interesting exception is Maeng (2019, Chapter 5), who also considers temporal dependence, but does not estimate which variables are affected (problem (P5)). An approach for detecting both sparse and dense changes in the mean of cross-correlated data that is computationally scalable and indicates which variables are affected is generally missing in the literature. We aim to fill this gap by a penalised cost approach in Paper IV.

Changes in the covariance matrix Assessing stability of the covariance matrix of multivariate observations has gained significant recent interest. One reason is that many methods for detecting changes in the mean assume that the covariance matrix is constant over time. The thorough analyst should therefore assess whether this assumption holds. Changes in the covariance matrix—or, equivalently, the precision matrix—may also be of independent interest. Kao et al. (2018), for instance, list several practical problems within finance and economics where this is the case.

Methods for detecting changes in the covariance matrix were first proposed for quality control purposes, e.g. the Gaussian likelihood ratio approach of Sullivan and Woodall (2000) or other control charts (see the review article of Yeh et al. (2005)). An early maximum likelihood treatment of the multiple changes in mean and covariance matrix problem is Maboudou-Tchao and Hawkins (2013), who additionally use the segment neighbourhood algorithm as their search method. Even though it is not connected to a specific publication, note that it is relatively straightforward to plug the Gaussian likelihood with unknown mean and covariance matrix and a linear penalty into the penalised cost (2.3) and optimise with PELT, for instance.

The CUSUM-based work of Aue et al. (2009) marks the starting point of the modern, more theoretically oriented line of research on offline covariance change detection methods. Their method and analysis is impressive as it also considers temporal dependence. Bai (2010) considers changes in the variances (in addition to the means), but not in a general covariance matrix. Later, CUSUM-based methods for covariance changes have been investigated by Cho and Fryzlewicz (2015), Kao et al. (2018), Wang et al. (2018) and Dette et al. (2020). All these methods assume that the mean is constant and the change is dense, except the very recent work of Dette et al. (2020), where potential sparsity is addressed.

Other recent approaches are proposed by Roy et al. (2017), who consider changes in sparse Markov random field models, which includes sparse precision matrices in Gaussian data as a special case, Avanesov and Buzun (2018), who offer a moving sum-based method applicable both in the offline and online setting, and Wang et al. (2019), who utilise U-statistics and self-normalisation to detect changes in both the mean and covariance matrix. Lastly, Grundy et al. (2020) propose a method for detecting changes in the means and variances of high-dimensional (Gaussian) data by mapping the data into two dimensions—one highlighting changes in mean, and the other highlighting changes in the variance.

Research on changes in high-dimensional covariance matrices is still on an infant stage compared to changes in the mean. The $p(p-1)/2$ parameters involved makes the problem much tougher computationally, and almost all published work has only considered the scenario of dense changes. In Paper I and Paper II we investigate how the classical principal component analysis can be used to alleviate the computational burden. We also consider sparse changes in the covariance matrix.

Changes in other features In many practical situations it can be hard to know both the distribution of the data as well as exactly what type of distributional change is of interest. Hence, deriving nonparametric methods for detecting changes in multivariate data is a hot topic. Needless to say, this is a hard problem in general, both theoretically and computationally, but even more so in high-dimensional settings where the curse of dimensionality kicks in. Be aware that nonparametric methods can be used for detecting the already discussed changes in mean and covariance matrix, but is expected to be less powerful compared to methods specifically made for a particular type of change.

Examples of contemporary multivariate nonparametric change detection methods are the approach based on hierachical clustering and distance measures of Matteson and James (2014), the kernel-based methods of Harchaoui and Cappe (2007), Arlot et al. (2019) and Padilla et al. (2020), the graph-based methods of Chen and Zhang (2015), Chu and Chen (2019) and Liu and Chen (2020), as well as Zhang et al. (2017), who use energy statistics and the Kolmogorov-Smirnov test. Note that all these methods assume that observations are independent in time, and no distinction is made between sparse and dense changes.

We also remark that detection of changes in the quite general class of vector autoregressive models is investigated in Kirch et al. (2015), Safikhani and Shojaie (2020) and Wang et al. (2020b). In addition, Liu et al. (2020) very recently proposed a framework based on U-statistics and CUSUMs for detecting a change in any high-dimensional parameter, with power against sparse and dense changes simultaneously.

2.3 Changepoint-based anomaly detection

One of the many applications of changepoint models is anomaly detection. That is, detecting significant deviations from some baseline behaviour of the data. For example, Olshen et al. (2004) use a changepoint model to detect DNA copy number variations, which might indicate cancer or other diseases; Fisch et al. (2019a) detect an exoplanet based on inferring changes in lightcurve data from a star; and we detect overheating of a ship's propulsion motor in Paper III.

The general changepoint models (2.1) or (2.2) are only useful for detecting certain types of anomalies. In the comprehensive review of Chandola et al. (2009), anomalies are divided into three classes: Global, contextual and collective (the names of the classes are from Fisch et al. (2019a)). Global and contextual anomalies are defined as single observations not conforming to either the global or local pattern of the data. E.g., a temperature measurement of 40°C in Oslo is a global anomaly as it would be a highly unusual temperature any time of the year, whereas a measurement of 10°C would only be a contextual anomaly during the winter. Following the terminology of Fisch et al. (2019a,b), we call both global and contextual anomalies *point anomalies* as they are both single outlying observations. Collective anomalies are collections of related observations that are anomalous only when viewed together. For example, an average temperature of 13°C during April in Oslo, compared to the normal of around 10°C. It is primarily collective anomalies the general changepoint models are capable of detecting, while the presence of point anomalies is known to cause trouble in the form of inaccurate additional changepoints being added (Fearnhead and Rigaill, 2019). In addition, the general changepoint model does not utilise the fact that there is a common baseline distribution for the data in many anomaly detection applications.

On the other hand, classical outlier detection techniques and many existing anomaly detection methods from the machine learning community are not suitable for detecting collective anomalies (Chandola et al., 2009). These methods are made with the aim of detecting point anomalies, and often does not consider the relatedness of observations, for example their time-ordering.

Based on these observations Fisch et al. (2019a,b) develop the penalised cost-based framework *collective and point anomalies* (CAPA) for jointly detecting both point and collective anomalies. The anomaly model first assumes that $\mathbf{x}_t$ has a baseline distribution $f(\mathbf{x}|\boldsymbol{\theta}_0)$. Each of the K anomalies are then modelled by two changepoints; one change from the baseline distribution at time s_k, and one change back at time e_k, where $\{(s_k, e_k]\}_{k=1}^{K}$ form non-overlapping intervals. Such changepoints are known as *epidemic* changepoints in the literature (Kirch et al., 2015). This model can be described by

$$\mathbf{x}_t \sim \begin{cases} f(\mathbf{x}|\boldsymbol{\theta}_1) & \text{for } t = s_1 + 1, \dots, e_1 \\ \vdots & \\ f(\mathbf{x}|\boldsymbol{\theta}_K) & \text{for } t = s_K + 1, \dots, e_K \\ f(\mathbf{x}|\boldsymbol{\theta}_0) & \text{otherwise,} \end{cases} \tag{2.9}$$

where $\boldsymbol{\theta}_k \neq \boldsymbol{\theta}_0$ for $k = 1, \dots, K$, $s_k < e_k$ and $s_{k+1} \geq e_k$. In this model, point anomalies are simply defined as anomalies of length 1, i.e., when $s_k = e_k$, while collective anomalies have length greater than 1; $e_k - s_k \geq 2$. To distinguish the two cases, let $\{(s_k, e_k)\}_{k=1}^K$ refer to the collective anomalies, while $O \subseteq [n]$ denotes the set of point anomaly locations. As in the general changepoint model, the aim is to estimate K, $\{(s_k, e_k)\}_{k=1}^K$ and O, as well as $\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_K$. The baseline parameter $\boldsymbol{\theta}_0$ is assumed to be known, but it is estimated robustly from the data in practice.

Inference on the positions of the anomalies from data is done by using a PELT type algorithm for efficiently solving

$$\max_{K, \{(s_k, e_k)\}_{k=1}^K, O} \left[\sum_{k=1}^K S(s_k, e_k) + \sum_{t \in O} S'(\mathbf{x}_t) \right], \tag{2.10}$$

subject to $e_k - s_k \geq 2$ and no overlap between the intervals specified by $\{(s_k, e_k)\}_{k=1}^K$ and O. In (2.10), $S(s, e)$ is the *penalised saving* for introducing an anomaly, defined as the cost-based test statistic

$$S(s, e) := C(\mathbf{x}_{(s+1):e}, \boldsymbol{\theta}_0) - \min_{\boldsymbol{\theta}} C(\mathbf{x}_{(s+1):e}, \boldsymbol{\theta}) - \beta, \tag{2.11}$$

where β is a penalty for adding an anomaly. $S'(\mathbf{x}_t)$ is the penalised saving for adding a point anomaly at t, and is defined as $S(t-1, t)$, but with a separate penalty β'. Note that maximising the penalised savings in (2.10) is equivalent to minimising the penalised cost. Also, $S(s, e)$ with the log-likelihood cost corresponds to the likelihood ratio test of whether $\mathbf{x}_{(s+1):e}$ has parameter $\boldsymbol{\theta}_0$ or not, with threshold β. Fisch et al. (2019a,b) derive penalties for collective and point anomalies based on controlling the false positive rate in independent Gaussian data. In practice, the penalty can be tuned to achieve a desired false positive rate on a training set consisting exclusively of baseline observations, if available.

The article of Fisch et al. (2019b) concerns anomaly detection in multivariate data, where it might be that only a sparse subset $\mathbf{J}_k \subseteq [p]$ of variables are anomalous, as in the general changepoint model. In this case, the penalty in (2.11) is switched with a penalty function $P(|\mathbf{J}|)$ such that the method becomes powerful for detecting both sparse and dense anomalies. In Paper IV, we extend their method by allowing explicit modelling of cross-dependence.

It should be noted that several other authors tackle the problem of detecting epidemic changes, for instance Olshen et al. (2004), Zhang et al. (2010), Kirch et al. (2015), Aston and Kirch (2018), and Zhao and Yau (2019). Methods from sparse mixture detection are also suitable for detecting epidemic changes, e.g. Jeng et al. (2013) who utilise the higher-criticism test of Donoho and Jin (2004). Yet other methods aim to be robust against outliers (Fearnhead and Rigaill, 2019), or include inference regarding point anomalies (Maeng and Fryzlewicz, 2019).

2.4 Other approaches and related problems

So far in this chapter, we have covered frequentist methodology for detecting abrupt changes in piecewise stationary data, where the changes are aligned across variables in the multivariate setting. We will finish by pointing to important related work outside this scope.

There are several directions of Bayesian changepoint analysis. One school of thought formulates the changepoint problem as a hidden Markov model with a fixed number of states, each state corresponding to a stationary segment between changes (Chib, 1998). Inference is done by Markov chain Monte Carlo (MCMC) and if the number of changepoints is unknown, reversible jump MCMC (Green, 1995) can be used to explore the model space. More recently, Ko et al. (2015) proposed to use a Dirichlet process prior on the transition probabilities of the hidden Markov model, avoiding the prespecification of the number of states, and allowing for uncertainty measures both on the number and locations of changepoints.

Another class of Bayesian changepoint methods uses the product-partition model, of which prominent examples are Barry and Hartigan (1993) and Fearnhead (2006). Here, the prior is put on the time between changepoints instead of the transition probabilities. These approaches seek to avoid the difficulties of setting up appropriate MCMC algorithms, and rather build models that allow for quick and exact simulation from the posterior distribution of the number and locations of changepoints. Bardwell and Fearnhead (2017) recently proposed such a Bayesian method for detecting possibly sparse anomalous segments. We will also mention a few examples of related Bayesian online methods at the end of Chapter 3.

Somewhere between frequentist and Bayesian statistics lie methods for constructing confidence distributions (Schweder and Hjort, 2016). That is, distributions over the parameter space that can be used to visualise confidence intervals at all confidence levels simultaneously. Cunen et al. (2018) propose a framework for constructing confidence distributions for a single changepoint. As the literature on obtaining uncertainty measures for changepoints outside the Bayesian school is scarce, such methods could prove to be valuable.

When it comes to detecting changes in other models than covered here and changes of different types, the literature is growing. Examples include detecting changes in the covariates of regression models (Maeng, 2019; Lee et al., 2016; Leonardi and Bühlmann, 2016), changes in network models (Zhao et al., 2019; Bhattacharjee et al., 2020), multivariate changes that does not align perfectly in time between variables (Fisch et al., 2019b; Bardwell et al., 2019; Eckley et al., 2020), as well as fitting piecewise linear models rather than piecewise constant ones (Fearnhead et al., 2019; Maeng and Fryzlewicz, 2019).

Chapter 3

Online change and anomaly detection

In the online mode of change detection, observations are processed *sequentially* as they arrive, as opposed to the offline setting where an entire data set is collected before analysed *retrospectively.* Looking back at problems (P1)-(P5) posed for offline methods, online methods are primarily concerned with updating inference regarding (P1)—testing whether a change has occured or not—for every new observation $\mathbf{x}_t$ given inference based on $\mathbf{x}_1, \ldots, \mathbf{x}_{t-1}$, potentially for $t \to \infty$. The aim is to detect that a true change has occured as quickly as possible, while controlling the rate of false alarms if not. When a change has been declared, offline methods can be used to answer the remaining questions (P2)-(P5). Nevertheless, online methods typically also output an estimate of the most recent changepoint and how the distribution has changed as a byproduct of testing for the presence of a change.

The vast majority of existing online change detection methods are constructed for solving some version of the following sequential hypothesis testing problem:

$$\begin{aligned} &H_0 : \mathbf{x}_t \sim F_0 \text{ for } t = 1, 2, \ldots. \\ &H_1 : \text{There is a } \tau \geq 0 \text{ such that} \\ &\qquad \mathbf{x}_t \sim F_0 \text{ for } t = 1, \ldots, \tau, \\ &\qquad \mathbf{x}_t \sim F_1 \text{ for } t = \tau + 1, \tau + 2, \ldots, \end{aligned} \tag{3.1}$$

where $\tau = 0$ refers to the alternative hypothesis of all observations stemming from F_1. Note that this is the same model as (2.1) with $K \in \{0, 1\}$ and $n \to \infty$. It is typically assumed that there is a training set of m observations known to be generated from F_0 available. Most commonly, this training set is used to pre-estimate F_0, before considering F_0 to be known in the sequential problem (3.1). Alternatively, F_0 is assumed unknown and its estimation brought into the sequential problem to account for its estimation uncertainty, in which case the training set is taken as the first m observations in (3.1) and the restriction $\tau \geq m$ added to H_1. F_1 can also be modelled as either known or unknown, depending on the situation. As in the offline chapter, we primarily concentrate on the parametric problem where F_k has a parametric density $f(\mathbf{x}|\boldsymbol{\theta}_k)$, $k = 0, 1$.

We remark that in the online context, the difference between an anomaly and a change introduced in Section 2.3 is not as useful due to F_0 being thought of as a baseline distribution in either case. Thus, when we use “changes” in this chapter, we might just as well have used “anomalies”.

A sequential or online change detection method for solving (3.1) is a stopping time $N \in \mathbb{N} \cup \{\infty\}$. All methods we consider are of the form

$$N = \inf\{t \geq 1 : T(\mathbf{x}_{1:t}) > b_t\}, \tag{3.2}$$

where b_t is a threshold function governing whether a test for a change at time t, $T(\mathbf{x}_{1:t})$, is significant or not.

To specify what is meant by “controlling false alarms” and “quick detection”, let P^τ and E^τ denote probability and expectation under the model (3.1) when there is a true changepoint at τ. In particular, P^∞ and E^∞ mean that there is no changepoint and correspond to probability and expectation under H_0. A typical goal for a sequential method N is to find b_t such that the *average run length* (ARL) $E^\infty[N]$ is controlled at a user-specified level γ, and rank methods based on their (worst-case) *expected detection delay* (EDD), given by

$$\bar{E}^\tau[N] := \sup_\tau E^\tau[N - \tau | N > \tau]. \tag{3.3}$$

The lower EDD or response time, the better. The ARL can be viewed as the analog to controlling Type I error in the offline setting, while minimising EDD corresponds to maximising power. It is also a common goal to minimise the worst-worst-case EDD, due to Lorden (1971), defined as

$$\sup_\tau \operatorname*{ess\,sup}_{\mathbf{x}_1,\ldots,\mathbf{x}_\tau} E^\tau[(N - \tau)^+ | \mathbf{x}_1, \ldots, \mathbf{x}_\tau]. \tag{3.4}$$

However, it is often overly conservative and difficult to work with analytically, so the EDD in (3.3) has become more popular. Polunchenko and Tartakovsky (2012) can be consulted for a discussion on most classical performance measures.

A naive way of constructing a method for the online problem would be to apply one of the offline methods from Chapter 2 to the entire batch of data for every new observation. However, doing so results in a highly dependent and complicated multiple testing task, and as the sample size potentially goes to infinity, it is not feasible computationally. Thus, in addition to detecting changes quickly, an algorithm for online change detection should have computational complexity not depending on the current sample size t when updating inference from one observation to the next (Chen et al., 2020).

In the rest of this chapter, a brief overview of online change detection methods is given. The literature on online change detection is far sparser than its offline counterpart. Nevertheless, useful recent surveys include Aminikhanghahi and Cook (2017) and Polunchenko and Tartakovsky (2012), and the two books Siegmund (1985) and Basseville and Nikiforov (1993) give a thorough introduction to classical sequential methods. Our main focus is on methods that are related to the work in the papers of this thesis and fit within the online change detection framework just described. Section 3.1 introduces the most popular classical online methods in the univariate setting, before the multivariate setting is covered in Section 3.2. Section 3.3 provides pointers to recent research on related problems and methods outside the current scope.

3.1 Classical methods—univariate data

CUSUM methods CUSUM statistics play an equally important role in online as in offline change detection. As mentioned in Section 2.1, around (2.8), the CUSUM referred to in the online literature is not the same as in the offline literature, but they have a lot in common. Most importantly, both can be written in terms of cumulative sums and arise from likelihood ratio tests. The offline CUSUM originates from a likelihood ratio test between two unknown means in Gaussian data with known variance, whereas the online CUSUM of Page (1954) arises from a likelihood ratio test between two simple hypotheses;

$$T(x_{1:t}) = \max_{k<t} \sum_{i=k+1}^{t} \log \frac{f(x_i|\boldsymbol{\theta}_1)}{f(x_i|\boldsymbol{\theta}_0)}, \tag{3.5}$$

where $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_1$ are fixed pre- and post-change parameters. An online CUSUM method is then obtain by plugging (3.5) into (3.2), together with a constant threshold $b_t = b$ tuned to achieve an appropriate ARL.

A major contributor to the CUSUM's popularity is the fact that it can be written in the following recursive form:

$$T(x_{1:t}) = S_t = \left(S_{t-1} + \log \frac{f(x_t|\boldsymbol{\theta}_1)}{f(x_t|\boldsymbol{\theta}_0)} \right)^+, \tag{3.6}$$

where $S_0 = 0$ and $(\cdot)^+ := \max(0, \cdot)$. This recursion is obtained by viewing the CUSUM (3.5) as a repeated sequential probability ratio test with lower boundary 0 and upper boundary b (Basseville and Nikiforov, 1993, p. 38). Every time $T(x_{1:(t-1)})$ is below 0—i.e., the null hypothesis of no change is accepted—the test is restarted. In addition, the CUSUM's simple form facilitates theoretical analysis. As $t \to \infty$ the CUSUM behaves like a Brownian motion (Siegmund, 1985), which can guide the selection of the threshold b. It has also been proven that the CUSUM is optimal in terms of minimising the worst-worst-case EDD (3.4) asymptotically as the ARL $\gamma \to 0$ (Lorden, 1971), and for every $\gamma > 0$ (Moustakides, 1986).

The most problematic aspect of Page's CUSUM is that it not only assumes the pre-change distribution to be known, but also the post-change distribution, which is rarely the case in practice. A number of tweaks to the CUSUM have therefore been proposed since its initial release, aiming at adapting to unknown distributions while retaining the simple computational form. In Paper III, we use the post-change adapting CUSUM of Lorden and Pollak (2008) for detecting overheating in ship engines. Other examples of CUSUMs adapting to unknown pre- or post-change parameters are Pollak and Siegmund (1991) and McDonald (1990).

Generalised likelihood ratio methods An alternative class of online change detection methods for handling unknown parameters in both the pre- and post-change distribution are generalised likelihood ratio (GLR) methods. They incorporate maximum likelihood estimation of the unknown parameters. Hence, in the case of known pre-change parameter and unknown post-change parameter, GLR methods are defined by test statistics of the form

$$T(x_{1:t}) = \max_{k<t} \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \left[\sum_{i=k+1}^{t} \log \frac{f(x_i|\boldsymbol{\theta})}{f(x_i|\boldsymbol{\theta}_0)} \right], \tag{3.7}$$

where $\boldsymbol{\Theta}$ is a subset of the parameter space (see Basseville and Nikiforov (1993) or Lai (1995)). In the case of exponential families and composite alternative hypotheses, such statistics are optimal in the sense of Lorden (1971). Additionally assuming an unknown pre-change parameter brings us back to a statistic of the form (2.7) with log-likelihood cost (2.4), maximised over all $\tau < t$ for each new observations x_t. Unfortunately, in either case, the maximisation over the parameter space for each t and $k < t$ implies that plain GLR methods have computational complexity growing to infinity with the sample size.

Several solutions to alleviate the computational burden of GLR methods have been proposed, of which some are listed in the introduction of Lai (1995). The perhaps most widely used solution is to restrict the maximisation over candidate changepoints k to a set $\mathcal{K} \subseteq [t-1]$, for example a window of length w; $\mathcal{K} = \{k \geq 0 : t - w < k < t\}$. The effect of using a window is that only changes of a certain minimum size can be detected, with wider windows allowing for detectability of smaller changes and vice versa. Lai (1995) discusses how $\mathcal{K}$ can be constructed such that a vanishingly small amount of performance is lost. Such tricks bound the number of operations GLR methods need to update inference from one observation to the next, but the computational burden remains significantly larger than for CUSUM methods.

Moreover, note that it is significantly more complicated to evaluate the distribution of a GLR stopping time (3.2) than one based on a CUSUM. This is the case even for a change in mean in Gaussian data with known variance and pre-change mean, although Siegmund and Venkatraman (1995) derive approximations to the ARL that are quite accurate.

Other methods Several other methods have frequently been used for change detection, many originating from statistical process control. Two prominent examples are the Shewart's chart (Shewhart, 1925) and the exponentially weighted moving average chart (Hunter, 1986).

An alternative to the GLR statistic for an unknown post-change parameter is the so-called "mixture" or "weighted" likelihood ratio approach of Pollak and Siegmund (1975). Rather than maximising over the unknown parameter in (3.7), the mixture likelihood ratio approach integrates the likelihood ratio with respect to some probability distribution of the post-change $\boldsymbol{\theta}$.

The final classical method we mention is the Shiryaev-Roberts chart, due to Shiryaev (1963) and Roberts (1966). The Shiryaev-Roberts chart is a Bayesian analog to Page's CUSUM, and is given by exchanging the maximisation with summation in (3.5). It is a rather popular method as it is provably optimal in a certain Bayesian sense (see Polunchenko and Tartakovsky (2012, Section 4)).

3.2 Multivariate methods

We now present some important contributions to online change detection in multivariate data. As in the offline setting, the point of taking a multivariate approach is to be able to detect smaller changes more reliably than would be possible by a set of univariate methods. The distinction between sparse and dense changes is just as relevant in the online setting, as well as the additional challenge (P5) of identifying which variables are changing.

Changes in the mean For the prototypical change in mean setting, a major line of research on multivariate methods considers different ways of aggregating sequential changepoint tests applied to each univariate time series $x_{1:t}^{(j)}$, for $j = 1, \dots, p$. Roughly, Tartakovsky et al. (2006), Siegmund and Yakir (2008) and Mei (2010) propose aggregation-based tests for dense changes, while Xie and Siegmund (2013), Liu et al. (2017) and Chan (2017) focus on sparse changes. All these works consider individual tests of either CUSUM, GLR or Shiryaev-Roberts type, except Liu et al. (2017) who consider aggregation of any individual test of choice. Chan (2017) proves that his GLR-based method is optimal for detecting positive mean changes in the worst-worst-case sense of Lorden (1971). Alternative methods include the higher-criticism-based method of Zou et al. (2014a), the sketching- and dimension reduction-based method of Cao et al. (2019), as well as the recently proposed method of Chen et al. (2020), who combine CUSUMs both over variables and different post-change sizes of the means.

Changes in the covariance matrix Online detection of changes in the covariance matrix has yet to receive sufficient attention in the modern literature. An overview of methods for this problem from statistical process control is given by Yeh et al. (2005), and Sullivan and Woodall (2000) as well as Hawkins and Zamba (2009) study the GLR for detecting general changes in the mean and/or covariance matrix of multivariate normal data. Recent contributions are the moving sum-based approach of Avanesov and Buzun (2018) and the CUSUM-based method of Xie et al. (2018) for detecting changes in a spiked covariance matrix model. To the best of my knowledge, all existing methods are constructed to be efficient for dense alternatives. Detecting sparse changes in the covariance matrix is a problem we investigate in Paper II.

Changes in other features For sequentially detecting changes in other features than the mean or covariance matrix, one strategy is to decide on a likelihood for the data and construct a multivariate CUSUM or GLR test in a similar way as described for the univariate case in Section 3.1. Alternatively, one of the aggregation strategies for changes in the mean can be applied to any univariate or lower dimensional likelihoods of choice, as suggested by Liu et al. (2017). Recent nonparametric methods are the kernel-based method of Li et al. (2015) and the method based on windowed Kolmogorov-Smirnov tests of Madrid Padilla et al. (2019).

3.3 Other approaches

Not all the online literature fall nicely within the framework of controlling the ARL and minimising EDD (3.3). One deficiency of the classical methods is that the probability of declaring a change goes to one as t goes to infinity, i.e., a false alarm will eventually be raised. As a remedy, Chu et al. (1996) propose a different framework enabling control of $P^{\infty}(T < \infty)$ at a chosen significance level α under the asymptotic regime of m—the size of the training set—going to infinity. This approach has gained popularity in recent years, with methodology applicable in very general data scenarios being put forward by e.g. Aue et al. (2012), Kirch and Tadjuidje Kamgaing (2015) and Gösmann et al. (2020).

As online methods aim to update inference incrementally as data arrive, a Bayesian formulation in terms of updating the posterior distribution for every new observation seems a very natural one. Adams and MacKay (2007) and Fearnhead and Liu (2007) initialised the line of research on such methods. They utilise the product partition model as in the offline setting, and put a prior on the length between successive changepoints. These Bayesian methods are closer in spirit to offline methods as they aim to estimate the number and locations of changepoints, but in an online fashion, rather than detecting changes as quickly as possible. In addition, they have the advantage of providing uncertainty quantification of all unknown parameters, given the prior. Recent contributions to this class of methods include Ruggieri and Antonellis (2016), who introduce less informative priors, the multivariate anomaly detector of Bardwell and Fearnhead (2017) and the outlier-robust methodology of Knoblauch et al. (2018).

Chapter 4

Summaries of the papers

4.1 Paper I

Tveten, M. (2019). Which principal components are most sensitive in the change detection problem? *Stat*, 8(e252).

Principal component analysis (PCA) is arguably the most common method for reducing the dimensionality of multivariate data. It has been used for numerous applications both in statistics and machine learning, and it is therefore no surprise that it also forms the basis of many multivariate anomaly detection methods. In this short article, the behaviour of PCA within the change detection problem is investigated through a notion of each pre-change principal component's *sensitivity* to a change.

To be precise, consider the single changepoint setup where $\mathbf{x}_t \sim N(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$ for $t = 1, \ldots, \tau$, and $\mathbf{x}_t \sim N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ for $t = \tau + 1, \ldots, n$. Now let $\{\lambda_j, \mathbf{v}_j\}_{j=1}^p$ denote the normalised eigensystem of the pre-change $\boldsymbol{\Sigma}_0$, ordered decreasingly in λ_j. Our objects of interest are the *pre-change principal components* $y_{j,t} = \mathbf{v}_j^\intercal \mathbf{x}_t$. Before a change, the distribution of $y_{j,t}$ is $p(y) = N(y|0, \lambda_j)$, while after a change, the distribution of y_j is $q(y) = N(y|\mathbf{v}_j^\intercal \boldsymbol{\mu}_1, \mathbf{v}_1^\intercal \boldsymbol{\Sigma}_1 \mathbf{v}_1)$, where it is assumed without loss of generality that $\boldsymbol{\mu}_0 = \mathbf{0}$. The sensitivity to a change of the j'th pre-change principal component is then defined as the Hellinger distance between its marginal distribution before and after a change, given by $H(p_j, q_j)$.

The main contribution of this paper is to prove that for bivariate normal data, the least varying pre-change principal component, $y_{2,t}$, is the most sensitive for a range of pre-change covariance matrices $\boldsymbol{\Sigma}_0$, and changes $\boldsymbol{\mu}_1$ and $\boldsymbol{\Sigma}_1$. Most notably, $y_{2,t}$ is almost always the most sensitive if only a single parameter of the original distribution changes, i.e., in cases where one of the means, one of the variances or the correlation parameter of the original data change. This result suggests that the least varying pre-change components should be used for detecting sparse distributional changes in higher dimensional data as well.

4.2 Paper II

Tveten, M. and Glad, I. K. (2019). Online detection of sparse changes in high-dimensional data streams using tailored projections. *Manuscript.*

This article builds on the insights from Paper I to propose a method for tailoring the choice of pre-change principal components to a specific change or anomaly detection problem. We call this method tailored PCA (TPCA), and it is implemented in the accompanying R package `tpca`. In addition, we combine TPCA with an extension of the online change detection scheme of Xie and Siegmund (2013) to create a method for detecting changes both in the mean and the covariance matrix of potentially high-dimensional data. As mentioned in Chapter 3.2, online detection of changes in variance and correlation is an understudied subject. Note that TPCA can also be used in the offline setting. It is especially suitable for anomaly detection because of the explicit assumption of a baseline parameter that the choice of pre-change principal components can be based upon.

To select pre-change principal components by TPCA, the following ingredients are needed: A pre-change covariance matrix, $\mathbf{\Sigma}_0$, a divergence measure, a distribution over the post-change parameter space called the *change distribution* and a cutoff value $c \in [0, 1]$. In the paper, we use the Hellinger distance throughout in agreement with Paper I, but the `tpca` R-package allows any measure of divergence to be used. Using the notation of Section 4.1, we aim to rank the principal components' sensitivity to changes by

$$P_j := \mathbb{P}\left(\operatorname*{argmax}_{1\leq i\leq p} H(p_i, q_i) = j|\mathbf{\Sigma}_0\right) \tag{4.1}$$

for $j = 1, \ldots, p$, where the probability is taken with respect to the change distribution. In practice, simulations from the change distribution is used to estimate P_j. TPCA selects the pre-change principal components indexed by

$$\mathcal{J} = \min_{\mathcal{I}\subseteq\{1,\ldots,p\}} \sum_{j\in\mathcal{I}} P_j \geq c. \tag{4.2}$$

In our simulated test scenarios, $\mathcal{J}$ almost always corresponds to a small subset of the least varying pre-change principal components, often facilitating a dimension reduction of $80 - 98\%$ for $c \in [0.8, 0.999]$.

In the simulations for assessing the performance of our TPCA-based online change detection method, we focus on detecting both sparse and dense changes in the mean, variance and correlation. If the correlation coefficients in $\mathbf{\Sigma}_0$ is sufficiently large, we find evidence of our method being able to detect changes quicker from a small set of principal components than the baseline method of Xie and Siegmund (2013). I.e., we observe quicker detection and computation simultaneously. For weaker pre-change cross-correlation, this clear advantage is not present, but significant dimension reduction is still possible without a great loss in performance.

At the end of the paper, we illustrate how our method can be used on time-dependent data by using dynamic PCA in place of the classic PCA, and compare our method to dynamic PCA as used within stochastic process control. This illustration is performed on a realistically simulated dataset of the Tennessee Eastman Process. We find that, in settings where there is no extra validation set for tuning the detection threshold, our method is superior to the classical dynamic PCA method.

4.3 Paper III

Hellton, K. H., Tveten, M., Stakkeland, M., Engebretsen, S., Haug, O. and Aldrin, M. (2020). Real-time prediction of propulsion motor overheating using machine learning. ***Submitted for publication.***

In this paper, online change detection methodology is applied to predict overheating in electrical propulsion motors onboard marine vessels. Technology that protects the motors from overheating is obviously critical for the safety of a ship and those on board. The data used in this study contain observations from four vessels, each with three motors and six temperature sensors at various locations per motor, over time periods ranging from 80 to 294 days.

Almost all of the data is collected during normal operating conditions, but there is one known overheating event in one of the vessels' motors. The main contribution of this paper is to show that by using mostly basic statistical tools, the onset of similar overheating events can be detected reliably 60-90 minutes in advance, and thereby avoided in the future. Parts of the method have already been implemented as a new thermal protection function on several ships.

First, we construct a simple but general linear model for predicting the sensor-observed temperatures from other operating variables of the vessel under normal conditions—power and speed of the motors, for example. Then the six series of residuals of the actual temperature observations and the predictions are monitored simultaneously for large, positive changes in the mean by a combination of an adaptive version of the CUSUM (Lorden and Pollak, 2008) and the shrinkage-aggregation framework proposed by Liu et al. (2017). If a sufficiently large, positive change in the residuals' mean is detected, this is taken as an initial sign of overheating, and an alarm is raised.

In this application, it is not only important to be able to detect an emerging overheating event in a timely fashion, but also to keep false alarms to an absolute minimum. If false alarms are too frequent, the operators of the vessel is likely to put a piece of tape over the red lamp meant to indicate an impending fault, which, needless to say, could be catastrophic. Consequently, a methodological contribution of this article is an automatic tuning procedure for the change detection algorithm that takes as input the acceptable number of false alarms in the fault-free training data. This tuning procedure uses information about the known fault—making it a supervised method—and thus risks to overfit to the single observed overheating event. A mechanism for balancing early detection with a countering of overfitting is therefore also built in.

4.4 Paper IV

Tveten, M., Eckley, I. A. and Fearnhead, P. (2020). Scalable change-point and anomaly detection in cross-correlated data with an application to condition monitoring. ***Invited to submit a revision to Annals of Applied Statistics.***

We study and propose methods for the offline multiple anomaly and change detection problems in multivariate data when variables are cross-correlated and changes occur in an unknown subset of the mean components. In addition, we demonstrate the anomaly detection method's usefulness for sensor-based condition monitoring of an industrial process pump. The paper is accompanied by the R package `capacc`, providing efficient implementations of our methods.

The first main methodological contribution of the paper is the derivation of the penalised cost-based methods CAPA-CC (collective and point anomalies in cross-correlated data) and CPT-CC (changepoints in cross-correlated data) for solving each of these problems in a computationally efficient manner. Both methods are built on a particular approximation of the penalised saving (2.11) corresponding to a penalised Gaussian likelihood ratio tests for a single anomaly or change. Encapsulating these tests for a single change or anomaly, CPT-CC uses a binary segmentation type algorithm to detect multiple changes, while CAPA-CC uses a PELT type algorithm to detect multiple anomalies.

An approximation of the penalised saving is necessary for a moderately large p, as the exact maximum likelihood estimator of a subset mean in correlated data corresponds to a combinatorial optimisation problem, as far as we can see. The approximation we propose is motivated from the form of the maximum likelihood estimator and corresponds to what is known as an unconstrained *binary quadratic program*. Such binary quadratic programs are of the form

$$\max_{\mathbf{u}\in\{0,1\}^p} \mathbf{u}^\intercal \mathbf{A}\mathbf{u} + \mathbf{u}^\intercal \mathbf{b} + c, \tag{4.3}$$

where $\mathbf{A}$ is a real, symmetric, $(p\times p)$-dimensional matrix, $\mathbf{b}$ is a real, p-dimensional vector and c is a real scalar. A second major result in the paper, of possibly independent interest, is a dynamic programming algorithm requiring $O(p2^r)$ operations for obtaining an exact solution to (4.3) when $\mathbf{A}$ is r-banded. This algorithm is inspired by the optimal partitioning algorithm (2.5) in the way of proceeding recursively through the variables $d = 1, \dots, p$ and conditioning on the optimal penalised saving for variables $1, \dots, d-1$ at each d.

In our problems, $\mathbf{A}$ is banded if the precision matrix $\mathbf{Q}$ is banded. As a consequence, a banded estimate of $\mathbf{Q}$ is required for our methods to be scalable. To obtain an estimate of a desired band we utilise a robust version of the GLASSO algorithm. An important result from our simulation study is that our method performs advantageously compared to other methods in terms of power and estimation accuracy in a range of data settings, also when a truly dense precision matrix is approximated by a banded estimate.

The simulation study also points to interesting facts about which scenarios incorporating cross-correlations is favourable in the change or anomaly detection analysis compared to ignoring it. Surprisingly, if the change is dense and the changed mean components have similar values, ignoring cross-correlations results in a more powerful method.

Chapter 5

Discussion

In Chapters 2 and 3, we introduced the offline and online change detection problems, respectively, and briefly summarised parts of the statistical literature on these topics. The literature review is only meant to provide context for Papers I–IV—summarised in Chapter 4—and is by no means exhaustive. In this chapter, I discuss the papers critically, pointing to limitations and possible improvements not already mentioned in the papers. It is therefore advantageous to read the papers in full length in advance. The chapter is concluded by a discussion of some open challenges and future directions of the change detection field in general.

5.1 Discussion of the papers

Paper I In this paper, I used the Hellinger distance between distributions to define sensitivity to changes partly because it proved simple to work with. It would have been interesting to obtain similar results using the Kullback-Leibler divergence, however, as it is more directly linked to properties of change detection methods. For example, for online methods, Lorden (1971) showed that the optimal worst-worst-case detection delay (3.4) is governed by

$$\bar{D}(g, f) := \frac{\log \gamma}{I(g, f)}, \tag{5.1}$$

asymptotically as γ (the ARL) goes to infinity, where $I(g, f)$ is the Kullback-Leibler divergence from the pre-change distribution f to the post-change distribution g;

$$I(g, k) := \int \log \frac{g(x)}{f(x)} g(x) dx. \tag{5.2}$$

Thus, comparing the Kullback-Leibler divergences $I(p_j, q_j)$, where p_j and q_j are the pre-change and post-change distributions of principal component j as in Section 4.1, can be directly translated to *how much* quicker a particular change can be detected by each principal component. By using the Hellinger distance, we only get to know the ordering of which principal component will be the most efficient to monitor.

Paper II Our TPCA method is a tool for testing the usefulness of the knowledge and concepts from Paper I in practice. Empirically, it seems to work well, but unfortunately, we have little theory to support it. For instance, it would be beneficial to get some measure of uncertainty on the selected subset $\mathcal{J}$ in (4.2), and some guidance on the number of Monte Carlo simulations needed to approximate the distribution (4.1) well. For our chosen measure for ranking the axes (4.1)—the probability of a principal component being the most sensitive with respect to a distribution over changes—this is hard to obtain. Thus, in a future version of the manuscript, an option is to change the selection criterion for which principal components to monitor to one that can offer more in terms of guarantees on performance.

One such alternative selection criterion we have started to explore is based on the Kullback-Leibler divergence and its connection to the detection delay (5.1). The idea is to keep enough principal components such that a minimum of $c100\%$ of the information about changes occuring according to a change distribution is conserved with probability $1 - \alpha$, for chosen $c, \alpha \in (0, 1)$. This can be formalised as the problem of finding the minimal $\mathcal{J}$ such that

$$P\left(\sum_{j \in \mathcal{J}} I_j \Big/ \sum_{j=1}^{p} I_j \geq c\right) \geq 1 - \alpha \tag{5.3}$$

holds, where $I_j := I(p_j, q_j)$. Through this criterion, we can specify how much loss in detection speed is permissible at some probability $1 - \alpha$, as $\bar{D} \geq \log \gamma / (c \sum_{j=1}^{p} I_j)$ with probability $1 - \alpha$ when monitoring the (5.3)-selected principal components. Moreover, for a multivariate Gaussian change distribution for the mean, $\boldsymbol{\mu} \sim N(\boldsymbol{\theta}, \boldsymbol{\Gamma})$, combined with a Kullback-Leibler divergence between two Gaussians in I_j, the distribution (5.3) for a fixed $\mathcal{J}$ is possible to derive analytically; it can be expressed as the probability distribution of a quadratic form $\boldsymbol{\mu}^\mathsf{T} \mathbf{A} \boldsymbol{\mu}$, known to be distributed as a linear combination of independent non-central chi-square random variables. Motivated by the results of the minor principal components being the most sensitive, an approximate minimisation over $\mathcal{J}$ can be performed by starting from $\mathcal{J} = \{p\}$, and progressively adding more and more varying components until the criterion (5.3) is met. Results of this flavour could be useful as computationally efficient default settings, and to approximate more complicated change and data distributions.

More generally, we would like to address the choice of change distribution more thoroughly in the future. In the current manuscript, the change distribution used throughout represents little prior information, but it might seem quite arbitrary. As mentioned in the previous paragraph, finding a choice of change distribution that enables selection of the tailored principal components in a less brute force manner than Monte Carlo simulation would be highly beneficial. Such change distributions could then be studied under misspecified scenarios to assess the value of setting up a more complicated change distribution.

In the simulation study, we have divided results into classes of “low” and “high” correlation based on the value of the α_d parameter in the method of Joe (2006) for generating random correlation matrices being less than or greater than 1. The motivation for using this method was to obtain a large range of different correlation matrices. However, it is not that easy to interpret the size of the correlations in each class. Selecting a few simpler classes of correlation matrices as test beds, as we did in Paper IV, might therefore provide more informative results in terms of how strong the correlation must be for TPCA to perform better than the mixture procedure of Xie and Siegmund (2013).

Today, there are also more methods it would be relevant to compare performance with, especially the methods mentioned in the paragraph on detecting changes in the covariance matrix of Section 3.2.

Paper III The aim of this paper was to propose a method for detecting when a ship's motor is about to fail. Specifically, the method had to be able to predict an observed fault in a historical data set sufficiently early in advance with a minimal amount of false alarms, be generalisable to other ships and motors, as well as be simple conceptually and simple to implement on the on-board system of the ship. The two latter requirements lead us to using simple i.i.d. Gaussian models for the data, both when constructing the model for the motor temperature and when monitoring the residuals. These modelling assumptions were justified because the size of the change in mean signalling the observed fault was large enough to be detected early with few false alarms, despite the threshold having to absorb all aspects of the data not captured by the i.i.d. Gaussian model. The lesson here, from a practical point of view, is that much can be achieved by a very simple model.

However, other failures may not be equally pronounced as the one in our test set. In failure cases with smaller changes, more effort must be put on modelling the data. There are at least three improvements that would make detection of significantly smaller changes possible, if we disregard the requirement of implementational and conceptual simplicity. First, as mentioned in the discussion section of the paper, there is a consistent bias in the temperature residuals for each sensor. This is due to the model for generating the residuals being based on the average temperature over the six sensors, such that individual differences between the sensors are lost. One way of reducing the bias is thus to construct a temperature model for each sensor by including a training period for each motor. Note that a part of the bias is already handled by the parameter ρ in the adaptive CUSUM, but lowering ρ is also of interest to be able to detect smaller changes. A second improvement is to model the temporal dependence explicitly in the change detection method. The improvement is likely to be remarkable as the temperature residuals are very strongly auto-correlated as a consequence of the motor temperature being a slowly varying process relative to the once per second sampling rate. Thirdly, the spatial dependence between the sensors is also strong, so modelling it would further increase detection power (as the results in Paper IV show).

On the other hand, there will always be behaviour of the temperature sensor data not captured by even an extremely complex model. From the point of view of a change detection method, such deviations from the model will often be interpreted as evidence for a change. Thus, a reformulation of the change detection problem relevant to this application is to only detect *relevant changes*. The ρ-parameter in the adaptive CUSUM in practice filters out too small changes, but another alternative is to incorporate the relevant size of a change directly in the hypothesis testing problem. For a change in mean in univariate data, this means studying null hypotheses such as

$$H_0 : |\mu_0 - \mu_1| \leq \Delta,$$

where μ_0 and μ_1 are the pre- and post-change means. Initial work on change detection problems of this form has already been carried out by Dette and Gösmann (2018).

Paper IV Much of the discussion of Paper III also applies to the application of condition monitoring a process pump in Paper IV. Specifically, modelling of temporal dependence and a more sophisticated model for removing trends in the data associated with the operational state of the pump is likely to increase performance. By "operational state", I mean, for instance, the volume fractions of the different fluids being pumped, their flow rate, the power of the pump, and so forth.

An online version of CAPA-CC is needed to be able to monitor the pump in real-time. In this particular application, the current offline version is primarily useful for analysing historical data of the pump, either to prepare a training set for an online method, or to explore when the pump has been running suboptimally in the past, perhaps discovering previously unknown anomalous segments. Fisch et al. (2020) recently showed how the univariate CAPA method can be made sequential, and similar ideas can be used to create an online counterpart of CAPA-CC.

On the methodological side there are also numerous possibilities for extensions. In the penalised cost framework of our methods, we use a pointwise minimum between a linear and a constant penalty on the number of changing variables. Akin to the optimal partitioning algorithm in (2.5), the restriction to linear penalties in the sparse regime is what allows for quick computation of the penalised saving for a fixed changepoint or anomalous segment. There may, however, be scenarios where a non-linear penalty is preferred, and Fisch et al. (2019b) show that for intermediately sparse changes—that is, for $p^{-1/2} < |\mathbf{J}| \leq p^{-3/4}$—in cross-independent Gaussian data, a third, non-linear penalty regime is needed for optimal power. It is possible to accommodate for non-linear penalties in our method by deriving a segment neighbourhood analog to our optimal partitioning-inspired algorithm for computing the penalised saving. (The segment neighbourhood algorithm is described at the end of the paragraph on dynamic programming-based methods in Section 2.1.) By this, I mean that in addition to sequentially conditioning on the optimal penalised saving until variable $d \leq p$, one can also condition on the number of changing variables, starting from finding the single variable that increases the penalised saving the most, before proceeding recursively until a maximum number of changing variables, $\bar{J}$, is reached. Such an algorithm would scale quadratically in p if $\bar{J}$ grows linearly in p. This is prohibitive for large p, but for moderately sized p, as in our 5-dimensional pump data example, it may have practical value.

As CAPA-CC and CPT-CC are based on the multivariate Gaussian model, it is of course relevant to explore other costs, both likelihood-based costs and others. It would be interesting to seek more general models where our algorithm for solving binary quadratic programs can be used to approximate tests for subset changes, or if it can find use in other tasks involving variable selection.

I am open to suggestions on how to obtain stronger theoretical results on the quality of the approximate versus the exact penalised saving.

5.2 Open challenges in change detection

We conclude this introduction by discussing some interesting open challenges in the change detection field.

A major issue with the vast majority of frequentist change detection methods is that they only provide point estimates of changepoints, without a measure of confidence in these estimates. The quality of changepoint estimates is mainly assessed by proving their consistency and deriving convergence rates. As pointed out by Paul Fearnhead in the discussion on Frick et al. (2014), additional challenges with confidence intervals for changepoints arise when the the number of changes are unknown, as is often the case. Should the confidence intervals be constructed with respect to a fixed number of changepoints? How should uncertainty on the number of changepoints be incorporated? And how can confidence intervals for the number of changepoints be constructed? The confidence intervals of Frick et al. (2014) rely on their method consistently estimating the number of changepoints, and then asymptotic confidence intervals for the changepoints are constructed conditional on the estimated number of changes. Continuing to paraphrase Paul Fearnhead, it is not clear how to interpret such confidence intervals in many real data settings, as there is often significant uncertainty regarding the number of changes. The confidence distribution approach of Cunen et al. (2018) would face similar challenges as they assume there is maximally a single changepoint. Bayesian methods, on the other hand, are able to incorporate uncertainty on both the number and location of changepoints simultaneously, and may be the only option for full uncertainty quantification. However, such Bayesian inference is of course conditional on the often subjectively specified prior.

In many applied problems, including both the ship motor and pump monitoring problems of Paper III and Paper IV, the mean function of the data is not constant or linear between changepoints, but contains local fluctuations or trends of a complicated functional form. In our problems, a portion of these trends can be ascribed to a time-varying context of the machines; the temperature of the motor naturally increases as the motor's power increases, for example. Given the true relationship between the power and the temperature of the motor, this trend could be removed entirely. In practice, however, relationships of this sort have to be modelled and estimated, and variables explaining the trend might not always be recorded. Some trends or local fluctations will, consequently, often remain, no matter how hard one tries to remove them. Thus, change and anomaly detection methods that allow the mean function between changepoints to be smoothly time-varying or stochastic are likely to be useful in practice.

Combine this with modelling of temporal dependence and outlier-robustness, and the practical usefulness will increase even further. Initial work in this direction has been carried out by Romano et al. (2020) for univariate data. Multivariate and online versions are still to be explored.

Online or sequential change detection is still an underexplored problem compared to the offline problem, despite the origin of change detection being sequential. The current online field is mainly focused on the speed of detecting a single change. However, in several applied settings, it is more important that detection is reliable in terms of avoiding false alarms than quick, so long as detection is "quick enough". Consequently, a formulation of the online change detection problem starting from what is sufficiently quick detection before minimising the probability of false alarms might be fruitful. Moreover, constructing online versions of the algorithms for existing offline methods could be useful for adapting the online field to multiple change scenarios, thereby bridging the gap between the two settings.

Bibliography

Adams, R. P. and MacKay, D. J. C. (2007). Bayesian online changepoint detection. *arXiv:0710.3742 [stat.ML]*. arXiv: 0710.3742.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723. Conference Name: IEEE Transactions on Automatic Control.

Aminikhanghahi, S. and Cook, D. J. (2017). A survey of methods for time series change point detection. *Knowledge and Information Systems*, 51(2):339–367.

Arlot, S., Celisse, A., and Harchaoui, Z. (2019). A Kernel Multiple Change-point Algorithm via Model Selection. *Journal of Machine Learning Research*, 20:1–56.

Aston, J. A. D. and Kirch, C. (2018). High dimensional efficiency with applications to change point tests. *Electronic Journal of Statistics*, 12(1):1901–1947. Publisher: The Institute of Mathematical Statistics and the Bernoulli Society.

Aue, A. and Horváth, L. (2013). Structural breaks in time series. *Journal of Time Series Analysis*, 34(1):1–16. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-9892.2012.00819.x.

Aue, A., Horváth, L., Kühn, M., and Steinebach, J. (2012). On the reaction time of moving sum detectors. *Journal of Statistical Planning and Inference*, 142(8):2271–2288.

Aue, A., Hörmann, S., Horváth, L., and Reimherr, M. (2009). Break detection in the covariance structure of multivariate time series models. *Annals of Statistics*, 37(6B):4046–4087. Publisher: Institute of Mathematical Statistics.

Auger, I. E. and Lawrence, C. E. (1989). Algorithms for the optimal identification of segment neighborhoods. *Bulletin of Mathematical Biology*, 51(1):39–54.

Avanesov, V. and Buzun, N. (2018). Change-point detection in high-dimensional covariance structure. *Electronic Journal of Statistics*, 12(2):3254–3294. Publisher: The Institute of Mathematical Statistics and the Bernoulli Society.

Bai, J. (1995). Least Absolute Deviation Estimation of a Shift. *Econometric Theory*, 11(3):403–436. Publisher: Cambridge University Press.

Bai, J. (2010). Common breaks in means and variances for panel data. *Journal of Econometrics*, 157(1):78–92.

Baranowski, R., Chen, Y., and Fryzlewicz, P. (2019). Narrowest-over-threshold detection of multiple change points and change-point-like features. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 81(3):649–672. _eprint: https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/rssb.12322.

Bardwell, L. and Fearnhead, P. (2017). Bayesian Detection of Abnormal Segments in Multiple Time Series. *Bayesian Analysis*, 12(1):193–218.

Bardwell, L., Fearnhead, P., Eckley, I. A., Smith, S., and Spott, M. (2019). Most Recent Changepoint Detection in Panel Data. *Technometrics*, 61(1):88–98.

Barry, D. and Hartigan, J. A. (1993). A Bayesian Analysis for Change Point Problems. *Journal of the American Statistical Association*, 88(421):309–319. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/01621459.1993.10594323.

Basseville, M. and Nikiforov, I. V. (1993). *Detection of abrupt changes: theory and application*, volume 104. Prentice Hall, Englewood Cliffs.

Baudry, J.-P., Maugis, C., and Michel, B. (2012). Slope heuristics: overview and implementation. *Statistics and Computing*, 22(2):455–470.

Bhattacharjee, M., Banerjee, M., and Michailidis, G. (2019). Change Point Estimation in Panel Data with Temporal and Cross-sectional Dependence. *arXiv:1904.11101 [math.ST]*.

Bhattacharjee, M., Banerjee, M., and Michailidis, G. (2020). Change Point Estimation in a Dynamic Stochastic Block Model. *Journal of Machine Learning Research*, 21:1–59.

Birgé, L. and Massart, P. (2001). Gaussian model selection. *Journal of the European Mathematical Society*, 3(3):203–268.

Birgé, L. and Massart, P. (2007). Minimal penalties for Gaussian model selection. *Probability Theory and Related Fields*, 138(1-2):33–73.

Cai, T. T., Jeng, X. J., and Jin, J. (2011). Optimal detection of heterogeneous and heteroscedastic mixtures. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(5):629–662. _eprint: https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-9868.2011.00778.x.

Cao, Y., Thompson, A., Wang, M., and Xie, Y. (2019). Sketching for sequential change-point detection. *EURASIP Journal on Advances in Signal Processing*, 2019(1):42.

Chakar, S., Lebarbier, E., Lévy-Leduc, C., and Robin, S. (2017). A robust approach for estimating change-points in the mean of an AR(1) process. *Bernoulli*, 23(2):1408–1447. Publisher: Bernoulli Society for Mathematical Statistics and Probability.

Chan, H. P. (2017). Optimal sequential detection in multi-stream data. *The Annals of Statistics*, 45(6):2736–2763.

Chandola, V., Banerjee, A., and Kumar, V. (2009). Anomaly detection: A survey. *ACM Computing Surveys*, 41(3):15:1–15:58.

Chen, H. and Zhang, N. (2015). Graph-based change-point detection. *Annals of Statistics*, 43(1):139–176. Publisher: Institute of Mathematical Statistics.

Chen, J. and Gupta, A. K. (2011). *Parametric Statistical Change Point Analysis: With Applications to Genetics, Medicine, and Finance.* Birkhäuser, New York.

Chen, Y., Wang, T., and Samworth, R. J. (2020). High-dimensional, multiscale online changepoint detection. *arXiv:2003.03668 [stat.ME].* arXiv: 2003.03668.

Chib, S. (1998). Estimation and comparison of multiple change-point models. *Journal of Econometrics*, 86(2):221–241.

Cho, H. (2016). Change-point detection in panel data via double CUSUM statistic. *Electronic Journal of Statistics*, 10(2):2000–2038.

Cho, H. and Fryzlewicz, P. (2011). Multiscale interpretation of taut string estimation and its connection to Unbalanced Haar wavelets. *Statistics and Computing*, 21(4):671–681.

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

Chu, C.-S. J., Stinchcombe, M., and White, H. (1996). Monitoring Structural Change. *Econometrica*, 64(5):1045–1065. Publisher: [Wiley, Econometric Society].

Chu, L. and Chen, H. (2019). Asymptotic distribution-free change-point detection for multivariate and non-Euclidean data. *Annals of Statistics*, 47(1):382–414. Publisher: Institute of Mathematical Statistics.

Csörgő, M. and Horváth, L. (1988). 20 Nonparametric methods for changepoint problems. In *Handbook of Statistics*, volume 7 of *Quality Control and Reliability*, pages 403–425. Elsevier.

Cunen, C., Hermansen, G., and Hjort, N. L. (2018). Confidence distributions for change-points and regime shifts. *Journal of Statistical Planning and Inference*, 195:14–34.

Dette, H. and Gösmann, J. (2018). Relevant change points in high dimensional time series. *Electronic Journal of Statistics*, 12(2):2578–2636. Publisher: The Institute of Mathematical Statistics and the Bernoulli Society.

Dette, H., Pan, G. M., and Yang, Q. (2020). Estimating a Change Point in a Sequence of Very High-Dimensional Covariance Matrices. *Journal of the American Statistical Association*. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/01621459.2020.1785477.

Donoho, D. and Jin, J. (2004). Higher criticism for detecting sparse heterogeneous mixtures. *The Annals of Statistics*, 32(3):962–994.

Eckley, I., Kirch, C., and Weber, S. (2020). A novel change point approach for the detection of gas emission sources using remotely contained concentration data. *Annals of Applied Statistics*.

Eckley, I. A., Fearnhead, P., and Killick, R. (2011). Analysis of changepoint models. In Barber, D., Cemgil, A. T., and Chiappa, S., editors, *Bayesian Time Series Models*, pages 205–224. Cambridge University Press, Cambridge.

Edwards, A. W. F. and Cavalli-Sforza, L. L. (1965). A Method for Cluster Analysis. *Biometrics*, 21(2):362–375. Publisher: [Wiley, International Biometric Society].

Eichinger, B. and Kirch, C. (2018). A MOSUM procedure for the estimation of multiple random change points. *Bernoulli*, 24(1):526–564. Publisher: Bernoulli Society for Mathematical Statistics and Probability.

Enikeeva, F. and Harchaoui, Z. (2019). High-dimensional change-point detection under sparse alternatives. *Annals of Statistics*, 47(4):2051–2079. Publisher: Institute of Mathematical Statistics.

Fearnhead, P. (2006). Exact and efficient Bayesian inference for multiple changepoint problems. *Statistics and Computing*, 16(2):203–213.

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

Fearnhead, P., Maidstone, R., and Letchford, A. (2019). Detecting Changes in Slope With an L_0 Penalty. *Journal of Computational and Graphical Statistics*, 28(2):265–275. Publisher: Taylor & Francis.

Fearnhead, P. and Rigaill, G. (2019). Changepoint Detection in the Presence of Outliers. *Journal of the American Statistical Association*, 114(525):169–183.

Fisch, A. T. M., Bardwell, L., and Eckley, I. A. (2020). Real Time Anomaly Detection And Categorisation. *arXiv:2009.06670 [stat.ME]*.

Fisch, A. T. M., Eckley, I. A., and Fearnhead, P. (2019a). A linear time method for the detection of point and collective anomalies. *arXiv:1806.01947 [stat.ML]*.

Fisch, A. T. M., Eckley, I. A., and Fearnhead, P. (2019b). Subset Multivariate Collective And Point Anomaly Detection. *arXiv:1909.01691 [stat.ME]*.

Frick, K., Munk, A., and Sieling, H. (2014). Multiscale change point inference. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 76(3):495–580. Publisher: [Royal Statistical Society, Wiley].

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

Fryzlewicz, P. (2018). Tail-greedy bottom-up data decompositions and fast multiple change-point detection. *Annals of Statistics*, 46(6B):3390–3421. Publisher: Institute of Mathematical Statistics.

Fryzlewicz, P. (2020). Detecting possibly frequent change-points: Wild Binary Segmentation 2 and steepest-drop model selection. *Journal of the Korean Statistical Society*.

Gao, Z., Du, P., Jin, R., and Robertson, J. L. (2020). Surface temperature monitoring in liver procurement via functional variance change-point analysis. *Annals of Applied Statistics*, 14(1):143–159. Publisher: Institute of Mathematical Statistics.

Gombay, E. and Horvath, L. (1994). An application of the maximum likelihood test to the change-point problem. *Stochastic Processes and their Applications*, 50(1):161–171. Publisher: Elsevier.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4):711–732. Publisher: Oxford Academic.

Grundy, T., Killick, R., and Mihaylov, G. (2020). High-dimensional changepoint detection via a geometrically inspired mapping. *Statistics and Computing*, 30(4):1155–1166.

Gösmann, J., Kley, T., and Dette, H. (2020). A new approach for open-end sequential change point monitoring. *arXiv:1906.03225 [math.ST]*. arXiv: 1906.03225.

Harchaoui, Z. and Cappe, O. (2007). Retrospective Mutiple Change-Point Estimation with Kernels. In *2007 IEEE/SP 14th Workshop on Statistical Signal Processing*, pages 768–772, Madison, WI, USA. ISSN: 2373-0803.

Harchaoui, Z. and Lévy-Leduc, C. (2010). Multiple Change-Point Estimation With a Total Variation Penalty. *Journal of the American Statistical Association*, 105(492):1480–1493. Publisher: Taylor & Francis.

Hawkins, D. M. and Zamba, K. D. (2009). A Multivariate Change-Point Model for Change in Mean Vector and/or Covariance Structure. *Journal of Quality Technology*, 41(3):285–303.

He, H. and Severini, T. A. (2010). Asymptotic properties of maximum likelihood estimators in models with multiple change points. *Bernoulli*, 16(3):759–779. Publisher: Bernoulli Society for Mathematical Statistics and Probability.

Henderson, R. and Matthews, J. N. S. (1993). An Investigation of Changepoints in the Annual Number of Cases of Haemolytic Uraemic Syndrome. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 42(3):461–471. _eprint: https://rss.onlinelibrary.wiley.com/doi/pdf/10.2307/2986325.

Hinkley, D. V. (1970). Inference about the change-point in a sequence of random variables. *Biometrika*, 57(1):1–17.

Hinkley, D. V. (1971). Inference about the change-point from cumulative sum tests. *Biometrika*, 58(3):509–523. Publisher: Oxford Academic.

Hlávka, Z., Hušková, M., Kirch, C., and Meintanis, S. G. (2017). Fourier–type tests involving martingale difference processes. *Econometric Reviews*, 36(4):468–492.

Horváth, L. and Hušková, M. (2012). Change-point detection in panel data. *Journal of Time Series Analysis*, 33(4):631–648.

Horváth, L., Kokoszka, P., and Steinebach, J. (1999). Testing for Changes in Multivariate Dependent Observations with an Application to Temperature Changes. *Journal of Multivariate Analysis*, 68(1):96–119.

Hsu, D. A. (1977). Tests for Variance Shift at an Unknown Time Point. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 26(3):279–284. _eprint: https://rss.onlinelibrary.wiley.com/doi/pdf/10.2307/2346968.

Huber, P. J. (2004). *Robust Statistics*. John Wiley & Sons. Google-Books-ID: e62RhdqIdMkC.

Hušková, M. (2013). Robust Change Point Analysis. In Becker, C., Fried, R., and Kuhnt, S., editors, *Robustness and Complex Data Structures*, pages 171–190. Springer, Berlin, Heidelberg.

Hušková, M. and Slabý, A. (2001). Permutation tests for multiple changes. *Kybernetika*, 37(5):605–622. Publisher: Institute of Information Theory and Automation AS CR.

Hunter, J. S. (1986). The Exponentially Weighted Moving Average. *Journal of Quality Technology*, 18(4):203–210. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/00224065.1986.11979014.

Inclán, C. and Tiao, G. C. (1994). Use of Cumulative Sums of Squares for Retrospective Detection of Changes of Variance. *Journal of the American Statistical Association*, 89(427):913–923. Publisher: Taylor & Francis.

Jackson, B., Scargle, J., Barnes, D., Arabhi, S., Alt, A., Gioumousis, P., Gwin, E., Sangtrakulcharoen, P., Tan, L., and Tsai, T. T. (2005). An algorithm for optimal partitioning of data on an interval. *IEEE Signal Processing Letters*, 12(2):105–108. Conference Name: IEEE Signal Processing Letters.

Jandhyala, V., Fotopoulos, S., MacNeill, I., and Liu, P. (2013). Inference for single and multiple change-points in time series. *Journal of Time Series Analysis*, 34(4):423–446.

Jeng, X. J., Cai, T. T., and Li, H. (2013). Simultaneous discovery of rare and common segment variants. *Biometrika*, 100(1):157–172.

Jirak, M. (2015). Uniform change point tests in high dimension. *The Annals of Statistics*, 43(6):2451–2483.

Joe, H. (2006). Generating random correlation matrices based on partial correlations. *Journal of Multivariate Analysis*, 97(10):2177–2189.

Kao, C., Trapani, L., and Urga, G. (2018). Testing for instability in covariance structures. *Bernoulli*, 24(1):740–771. Publisher: Bernoulli Society for Mathematical Statistics and Probability.

Killick, R., Fearnhead, P., and Eckley, I. A. (2012). Optimal Detection of Changepoints With a Linear Computational Cost. *Journal of the American Statistical Association*, 107(500):1590–1598.

Kirch, C., Muhsal, B., and Ombao, H. (2015). Detection of Changes in Multivariate Time Series With Application to EEG Data. *Journal of the American Statistical Association*, 110(511):1197–1216.

Kirch, C. and Tadjuidje Kamgaing, J. (2015). On the use of estimating functions in monitoring time series for change points. *Journal of Statistical Planning and Inference*, 161:25–49.

Knoblauch, J., Jewson, J. E., and Damoulas, T. (2018). Doubly Robust Bayesian Inference for Non-Stationary Streaming Data with β-Divergences. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R., editors, *Advances in Neural Information Processing Systems 31*, pages 64–75. Curran Associates, Inc.

Ko, S. I. M., Chong, T. T. L., and Ghosh, P. (2015). Dirichlet Process Hidden Markov Multiple Change-Point Model. *Bayesian Analysis*, 10(2):275–296. Publisher: International Society for Bayesian Analysis.

Kovács, S., Li, H., Bühlmann, P., and Munk, A. (2020). Seeded Binary Segmentation: A general methodology for fast and optimal change point detection. *arXiv:2002.06633 [stat.ME].*

Lai, T. L. (1995). Sequential Changepoint Detection in Quality Control and Dynamical Systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(4):613–658.

Lee, S., Ha, J., Na, O., and Na, S. (2003). The Cusum Test for Parameter Change in Time Series Models. *Scandinavian Journal of Statistics*, 30(4):781–796. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1467-9469.00364.

Lee, S., Seo, M. H., and Shin, Y. (2016). The lasso for high dimensional regression with a possible change point. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, 78(1):193–210.

Leonardi, F. and Bühlmann, P. (2016). Computationally efficient change point detection for high-dimensional regression. *arXiv:1601.03704 [stat.ME]*. arXiv: 1601.03704.

Li, H., Munk, A., and Sieling, H. (2016). FDR-control in multiscale change-point segmentation. *Electronic Journal of Statistics*, 10(1):918–959. Publisher: The Institute of Mathematical Statistics and the Bernoulli Society.

Li, J., Xu, M., Zhong, P.-S., and Li, L. (2019). Change Point Detection in the Mean of High-Dimensional Time Series Data under Dependence. *arXiv:1903.07006 [stat.ME]*.

Li, S., Xie, Y., Dai, H., and Song, L. (2015). M-Statistic for Kernel Change-Point Detection. In Cortes, C., Lawrence, N. D., Lee, D. D., Sugiyama, M., and Garnett, R., editors, *Advances in Neural Information Processing Systems 28*, pages 3366–3374. Curran Associates, Inc.

Liu, B., Zhou, C., Zhang, X., and Liu, Y. (2020). A unified data-adaptive framework for high dimensional change point detection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(4):933–963. _eprint: https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/rssb.12375.

Liu, K., Zhang, R., and Mei, Y. (2017). Scalable SUM-Shrinkage Schemes for Distributed Monitoring Large-Scale Data Streams. *Statistica Sinica*, 29:1–22.

Liu, Y.-W. and Chen, H. (2020). A Fast and Efficient Change-point Detection Framework for Modern Data. *arXiv:2006.13450 [stat.ME]*. arXiv: 2006.13450.

Lorden, G. (1971). Procedures for reacting to a change in distribution. *The Annals of Mathematical Statistics*, 42(6):1897–1908.

Lorden, G. and Pollak, M. (2008). Sequential Change-Point Detection Procedures That are Nearly Optimal and Computationally Simple. *Sequential Analysis*, 27(4):476–512.

Lévy-Leduc, C. and Roueff, F. (2009). Detection and localization of change-points in high-dimensional network traffic data. *The Annals of Applied Statistics*, 3(2):637–662.

Maboudou-Tchao, E. M. and Hawkins, D. M. (2013). Detection of multiple change-points in multivariate data. *Journal of Applied Statistics*, 40(9):1979–1995. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/02664763.2013.800471.

Madrid Padilla, O. H., Athey, A., Reinhart, A., and Scott, J. G. (2019). Sequential Nonparametric Tests for a Change in Distribution: An Application to Detecting Radiological Anomalies. *Journal of the American Statistical Association*, 114(526):514–528. Publisher: Taylor & Francis.

Maeng, H. (2019). *Adaptive multiscale approaches to regression and trend segmentation*. PhD Thesis, The London School of Economics and Political Science.

Maeng, H. and Fryzlewicz, P. (2019). Detecting linear trend changes and point anomalies in data sequences. *arXiv:1906.01939 [stat.ME]*. arXiv: 1906.01939.

Maidstone, R. (2016). *Efficient analysis of complex changepoint problems*. PhD Thesis, Lancaster University.

Maidstone, R., Hocking, T., Rigaill, G., and Fearnhead, P. (2017). On optimal multiple changepoint algorithms for large data. *Statistics and Computing*, 27(2):519–533.

Matteson, D. S. and James, N. A. (2014). A Nonparametric Approach for Multiple Change Point Analysis of Multivariate Data. *Journal of the American Statistical Association*, 109(505):334–345. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/01621459.2013.849605.

McDonald, D. (1990). A cusum procedure based on sequential ranks. *Naval Research Logistics*, 37(5):627–646. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/1520-6750%28199010%2937%3A5%3C627%3A%3AAID-NAV3220370504%3E3.0.CO%3B2-F.

Mei, Y. (2010). Efficient scalable schemes for monitoring a large number of data streams. *Biometrika*, 97(2):419–433.

Mendiratta, V., Liu, Z., Bhattacharjee, M., and Zhou, Y. (2019). Detecting and Diagnosing Anomalous Behavior in Large Systems with Change Detection Algorithms. In *2019 IEEE International Symposium on Software Reliability Engineering Workshops (ISSREW)*, pages 47–52.

Moustakides, G. V. (1986). Optimal stopping times for detecting changes in distributions. *The Annals of Statistics*, 14(4):1379–1387.

Niu, Y. S., Hao, N., and Zhang, H. (2016). Multiple Change-Point Detection: A Selective Overview. *Statistical Science*, 31(4):611–623. Publisher: Institute of Mathematical Statistics.

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

Padilla, O. H. M., Yu, Y., Wang, D., and Rinaldo, A. (2020). Optimal nonparametric multivariate change point detection and localization. *arXiv:1910.13289 [math.ST]*. arXiv: 1910.13289.

Page, E. S. (1954). Continuous inspection schemes. *Biometrika*, 41(1/2):100–115.

Page, E. S. (1955). A test for a change in a parameter occurring at an unknown point. *Biometrika*, 42(3/4):523–527.

Pein, F., Sieling, H., and Munk, A. (2017). Heterogeneous change point inference. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, 79(4):1207–1227.

Pettitt, A. N. (1979). A Non-Parametric Approach to the Change-Point Problem. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 28(2):126–135. _eprint: https://rss.onlinelibrary.wiley.com/doi/pdf/10.2307/2346729.

Pollak, M. and Siegmund, D. (1975). Approximations to the Expected Sample Size of Certain Sequential Tests. *The Annals of Statistics*, 3(6):1267–1282. Publisher: Institute of Mathematical Statistics.

Pollak, M. and Siegmund, D. (1991). Sequential detection of a change in a normal mean when the initial value is unknown. *Annals of Statistics*, 19(1):394–416. Publisher: Institute of Mathematical Statistics.

Polunchenko, A. S. and Tartakovsky, A. G. (2012). State-of-the-Art in Sequential Change-Point Detection. *Methodology and Computing in Applied Probability*, 14(3):649–684.

Preuss, P., Puchstein, R., and Dette, H. (2015). Detection of Multiple Structural Breaks in Multivariate Time Series. *Journal of the American Statistical Association*, 110(510):654–668. Publisher: Taylor & Francis.

Rigaill, G. (2010). Pruned dynamic programming for optimal multiple change-point detection. *arXiv:1004.0887 [stat.CO]*.

Roberts, S. W. (1966). A Comparison of Some Control Chart Procedures. *Technometrics*, 8(3):411–430. Publisher: Taylor & Francis.

Romano, G., Rigaill, G., Runge, V., and Fearnhead, P. (2020). Detecting Abrupt Changes in the Presence of Local Fluctuations and Autocorrelated Noise. *arXiv:2005.01379 [stat.ME]*. arXiv: 2005.01379.

Roy, S., Atchadé, Y., and Michailidis, G. (2017). Change point estimation in high dimensional Markov random-field models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(4):1187–1206.

Ruggieri, E. and Antonellis, M. (2016). An exact approach to Bayesian sequential change point detection. *Computational Statistics & Data Analysis*, 97:71–86.

Safikhani, A. and Shojaie, A. (2020). Joint Structural Break Detection and Parameter Estimation in High-Dimensional Nonstationary VAR Models. *Journal of the American Statistical Association*. Publisher: Taylor & Francis.

Schwarz, G. (1978). Estimating the Dimension of a Model. *Annals of Statistics*, 6(2):461–464. Publisher: Institute of Mathematical Statistics.

Schweder, T. and Hjort, N. L. (2016). *Confidence, Likelihood, Probability*. Cambridge University Press. Google-Books-ID: vPx9CwAAQBAJ.

Scott, A. J. and Knott, M. (1974). A Cluster Analysis Method for Grouping Means in the Analysis of Variance. *Biometrics*, 30(3):507–512. Publisher: [Wiley, International Biometric Society].

Shewhart, W. A. (1925). The Application of Statistics as an Aid in Maintaining Quality of a Manufactured Product. *Journal of the American Statistical Association*, 20(152):546–548. Publisher: Taylor & Francis _eprint: https://www.tandfonline.com/doi/pdf/10.1080/01621459.1925.10502930.

Shiryaev, A. N. (1963). On Optimum Methods in Quickest Detection Problems. *Theory of Probability & Its Applications*, 8(1):22–46. Publisher: Society for Industrial and Applied Mathematics.

Siegmund, D. (1985). *Sequential Analysis: Tests and Confidence Intervals*. Springer, New York, USA.

Siegmund, D. and Venkatraman, E. S. (1995). Using the generalized likelihood ratio statistic for sequential detection of a change-point. *The Annals of Statistics*, 23(1):255–271.

Siegmund, D. and Yakir, B. (2008). Detecting the emergence of a signal in a noisy image. *Statistics and Its Interface*, 1(1):3–12.

Siegmund, D., Yakir, B., and Zhang, N. R. (2011). Detecting simultaneous variant intervals in aligned sequences. *The Annals of Applied Statistics*, 5(2A):645–668. Publisher: Institute of Mathematical Statistics.

Srivastava, M. S. and Worsley, K. J. (1986). Likelihood Ratio Tests for a Change in the Multivariate Normal Mean. *Journal of the American Statistical Association*, 81(393):199–204. Publisher: Taylor & Francis _eprint: https://www.tandfonline.com/doi/pdf/10.1080/01621459.1986.10478260.

Sullivan, J. H. and Woodall, W. H. (2000). Change-point detection of mean vector or covariance matrix shifts using multivariate individual observations. *IIE Transactions*, 32(6):537–549.

Tartakovsky, A. G., Rozovskii, B. L., Blažek, R. B., and Kim, H. (2006). Detection of intrusions in information systems by sequential change-point methods. *Statistical Methodology*, 3(3):252–293.

Tibshirani, R. (1996). Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288. _eprint: https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/j.2517-6161.1996.tb02080.x.

Tickle, S. (2020). *Changepoint detection for data intensive settings.* PhD Thesis, Lancaster University.

Tickle, S. O., Eckley, I. A., Fearnhead, P., and Haynes, K. (2020). Parallelization of a Common Changepoint Detection Method. *Journal of Computational and Graphical Statistics*, 29(1):149–161. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/10618600.2019.1647216.

Truong, C., Oudre, L., and Vayatis, N. (2020). Selective review of offline change point detection methods. *Signal Processing*, 167:107299.

Venkatraman, E. S. (1993). *Consistency results in multiple change-point problems.* PhD Thesis, Stanford University.

Vostrikova, L. J. (1981). Detecting disorder in multidimensional random processes. *Soviet Mathematics Doklady*, 24:55–59.

Wang, D., Yu, Y., and Rinaldo, A. (2018). Optimal Covariance Change Point Localization in High Dimension. *arXiv:1712.09912 [math.ST].* arXiv: 1712.09912.

Wang, D., Yu, Y., and Rinaldo, A. (2020a). Univariate mean change point detection: Penalization, CUSUM and optimality. *Electronic Journal of Statistics*, 14(1):1917–1961. Publisher: The Institute of Mathematical Statistics and the Bernoulli Society.

Wang, D., Yu, Y., Rinaldo, A., and Willett, R. (2020b). Localizing Changes in High-Dimensional Vector Autoregressive Processes. *arXiv:1909.06359 [math.ST].* arXiv: 1909.06359.

Wang, R. and Shao, X. (2020). Dating the Break in High-dimensional Data. *arXiv:2002.04115 [math.ST].* arXiv: 2002.04115.

Wang, R., Volgushev, S., and Shao, X. (2019). Inference for Change Points in High Dimensional Data. *arXiv:1905.08446 [math.ST].* arXiv: 1905.08446.

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

Westerlund, J. (2019). Common Breaks in Means for Cross-Correlated Fixed-T Panel Data. *Journal of Time Series Analysis*, 40(2):248–255.

Worsley, K. J. (1986). Confidence regions and tests for a change-point in a sequence of exponential family random variables. *Biometrika*, 73(1):91–104. Publisher: Oxford Academic.

Xie, L., Moustakides, G. V., and Xie, Y. (2018). First-Order Optimal Sequential Subspace Change-Point Detection. In *2018 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pages 111–115.

Xie, Y. and Siegmund, D. (2013). Sequential multi-sensor change-point detection. *The Annals of Statistics*, 41(2):670–692.

Yeh, A. B., Lin, D. K. J., and McGrath, R. N. (2005). Multivariate Control Charts for Monitoring Covariance Matrix: A Review. *Quality Technology & Quantitative Management*, 3(4):415–436.

Zhang, N. R. and Siegmund, D. O. (2007). A Modified Bayes Information Criterion with Applications to the Analysis of Comparative Genomic Hybridization Data. *Biometrics*, 63(1):22–32. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1541-0420.2006.00662.x.

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

Zhang, W., James, N. A., and Matteson, D. S. (2017). Pruning and Nonparametric Multiple Change Point Detection. In *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 288–295. ISSN: 2375-9259.

Zhao, Z., Chen, L., and Lin, L. (2019). Change-point detection in dynamic networks via graphon estimation. *arXiv:1908.01823 [stat.ME]*. arXiv: 1908.01823.

Zhao, Z. and Yau, C. Y. (2019). Alternating Pruned Dynamic Programming for Multiple Epidemic Change-Point Estimation. *arXiv:1907.06810 [stat.ME]*, 1907.

Zou, C., Wang, Z., Zi, X., and Jiang, W. (2014a). An Efficient Online Monitoring Method for High-Dimensional Data Streams. *Technometrics*, 57(3):374–387.

Zou, C., Yin, G., Feng, L., and Wang, Z. (2014b). Nonparametric maximum likelihood approach to multiple change-point problems. *Annals of Statistics*, 42(3):970–1002. Publisher: Institute of Mathematical Statistics.

Papers

Paper I

Which principal components are most sensitive in the change detection problem?

Martin Tveten

Published in *Stat*, December 2019, volume 8, issue e252, DOI: 10.1002/sta4.252.

Received: 12 August 2019 | Revised: 6 September 2019 | Accepted: 12 September 2019
DOI: 10.1002/sta4.252

WILEY

ORIGINAL ARTICLE

Which principal components are most sensitive in the change detection problem?

Martin Tveten

Department of Mathematics, University of Oslo, Oslo, Norway

Correspondence
Martin Tveten, Department of Mathematics, University of Oslo, Niels Henrik Abels hus, Moltke Moes vei 35, 0851 Oslo, Norway.
Email: martintv@math.uio.no

Funding information
Norges Forskningsråd, Grant/Award Number: 237718

Principal component analysis (PCA) is often used in anomaly detection and statistical process control tasks. For bivariate normal data, we prove that the minor projection (the least varying projection) of the PCA-rotated data is the most sensitive to distributional changes, where sensitivity is defined as the Hellinger distance between the projections' marginal distributions before and after a change. In particular, this is almost always the case if only one parameter of the bivariate normal distribution changes, that is, the change is sparse. Simulations indicate that the minor projections are the most sensitive for a large range of changes and pre-change settings in higher dimensions as well, including changes that are very sparse. This motivates using only a few of the minor projections for detecting sparse distributional changes in high-dimensional data.

KEYWORDS

machine learning, quality control, statistical process control

1 | INTRODUCTION

It is popular to use principal component analysis (PCA) for anomaly detection and stochastic process control (SPC). Using PCA in SPC goes back to the work of Jackson and Morris (1957) and Jackson and Mudholkar (1979), and its various extensions (see Ketelaere et al., 2015 and Rato et al., 2016, for an overview) have been successfully applied to many real data situations. Within the machine learning literature on anomaly detection, Mishin et al. (2014) use PCA for temperature monitoring at Johns Hopkins, Harrou et al. (2015) apply PCA-based anomaly detection to find segments with abnormal rates of patient arrivals at an emergency department, and Camacho et al. (2016) relate PCA-based monitoring in SPC to modern anomaly detection in statistical networks. PCA has also been studied in the setting of change detection in multivariate functional data with the aim of detecting faulty profiles in a forging manufacturing process (Wang et al., 2018). Pimentel et al. (2014) provide an extensive review of novelty detection techniques and applications, and it is pointed to PCA being very useful for detecting outliers in this setting, for a large range of real world examples, covering industrial monitoring, video surveillance, text mining, sensor networks, and IT security. Moreover, many authors (Huang et al., 2007; Lakhina et al., 2004; Pimentel et al., 2014) acknowledge that it is most often the residual subspace of PCA that is most useful for outlier detection. On a similar note, Kuncheva and Faithfull (2014) offer an interesting alternative way to use PCA for change detection problems.

Most PCA-based methods utilize PCA in the intended way of creating a model based on retaining a small number of the most varying projections onto eigenvectors of the covariance matrix. As a consequence, the data are split into a model subspace that explains most of the variance in the data and a residual subspace. It is not self-evident that this is the best way to use PCA as a dimension reduction tool for change detection, so Kuncheva and Faithfull (2014) pose the question of which projections are the most sensitive to distributional changes in the data. Sensitivity is measured by a statistical divergence between the marginal distributions of projections before and after a change. They give a brief two-dimensional theoretical example that motivates monitoring the minor projections (the least varying projections) to detect anomalies that manifest in the form of sustained changes in the distribution of the data. An important feature of such an approach is that it can potentially be used to choose a subspace based on criteria linked to change detection, rather than on retaining data variance, hopefully yielding a better change and anomaly detection methods. The goal of this article is to give a more complete treatment of and extend the bivariate problem of Kuncheva and Faithfull (2014) in order to better understand the projections' sensitivity to changes under a simple setup and then study how these results carry over to higher dimensions by simulations.

There are three main differences between our approach and the approach of Kuncheva and Faithfull (2014). First, we express the projections' sensitivity to changes as functions of the parameters of the original data rather than of the parameters of the projections. The reason for this choice is that the original data are the object of the main interest, whereas the projections are ancillary. Our approach allows one to change individual parameters of the original data independently and see how this affects the marginal distributions of the projections as a consequence. We argue that this is more informative. Second, we study a much larger space of possible changes, including changes in only one parameter at a time. Such change scenarios where only a few of the dimensions change are called *sparse changes*, and they are the subject of much current interest (Chan, 2017; Liu et al., 2017; Wang et al., 2018; Wang & Samworth, 2018; Xie & Siegmund, 2013). Third, we measure sensitivity by the normal Hellinger distance between the marginal distributions of projections before and after a change, whereas Kuncheva and Faithfull (2014) use the normal Bhattacharyya distance. See Section 2 for an explanation of this choice.

In short, we find the following. For bivariate data, we prove that if only one of the two components' means changes in any direction, one component's variance increases, or the correlation between the components changes, the minor projection is the most sensitive. The principal projection is the most sensitive if one of the components' variance decreases and the correlation is not too close to 1. Lastly, if both means change, which projection is the most sensitive depends on the relative directions and sizes of change, and when both variances change by an equal amount, both projections are equally sensitive. Thus, on average (with all change scenarios up to a certain size equally likely), the minor projection is the most sensitive, mainly due to the sparse change scenarios. Our simulations confirm that the trend of the minor projections being more sensitive on average also holds for higher dimensions. Moreover, and most importantly, the minor projections seem to be quite sensitive even to very sparse changes. This knowledge carries large potential for creating more efficient change or anomaly detection methods.

The rest of the article is organized as follows: Section 2 formulates the problem precisely, Section 3 contains the theoretical results about sensitivity to changes in two dimensions, and in Section 4, we explore sensitivity in higher dimensions by simulations. The proofs are found in Appendix A.

2 | PROBLEM FORMULATION

Consider independent observations $x_t \in \mathbb{R}^D$, $t = 1, \ldots, n$, and let $\kappa \in \{1, \ldots, n-1\}$ be a change-point. For $t \leq \kappa$, the observations have mean $\boldsymbol{\mu_0}$ and covariance matrix Σ_0, whereas for $t > \kappa$, the data have mean μ_1 and covariance matrix Σ_1. Assume without loss of generality that the data are standardized with respect to the pre-change parameters, so that $\boldsymbol{\mu_0} = \mathbf{0}$ and Σ_0 is a correlation matrix with correlation parameter ρ. For $D = 2$, the changed mean is given by $\boldsymbol{\mu_1} = (\mu_1, \mu_2)^t$, and the changed covariance matrix can be expressed in terms of Σ_0 and parameter-wise multiplicative change factors as

$$\Sigma_1 = \begin{pmatrix} a_{11}^2 & a_{11}a_{22}a_{12}\rho \\ a_{11}a_{22}a_{12}\rho & a_{22}^2 \end{pmatrix},$$

where

$$-1 < \rho, a_{12}\rho < 1 \text{ and } \rho \neq 0. \tag{1}$$

For example, if $a_{11} = 2$, it means that the standard deviation of the first component has doubled compared with what it was originally in Σ_0. Similarly, $a_{12} = 0.5$ means that the correlation is half as strong after the change. Note that we exclude the degenerate cases of correlations equal to -1 and 1.

Next, let $\{\lambda_j, v_j\}_{j=1}^D$ be the normalized eigensystem of Σ_0, ordered by $\lambda_1 \geq \ldots \geq \lambda_D$. The orthogonal projections $y_{j,t} = v_j^t x_t$, with progressively decreasing variances λ_j, are our main objects of interest.

The general problem is to find out which of the D projections are the most sensitive to different distributional changes defined by $(\boldsymbol{\mu_1}, \Sigma_1)$, for each pre-change correlation matrix Σ_0. In the bivariate case, $(\Sigma_0, \boldsymbol{\mu_1}, \Sigma_1)$ is fully specified by $(\rho, \mu_1, \mu_2, a_{11}, a_{12}, a_{22})$. Note that a collection of the most and least varying $y_{j,t}$'s is referred to as the *principal projections* and *minor projections*, respectively.

We define sensitivity to changes as the normal Hellinger distance between the marginal distribution of a projection before and after a change. The squared Hellinger distance between two normal distributions $p(x) = N(x|\xi_1, \sigma_1^2)$ and $q(x) = N(x|\xi_2, \sigma_2^2)$ is given by

$$H^2(p,q) = 1 - \sqrt{\frac{2\sigma_1\sigma_2}{\sigma_1^2 + \sigma_2^2}} \exp\left\{-\frac{1}{4}\frac{(\xi_1 - \xi_2)^2}{\sigma_1^2 + \sigma_2^2}\right\}.$$

The formal definition of sensitivity to changes is contained in Definition 1.

Definition 1. For $j = 1, \ldots, D$, let p_j and q_j denote the marginal pre- and post-change density functions of $y_{j,t}$, respectively, given by

$$\begin{aligned} p_j(y) &= N(y \mid v_j^T \boldsymbol{\mu_0}, v_j^T \Sigma_0 v_j) = N(y|0, \lambda_j), \\ q_j(y) &= N(y \mid v_j^T \boldsymbol{\mu_1}, v_j^T \Sigma_1 v_j). \end{aligned}$$

The sensitivity of the jth projection based on Σ_0 to the change specified by $(\boldsymbol{\mu_1}, \Sigma_1)$ is defined as $H(p_j, q_j)$, abbreviated by H_j or $H_j(\Sigma_0, \boldsymbol{\mu_1}, \Sigma_1)$.

Our aim in the next section is to determine which pre-change parameters and changes the inequality $H_2 > H_1$ holds for when $D = 2$ in light of Definition 1.

Remark

(i) Kuncheva and Faithfull (2014) also define sensitivity as a divergence between distributions before and after a change but use the Bhattacharyya distance. The closely related Hellinger distance was chosen here because it turns out to be simpler to prove the sensitivity propositions because of Lemma 1 (see Appendix A). It is also an advantageous feature of the Hellinger distance that it is a true metric and takes values in $[0, 1]$. That it is a true metric implies for instance that a change in variance from 1 to $a > 1$ is an equally large change as from 1 to $1/a$ for the normal distribution. We find this an appealing feature because it is also a property of the generalized likelihood ratio test for a change in the mean and/or variance of normal data (see Hawkins & Zamba, 2005, for the corresponding test statistic).

(ii) One of the differences between our approach and the work of Kuncheva and Faithfull (2014) can now be stated more precisely. Our aim is to study the sensitivity of the $y_{j,t}$'s as functions of parameters of the original data x_t. Kuncheva and Faithfull (2014), on the other hand, study (additive) changes in the parameters of y_t directly; for instance, λ_j changing to $\lambda_j + a$ for all j, but without relating this a back to which Σ_1's this change corresponds to.

3 | BIVARIATE RESULTS

This section contains all the bivariate results about sensitivity to changes. The detailed proofs are given in Appendix A.

For changes in the mean in two-dimensional data, Proposition 1 gives the condition for determining which projection is the most sensitive, as well as the results for some special cases.

Proposition 1. *Let $a_{11} = a_{22} = a_{12} = 1$ and $\mu_1, \mu_2 \in \mathbb{R}$ while not both being 0 simultaneously (only the mean changes). $H_2 > H_1$ if and only if $(\mu_1 - \mu_2)^2/(\mu_1 + \mu_2)^2 > (1 - |\rho|)/(1 + |\rho|)$.*

In particular, for all $|\rho| \in (0, 1)$,

1. *$H_2 > H_1$ if one of μ_1 and μ_2 is 0 whereas the other is not (one mean changes).*
2. *$H_2 > H_1$ if $\mu_1 = -\mu_2 = \mu \neq 0$ (equal changes in opposite directions).*
3. *$H_2 < H_1$ if $\mu_1 = \mu_2 = \mu \neq 0$ (equal changes in the same direction).*

When both variances change by the same amount, Proposition 2 tells us that both projections are equally sensitive no matter what the pre-change correlation or size of the change is.

Proposition 2. *Let $\mu_1 = \mu_2 = 0$, $a_{12} = 1$ and $a_{11} = a_{22} = a \neq 1$ (both variances change equally). For any $|\rho| \in (0, 1)$ and $a > 0$, $H_2 = H_1$.*

The picture becomes more complicated when only one variance changes (Proposition 3). If the variance increases, the minor projection is always the most sensitive. On the other hand, if the variance decreases, the principal projection is mostly the most sensitive but not always if the pre-change correlation is high (greater than $\sqrt{3}/2$). In total, this gives a slight edge to the minor projection.

Proposition 3. *Let $\mu_1 = \mu_2 = 0$, $a_{12} = 1$, and either $a_{11} = 1$ and $a_{22} = a \neq 1$, or $a_{11} = a$ and $a_{22} = 1$, where $a > 0$ (one variance changes).*

1. *For any $|\rho| \in (0, 1)$ and $a > 1$ (variance increase), $H_2 > H_1$.*
2. *When $|\rho| \in (0, 1)$ and $a \in (0, 1)$ (variance decrease), $H_2 < H_1$ in most cases. The only exception is if $|\rho| \in (\sqrt{3}/2, 1)$ and $a \in (0, \sqrt{4\rho^2 - 3})$, where $H_2 > H_1$.*

Finally, for a change in correlation, the minor projection is the most sensitive in most cases (Proposition 4). Only if the correlation changes direction and becomes stronger is the principal projection more sensitive.

Proposition 4. *Let $\mu_1 = \mu_2 = 0$, $a_{11} = a_{22} = 1$ and $a_{12} = a \neq 1$ such that (1) holds (the correlation changes). Then $H_2 > H_1$ for any $|\rho| \in (0, 1)$ and $a > -1$.*

4 | EXPLORING HIGHER DIMENSIONS

In the two-dimensional case, we saw that which projection is the most sensitive depends both on the change $(\boldsymbol{\mu_1}, \Sigma_1)$ and on the pre-change correlation matrix Σ_0. For a higher dimension D, solving inequalities like above for all the parameters in $(\Sigma_0, \boldsymbol{\mu_1}, \Sigma_1)$ quickly becomes tedious and uninformative. Therefore, we use simulation to obtain Monte Carlo estimates $E[H_j(\Sigma_0, \boldsymbol{\mu_1}, \Sigma_1)]$ instead, where we vary which parameters that change, the size of the changes, and the sparsity of the change (the number of dimensions that change). Let $\rho_{i,d}$ for $i \neq d$ denote the off-diagonal elements of Σ_0, μ_d be the dth element of $\boldsymbol{\mu_1}$, and σ_d be the dth diagonal element of Σ_1. Then our simulation protocol to get such estimates is as follows:

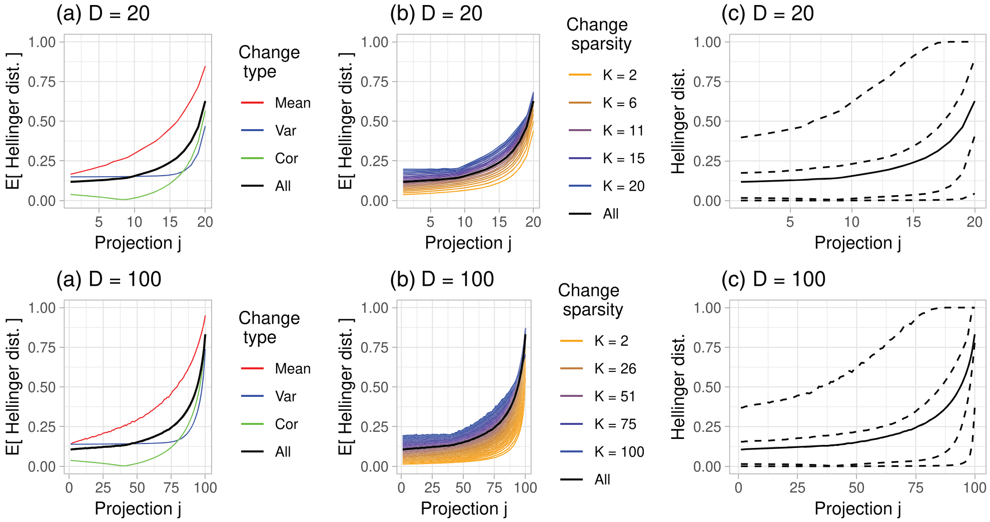

FIGURE 1 A summary of the sensitivity results obtained by the simulation protocol for $D = 20$ for $D = 100$. (a) Monte Carlo estimates of $E[H_j]$ for uniformly drawn changes in the mean, variance, and (decreases in) correlation, as well as uniformly drawn pre-change correlation matrices Σ_0. (b) Same as (a), but now the average sensitivity is conditional on the sparsity of the change, rather than the type of parameter. (c) 0.05, 0.25, 0.75, and 0.95 percentiles (the dashed lines, from bottom to top) of the distribution of H_j, together with $E[H_j]$ (solid line). Note that the percentiles are over Σ_0, $\boldsymbol{\mu}_1$, and Σ_1 simultaneously

1. Draw a correlation matrix Σ_0 uniformly from the space of correlation matrices by the method of Joe (2006) (clusterGeneration::rcorrmatrix in R).
2. Draw a change sparsity $K \sim \text{Unif}\{2, \ldots, D\}$.
3. Draw a random subset $\mathcal{D} \subseteq \{1, \ldots, D\}$ of size K.
4. Draw an additive change in mean $\mu \sim \text{Unif}(-3, 3)$, and set $\mu_d = \mu$ for $d \in \mathcal{D}$, whereas $\Sigma_1 = \Sigma_0$.
5. Draw a multiplicative change in standard deviation $\sigma \sim \frac{1}{2}\text{Unif}(1/3, 1) + \frac{1}{2}\text{Unif}(1, 3)$ (equal probability of decrease and increase in standard deviation) and set $\sigma_d = \sigma$ for $d \in \mathcal{D}$, keeping the remaining parameters constant.
6. Draw a multiplicative change in correlation $a \sim \text{Unif}(0, 1)$ and change $\rho_{i,d}$ to $a\rho_{i,d}$ for all $i \neq d \in \mathcal{D}$. The other parameters are kept constant.
7. For each of the three change scenarios 4–6, calculate $H_j(\Sigma_0, \boldsymbol{\mu}_1, \Sigma_1)$, $j = 1, \ldots, D$.
8. Repeat 2–7 10^3 times.
9. Repeat 1–8 10^3 times.

Averaging the simulated H_js yields estimates of $E[H_j]$, and we can condition on the type of parameter that changes and the change sparsity to see what the sensitivity is expected to be for different classes of changes. (Note that we only consider decreases in correlation. This is to avoid getting too many indefinite Σ_1's. If indefinite Σ_1's still occur, we find the closest positive-definite one by Higham's algorithm (Higham, 2002), implemented in the Matrix::nearPD-function in R.

Figure 1 shows that the trend of the minor components being the most sensitive continues for $D = 20$ and $D = 100$. This holds for changes in the mean, variance, and correlation (a) as well as all the different change sparsities (b). From the quantile plots (c), however, observe that a lot of variation is hidden in these averages, meaning that which projection is the most sensitive will depend on the specific Σ_0 and change $(\boldsymbol{\mu}_1, \Sigma_1)$, as in the bivariate case.

5 | CONCLUDING REMARKS

We have presented bivariate theory demonstrating that the minor projection of PCA-rotated data is usually the most sensitive to changes, especially if the change is sparse. Simulations confirm this to be the case on average for higher dimensions as well, but, in general, the sensitivity strongly varies with the pre-change correlation matrix and the specific change.

In future work, we aim to exploit these insights for creating computationally efficient change detection methods for high-dimensional data. The most promising and surprising part of our results is that even very sparse changes seem to be quite noticeable in the minor projections. This is important for change detection in high-dimensional data because a change rarely affects all dimensions or parameters at once. Most often, only a few parameters among many will change, and therefore, the problem of sparse changes will be the most relevant. One interpretation of the results presented here is that for detecting sparse changes in the mean vector and/or covariance matrix of a high-dimensional data set or of a sequentially arriving data stream, it is potentially sufficient to search for changes in a few selected minor projections. This might lead to major improvements, not only computationally but also in terms of detection accuracy or speed. Choosing which minor projections to use for a specific change detection problem is the subject of ongoing work.

This work is funded by the Norwegian Research Council centre Big Insight, Project 237718. The author would also like to thank Ingrid Glad for useful input on the presentation of the material.

SUPPORTING INFORMATION AND DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available as part of the supporting information for this online article: **R code** .R file with the code for reproducing (and easily extending) the simulation study and Figure 1.

ORCID

Martin Tveten https://orcid.org/0000-0002-4236-633X

REFERENCES

Camacho, J., Pérez-Villegas, A., García-Teodoro, P., & Maciá-Fernández, G. (2016). PCA-based multivariate statistical network monitoring for anomaly detection. *Computers & Security*, *59*, 118–137. https://doi.org/10.1016/j.cose.2016.02.008

Chan, H. P. (2017). Optimal sequential detection in multi-stream data. *The Annals of Statistics*, *45*(6), 2736–2763. https://doi.org/10.1214/17-AOS1546

Harrou, F., Kadri, F., Chaabane, S., Tahon, C., & Sun, Y. (2015). Improved principal component analysis for anomaly detection: Application to an emergency department. *Computers & Industrial Engineering*, *88*, 63–77. https://doi.org/10.1016/j.cie.2015.06.020

Hawkins, D. M., & Zamba, K. D (2005). Statistical process control for shifts in mean or variance using a changepoint formulation. *Technometrics*, *47*(2), 164–173. https://doi.org/10.1198/004017004000000644

Higham, N. J. (2002). Computing the nearest correlation matrix—A problem from finance. *IMA Journal of Numerical Analysis*, *22*(3), 329–343. https://doi.org/10.1093/imanum/22.3.329

Huang, L., Nguyen, X., Garofalakis, M., Jordan, M. I., Joseph, A., & Taft, N. (2007). In-network PCA and anomaly detection. In Schölkopf, B., Platt, J. C., & Hoffman, T. (Eds.), *Advances in Neural Information Processing Systems 19*. MA, USA: MIT Press, pp. 617–624.

Jackson, J. E., & Morris, R. H. (1957). An application of multivariate quality control to photographic processing. *Journal of the American Statistical Association*, *52*(278), 186–199.

Jackson, J. E., & Mudholkar, G. S. (1979). Control procedures for residuals associated with principal component analysis. *Technometrics*, *21*(3), 341–349. https://doi.org/10.1080/00401706.1979.10489779

Joe, H. (2006). Generating random correlation matrices based on partial correlations. *Journal of Multivariate Analysis*, *97*(10), 2177–2189. https://doi.org/10.1016/j.jmva.2005.05.010

Ketelaere, B. D., Hubert, M., & Schmitt, E. (2015). Overview of PCA-based statistical process-monitoring methods for time-dependent, high-dimensional data. *Journal of Quality Technology*, *47*(4), 318–335. https://doi.org/10.1080/00224065.2015.11918137

Kuncheva, L. I., & Faithfull, W. J. (2014). PCA Feature Extraction for Change Detection in Multidimensional Unlabeled Data. *IEEE transactions on neural networks and learning systems*, *25*(1), 69–80. https://doi.org/10.1109/TNNLS.2013.2248094

Lakhina, A., Crovella, M., & Diot, C. (2004). Diagnosing network-wide traffic anomalies. In *Proceedings of the 2004 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications*, ACM, New York, USA, pp. 219–230. https://doi.org/10.1145/1015467.1015492

Liu, K., Zhang, R., & Mei, Y. (2017). Scalable SUM-shrinkage schemes for distributed monitoring large-scale data streams. *Statistica Sinica*, *29*, 1–22. https://doi.org/10.5705/ss.202015.0316

Mishin, D., Brantner-Magee, K., Czako, F., & Szalay, A. S. (2014). Real time change point detection by incremental PCA in large scale sensor data. In *2014 IEEE High Performance Extreme Computing Conference (HPEC)*, pp. 1–6. https://doi.org/10.1109/HPEC.2014.7040959

Pimentel, M. A. F., Clifton, D. A., Clifton, L., & Tarassenko, L. (2014). A review of novelty detection. *Signal Processing*, *99*, 215–249. https://doi.org/10.1016/j.sigpro.2013.12.026

Rato, T., Reis, M., Schmitt, E., Hubert, M., & De Ketelaere, B. (2016). A systematic comparison of PCA-based statistical process monitoring methods for high-dimensional, time-dependent processes. *AIChE Journal*, *62*(5), 1478–1493. https://doi.org/10.1002/aic.15062

Wang, Y., Mei, Y., & Paynabar, K. (2018). Thresholded multivariate principal component analysis for phase I multichannel profile monitoring. *Technometrics*, *60*(3), 360–372. https://doi.org/10.1080/00401706.2017.1375993

Wang, T., & Samworth, R. J. (2018). High dimensional change point estimation via sparse projection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *80*(1), 57–83. https://doi.org/10.1111/rssb.12243

Xie, Y., & Siegmund, D. (2013). Sequential multi-sensor change-point detection. *The Annals of Statistics*, *41*(2), 670–692. https://doi.org/10.1214/13-AOS1094

How to cite this article: Tveten M. Which principal components are most sensitive in the change detection problem?. *Stat*. 2019;8:e252. https://doi.org/10.1002/sta4.252

APPENDIX A: PROOFS

Before turning to the proofs of the propositions in Section 3, the expressions for the pre- and post-change means and variances of each projection are needed. The normalized eigenvectors (principal axes) and corresponding eigenvalues (variance in the data along a given principal axis) of Σ_0 are quickly verified to be

$$\begin{aligned} \lambda_1 &= 1+\rho, \quad \mathbf{v}_1 = \frac{1}{\sqrt{2}}\begin{pmatrix}1\\1\end{pmatrix}, \\ \lambda_2 &= 1-\rho, \quad \mathbf{v}_2 = \frac{1}{\sqrt{2}}\begin{pmatrix}-1\\1\end{pmatrix}. \end{aligned} \tag{A1}$$

Note that which principal axis is the dominant one depends on the sign of ρ. If ρ is positive, $\mathbf{v}_1$ is the dominant one, but $\mathbf{v}_2$ is dominant if ρ is negative.

From the projections in (A1), the parameters of the projections before and after a change can be expressed as functions of the original correlation matrix and multiplicative change factors. For the principal component, the original and changed variances become as follows, respectively:

$$\begin{aligned} o_1^2 &= 1+\rho, \\ c_1^2 &= \frac{1}{2}a_{11}^2 + \frac{1}{2}a_{22}^2 + a_{11}a_{22}a_{12}\rho. \end{aligned} \tag{A2}$$

The expressions for the variances of the minor component are identical up to one switched sign:

$$\begin{aligned} o_2^2 &= 1-\rho, \\ c_2^2 &= \frac{1}{2}a_{11}^2 + \frac{1}{2}a_{22}^2 - a_{11}a_{22}a_{12}\rho. \end{aligned} \tag{A3}$$

Observe that if $\rho < 0$, then o_2 and c_2 would be equal to o_1 and c_1 with positive ρ, and vice versa. Thus, for $\rho \in (-1, 1)$, the general expressions are obtained by replacing ρ with $|\rho|$. Lastly, the changed mean components are given by

$$\begin{aligned} m_1 &= \frac{1}{\sqrt{2}}(\mu_1 + \mu_2), \\ m_2 &= \frac{1}{\sqrt{2}}(\mu_1 - \mu_2). \end{aligned} \tag{A4}$$

We first prove Proposition 1 for changes in the mean.

Proof of Proposition 1. Let $p_1(x) = N(x \mid 0, o_1^2)$, $q_1(x) = N(x \mid m_1, o_1^2)$, $p_2(x) = N(x \mid 0, o_2^2)$, and $q_2(x) = N(x \mid m_2, o_2^2)$, where m_i, o_i are as in (A2), (A3), and (A4), with ρ replaced by $|\rho|$ as noted above. The Hellinger distances between the distributions before and after a change along each principal axis are given by for $j = 1, 2$

$$H_j^2 = H^2(p_j, q_j) = 1 - \exp\left\{-\frac{1}{8o_j^2}m_j^2\right\}.$$

Then some algebra results in the inequality we needed to prove:

$$\begin{gathered} H_2 > H_1 \\ \frac{1}{8(1-|\rho|)}\frac{(\mu_1-\mu_2)^2}{2} > \frac{1}{8(1+|\rho|)}\frac{(\mu_1+\mu_2)^2}{2} \\ \frac{(\mu_1-\mu_2)^2}{(\mu_1+\mu_2)^2} > \frac{1-|\rho|}{1+|\rho|} \end{gathered}$$

From this inequality, the three special cases (i), (ii), and (iii) are immediately given. □

In the proofs concerning changes in the covariance matrix, we will make use of the following lemma. It reduces the inequality of Hellinger distances to a simpler inequality of ratios of variances.

Lemma 1. *Let p_1, q_1, p_2, q_2 be 0-mean normal distribution functions with variances $\sigma^2_{p_1}, \sigma^2_{q_1}, \sigma^2_{p_2}$, and $\sigma^2_{q_2}$, respectively. Furthermore, let*

$$\log r_j = \left|\log \frac{\sigma^2_{q_j}}{\sigma^2_{p_j}}\right|, \quad j = 1, 2.$$

Then $H(p_2, q_2) > H(p_1, q_1)$ if and only if $\log r_2 > \log r_1$.

Proof. First observe that when the means are 0, then we can write the Hellinger distance between two normal distributions as the following.

$$\begin{aligned} H^2(p,q) &= 1 - \left(\frac{2\sigma_p\sigma_q}{\sigma_p^2 + \sigma_q^2}\right)^{1/2} \\ &= 1 - \sqrt{2}\left(\frac{\sigma_p}{\sigma_q} + \frac{\sigma_q}{\sigma_p}\right)^{-1/2} \\ &= 1 - \sqrt{2}\left(\frac{\sigma_p^2}{\sigma_q^2} + \frac{\sigma_q^2}{\sigma_p^2} + 2\right)^{-1/4}. \end{aligned}$$

This gives us the inequality

$$\begin{gathered} H(p_2, q_2) > H(p_1, q_1) \\ \frac{\sigma^2_{p_2}}{\sigma^2_{q_2}} + \frac{\sigma^2_{q_2}}{\sigma^2_{p_2}} > \frac{\sigma^2_{p_1}}{\sigma^2_{q_1}} + \frac{\sigma^2_{q_1}}{\sigma^2_{p_1}}. \end{gathered}$$

By setting $r_2 = \sigma^2_{p_2}/\sigma^2_{q_2}$ and $r_1 = \sigma^2_{p_1}/\sigma^2_{q_1}$, the inequality can be written as

$$r_2 + r_2^{-1} > r_1 + r_1^{-1}.$$

Now assume first that $r_1, r_2 > 1$, that is, $\sigma^2_{p_j} > \sigma^2_{q_j}$. Then we see that

$$\begin{aligned} r_2 + r_2^{-1} &> r_1 + r_1^{-1} \\ r_2 - r_1 + \frac{r_1 - r_2}{r_1 r_2} &> 0 \\ (r_2 - r_1)\left(1 - \frac{1}{r_1 r_2}\right) &> 0. \end{aligned}$$

By the assumption that $r_1, r_2 > 1$, this inequality holds if and only if $r_2 > r_1$.

Finally, note that by interchanging $\sigma^2_{p_j}$ and $\sigma^2_{q_j}$, the same result is obtained when $\sigma^2_{q_j} \geq \sigma^2_{p_j}$. Thus, to make the result hold in general, we can set

$$r_j = \exp\left\{\left|\log \frac{\sigma^2_{q_j}}{\sigma^2_{p_j}}\right|\right\}, \quad j = 1, 2,$$

which is an expression for the ratio between variances where the largest of the variances is always in the numerator. Therefore, we get that $\log r_2 > \log r_1$ is equivalent to $H_2 > H_1$. □

The rest of this article contains the individual proofs of the remaining propositions in the main body of the text.

Proof of Proposition 2. Let $\log r_j$ for $j = 1, 2$ be defined as in Lemma 1. When assuming that $a_{12} = 1$ and $a_{11} = a_{22} = a \neq 1$, we get that

$$\log r_2 = \left|\log \frac{a^2/2 + a^2/2 - |\rho|a^2}{1 - |\rho|}\right| = |\log a^2|,$$

and

$$\log r_1 = \left|\log \frac{a^2/2 + a^2/2 + |\rho|a^2}{1 + |\rho|}\right| = |\log a^2|.$$

Hence, by arguments along the lines of the proof of Lemma 1, we see that $H_2 = H_1$ no matter what $|\rho|$ or a is. □

Proof of Proposition 3. Using the formulas for the variances of the projections (A2) and (A3), the inequality we have to study according to Lemma 1 becomes the following:

$$\begin{gathered}\left|\log\frac{a^2-2a|\rho|+1}{2(1-|\rho|)}\right| > \left|\log\frac{a^2+2a|\rho|+1}{2(1+|\rho|)}\right| \\ \left|\log\left[\frac{(1-a)^2}{2(1-|\rho|)}+a\right]\right| > \left|\log\left[\frac{(1-a)^2}{2(1+|\rho|)}+a\right]\right|.\end{gathered} \tag{A5}$$

First, we have to find the sign of the expressions inside the absolute values for each a and $|\rho|$. For the left-hand side, we get

$$\begin{gathered}\frac{(1-a)^2}{2(1-|\rho|)}+a=1 \\ a=1 \text{ and } a=2|\rho|-1.\end{gathered}$$

Thus, for $a>1$ and $a<2|\rho|-1$, the left-hand side is positive, whereas negative in between. For the right-hand side, the expression inside the absolute value signs are positive for $a>1$ and $a<-(1+2|\rho|)$. Because $a>0$, however, the relevant root for the right-hand side is only $a=1$. In total, this gives us three regions of $(a,|\rho|)$-values to check inequality (A5): $a>1$ and $|\rho|\in(0,1)$, $a\in(2|\rho|-1,1)$ and $|\rho|\in(0,1)$, and $a\in(0,2|\rho|-1)$ and $|\rho|\in(1/2,1)$.

$a>1$ and $|\rho|\in(0,1)$:
The absolute value signs can now be dissolved, so that inequality (A5) becomes

$$\frac{(1-a)^2}{(1-|\rho|)} > \frac{(1-a)^2}{(1+|\rho|)}.$$

Because $|\rho|\in(0,1)$, we see that the inequality holds for any $a>1$. Hence, $H_2>H_1$ in this scenario, when the variance increases.

$a\in(2|\rho|-1,1)$ and $|\rho|\in(0,1)$:
In this case, inequality (A5) becomes

$$\frac{(1-a)^2}{(1-|\rho|)} < \frac{(1-a)^2}{(1+|\rho|)}.$$

That is, it does not hold for any of the a's or $|\rho|$'s within the relevant region. Note that when $|\rho|<1/2$, a is kept between $(0,1)$.

$a\in(0,2|\rho|-1)$ and $|\rho|\in(1/2,1)$:
Now we get the inequality

$$\frac{(1-a)^2}{2(1-|\rho|)}+a > \left(\frac{(1-a)^2}{2(1+|\rho|)}+a\right)^{-1},$$

which is equivalent to

$$a^4-a^2(4\rho^2-2)+4\rho^2-3>0. \tag{A6}$$

The roots of the function on the left-hand side are $a=\pm1$ and $a=\pm\sqrt{4\rho^2-3}$, but the only relevant root for $a\in(0,2|\rho|-1)$ and $|\rho|\in(1/2,1)$ is $a_0:=\sqrt{4\rho^2-3}$.

Next, for $|\rho|<\sqrt{3}/2$, the root a_0 moves into the complex plane, and the function on the left-hand side of (A6) is always less than 0 for the relevant a's. That is, $H_2<H_1$ in this case. If $|\rho|>\sqrt{3}/2$, on the other hand, then (A6) holds for $a\in(0,a_0)$, but not for $a\in(a_0,2|\rho|-1)$.□

Proof of Proposition 4. In this scenario, the inequality to check due to Lemma 1 and expressions (A2) and (A3) is

$$\left|\log\frac{1-a|\rho|}{1-|\rho|}\right| > \left|\log\frac{1+a|\rho|}{1+|\rho|}\right|. \tag{A7}$$

To dissolve the absolute value signs, we first have to see for which values of a and $|\rho|$ the expressions inside are positive or negative. It is easily verified that the expression inside the left-hand side absolute value is positive for $a<1$, whereas the right-hand side is positive if $a>1$, both being negative otherwise.

First assume that $a<1$. Then inequality (A7) becomes

$$\begin{gathered}\frac{1-a|\rho|}{1-|\rho|} > \frac{1+|\rho|}{1+a|\rho|} \\ 1-(a\rho)^2 > 1-\rho^2 \\ a^2<1.\end{gathered}$$

Hence, $a \in (-1, 1)$ yields $H_2 > H_1$. On the other hand, if $a > 1$, we obtain

$$\frac{1-|\rho|}{1-a|\rho|} > \frac{1+a|\rho|}{1+|\rho|}$$
$$a^2 > 1,$$

which is always true. Thus, in total, $H_2 < H_1$ only if $a < -1$. □