Electronic Journal of Statistics
Vol. 12 (2018) 1901–1947
ISSN: 1935-7524
https://doi.org/10.1214/18-EJS1442

High dimensional efficiency with applications to change point tests*

John A.D. Aston[1] **Claudia Kirch**[2,3]

[1]*Statistical Laboratory, DPMMS, University of Cambridge, Cambridge, CB3 9HD, UK*
e-mail: j.aston@statslab.cam.ac.uk

[2]*Institute of Mathematical Stochastics, Department of Mathematics, Otto-von-Guericke University, Magdeburg, Germany*
e-mail: claudia.kirch@ovgu.de

[3]*Center for Behavioral Brain Sciences (CBBS), Magdeburg, Germany*

Abstract: This paper rigourously introduces the asymptotic concept of high dimensional efficiency which quantifies the detection power of different statistics in high dimensional multivariate settings. It allows for comparisons of different high dimensional methods with different null asymptotics and even different asymptotic behavior such as extremal-type asymptotics. The concept will be used to understand the power behavior of different test statistics as the performance will greatly depend on the assumptions made, such as sparseness or denseness of the signal. The effect of misspecification of the covariance on the power of the tests is also investigated, because in many high dimensional situations estimation of the full dependency (covariance) between the multivariate observations in the panel is often either computationally or even theoretically infeasible. The theoretic quantification by the theory is accompanied by simulation results which confirm the theoretic (asymptotic) findings for surprisingly small samples. The development of this concept was motivated by, but is by no means limited to, high-dimensional change point tests. It is shown that the concept of high dimensional efficiency is indeed suitable to describe small sample power.

MSC 2010 subject classifications: 62F05, 62M10, 62G10.
Keywords and phrases: CUSUM, high dimensional efficiency, model misspecification, panel data, projections.

Received November 2017.

1. Introduction

There has recently been a renaissance in research for statistical methods for change point problems (Horváth and Rice, 2014). This has been driven by applications where non-stationarities in the data can often be best described as change points in the data generating process (Eckley et al., 2011; Frick et al., 2014; Aston and Kirch, 2012b). However, data sets are now routinely considerably more complex than univariate time series classically studied in change point problems (Page, 1954; Robbins et al., 2011; Aue and Horváth, 2013; Horváth and Rice, 2014), and as such methodology for detecting and estimating change

*This is an original survey paper

points in a wide variety of settings, such as multivariate (Horváth et al., 1999; Ombao et al., 2005; Aue et al., 2009b; Kirch et al., 2015) functional (Berkes et al., 2009; Aue et al., 2009a; Hörmann and Kokoszka, 2010; Aston and Kirch, 2012a; Torgovitski, 2015) and high dimensional settings (Bai, 2010; Horváth and Hušková, 2012; Chan et al., 2012; Enikeeva and Harchaoui, 2013; Cho and Fryzlewicz, 2015) have recently been proposed. In panel data settings, these include methods based on taking maxima statistics across panels coordinate-wise (Jirak, 2015), use of scan statistic approaches (Enikeeva and Harchaoui, 2013), using sparsified binary segmentation for multiple change point detection (Cho and Fryzlewicz, 2015), uses of double CUSUM procedures (Cho, 2015), as well as those based on structural assumptions such as sparsity (Wang and Samworth, 2016).

In this paper, we develop a theoretic framework to understand and compare the power behavior of simple mean change tests in high dimensional settings. As benchmarks we investigate a class of tests based on projections, where the optimal (oracle) projection test is closely related to the likelihood ratio test under the knowledge of the direction of the change giving an upper benchmark. As a lower benchmark we consider a projection into a random direction. Secondly, we closely examine the power behavior of a universal change point test in this setting that has been introduced by Horváth and Hušková (2012). Here, we take universal to mean that its power behaviour does not depend on how sparse or dense the change is across the multivariate vector. The results and techniques, we introduce, can subsequently be extended to more complex change point settings as well as different statistical frameworks, such as two sample tests. In fact, Cho (2015) has already extended the findings from a preprint of this paper to some additional change point tests that have recently been proposed. We make use of the following two key concepts: Firstly, we consider contiguous changes where the size of the change tends to zero as the sample size and with it the number of dimensions increases leading to the notion of high dimensional efficiency. This concept is closely related to Asymptotic Relative Efficiency (ARE) (see Lehmann (1999, Sec. 3.4) and Lopes et al. (2011) where ARE is used in a high dimensional setting).

Optimal power in the sense of the oracle projection is only achieved if information about the direction of the change are known, where known can include assumptions such as sparse or balanced changes, meaning that there exists a small change of similar magnitude in each component. However, such procedures typically break down to the power of a random projection henceforth called tolerable power if those assumptions are not met. In addition, inherent misspecification in other parts of the model, such as the covariance structure, will have a detrimental effect on detection, which can result in procedures having no better than tolerable power.

We will consider a simple setup for our analysis, although one which is inherently the base for most other procedures, and one which can easily be extended to complex time dependencies and change point definitions using corresponding results from the literature (Kirch and Kamgaing, 2015, 2016). For a set of observations $X_{i,t}$, $1 \leqslant i \leqslant d = d_T, 1 \leqslant t \leqslant T$, the change point model is defined to

be

$$X_{i,t} = \mu_i + \delta_{i,T}\, g(t/T) + e_{i,t}, \quad 1 \leqslant i \leqslant d = d_T, 1 \leqslant t \leqslant T, \tag{1.1}$$

where $\mathrm{E}\, e_{i,t} = 0$ for all i and t with $0 < \sigma_i^2 = \operatorname{var} e_{i,t} < \infty$ and $g : [0,1] \to \mathbb{R}$ is a Riemann-integrable function. Here $\delta_{i,T}$ indicates the size of the change for each component. This setup incorporates a wide variety of possible changes by the suitable selection of the function g, as will be seen below. For simplicity, for now it is assumed that $\{e_{i,t} : t \in \mathbb{Z}\}$ are independent, i.e. we assume independence across time but not location. If the number of dimensions d is fixed, the results readily generalise to situations where a multivariate functional limit theorem exists as is the case for many weak dependent time series. If d can increase to infinity with T, then generalizations are possible if the $\{e_{i,t} : 1 \leqslant t \leqslant T\}$ form a linear process in time but the errors are independent between components (dependency between components will be discussed in detail in the next section). Existence of moments strictly larger than two is needed in all cases.

The change (direction) is given by $\boldsymbol{\Delta}_d = (\delta_{1,T}, \ldots, \delta_{d,T})'$ and the type of alternative is given by the function g in rescaled time. While g is defined in a general way, it includes as special cases most of the usual change point alternatives, for example,

- At most one change (AMOC): $g(u) = \begin{cases} 0 & 0 \leq u \leq \theta \\ 1 & \theta < u \leq 1 \end{cases}$
- Epidemic change (EC): $g(u) = \begin{cases} 0 & 0 \leq u \leq \theta_1 \\ 1 & \theta_1 < u < \theta_2 \\ 0 & \theta_2 < u \leq 1 \end{cases}$

The form of g will influence the choice of test statistic to detect the change point. As in the above two examples in the typical definition of change points the function g is modelled by a step function (which can approximate many smooth functions well). In such situations, test statistics based on partial sums of the observations have been well studied (Csörgő and Horváth, 1997). We focus on test statistics for the AMOC situation and show these statistics are robust (in the sense of still having non-zero power) to a wide variety of g. We derive the asymptotic theory for these partial sum processes and the results readily carry over to other statistics based on change point tests such as the ones for epidemic change points.

The model in (1.1) is defined for univariate ($d = 1$), multivariate (d fixed) or panel data ($d \to \infty$). The panel data (also known as "small n large p" or "high dimensional low sample size") setting is able to capture the small sample properties very well in situations where d is comparable or even larger than T using asymptotic considerations. In this asymptotic framework the detection ability or efficiency of various tests can be defined by the rates at which vanishing alternatives can still be detected. However, many of our results, particularly for the proposed projection tests, are also qualitatively valid in the multivariate or d fixed setting.

The paper proceeds as follows. In Section 2, the concept of high dimensional efficiency as a way of comparing the power of high dimensional tests is intro-

duced. Also in Section 2, we derive the high dimensional efficiency for the panel based change point statistics already suggested in Horváth and Hušková (2012). This will be done for a correctly specified covariance structure as well as if the covariance assumptions are violated. In Section 3 we develop the asymptotic theory for projection statistics which will act as a lower and upper benchmark for the panel change point test from the previous section. Here, too, misspecification will be taken into account. In Section 3.4 we summarize and interpret the high dimensional efficiency for a wide range of high dimensional change point tests recently proposed in the literature based on results obtained by Cho (2015). Section 4 provides a short illustrative example with respect to multivariate market index data. Section 5 concludes with some discussion of the different statistics proposed. The proofs in addition to some further illustrative material is given in an appendix. In addition, rather than a separate simulation section, simulations will be interspersed throughout the theory. They complement the theoretic results, confirming that the conclusion are already valid for small samples, thus verifying that the concept of high-dimensional efficiency is indeed suitable to understand the power behavior of different test statistics. In all cases the simulations are based on 1000 repetitions of i.i.d. normally distributed data for each set of situations, and unless otherwise stated the number of time points is $T = 100$ with the change (if present) occurring half way through the series. Except in the simulations concerning size itself, all results are empirically size corrected to account for the size issues for the multivariate (panel) statistic that will be seen in Figure 3.1.

2. High dimensional efficiency and a universal panel mean change test

In this section, we will first derive a theoretic framework called high dimensional efficiency – an asymptotic concept to compare the power of several high dimensional tests. Secondly, we will calculate this high dimensional efficiency for universal panel CUSUM tests (with $d \to \infty$) introduced by Horváth and Hušková (2012) extending a multivariate setting with d fixed (Horváth et al., 1999). Since we do not assume Gaussianity in order to obtain the corresponding limits it is necessary to assume independence between components, because the proofs are based on a central limit theorem across components. As such they cannot be generalized to uncorrelated (but dependent) data unless in the Gaussian case. For this reason, we cannot easily derive the asymptotic theory after standardization of the data. This is different from the multivariate situation, where this can easily be achieved. This test is related to a test in the high-dimensional two-sample situation by Srivastava and Du (2008) who consider some kind of misspecification of the covariance structure but under the stronger assumption of Gaussianity of the data.

We are interested in a comparison of the high dimensional efficiency under different covariance structures, which yield weighting matrices $\boldsymbol{A}$, for example, the correctly specified covariance, i.e. $\boldsymbol{A} = \Sigma^{-1}$, in addition to a comparison in

the misspecified case, $\boldsymbol{A} = \boldsymbol{M}^{-1}$, for some $\boldsymbol{M}$ not equal to the true covariance. The latter has already been discussed in one particular case by Horváth and Hušková (2012). To be precise, a common factor is introduced and the limit of the statistic (with $\boldsymbol{A} = \Lambda^{-1}$) under the assumption that the components are independent (i.e. Λ being a diagonal matrix) is considered. Because of the necessity to estimate the unknown covariance structure for practical purposes, the same qualitative effects as discussed here can be expected if a statistic and corresponding limit distribution were available for the covariance matrix Σ.

2.1. High dimensional efficiency

As the main focus of this paper is to compare several test statistics with respect to their detection power, we introduce a new asymptotic concept that allows us to understand this detection power in a high dimensional context. In the subsequent sections, simulations accompanying the theoretic results will show that this concept is indeed able to give insight into the small sample detection power. Thus this concept provides a theoretic tool for a power comparison, which – unlike simulation studies – gives a simultaneous insight into a large variety of situations.

Consider a typical testing situation, where (possibly after reparametrization) we test

$$H_0 : \mathbf{v}_d = 0, \quad \text{against} \quad H_1 : \mathbf{v_d} \neq 0, \tag{2.1}$$

for some parameter vector $\mathbf{v}_d \in \mathbb{R}^{l_d}$. In this paper, this vector will be the change, i.e. $\mathbf{v}_d = \boldsymbol{\Delta}_d = (\delta_{1,T}, \ldots, \delta_{d,T})'$. However, it could also be the mean vector in a one-sample location problem ($l_d = d$), or the difference of the mean vectors in a two-sample model ($l_d = d$). For change point testing in parametric time series models it could be the difference of the corresponding parameter vectors, where l_d is the effective dimension given by the number of unknown parameters in the situation where d-dimensional data is observed.

To understand the small sample power of different statistics we consider local or contiguous alternatives with $\mathbf{v}_d = \mathbf{v}_{d,T} \to 0$ (as $T \to \infty$). For a panel setting we define:

Definition 2.1. *Consider the testing situation (2.1) with sample size $T \to \infty$ and sample dimension $d = d_T \to \infty$. The (absolute)* **high dimensional efficiency** *$\mathcal{E} = \mathcal{E}^{(d)}$ of a test statistic $\mathcal{T}(\mathbf{X}_1, \ldots, \mathbf{X}_T)$ is a sequence of functions*

$$\mathcal{E}^{(d)} : \mathbb{R}^{l_d} \to \mathbb{R}_+ : \ \mathbf{v}_d \mapsto \mathcal{E}^{(d)}(\mathbf{v}_d),$$

such that

(i) $\mathcal{T}(\mathbf{X}_1, \ldots, \mathbf{X}_T) \xrightarrow{\mathcal{L}} L$ *for some non-degenerate limit distribution L under H_0,*

(ii) $\mathcal{T}(\mathbf{X}_1, \ldots, \mathbf{X}_T) \xrightarrow{P} \infty$ *if* $\sqrt{T}\,\mathcal{E}(\mathbf{v}_d) \to \infty$,

(iii) $\mathcal{T}(\mathbf{X}_1, \ldots, \mathbf{X}_T) \xrightarrow{\mathcal{L}} L$ *if* $\sqrt{T}\,\mathcal{E}(\mathbf{v}_d) \to 0$.

Obviously, $\mathcal{E}$ is only defined up to multiplicative constants, and has to be understood as a representative of the class $\mathcal{E}_1 \sim \mathcal{E}_2$ iff

$$0 < c \leqslant \liminf_{T\to\infty} \frac{\mathcal{E}_1(\mathbf{v}_d)}{\mathcal{E}_2(\mathbf{v}_d)} \leqslant \limsup_{T\to\infty} \frac{\mathcal{E}_1(\mathbf{v}_d)}{\mathcal{E}_2(\mathbf{v}_d)} \leqslant C < \infty$$

for all sequences of alternatives $\mathbf{v}_d$ and some constants c, C only depending on $\mathcal{E}_1$ and $\mathcal{E}_2$.

Remark 2.1. The above definition has the following connection to minimax optimality: If $\sqrt{T}\mathcal{E}(\mathbf{v}_d)$ is equal to the minimax separation rate in the sense of (Ingster and Suslina, 2012) (for a given norm), then the corresponding test is in fact minimax optimal in that sense (with respect to that norm).
Additionally, the above notion allows us to compare the power behavior for particular types of alternatives (such as e.g. sparse alternatives) of different test statistics leading to the notion of **relative high dimensional efficiency**, where it is not constants that are of interest but dependence on d. For example a high dimensional efficiency of $2d\|\mathbf{v}_d\|$ is a factor $\sqrt{d}$ better than one of $10\sqrt{d}\|\mathbf{v}_d\|$ resulting in a relative high dimensional efficiency of $\sqrt{d}$ for the first test versus the second one. As in the classical interpretation this means that the magnitude of $\|\mathbf{v}_d\|$ can be a factor $1/\sqrt{d}$ (again up to constants) smaller for the first test and still have the same detection power as the second test.

In particular a test has asymptotic power one for a sequence of alternatives if $\sqrt{T}\,\mathcal{E}(\mathbf{v}_d) \to \infty$, but a power equal to the level if $\sqrt{T}\,\mathcal{E}(\mathbf{v}_d) \to 0$. Typically, for $\sqrt{T}\,\mathcal{E}(\mathbf{v}_d) \to \alpha \neq 0$ it holds $\mathcal{T}(\mathbf{X}_1, \ldots, \mathbf{X}_T) \xrightarrow{\mathcal{L}} L(\alpha) \overset{\mathcal{D}}{\neq} L$, usually resulting in an asymptotic power strictly between the level and one. In the classic notion (with d constant) of absolute relative efficiency (ARE, or Pitman Efficiency) for test statistics with a standard normal limit it is the additive shift between $L(\alpha)$ and L (see Lehmann, 1999, Sec 3.4) that shows power differences for different statistics. Consequently, this shift has been used to define asymptotic efficiency. This idea has been considered in high dimensional settings by Lopes et al. (2011) as well as Wang et al. (2015) in a two-sample setting and by Srivastava and Du (2008) in the one-sample setting where the tests considered all converge to a standard normal limit – an assumption that is not true for most change point tests.

It turns out that the distinction made by the rates as captured by the high dimensional efficiency is already sufficient to compare the power behavior of the change point tests in this paper. In fact, if those rates differ, then the classic asymptotic relative efficiency is not defined (or rather yields 0 or ∞). It is only in situations, where the high dimensional efficiency of two tests is equal as e.g. in Lopes et al. (2011); Wang et al. (2015); Srivastava and Du (2008), that the constants as in the classic notion of efficiency become important to understand the differences in efficiency.

For standard test statistics exhibiting the usual distributional asymptotics, the above definition guarantees that the high dimensional efficiency $\mathcal{E}(\mathbf{v}_d)$ only depends on the type of alternatives and the dimension d but not on the sample

size T. The reason is that the rate with respect to the sample size of contiguous alternatives is the same as in classical testing namely $\sqrt{T}$. Nevertheless, the above concept also allows the investigation of test statistics exhibiting e.g. extreme-value behavior, which is often the case in change point analysis and appears for high dimensional change point tests see e.g. Chan et al. (2012) or Jirak (2015). In these examples, the high dimensional efficiency will now typically also depend on T as due to the extremal behavior the rate will no longer be $\sqrt{T}$ but $\sqrt{T/\log\log T}$. However, since the Logarithm converges very slowly, the dependence on d may be much more important. As illustrative example the maximum-likelihood CUSUM statistic which is an extreme-value-type change point test will be compared to a differently weighted CUSUM statistic with standard distributional asymptotics for $d = 1$ in Section A in the appendix.

2.2. Illustrative examples of spatial dependencies

In order to be able to prove asymptotic results for change point statistics even if $d \to \infty$, we need to make the following assumptions on the underlying error structure. This is much weaker than the independence assumption of the universal panel statistic from the previous section as considered by Horváth and Hušková (2012). Furthermore, we do not need to restrict the rate with which d grows. If we do have restrictions on the growth rate in particular for the multivariate setting with d fixed, these assumptions can be relaxed and more general error sequences can be allowed.

Assumption $\mathcal{A}$.1. Let $\eta_{1,t}(d), \eta_{2,t}(d), \ldots$ be independent with $\mathrm{E}\,\eta_{i,t}(d) = 0$, $\mathrm{var}\,\eta_{i,t}(d) = 1$ and $\mathrm{E}\,|\eta_{i,t}(d)|^{\nu} \leqslant C < \infty$ for some $\nu > 2$ and all i and d. For $t = 1, \ldots, T$ we additionally assume for simplicity that $(\eta_{1,t}(d), \eta_{2,t}(d), \ldots)$ are identically distributed (leading to data which is identically distributed across time). The errors within the components are then given as linear processes of these innovations:

$$e_{l,t}(d) = \sum_{j \geqslant 1} a_{l,j}(d)\eta_{j,t}(d), \quad l = 1, \ldots, d, \quad \sum_{j \geqslant 1} a_{l,j}(d)^2 < \infty$$

or equivalently in vector notation $e_t(d) = (e_{1,t}(d), \ldots, e_{d,t}(d))'$ and $\boldsymbol{a}_j(d) = (a_{1,j}(d), \ldots, a_{d,j}(d))'$

$$\mathbf{e}_t(d) = \sum_{j \geqslant 1} \mathbf{a}_j(d)\eta_{j,t}(d).$$

These assumptions allow us to consider many varied dependency relationships between the components (and we will concentrate on within the component dependency at this point, as temporal dependency adds multiple layers of notational difficulties, but little in the way of insight as almost all results generalise simply for weakly dependent and linear processes including the particular cases we will now discuss).

The following three cases of different spatial dependency structures are very helpful in understanding the effect of misspecification of the covariance structure on the high dimensional efficiency. They will be used as examples throughout the paper:

Case $\mathcal{C}.1$ (Independent Components)**.** The components are independent, i.e. $\mathbf{a}_j = (0, \ldots, s_j, \ldots, 0)'$ the vector which is $s_j > 0$ at point j and zero everywhere else, $j \leqslant d$, and $\mathbf{a}_j = \mathbf{0}$ for $j \geqslant d+1$. In particular, each channel has variance

$$\sigma_j^2 = s_j^2.$$

Case $\mathcal{C}.2$ (Fully Dependent Components)**.** There is one common factor to all components, leading to completely dependent components, i.e. $\mathbf{a}_1 = \mathbf{\Phi}_d = (\Phi_1, \ldots, \Phi_d)'$, $\mathbf{a}_j = \mathbf{0}$ for $j \geqslant 2$. In this case,

$$\sigma_j^2 = \Phi_j^2.$$

This case, while being somewhat pathological, is useful for gaining intuition into the effects of possible dependence and also helps with understanding the next case.

Case $\mathcal{C}.3$ (Mixed Components)**.** The components contain both an independent and dependent term. Let $\mathbf{a}_j = (0, \ldots, s_j, \ldots, 0)'$ the vector which is $s_j > 0$ at point j and zero everywhere else, and $\mathbf{a}_{d+1} = \mathbf{\Phi}_d = (\Phi_1, \ldots, \Phi_d)'$, $\mathbf{a}_j = \mathbf{0}$ for $j \geqslant d+2$. Then

$$\sigma_j^2 = s_j^2 + \Phi_j^2$$

This mixed case allows consideration of dependency structures between cases $\mathcal{C}.1$ and $\mathcal{C}.2$. It is used in the simulations with $\Phi_d = \Phi(1, \ldots, 1)'$, where $\Phi = 0$ corresponds to $\mathcal{C}.1$ and $\Phi \to \infty$ corresponds to $\mathcal{C}.3$. We also use this particular example for the universal panel statistic in this section to quantify the effect of misspecification.

Of course, many other dependency structures are possible, but these three cases give insight into the cases of no, complete and some dependency respectively.

2.3. Efficiency for universal change point test for independent panels

Multivariate CUSUM statistics have been adapted to the panel data setup under the assumption of independent components by Bai (2010) for estimation as well as Horváth and Hušková (2012) for testing. Those statistics are obtained as

weighted maxima or sum of the following (univariate) partial sum process

$$V_{d,T}(x) = \frac{1}{\sqrt{d}} \sum_{i=1}^{d} \left(\frac{1}{\sigma_i^2} Z_{T,i}^2(x) - \frac{\lfloor Tx \rfloor (T - \lfloor Tx \rfloor)}{T^2} \right), \tag{2.2}$$

$$\text{where } Z_{T,i}(x) = \frac{1}{T^{1/2}} \left(\sum_{t=1}^{\lfloor Tx \rfloor} X_{i,t} - \frac{\lfloor Tx \rfloor}{T} \sum_{t=1}^{T} X_{i,t} \right) \tag{2.3}$$

with $\mathbf{X}_d(t) = (X_{1,1}, \ldots, X_{d,T})'$ and $\sigma_i^2 = \operatorname{var} e_{i,1}$.

Theorem B.1 in the appendix gives a central limit theorem for errors as in Case $\mathcal{C}.1$ for this partial sum process (under the null) from which null asymptotics of the corresponding statistics can be derived if $\frac{d}{T^2} \to 0$. This was proven by Horváth and Hušková (2012, Theorem 1) who did allow for a linear process structure across time. However, the independence across components cannot be dropped, which has some effects on the high dimensional efficiency that we will investigate in Section 2.4. For the corresponding Darling-Erdösz-type theorem as discussed in Chan et al. (2012) the quite restrictive assumption $\frac{d}{T^2} \to 0$ can be dropped. The corresponding test is related to the weighted CUSUM test M_2 as discussed in Appendix A for the univariate case, which also exhibits a Darling-Erdösz-type asymptotic. As in the discussion there, this Darling-Erdösz-test has similar high dimensional efficiency as $\max_{0 \leqslant x \leqslant 1} V_{d,T}(x)$ up to an additional log-term, but will not be discussed here in detail.

The following theorem derives the high dimensional efficiency in this setting for the universal panel statistics such as $\max_{0 \leqslant x \leqslant 1} V_{d,T}(x)$, which we use in simulations with both known and estimated standard deviations, or $\int_0^1 V_{d,T}(x)$.

Theorem 2.1. *Let Case $\mathcal{C}.1$ hold with $\sigma_i^2 = s_i^2 \geqslant c > 0$ for all i, which implies in particular that $\Sigma = diag(\sigma_1^2, \ldots, \sigma_d^2)$, and $\limsup_{d \to \infty} \frac{1}{d} \sum_{i=1}^{d} \mathrm{E} |e_{i,t}|^{\nu} < \infty$ for some $\nu > 4$. Furthermore, let $\frac{d}{T^2} \to 0$. Then, the high dimensional efficiency of the universal panel statistic tests is given by*

$$\mathcal{E}_1(\boldsymbol{\Delta}_d) = \frac{1}{d^{1/4}} \|\Sigma^{-1/2} \boldsymbol{\Delta}_d\|,$$

where $\|\cdot\|$ refers to the Euclidean norm.

Most notably the efficiency of this test statistic unlike test statistics particularly developed for sparse changes as discussed in Section 3.4 only depends on the magnitude of the covariance scaled change but not on how the mass of the change is placed within the vector (i.e. if it is balanced across the vector or only sparsely in a few components). Proposition 1 in Baraud et al. (2002) shows that the minimax separation rate in the L_2-norm for the signal detection problem (which also provides lower bounds for the present change point problem) is given by $d^{1/4}/\sqrt{T}$, i.e. no uniform test exists in the above change point situation when $\sqrt{T}\mathcal{E}_1(\boldsymbol{\Delta}_d) \to 0$. Consequently, the test by Horváth and Hušková (2012) achieves L_2-minimax optimality in the sense of (Ingster and Suslina, 2012) and cannot be improved uniformly over all $\boldsymbol{\Delta}$.

For constant magnitude of the change the efficiency is given by $d^{-1/4}$ and as such will turn out to be a magnitude of $d^{-1/4}$ worse than oracle efficiency but a magnitude of $d^{-1/4}$ better than tolerable efficiency (as obtained by a random projection).

We can see the finite sample nature of this phenomena in Figure 3.2 (a).

2.4. *Efficiency of universal change point tests under dependence between components*

We now turn again to the misspecified situation, where we use the above statistic in a situation where components are not uncorrelated. Following Horváth and Hušková (2012), we consider the mixed case $\mathcal{C}.3$ for illustration. The next proposition derives the null limit distribution for that special case. It turns out that the limit as well as convergence rates depend on the strength of the contamination by the common factor.

Lemma 2.2. *Let Case $\mathcal{C}.3$ hold with $\nu > 4$, $0 < c \leqslant s_i \leqslant C < \infty$ and $\Phi_i^2 \leqslant C < \infty$ for all i and some constants c, C and consider $V_{d,T}(x)$ defined as in (2.2), where $\sigma_i^2 = \operatorname{var} e_{i,1}$ but the rest of the dependency structure is not taken into account. The asymptotic behavior of $V_{d,T}(x)$ then depends on the behavior of*

$$A_d := \sum_{i=1}^{d} \frac{\Phi_i^2}{\sigma_i^2}.$$

a) If $A_d/\sqrt{d} \to 0$, then the dependency is negligible, i.e.

$$V_{d,T}(x) \xrightarrow{D[0,1]} \sqrt{2}(1-x)^2 W\left(\frac{x^2}{(1-x)^2}\right),$$

where $W(\cdot)$ is a standard Wiener process.

b) If $A_d/\sqrt{d} \to \xi$, $0 < \xi < 1$, then

$$V_{d,T}(x) \xrightarrow{D[0,1]} \sqrt{2}(1-x)^2 W\left(\frac{x^2}{(1-x)^2}\right) + \xi\left(B^2(x) - x(1-x)\right),$$

where $W(\cdot)$ is a standard Wiener process and $B(\cdot)$ is a standard Brownian bridge.

c) If $A_d/\sqrt{d} \to \infty$, then

$$\frac{\sqrt{d}\, V_{d,T}(x)}{A_d} \xrightarrow{D[0,1]} B^2(x) - x(1-x),$$

where $\{B(x) : 0 \leqslant x \leqslant 1\}$ is a standard Brownian bridge.

Because A_d in the above theorem cannot feasibly be estimated, this result cannot be used to derive critical values for panel test statistics. Consequently,

the exact shape of the limit distribution in the above lemma is not important. However, the lemma is necessary to derive the high dimensional efficiency of the panel statistics in this misspecified case. Furthermore, it indicates that using the limit distribution from the previous section to derive critical values will result in asymptotically wrong sizes if a strong contamination by a common factor is present. The simulations in Figure 3.1 also confirm this fact and show that the size distortion can be enormous. It does not matter whether the variance of the components in the panel statistic takes into account the dependency or simply uses the noise variance (Figure 3.1(a)), or whether a change is accounted for or not in the estimation (Figure 3.1(b)-(c)). This illustrates, that the full panel statistic is very sensitive with respect to deviations from the assumed underlying covariance structure in terms of size.

In the situation of a) and b) above, the dependency structure introduced by the common factor is still small enough asymptotically to not change the high dimensional efficiency as given in Theorem 2.1, which is analogous to the proof of Theorem 2.1. Therefore, we will now concentrate on situation c), which is the case where the noise coming from the common factor does not disappear asymptotically.

Theorem 2.3. *Let the assumptions of Lemma 2.2 on the errors be fulfilled and* $A_d/\sqrt{d} \to \infty$*, then the corresponding panel statistics have high dimensional efficiency*

$$\mathcal{E}_2(\mathbf{\Delta}_d) = \frac{1}{\sqrt{A_d}} \sqrt{\mathbf{\Delta}_d' \, diag\left(\frac{1}{s_1^2 + \Phi_1^2}, \ldots, \frac{1}{s_d^2 + \Phi_d^2}\right) \mathbf{\Delta}_d}.$$

Corollary 3.8 below will show that the efficiency of the universal panel test becomes as bad as the tolerable efficiency if $A_d/d \to A > 0$, which is typically the case if the dependency is non-sparse and non-negligible.

3. Change points and projections

3.1. Projections

We now describe how projections can be used to obtain change point statistics in high dimensional settings, which will be used as an upper (in the form of an oracle projection) and a lower benchmark (in the form of a random projection) statistics for other change point tests.

In model (1.1), the change $\mathbf{\Delta}_d = (\delta_{1,T}, \ldots, \delta_{d,T})'$ (as a direction) is always a rank one (vector) object no matter the number of components d. This observation suggests that knowing the direction of the change $\mathbf{\Delta}_d$ in addition to the underlying covariance structure can significantly increase the signal-to-noise ratio. Furthermore, for μ and $\mathbf{\Delta}_d/\|\mathbf{\Delta}_d\|$ (but not $\|\mathbf{\Delta}_d\|$) known with i.i.d. normal errors, one can easily verify that the corresponding likelihood ratio statistic is obtained as a projection statistic with projection vector $\Sigma^{-1}\mathbf{\Delta}_d$, which can also

be viewed as an oracle projection. Under (1.1) it holds

$$\langle \mathbf{X}_d(t), \mathbf{p}_d \rangle = \langle \boldsymbol{\mu}, \mathbf{p}_d \rangle + \langle \boldsymbol{\Delta}_d, \mathbf{p}_d \rangle g(t/T) + \langle \mathbf{e}_t, \mathbf{p}_d \rangle,$$

where $\mathbf{X}_d(t) = (X_{1,t}, \ldots, X_{d,T})'$, $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_d)'$ and $\mathbf{e}_t = (e_{1,t}, \ldots, e_{d,t})'$. The projection vector $\mathbf{p}_d$ plays a crucial role in the following analysis and will be called the search direction. Because multiplicative constants do not change the signal-to-noise ratio nor the high dimensional efficiency, we will always use the normalized vector $\|p_d\| = 1$ for simplicity. This representation shows that the projected time series exhibits the same behavior as before as long as the change is not orthogonal to the projection vector. Furthermore, the power is better the larger $\langle \boldsymbol{\Delta}_d, \mathbf{p}_d \rangle$ and the smaller the variance of $\langle \mathbf{e}_t, \mathbf{p}_d \rangle$. Consequently, an optimal projection in terms of power depends on $\boldsymbol{\Delta}_d$ as well as $\Sigma = \operatorname{var} \mathbf{e}_1$.

3.2. Efficiency of change point tests based on projections

In this section, we derive the efficiency of change point tests based on projections under rather general assumptions. Furthermore, we will see that the size behavior is very robust with respect to deviations from the assumed underlying covariance structure. The power on the other hand turns out to be less robust but more so than statistics taking the full multivariate information into account.As special cases, we then obtain our benchmark efficiencies, the oracle efficiency and the tolerable efficiency (obtained from random projections) in Section 3.3.

Standard statistics such as the CUSUM statistic are based on partial sum processes, so in order to quantify the possible power gain by the use of projections we will consider the partial sum process of the projections, i.e.

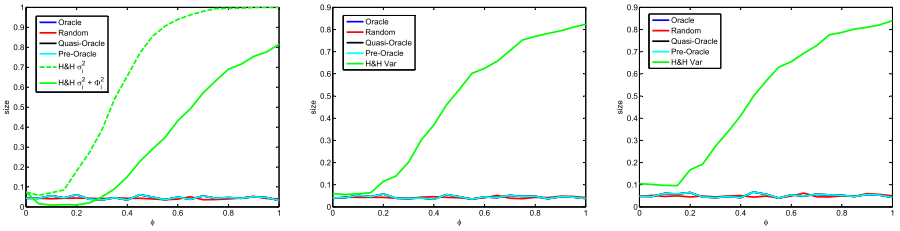
$$U_{d,T}(x) = \langle \mathbf{Z}_T(x), \mathbf{p}_d \rangle = \frac{1}{\sqrt{T}} \sum_{t=1}^{\lfloor Tx \rfloor} \left(\langle \mathbf{X}_d(t), \mathbf{p}_d \rangle - \frac{1}{T} \sum_{j=1}^{T} \langle \mathbf{X}_d(j), \mathbf{p}_d \rangle \right), \quad (3.1)$$

where $Z_{T,i}$ is as in (2.3). Different test statistics can be defined for a range of g (see Section C.1 in the appendix for more details). One popular test statistic designed for the at-most-one-change (but with power against any non-constant g) is the max-type statistic, analogous to that for the universal panel test given above.

In this section we first derive a functional central limit theorem for the process $U_{d,T}(x)$, which implies the asymptotic null behavior for these tests. Then, we derive the asymptotic behavior of the partial sum process under contiguous alternatives to obtain the high dimensional efficiency for projection statistics.

3.2.1. Null asymptotics

As a first step towards the efficiency of projection statistics, we derive the null asymptotics. This is also of independent interest if projection statistics are applied to a given data set in order to find appropriate critical values. In the



(a) Known variance (b) Estimated variance as in (C.5) (c) Estimated variance as in (C.6)

Fig 3.1: Size of tests as the degree of dependency between the components increases. As can be seen, all the projection methods, Oracle, Quasi-Oracle, Pre-Oracle and Random projections defined in Section 3.3 maintain the size of the tests. Those based on using the full information, the universal test (indicated as H&H) as described in Section 2 have size problems as the degree of dependency increases. The simulations correspond to Case $\mathcal{C}.3$ with $s_j = 1, \Phi_j = \phi$, $j = 1, \ldots, d$ with $d = 200$, where ϕ is given on the x-axis).

following theorem d can be fixed but it is also allowed that $d = d_T \to \infty$, where no restrictions on the rate of convergence are necessary.

Theorem 3.1. *Let model (1.1) hold. Let $\mathbf{p}_d$ be a possibly random projection independent of $\{e_{i,t} : 1 \leqslant t \leqslant T, 1 \leqslant i \leqslant d\}$. Furthermore, let $\tau^2(\mathbf{p}_d) \neq 0$ (almost surely), which means that the projected data is not degenerate with probability one. Under Assumption $\mathcal{A}.1$ and if $\{\mathbf{p}_d\}$ is independent of $\{\eta_{i,t}(d) : i \geqslant 1, 1 \leqslant t \leqslant T\}$, then it holds under the null hypothesis*

$$\left\{ \frac{U_{d,T}(x)}{\tau(\mathbf{p}_d)} : 0 \leqslant x \leqslant 1 \,|\, \mathbf{p}_d \right\} \xrightarrow{D[0,1]} \{B(x) : 0 \leqslant x \leqslant 1\} \qquad a.s., \tag{3.2}$$

where $B(\cdot)$ is a standard Brownian bridge. The assertions remain true if $\tau^2(\mathbf{p}_d)$ is replaced by $\widehat{\tau}^2_{d,T}$ such that for all $\epsilon > 0$

$$P\left(\left| \frac{\widehat{\tau}^2_{d,T}}{\tau^2(\mathbf{p}_d)} - 1 \right| > \epsilon \,|\, \mathbf{p}_d \right) \to 0 \qquad a.s. \tag{3.3}$$

Assumption $\mathcal{A}.1$ can be replaced by a different assumption which is always fulfilled in the multivariate case but often too restrictive in the panel situation (see Theorem C.1 in the appendix for more details). Lemma C.2 shows that the following estimators fulfill (3.3):

$$\widehat{\tau}^2_{1,d,T}(\mathbf{p}_d) = \frac{1}{T} \sum_{j=1}^{T} \left(\mathbf{p}_d' \mathbf{e}_t(d) - \frac{1}{T} \sum_{i=1}^{T} \mathbf{p}_d' \mathbf{e}_t(d) \right)^2$$

$$\widehat{\tau}^2_{2,d,T}(\mathbf{p}_d) = \frac{1}{T}\left(\sum_{j=1}^{\widehat{k}_{d,T}}\left(\mathbf{p}_d'\mathbf{e}_j(d) - \frac{1}{T}\sum_{i=1}^{\widehat{k}_{d,T}}\mathbf{p}_d'\mathbf{e}_i(d)\right)^2 + \sum_{j=\widehat{k}_{d,T}+1}^{\prime}\left(\mathbf{p}_d'\mathbf{e}_t(d) - \frac{1}{T}\sum_{i=\widehat{k}_{d,T}+1}^{T}\mathbf{p}_d'\mathbf{e}_i(d)\right)^2\right),$$

where $\widehat{k}_{d,T} = \arg\max_{t=1,\ldots,T} U_{d,T}(t/T)$.

The second estimator is typically used in an at-most-one-change setting as it usually leads to a small-sample power improvement for the corresponding tests as it is also consistent under the at-most-one-change alternative.

From Theorem 3.1 one can easily derive the null asymptotics for standard change point tests such as the max-type and sum-type tests (see Corollary C.3 in the appendix for more details). As can be seen in Figure 3.1, regardless of whether the variance is known or estimated, the projection methods all maintain the correct size even when there is a high degree of dependence between the different components (the specific projection methods will be characterised in Section 3.3 below).

3.2.2. Absolute high dimensional efficiency

We are now ready to derive the high dimensional efficiency of projection statistics. Furthermore, we show that a related estimator for the location of the change is asymptotically consistent.

Theorem 3.2. *Let the assumptions of Theorem 3.1 be fulfilled. Then, the max-type statistic based on 3.1 has the following absolute high dimensional efficiency:*

$$\mathcal{E}_3(\boldsymbol{\Delta}_d, \mathbf{p}_d) := \frac{\|\boldsymbol{\Delta}_d\|\|\mathbf{p}_d\|\cos(\alpha_{\boldsymbol{\Delta}_d,\mathbf{p}_d})}{\tau(\mathbf{p}_d)} = \frac{|\langle\boldsymbol{\Delta}_d, \mathbf{p}_d\rangle|}{\tau(\mathbf{p}_d)}, \tag{3.4}$$

where $\tau^2(\mathbf{p}_d)$ is as in Theorem 3.1 and $\alpha_{\mathbf{u},\mathbf{v}}$ is the (smallest) angle between $\mathbf{u}$ and $\mathbf{v}$. In addition, the asymptotic power increases with increasing multiplicative constant.

The assertion remains true under the assumptions of Theorem C.1 as well as for the max and sum-type statistics with a weight function $w(\cdot)$ as in Corollary C.3 fulfilling $w^2(x)\left(\int_0^x g(t)\,dt - x\int_0^1 g(t)\,dt\right)^2 \neq 0$.

In the following, $\mathcal{E}_3(\boldsymbol{\Delta}_d, \mathbf{p}_d)$ is fixed to the above representative of the class, so that different projection procedures with the same rate but with different constants can be compared.

Remark 3.1. For random projections the high dimensional efficiency is a random variable. The convergences in Definition 2.1 is understood given the projection vector $\mathbf{p}_d$, where we get either $a.s.$-convergence or P-stochastic convergence

depending on whether $\sqrt{T}\,\mathcal{E}_3(\mathbf{\Delta}_d, \mathbf{p}_d)$ converges *a.s.* or in a P-stochastic sense (in the latter case the assertion follows from the subsequence-principle).

The above result shows in particular that sufficiently large changes (as defined by the high dimensional efficiency) are detected asymptotically with power one. For such changes in the at-most-one-change situation, one can easily derive that the corresponding change point estimator is consistent in rescaled time (see Corollary C.4 in the appendix).

Remark 3.2. The proof shows in particular that all deviations from a stationary mean are detectable with asymptotic power one as $\int_0^x g(t)\,dt - x\int_0^1 g(t)\,dt \neq 0$ for non-constant g. It is this g function which determines which weight function gives best power.

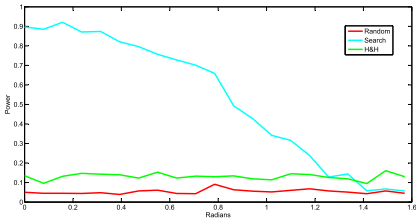
Remark 3.3. We derive the high dimensional efficiency for a given g and disappearing magnitude of the change $\|\mathbf{\Delta}_d\|$. For an epidemic change situation with $g(x) = 1_{\{\vartheta_1 < x \leqslant \vartheta_2\}}$ for some $0 < \vartheta_1 < \vartheta_2 < 1$, this means that the duration of the change is relatively large but the magnitude relatively small with respect to the sample size. Alternatively, one could also consider the situation, where the duration gets smaller asymptotically (see e.g. Frick et al. (2014)) resulting in a different high dimensional efficiency, which is equal for both the projection as well as the multivariate or panel statistic, as long as the same weight function and the same type of statistic (sum/max) is used. Some preliminary investigations suggest that in this case using projections based on principle component analysis similar to Aston and Kirch (2012a) can be advantageous, however this is not true for the setting discussed in this paper.

In the next section we will further investigate the high dimensional efficiency and see that the power depends essentially on the angle between $\Sigma^{1/2}\mathbf{p}_d$ and the 'standardized' change $\Sigma^{-1/2}\mathbf{\Delta}$ if Σ is invertible. In fact, the smaller the angle the larger the power. Some interesting insight can also come from the situation where Σ is not invertible by considering case $\mathcal{C}.2$ above (and this is given in section C.4 of the appendix).

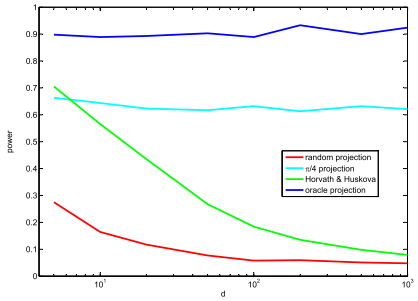
3.3. High dimensional efficiency of oracle and random projections

In this section, we will further investigate the high dimensional efficiency of certain particular projections that can be viewed as benchmark projections. In particular, we will see that the efficiency depends only on the angle between the projection and the change both properly scaled with the underlying covariance structure.

The highest efficiency is obtained by $\mathbf{o} = \Sigma^{-1}\mathbf{\Delta}_d$ as the next theorem shows, which will be called the oracle projection. This oracle is equivalent to a projection after first standardizing the data on the 'new' change $\Sigma^{-1/2}\mathbf{\Delta}_d$. The corresponding test is related to the likelihood ratio statistic for i.i.d. normal innovations, where both the original mean and the direction (but not magnitude) of the change are known. This oracle is effectively the optimal linear classifier



(a) Power of tests as the angle between the search direction and the oracle increases: $\Sigma_d = I_d$, $d = 200$, and $\boldsymbol{\Delta}_d = 0.05\ \mathbf{1_d}$, corresponding to Case $\mathcal{C}$.1



(b) Power of the tests as d increases with a fixed sample size ($T = 100$): $\|\boldsymbol{\Delta}_d\| =$ const.and $\Sigma_d = I_d$, i.e. $\|\Sigma^{-1/2}\boldsymbol{\Delta}_d\| =$ const., corresponding to Case $\mathcal{C}$.1

Fig 3.2: As can be seen in (a), the search projection method decreases similarly to cosine of the angle, while the random projection and universal panel tests (H&H) as introduced in Section 2 are given for comparison. In (b) as predicted by theory we get roughly constant power for fixed angle projection tests (as $\|\boldsymbol{\Delta}_d\|$ is constant), while results in decreasing power for both the panel statistic test and random projections as predicted by theory.

as proposed by Delaigle and Hall (2012). As a lower (worst case) benchmark we consider a scaled random projection $\mathbf{r}_{d,\Sigma} = \Sigma^{-1/2}\mathbf{r}_d$, where $\mathbf{r}_d$ is a random projection on the d-dimensional unit sphere. This is equivalent to a random projection onto the unit sphere after standardizing the data. Both projections depend on Σ which is usually not known so that it needs to be estimated. The latter is rather problematic in particular in high dimensional settings without additional parametric or sparsity assumptions (see Zou et al. (2006), Bickel and Levina (2008) and Fan et al. (2013) including related discussion, and Cai and Liu (2011) for a case where the assumption of sparsity can be used to facilitate direct estimation of the vector of interest without full covariance estimation). Furthermore, it is actually the inverse that needs to be estimated which results in additional numerical problems if d is large. For this reason we check the robustness of the procedure with respect to not knowing or misspecifying Σ in a second part of this section.

3.3.1. Correctly scaled projections

In this section we characterize which projection yields an optimal high dimensional efficiency associated with the highest power if the covariance matrix Σ is invertible. In Section C.4 in the appendix, we look at the situation if Σ is not invertible.

Proposition 3.3. *If Σ is invertible, then*

$$\mathcal{E}_3(\boldsymbol{\Delta}, \mathbf{p}_d) = \|\Sigma^{-1/2}\boldsymbol{\Delta}_d\| \cos(\alpha_{\Sigma^{-1/2}\boldsymbol{\Delta}_d, \Sigma^{1/2}\mathbf{p}_d}). \tag{3.5}$$

Proposition 3.3 shows in particular, that after standardizing the data, i.e. for $\Sigma = I_d$, the power depends solely on the cosine of the angle between the oracle and the projection (see Figure 3.2 (a)).

From the representation in this proposition it follows immediately that the 'oracle' choice for the projection to maximize the high dimensional efficiency is $\boldsymbol{o} = \Sigma^{-1}\boldsymbol{\Delta}_d$ as it maximizes the only term which involves the projection namely $\cos(\alpha_{\Sigma^{-1/2}\boldsymbol{\Delta}_d, \Sigma^{1/2}\mathbf{p}_d})$. Therefore, we define:

Definition 3.1. *The projection* $\mathbf{o} = \Sigma^{-1}\boldsymbol{\Delta}_d$ *is called* **oracle** *if* Σ^{-1} *exists. Since the projection procedure is invariant under multiplications with non-zero constants of the projected vector, all non-zero multiples of the oracle have the same properties, so that they correspond to a class of projections.*

By Proposition 3.3 the oracle choice leads to a high dimensional efficiency of $\mathcal{E}_3(\boldsymbol{\Delta}_d, \mathbf{o}) = \|\Sigma^{-1/2}\boldsymbol{\Delta}_d\|$.

Another way of understanding the Oracle projection is the following: If we first standardize the data, then for a projection on a unit (w.l.o.g.) vector the variance of the noise is constant and the signal is given by the scalar product of $\Sigma^{-1/2}\boldsymbol{\Delta}$ and the (unit) projection vector, which is obviously maximized by a projection with $\Sigma^{-1/2}\boldsymbol{\Delta}/\|\Sigma^{-1/2}\boldsymbol{\Delta}\|$ which is equivalent to using $\mathbf{p}_d = \Sigma^{-1}\boldsymbol{\Delta}$ as a projection vector for the original non-standardized version.

So, if we know Σ and want to maximize the efficiency respectively power close to a particular search direction $\mathbf{s}_d$ of our interest, we should use the **scaled search direction** $\mathbf{s}_{\Sigma,d} = \Sigma^{-1}\mathbf{s}_d$ as a projection.

Because the cosine falls very slowly close to zero, the efficiency will be good if the search direction is not too far off the true change. From this, one could get the impression that even a scaled random projection $\mathbf{r}_{\Sigma,d} = \Sigma^{-1/2}\mathbf{r}_d$ may not do too badly, where $\mathbf{r}_d$ is a uniform random projection on the unit sphere. This is equivalent to using a random projection on the unit sphere after standardizing the data, which also explains the different scaling as compared to the oracle or the scaled search direction, where the change $\boldsymbol{\Delta}_d$ is also transformed to $\Sigma^{-1/2}\boldsymbol{\Delta}_d$ by the standardization. However, since for increasing d the space covered by the far away angles is also increasing, the high dimensional efficiency of the scaled random projection is not only worse than the oracle by a factor $\sqrt{d}$ but also by a factor $d^{1/4}$ than the universal statistic discussed in Section 2.

The following theorem shows the high dimensional efficiency of the scaled random projection.

Theorem 3.4. *Let the alternative hold, i.e.* $\|\boldsymbol{\Delta}_d\| \neq 0$. *Let* $\mathbf{r}_d$ *be a random uniform projection on the d-dimensional unit sphere and* $\mathbf{r}_{\Sigma,d} = \Sigma^{-1/2}\mathbf{r}_d$, *then for all* $\epsilon > 0$ *there exist constants* $c, C > 0$ *(not depending on the dimension d), such that*

$$P\left(c \leqslant \mathcal{E}_3^2(\boldsymbol{\Delta}_d, \mathbf{r}_{\Sigma,d})\,\frac{d}{\|\Sigma^{-1/2}\,\boldsymbol{\Delta}_d\|^2} \leqslant C\right) \geqslant 1 - \epsilon.$$

Such a random projection on the unit sphere can be obtained as follows: Let $X_1, \ldots, X_d$ be i.i.d. N(0,1), then $\mathbf{r}_d = (X_1, \ldots, X_d)'/\|(X_1, \ldots, X_d)'\|$ is uniform on the d-dimensional unit sphere (Marsaglia, 1972).

Comparing the high dimensional efficiency of the scaled random projection with the one obtained for the oracle projection (confer Proposition 3.3) it becomes apparent that we lose an order $\sqrt{d}$. The universal panel statistic taking the full multivariate information into account has a high dimensional efficiency between those two losing $d^{1/4}$ in comparison to the oracle but gaining $d^{1/4}$ in comparison to a scaled random projection. From these results one obtains a cone around the search direction such that the projection statistic has higher power than the universal panel statistic, if the true change falls within this cone.

Figure 3.2 (b) shows the results of some simulations showing that a change that can be detected for the oracle with constant power as d increases rapidly loses power for the panel statistic as predicted by its high dimensional efficiency in Section 2 as well as for the random projection. This and the following simulations show clearly that the concept of high dimensional efficiency is indeed capable of explaining the small sample power of a statistic very well.

3.3.2. Misscaled projections with respect to the covariance structure

The analysis in the previous section requires the knowledge or a precise estimate of the inverse (structure) of Σ. However, in many situations such an estimate may not be feasible or too imprecise due to one or several of the below reasons, where the problems get worse due to the necessity for inversion.

- If d is large in comparison to T statistical estimation errors can accumulate and identification may not even be possible (Bickel and Levina, 2008).
- The theory can be generalized to time series errors but in this case the covariance matrix has to be replaced by the long-run covariance (which is proportional to the spectrum at 0) and is much more difficult to estimate (Aston and Kirch, 2012b; Kirch and Tadjuidje Kamgaing, 2012).
- Standard covariance estimators will be inconsistent under alternatives as they are contaminated by the change points. Consequently, possible changes have to be taken into account, but even in a simple at most one change situation it is unclear how best to generalize the standard univariate approach as in (C.6) as opposed to (C.5) to a multivariate situation as the estimation of a joint location already requires an initial weighting for the projection (or the multivariate statistic). Alternatively, componentwise univariate estimation of the change points could be done but require a careful asymptotic analysis in particular in a setting with $d \to \infty$.
- If d is large, additional numerical errors may arise when inverting the matrix (Higham, 2002, Ch 14).

We will now investigate the influence of misspecification or estimation errors on the high dimensional efficiency of a **misscaled oracle** $\mathbf{o}_M = \mathbf{M}^{-1}\mathbf{\Delta}_d$ in comparison to the **misscaled random projection** $\mathbf{r}_{\mathbf{M},d} = \mathbf{M}^{-1/2}\mathbf{r}_d$, where we only assume that the assumed covariance structure $\mathbf{M}$ is symmetric and positive definite and assumption $\mathcal{A}.1$ is fulfilled.

Theorem 3.5. *Let the alternative hold, i.e.* $\|\boldsymbol{\Delta}_d\| \neq 0$. *Let* $\mathbf{r}_d$ *be a random projection on the* d*-dimensional unit sphere and* $\mathbf{r}_{\mathbf{M},d} = \mathbf{M}^{-1/2}\mathbf{r}_d$ *be the misscaled random projection. Then, there exist for all* $\epsilon > 0$ *constants* $c, C > 0$, *such that*

$$P\left(c \leqslant \mathcal{E}_3^2(\boldsymbol{\Delta}_d, \mathbf{r}_{\mathbf{M},d}) \frac{tr\left(\mathbf{M}^{-1/2}\Sigma\mathbf{M}^{-1/2}\right)}{\|\mathbf{M}^{-1/2}\boldsymbol{\Delta}_d\|^2} \leqslant C\right) \geqslant 1 - \epsilon,$$

where tr denotes the trace.

We are now ready to prove the main result of this section stating that the high dimensional efficiency of a misscaled oracle can never be worse than the corresponding misscaled random projection.

Theorem 3.6. *Let Assumption* $\mathcal{A}.1$ *hold. Denote the misscaled oracle by* $\mathbf{o}_M = \mathbf{M}^{-1}\boldsymbol{\Delta}_d$, *then*

$$\mathcal{E}_3^2(\boldsymbol{\Delta}_d, \mathbf{o}_M) \geqslant \frac{\|\mathbf{M}^{-1/2}\boldsymbol{\Delta}_d\|^2}{tr(\mathbf{M}^{-1/2}\Sigma\mathbf{M}^{-1/2})}$$

where tr denotes the trace and equality holds iff there is only one common factor which is weighted proportional to $\boldsymbol{\Delta}_d$,

Because it is often assumed that components are independent and it is usually feasible to estimate the variances of each component, we consider the correspondingly misscaled oracles, which are scaled with the identity matrix (pre-oracle) respectively with the diagonal matrix of variances (quasi-oracle). The quasi-oracle is of particular importance as it uses the same type of misspecification as the universal panel statistic discussed in Section 2.

Definition 3.2. *(i) The projection* ${}_p\mathbf{o} = \boldsymbol{\Delta}_d$ *is called* **pre-oracle**.
(ii) The projection ${}_q\mathbf{o} = \Lambda_d^{-1}\boldsymbol{\Delta}_d = (\delta_1/\sigma_1^2, \ldots, \delta_d/\sigma_d^2)'$, $\Lambda_d = diag(\sigma_1^2, \ldots, \sigma_d^2)$ *is called* **quasi-oracle**, *if* $\sigma_j^2 > 0$, $j = 1, \ldots, d$.
As with the oracle, these projections should be seen as representatives of a class of projections.

The following proposition shows that in the important special case of uncorrelated components, the (quasi-)oracle and pre-oracle have an efficiency of same order if the variances in all components are bounded and bounded away from zero. The latter assumption is also needed for the panel statistic below and means that all components are on similar scales. In addition, the efficiency of the quasi-oracle is even in the misspecified situation always better than an unscaled random projection.

Proposition 3.7. *Assume that all variances are on the same scale, i.e. there exist* c, C *such that* $0 < c \leqslant \sigma_i^2 < C < \infty$ *for* $i = 1, \ldots, d$.

a) Let $\Sigma = diag(\sigma_1^2, \ldots, \sigma_d^2)$, *then*

$$\frac{c^2}{C^2}\mathcal{E}_3^2(\boldsymbol{\Delta}_d, {}_q\mathbf{o}) \leqslant \mathcal{E}_3^2(\boldsymbol{\Delta}_d, {}_p\mathbf{o}) \leqslant \mathcal{E}_3^2(\boldsymbol{\Delta}, {}_q\mathbf{o}) = \|\Sigma^{-1/2}\boldsymbol{\Delta}_d\|^2.$$

b) Under Assumption A.1, it holds

$$\mathcal{E}_3^2(\boldsymbol{\Delta}_d, {}_q\mathbf{o}) \geqslant \frac{c^2}{C^2} \frac{\|\boldsymbol{\Delta}_d\|^2}{tr(\Sigma)}.$$

The next corollary shows that the efficiency of the quasi oracle (which is scaled with diag $\left(\frac{1}{s_1^2+\Phi_1^2}, \ldots, \frac{1}{s_d^2+\Phi_d^2}\right)$ analogously to the panel statistic) is always at least as good as the efficiency of the universal panel statistic. Additionally, the efficiency of the universal panel statistic becomes as bad as the efficiency of the corresponding (diagonally) scaled random projection (tolerable efficiency) if $A_d/d \to A > 0$, which is typically the case if the dependency is non-sparse and non-negligible.

Corollary 3.8. *Let the assumptions of Lemma 2.2 on the errors be fulfilled, then the following assertions hold:*

a) The high dimensional efficiency of the quasi-oracle is always at least as good as the one of the misspecified panel statistic, i.e. with $\Sigma = diag(\sigma_1^2, \ldots, \sigma_d^2) + \boldsymbol{\Phi}\boldsymbol{\Phi}'$, $\Lambda_d = diag(\sigma_1^2, \ldots, \sigma_d^2)$, *it holds*

$$\mathcal{E}_3^2(\boldsymbol{\Delta}_d, {}_q\mathbf{o}) \geqslant \frac{\boldsymbol{\Delta}_d' \Lambda_d^{-1} \boldsymbol{\Delta}_d}{1 + A_d},$$

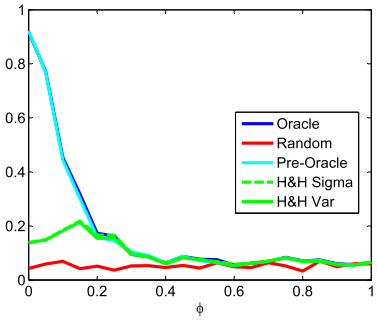
where equality holds iff $\boldsymbol{\Delta}_d \sim \boldsymbol{\Phi}$.

b) If $A_d/d \to A > 0$, *then the high dimensional efficiency of the panel statistic is as bad as a randomly scaled projection, i.e.*

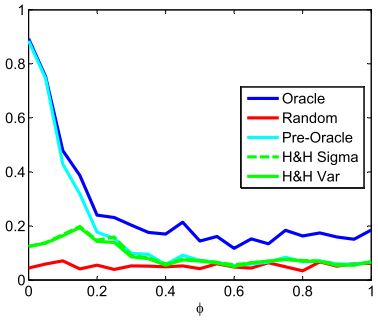
$$\mathcal{E}_2^2(\boldsymbol{\Delta}_d) = \frac{\boldsymbol{\Delta}_d' \Lambda_d^{-1} \boldsymbol{\Delta}_d}{d} (A_d + o(1)).$$

In particular, for $A_d/d \to A > 0$ the efficiency of the misscaled panel statistic is always as bad as the efficiency of the random projection, this only holds for the misscaled (quasi-) projection if $\boldsymbol{\Delta}_d \sim \boldsymbol{\Phi}$. This effect can be clearly see in Figures 3.3 and 3.4, where in all cases H&H Sigma refers to the panel statistic using known variance, and H&H Var uses an estimated variance, showing again that this concept of efficiency is very well suited to understand the small sample power behavior of the corresponding statistics. Additionally, the following assertions are confirmed by the simulations:

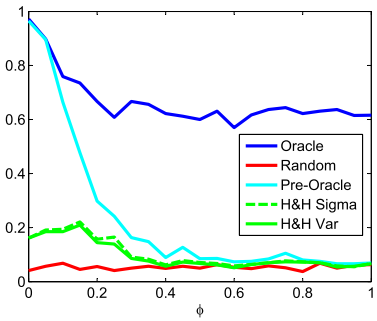
1) The power of the pre- and quasi-oracle is always better than the power of the misscaled random projection (the random projection assumes an identity covariance structure).
2) The power of the (correctly scaled) oracle can become as bad as the power of the (misscaled) random projection but only if $\boldsymbol{\Phi}_d \sim \boldsymbol{\Delta}_d$. In this case the power of the misscaled panel statistic (i.e. where the statistic but not the critical values are constructed under the wrong assumption of independence between components) is equally bad.



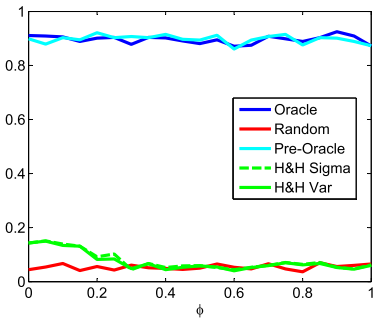
(a) Angle between $\mathbf{\Delta}_d$ and $\Phi = 0$ radians



(b) Angle between $\mathbf{\Delta}_d$ and $\Phi = \pi/8$ radians



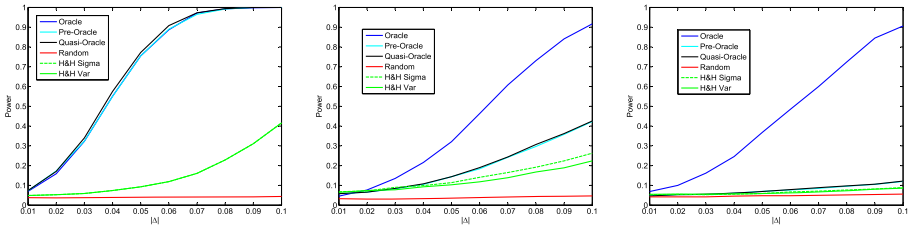
(c) Angle between $\mathbf{\Delta}_d$ and $\Phi = \pi/4$ radians



(d) Angle between $\mathbf{\Delta}_d$ and $\Phi = \pi/2$ radians

Fig 3.3: Power of tests as the angle between the change and the direction of dependency increases. As can be seen, if the change lies in the direction of dependency, then all methods struggle, which is in line with the theory of Section 3.3. However, if the change is orthogonal to the dependency structure the projection method works regardless of whether the dependency is taken into account or not. H&H Sigma and Var as in Section 2 represent the universal panel tests taking into account the true or estimated variances of the components. All results are empirically size corrected to account for the size issues seen in Figure 3.1. ($s_j = 1$, $\Phi_j = \phi$, $j = 1, \ldots, d$ with $d = 200$, $\|\mathbf{\Delta}_d\| = 0.05\sqrt{d}$, corresponding to Case $\mathcal{C}.3$), with ϕ as given on the x-axis.

3) While the power of the (misscaled) panel statistic becomes as bad as the power of the (misscaled) random projection for $\phi \to \infty$ irrespective of the angle between $\mathbf{\Delta}_d$ and $\mathbf{\Phi}_d$, it can be significantly better for the pre- and quasi-oracle. In fact, we saw above that the high dimensional efficiency of the misspecified panel statistic will be of the same order as a random projection for any choice $\mathbf{\Phi}_d$ with $\mathbf{\Phi}_d'\mathbf{\Phi}_d \sim d$, irrespective of the direction of any change that might be present.



(a) No Dependency - $\phi = 0$ (b) $\phi = 0.5$ (c) $\phi = 1$

Fig 3.4: Power of tests as the dependency increases. The covariance structure becomes closer to degenerate across the three graphs, but in all cases the pre-oracle and quasi-oracle still outperform random projections, although they become closer as the degeneracy increases. Here different variances are used across components, namely $s_i = 0.5 + i/d$, $\Phi_i = \phi_i$, $i = 1, \ldots, d$, $d = 200$, angle$(\Phi, \boldsymbol{\Delta}_d)$=$\pi/4$, corresponding to Case $\mathcal{C}.3$, and size of change as given on the x-axis (multiplied by $\sqrt{d}$).

We will now have a closer look at the three standard examples in order to understand the behavior in the simulations better (Case $\mathcal{C}.1$ is included in the simulations for $\Phi = 0$, while $\mathcal{C}.3$ is the limiting case for $\Phi \to \infty$).

Case $\mathcal{C}.1$ (Independent components)**.** If the components are uncorrelated, each with variance σ_i^2, i.e. $\Sigma_1 = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_d^2)$, we get

$$\mathrm{tr}(\Sigma_1) = \sum_{j=1}^{d} \sigma_j^2,$$

which is of order d if $0 < c \leqslant \sigma_j^2 \leqslant C < \infty$. Proposition 3.7, Theorem 3.4 and Theorem 3.5 show that in this situation both the high dimensional efficiency of the pre- and (quasi-)oracle are of an order $\sqrt{d}$ better than the correctly scaled and unscaled random projection.

The second case shows that high dimensional efficiency of misscaled oracles can indeed become as bad as a random projection and helps in the understanding of the mixed case:

Case $\mathcal{C}.2$ (Fully dependent components)**.** As noted in section C.4 of the appendix, we have to distinguish two cases:

(i) If $\boldsymbol{\Delta}_d$ is not a multiple of $\boldsymbol{\Phi}_d$, then the power depends on the angle of the projection with $\boldsymbol{\Phi}_d$ with maximal power for an orthogonal projection. So the goodness of the oracles depends on their angle with the vector $\boldsymbol{\Phi}_d$.

(ii) If $\boldsymbol{\Delta}_d$ is a multiple of $\boldsymbol{\Phi}_d$, the pre- and quasi-oracle are not orthogonal to the change, hence they share the same high dimensional efficiency with any scaled random projection as all random projections are not orthogonal to $\boldsymbol{\Phi}_d$ with probability 1.

We can now turn to the mixed case that is also used in the simulations.

Case $\mathcal{C}.3$ (Mixed case)**.** Let $\mathbf{a}_j = (0, \ldots, s_j, \ldots, 0)'$ the vector which is $s_j > 0$ at point j and zero everywhere else, and $\mathbf{a}_{d+1} = \mathbf{\Phi}_d = (\Phi_1, \ldots, \Phi_d)'$, $\mathbf{a}_j = \mathbf{0}$ for $j \geqslant d+2$. Then $\Sigma_3 = \operatorname{diag}(s_1^2, \ldots, s_d^2) + \mathbf{\Phi}_d \mathbf{\Phi}_d'$ and

$$\operatorname{tr}(\Sigma_3) = \sum_{j=1}^{d} s_j^2 + \sum_{j=1}^{d} \Phi_j^2. \tag{3.6}$$

The high dimensional efficiency of the pre-oracle can become as bad as for the random projection if the change $\mathbf{\Delta}_d$ is a multiple of the common factor $\mathbf{\Phi}_d$ and there is a substantial common effect. This is similar to Case $\mathcal{C}.2$ (which can be seen as a limiting case for increasing $\|\mathbf{\Phi}_d\|$). Intuitively, the problem is the following: By projecting onto the change, we want to maximize the signal i.e. the change in the projected sequence while minimizing the noise. In this situation however, the common factor dominates the noise in the projection as it essentially adds up in a linear manner, while the uncorrelated components add up only in the order of $\sqrt{d}$ (CLT). Now, projecting onto $\mathbf{\Delta}_d = \mathbf{\Phi}_d$ maximizes not only the signal but also the noise, which is why we cannot gain anything (but this also holds true for other procedures such as the universal panel tests).

A different interpretation is the following one: In situation $\mathcal{C}.3$, each component has a common factor $\{\eta_t\}$ weighted according to $\mathbf{\Phi}_d$ plus some independent noise. If a change occurs in sync with the common factor it will be difficult to detect as in order to get the correct size, we have to allow for the random movements of $\{\eta_t\}$ thus increasing the critical values in that direction. In directions orthogonal to it, we only have to take the independent noise into account which yields comparably smaller noise in the projection. In an economic setting, this driving factor could for example be thought of as an economic factor behind certain companies (e.g. ones in the same industry). If a change occurs in those companies proportional to this driving factor it will be difficult to distinguish a different economic state of this driving factor from a mean change that is proportional to the influence of this factor.

A mathematical analysis is given in Section C.5 in the appendix.

3.4. Data driven projections and high dimensional efficiency of some sparse change point tests

When using data-driven projections one has to be very careful as the projection will typically have an effect on the null asymptotic of the projection test requiring larger critical values. The reason is that in high-dimensional settings there are always directions in which the CUSUM statistic of the projected time series will become very large by chance. This effect can be made smaller by requiring additional assumptions such as sparsity.

In fact, most current change point tests for high dimensional data assume sparsity of the change point (Jirak, 2015; Cho and Fryzlewicz, 2015; Wang and Samworth, 2016) as well as possibly likelihood based considerations Chan and

Walther (2015). Some of these tests are effectively based on projections. For example, Jirak (2015) uses the maximum (resulting in an extreme-value behavior) of all projections on unit vectors consisting of all zeroes and just one one. Cho and Fryzlewicz (2015) use thresholding, which can also be viewed as a data-driven projection into a lower dimensional space. Most notably, Wang and Samworth (2016) use a data-driven projection based on a sparse singular value decomposition of the high-dimensional CUSUM matrix. Due to the sparseness assumption the noise level of the projection can be kept at bay, which is no longer the case if an unconstrained singular value decomposition is used.

Furthermore, using a preprint version of this paper, Cho (2015) derived the high dimensional efficiency for a number of tests including the tests by Jirak (2015), Enikeeva and Harchaoui (2013), Cho and Fryzlewicz (2015) as well as the Double CUSUM statistic introduced in that paper (see Table 1 in Cho (2015)). It turns out, that the tests by Jirak (2015), Cho and Fryzlewicz (2015) as well as the scan statistic by Enikeeva and Harchaoui (2013) achieve oracle efficiency (up to a log-term) for sparse changes but only tolerable efficiency for balanced changes. The linear test statistic introduced by Enikeeva and Harchaoui (2013) has the same power behavior as the universal panel test statistic discussed in this paper. The efficiency of the double CUSUM statistic introduced in Cho (2015) depends on the number of components with a mean change in addition to a parameter choice of the statistic. Depending on the combination of choice of this parameter and the number of components contaminated it can achieve both oracle efficiency and tolerable efficiency.

This discussion shows that considering the high dimensional efficiency yields understanding about for which change alternatives a given test has particularly good power and at what cost this comes with respect to other changes.

4. Data example

As an illustrative example which shows the small sample behaviour of the statistics illustrated above also apply in real data, we examine the stability of change points detected by different methods in several world stock market indices. More specifically, the Fuller Log Squared Returns (Fuller, 1996, p 496) of the FTSE, NASDAQ, DAX, NIKKEI, Hang Seng and CAC [1] indices for the year 2015 were examined for change points. Tests based on the multivariate statistics using full covariance estimates, a multivariate statistic using only variance estimates (i.e., a diagonal covariance structure), a projection statistic in the average direction $(1,1,1,1,1,1)'$, and a projection statistic in the direction of European countries vs non-European countries $(1,-1,1,-1,-1,1)'$ (orthogonal to the average direction) were carried out. Given the considerable dependence between the different components, we would expect economies to likely rise and fall together, justifying the use of the former projection direction. However, we think it unlikely that there will be changes of the kind that when European markets goes

[1] We only use a small number of series to allow reliable estimates for the covariance to be used in the full multivariate statistic.

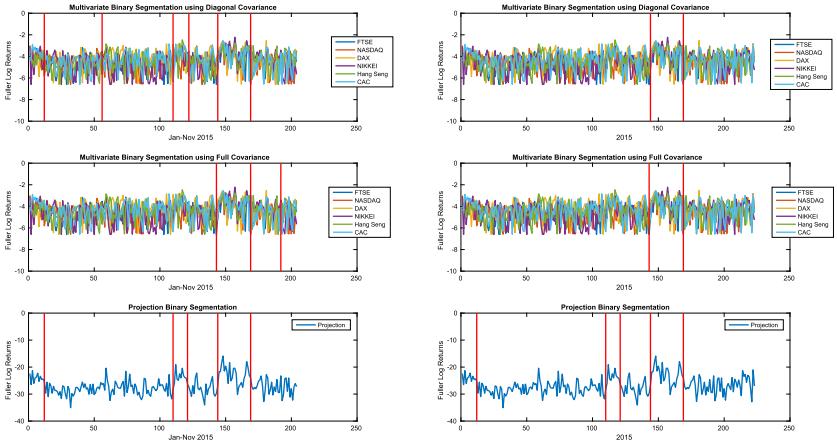


Fig 4.1: Estimated change point locations for the market indices from binary segmentation based on different test statistics and different spans of data. First Column: Data from Jan-Nov 2015, Second Column: Data from all of 2015. First Row: Multivariate statistic with full covariance estimation; Second Row: Multivariate Statistic with Diagonal Variance Estimate; Third Row: Projection Statistic in direction $(1, 1, 1, 1, 1, 1)'$. Red vertical lines indicate changes deemed to be significant at 5% level.

up, non-European markets go down, and visa versa, so take this projection as an example of direction where no change is likely. It should be noted at this point that the multivariate statistic treats both of these alternatives as equally likely. As there are possible multiple changes points in this data, we examine stability by performing binary segmentation using the proposed tests, firstly on data from January to November 2015, and then subsequently adding in the data from December 2015.

As can be seen in Figure 4.1, the multivariate test statistic is considerably less robust than the average projection based statistic, both to the length of the data, as well as to the choice of the covariance estimate. The major cause of this instability was that the CUSUM statistic over time had two peaks, but the location of the maximal peak differed from one to the other when further data was added. This caused knock-on effects in the entire binary segmentation. Here, in all cases, independence in time was assumed as once the changes were accounted for, there was little evidence of temporal dependence in the data. However, even if time series dependence is accounted for by using an estimate of the long run covariance in place of the independent covariance estimate, there is no difference in the qualitative conclusions (although the change points themselves varied considerably in all cases depending on the parameters chosen in the long run covariance estimation procedure (Politis, 2005)). In addition, the projection estimate was robust to whether the direction was scaled by the full covariance, the diagonal of the covariance or not scaled at all, as well as to

TABLE 4.1
Location, Statistic and p-value for the changes found in the 2015 market index data. (Limit Distributions: Multivariate: sum of six independent Brownian Bridges; Projection: Single Brownian Bridge)

Multivariate: Full Covariance					
Day				143	169
Statistic Value				6.8541	5.1581
p				0.0012	0.0173
Multivariate: Diagonal Covariance					
Day				144	169
Statistic Val				9.9995	11.7030
p-value				$< 10^{-5}$	$< 10^{-5}$
Projection: scaled $(111111)'$					
Day	12	110	121	144	169
Statistic Value	2.1307	3.5390	2.9518	3.3173	2.0900
p	0.0285	0.0017	0.0057	0.0027	0.0307

increasing the length of the data.

The p-values for the changes on full year's data are given in Table 4.1. While it can be seen that the projection p-vals are larger for the two common change points than in the multivariate case, the same changes are detected with all methods. However, additional changes are found with the projection method, and the p-vals are well below the critical value of 5%. This shows that having knowledge of the likely direction of change can allow further changes to be found beyond those in an unrestricted multivariate search. As expected though, using an unlikely direction does not find change points, with the hypothesis that there are no changes which affect European markets differently to non-European markets being accepted (p=0.18).

5. Conclusions

The primary aims of this paper were to introduce a theoretic method to compare the small sample behavior of different high dimensional tests by asymptotic methods. The new concept of high dimensional efficiency allows a comparison of the magnitude of changes that can be detected asymptotically as the number of dimensions increases. Both, the simulations as well as the data example confirm the assertions obtained from that theoretic concept indicating it is in fact a useful tool to analyse high dimensional tests.

As a benchmark, projection tests were investigated, including as an upper benchmark an oracle projection as well as as a lower benchmark a random projection.

In summary, the following assertions were obtained: The panel statistic (Bai, 2010; Horváth and Hušková, 2012) test works well in situations where the panels are independent across components, in particular if there is little to no information about the direction or properties of the change such as whether it is sparse or balanced. However, as soon as dependency is present, the size properties of these statistics become difficult and their high dimensional efficiencies mimic those of random projections. Unfortunately, this problem cannot even be

solved if the covariance structure is completely known unless under normality assumptions. Misspecification of the covariance structure can be problematic for all tests even projection tests with the correct change structure. Nevertheless, misscaled oracle tests (if accessible) are preferable to the misscaled panel statistic.

An investigation of Cho (2015) based on a preprint version of the present paper indicates that change point tests constructed for sparse alternatives will achieve approximately oracle power if the sparseness assumption is correct. However, they will achieve only tolerable power if the change is balanced (i.e. not sparse), so that both benchmarks are in fact important to understand the power behavior of recent change point tests in high dimensional settings.

The results in this paper raise many questions for future work. It would be of considerable interest to determine whether projections can be derived using data driven techniques, such as sparse PCA, for example, and whether such projections would be better than random projections. Preliminary work suggests that this may be so in some situations but not others, and a nice procedure by Wang and Samworth (2016) investigates a related technique.

While it is very unlikely that data-driven methods will be able to improve upon the behavior of the panel statistic without additional structural assumptions on the change, the question remains whether one can get close to the universal panel statistic's power properties while at the same time being more robust with respect to size. However, the framework here allows this question to be rigorously posed, and different approaches to be compared.

Appendix A: Comparing the power of two univariate CUSUM tests

We will illustrate how the concept of 'high dimensional' efficiency can be used even if very different asymptotics are involved. To this end we consider the following univariate change point setup

$$X_t = \mu + \delta_T 1_{\{t>k_T^*\}} + e_t,$$

with $\{e_t\}$ i.i.d. with $\mathrm{E}\, e_1 = 0$, $\operatorname{var} e_1 = 1$ (merely for simplicity) and $\mathrm{E}\,|e_1|^\nu < \infty$ for some $\nu > 0$. For $k_T^* = \lfloor \theta T \rfloor$ we have the (univariate) AMOC situation from the present paper, but here we allow for arbitrary changes k_T^*. The goal is now to compare the power behavior of the following two CUSUM statistics

$$M_1 = \max_{1\leqslant k\leqslant T} \frac{1}{\sqrt{T}} \left| \sum_{j=1}^{k} (X_j - \bar{X}_n) \right|, \qquad M_2 = \max_{1\leqslant k\leqslant T} \sqrt{\frac{T}{k(T-k)}} \left| \sum_{j=1}^{k} (X_j - \bar{X}_n) \right|.$$

Both statistics are well known in the change point community and very often accompagnied by statements such as 'statistic T_2 detects early and late changes better while the statistic T_1 detects changes in the middle of the observation period better'.

We will now demonstrate that the use of efficiency as defined in the present paper helps to make this statement precise. To this end, we adapt Definition 2.1

slightly by considering $\mathcal{E}(k_T^*, \delta_T)$, which will now depend on T, k_T^* and δ_T (and obviously drop the assumption $d \to \infty$ as we consider the univariate case $d = 1$). Furthermore, we identify the efficiency of M_2 with the one of $\tilde{M}_2$ (as they yield the same test) defined by

$$\tilde{M}_2 = \sqrt{2 \log\log T}\, M_2 - 2 \log\log T - \frac{1}{2} \log\log\log T + \frac{1}{2} \log \pi \xrightarrow{\mathcal{L}} G_2$$

under H_0, where G_2 has a Gumble extreme value distribution, i.e. $P(G_2 \leqslant x) = \exp(-2\exp(-x))$ (see e.g. the book by Csörgő and Horváth (1997)). The statistic M_1 on the other hand has the following standard null asymptotics (that follow immediately from the functional central limit theorem)

$$M_1 \xrightarrow{\mathcal{L}} \sup_{0 \leqslant t \leqslant 1} |B(t)| \quad \text{(under } H_0\text{)}.$$

Consequently, (i) of Definition 2.1 is fulfilled but with very different limit distributions (and in fact a very different limit behavior).

We will now show that the following efficiencies hold:

$$\begin{aligned} \mathcal{E}_{M_1}(k^*, \delta) &= \frac{\min(k^*, T - k^*)}{T} |\delta|, \\ \mathcal{E}_{M_2} = \mathcal{E}_{\tilde{M}_2}(k^*, \delta) &= \sqrt{\frac{\min(k^*, T - k^*)}{T \log\log T}} |\delta|. \end{aligned} \tag{A.1}$$

To see this note that

$$\sum_{j=1}^{k} (X_j - \bar{X}_n) = \sum_{j=1}^{k} (e_j - \bar{e}_n) + \delta \left((k - k^*)_+ - k \, \frac{(T - k^*)}{T} \right).$$

Assumption (ii) of Definition 2.1 with $\mathcal{E}_{M_1}$ and $\mathcal{E}_{\tilde{M}_2}$ as in (A.1) follows from this decomposition by using the partial sum process at $k = k^*$ as lower bound for the statistics. On the other hand Assumption (iii) follows from this decomposition because uniformly in k

$$\begin{aligned} &(k - k^*)_+ - k \, \frac{(T - k^*)}{T} = O\left(\frac{\min(k^*, T - k^*)}{T} \right), \\ &\sqrt{\frac{T^2}{k(T-k)}} \left((k - k^*)_+ - k \, \frac{(T - k^*)}{T} \right) = O\left(\sqrt{\frac{\min(k^*, T - k^*)}{T}} \right). \end{aligned}$$

From (A.1) we can see, that the efficiency of M_1 is an order $\sqrt{\log\log T}$ better than the efficiency of M_2 if $k^* = \lfloor \lambda T \rfloor$ but a lot worse for early and late changes such as $k^* = \lfloor \log T \rfloor$ or $k^* = T - \lfloor \log T \rfloor$.

Appendix B: Central limit theorem for universal panel statistics

The following theorem gives a central limit theorem for the partial sum process $V_{d,T}(\cdot)$ (under the null) from which null asymptotics of the corresponding statistics can be derived. It was proven by Horváth and Hušková (2012, Theorem 1), under somewhat more general assumptions allowing in particular for time series errors (in the form of linear processes). While this makes estimation of the covariances more difficult and less precise as long-run covariances need to be estimated, it has no effect on the high dimensional efficiency. Therefore, we will concentrate on the i.i.d. (across time) situation in this work to keep things simpler purely in terms of the calculations.

Theorem B.1. *Let Model (1.1) hold with* $\{e_{i,t} : i, t\}$ *independent (where the important assumption is the independence across components) such that* $\operatorname{var} e_{i,t} \geqslant c > 0$ *for all* i *and* $\limsup_{d\to\infty} \frac{1}{d}\sum_{i=1}^{d} \mathrm{E}\,|e_{i,t}|^{\nu} < \infty$ *for some* $\nu > 4$. *Furthermore, let* $\frac{d}{T^2} \to 0$. *Then, it holds under the null hypothesis of no change*

$$V_{d,T}(x) \xrightarrow{D[0,1]} \sqrt{2}(1-x)^2 W\left(\frac{x^2}{(1-x)^2}\right),$$

where $W(\cdot)$ *is a standard Wiener process.*

Appendix C: Projections

C.1. Change point statistics

Standard statistics such as the CUSUM statistic are based on partial sum processes, so in order to quantify the possible power gain by the use of projections we will consider the partial sum process of the projections as given in (2.3).

Different test statistics can be defined for a range of g in (1.1), however, assuming that $g \not\equiv 0$, the hypothesis of interest is

$$H_0 : \boldsymbol{\Delta}_d = \mathbf{0}$$

versus the alternative

$$H_1 : \boldsymbol{\Delta}_d \neq \mathbf{0}.$$

Test statistics are now defined in order to give good power characteristics for a particular g function. For example, the classic AMOC statistic for univariate and multivariate change point detection is based on $U_{d,T}(x)/\tau(\mathbf{p}_d)$, with

$$\tau^2(\mathbf{p}_d) = \mathbf{p}_d' \operatorname{var}\left(\mathbf{e}_1(d)\right)\mathbf{p}_d. \tag{C.1}$$

Typically, either the following max or sum type statistics are used:

$$\max_{1\leqslant k\leqslant T} w(k/T)\left|\frac{U_{d,T}(k/T)}{\tau(\mathbf{p}_d)}\right|, \qquad \frac{1}{T}\sum_{k=1}^{T} w(k/T)\left|\frac{U_{d,T}(k/T)}{\tau(\mathbf{p}_d)}\right|,$$

where $w \geqslant 0$ is continuous (which can be relaxed) and fulfills (C.7) (confer e.g. the book by Csörgő and Horváth (1997)). The choice of weight function $w(\cdot)$ can increase power for certain locations of the change points (Kirch et al., 2015).

For the epidemic change, typical test statistics are given by

$$\max_{1\leqslant k_1<k_2\leqslant T} \frac{1}{\tau(\mathbf{p}_d)} |U_{d,T}\left(k_2/T\right) - U_{d,T}\left(k_1/T\right)|,$$

$$\text{or} \qquad \frac{1}{T^2} \sum_{1\leqslant k_1<k_2\leqslant T} \frac{1}{\tau(\mathbf{p}_d)} |U_{d,T}\left(k_2/T\right) - U_{d,T}\left(k_1/T\right)|.$$

In the next section we first derive a functional central limit theorem for the process $U_{d,T}(x)$, which implies the asymptotic null behavior for the above tests. Then, we derive the asymptotic behavior of the partial sum process under contiguous alternatives to obtain the high dimensional efficiency for projection statistics.

Similarly, a multivariate change point statistic (using the full multivariate information and no additional knowledge about the change) for the at most one mean change is given as a weighted maximum or sum of the following quadratic form

$$V_d^M(x) = \mathbf{Z}_T(x)' \boldsymbol{A} \mathbf{Z}_T(x) \tag{C.2}$$

where $\mathbf{Z}_T(x) = (Z_{T,1}(x), \ldots, Z_{T,d}(x))'$ is defined as in (2.3). The usual choice is $\boldsymbol{A} = \Sigma^{-1}$, where Σ is the covariance matrix of the multivariate observations. The weighting with Σ^{-1} has the advantages that it (a) leads to a pivotal limit and (b) the statistic can detect all changes no matter what the direction. The second remains true for any positive definite matrix $\mathbf{A}$, the first also remains true for lower rank matrices with a decorrelation property of the errors, where this latter approach is essentially a projection (into a lower-dimensional space) as discussed in the previous sections. For an extensive discussion of this issue for the example of changes in the autoregressive structure of time series we refer to Kirch et al. (2015). The choice $\boldsymbol{A} = \Sigma^{-1}$ corresponds to the correctly scaled case, while the misscaled case corresponds to the choice $\boldsymbol{A} = \boldsymbol{M}^{-1}$.

However, this multivariate setup is not very suitable for the theoretic power comparison we are interested in because the limit distribution (a sum of d squared Brownian bridges with covariance matrix $\Sigma^{1/2}\boldsymbol{A}\Sigma^{1/2}$) still depends on d as well as the possible misspecification. Therefore, a comparison needs to take both the rates, the additive term and the noise level (which depends also on the misspecification of the covariance) present in the limit distribution into account. The panel data settings on the other hand, allows for an analysis by means of the high-dimensional efficiency as introduced in this paper. Furthermore, the panel statistic is strongly related to the multivariate statistic so that the same qualitative statements can be expected, which is confirmed by simulations (results not shown).

C.2. Null asymptotics

In this section, we give some additional insights for projection statistics under the null hypothesis.

Theorem C.1. *Let model (1.1) hold. Let* $\mathbf{p}_d$ *be a possibly random projection independent of* $\{e_{i,t} : 1 \leqslant t \leqslant T, 1 \leqslant i \leqslant d\}$. *Furthermore, let* $\mathbf{p}_d' \operatorname{cov}(\mathbf{e}_1(d))\mathbf{p}_d \neq 0$ *(almost surely), which means that the projected data is not degenerate with probability one.*
For i.i.d. error sequences $\{\mathbf{e}_t(d) : t = 1, \ldots, d\}$, $\mathbf{e}_t(d) = (e_{1,t}(d), \ldots, e_{d,t}(d))'$ *with an arbitrary dependency structure across components, and if* $\mathrm{E}\,|e_{1,t}(d)|^\nu \leqslant C < \infty$ *for all* t *and* d *as well as*

$$\frac{\|\mathbf{p}_d\|_1^2}{\mathbf{p}_d' \operatorname{cov}(\mathbf{e}_t)\mathbf{p}_d'} = o(T^{1-2/\nu}) \quad a.s., \tag{C.3}$$

where $\|\mathbf{a}\|_1 = \sum_{j=1}^d |a_j|$, *then (3.2) holds. The assertions remains true if* $\tau^2(\mathbf{p}_d)$ *is replaced by* $\widehat{\tau}_{d,T}^2$ *such that for all* $\epsilon > 0$

$$P\left(\left|\frac{\widehat{\tau}_{d,T}^2}{\tau^2(\mathbf{p}_d)} - 1\right| > \epsilon\right) \to 0 \qquad a.s. \tag{C.4}$$

Assumption (C.3) is always fulfilled for the multivariate situation with d fixed or if d is growing sufficiently slowly with respect to T as the left hand side of (C.3) is always bounded by $\sqrt{d}$ if $\mathbf{p}_d' \operatorname{cov}(e)\mathbf{p}_d / \|p_d\|^2$ is bounded away from zero. Otherwise, the assumption may hold for certain projections but not others. However, in this case, it is possible to put stronger assumptions on the error sequence such as in a), which are still much weaker than the usual assumption for panel data, that components are independent.

The following lemma shows that the following two different stimators for $\tau(\mathbf{p}_d)$ under the null hypothesis are both consistent. The second one is typically still consistent in the presence of one mean change which usually leads to a power improvement in the test for small samples. An analogous version can be defined for the epidemic change situation. However, it is much harder to get an equivalent correction in the multivariate setting because the covariance matrix determines how different components are weighted, which in turn has an effect on the location of the maximum. This problem does not arise in the univariate situation, because the location of the maximum does not depend on the variance estimate.

Lemma C.2. *Consider*

$$\widehat{\tau}_{1,d,T}^2(\mathbf{p}_d) = \frac{1}{T}\sum_{j=1}^T \left(\mathbf{p}_d' \mathbf{e}_t(d) - \frac{1}{T}\sum_{i=1}^T \mathbf{p}_d' \mathbf{e}_t(d)\right)^2 \tag{C.5}$$

as well as

$$\widehat{\tau}^2_{2,d,T}(\mathbf{p}_d) = \frac{1}{T}\left(\sum_{j=1}^{\widehat{k}_{d,T}}\left(\mathbf{p}_d'\mathbf{e}_j(d) - \frac{1}{T}\sum_{i=1}^{\widehat{k}_{d,T}}\mathbf{p}_d'\mathbf{e}_i(d)\right)^2 + \sum_{j=\widehat{k}_{d,T}+1}^{T}\left(\mathbf{p}_d'\mathbf{e}_t(d) - \frac{1}{T}\sum_{i=\widehat{k}_{d,T}+1}^{T}\mathbf{p}_d'\mathbf{e}_i(d)\right)^2\right), \tag{C.6}$$

where $\widehat{k}_{d,T} = \arg\max_{t=1,\dots,T} U_{d,T}(t/T)$.

a) Under the assumptions of Theorem 3.1 a) both estimators (C.5) as well as (C.6) fulfill (3.3).

b) Under the assumptions of Theorem 3.1 b), then estimator (C.5) fulfills (3.3) under the assumption

$$\frac{\|\mathbf{p}_d\|_1^2}{\mathbf{p}_d'\operatorname{cov}(\mathbf{e}_t)\mathbf{p}_d'} = o(T^{1-2/\min(\nu,4)}) \qquad a.s.,$$

while estimator (C.6) fulfills it under the assumption

$$\frac{\|\mathbf{p}_d\|_1^2}{\mathbf{p}_d'\operatorname{cov}(\mathbf{e}_t)\mathbf{p}_d'} = o(T^{1-2/\min(\nu,4)}(\log T)^{-1}) \qquad a.s.,$$

The following theorem gives the null asymptotic for the simple CUSUM statistic for the at most one change, other statistics as given in Section C.1 can be dealt with along the same lines.

Corollary C.3. *Let the assumptions of Theorem 3.1 be fulfilled and* $\widehat{\tau}(\mathbf{p}_d)$ *fulfill (3.3) under the null hypothesis, then for all* $x \in \mathbb{R}$ *it holds under the null hypothesis*

$$P\left(\max_{1\leqslant k\leqslant T} w^2(k/T)\frac{U^2_{d,T}(k/T)}{\widehat{\tau}^2(\mathbf{p}_d)} \leqslant x \,\middle|\, \mathbf{p}_d\right) \to P\left(\max_{0\leqslant t\leqslant 1} w^2(t)B^2(t) \leqslant x\right) \quad a.s.$$

$$P\left(\frac{1}{T}\sum_{1\leqslant k\leqslant T} w^2(k/T)\frac{U^2_{d,T}(k/T)}{\widehat{\tau}^2(\mathbf{p}_d)} \leqslant x \,\middle|\, \mathbf{p}_d\right) \to P\left(\int_0^1 w^2(t)B^2(t)\,dt \leqslant x\right) \quad a.s.$$

for any continuous weight function $w(\cdot)$ *with*

$$\lim_{t\to 0} t^\alpha w(t) < \infty, \quad \lim_{t\to 1}(1-t)^\alpha w(t) \quad \text{for some } 0 \leqslant \alpha < 1/2,$$
$$\sup_{\eta\leqslant t\leqslant 1-\eta} w(t) < \infty \qquad \text{for all } 0 < \eta \leqslant \frac{1}{2}. \tag{C.7}$$

C.3. Consistency of the AMOC change point estimator

The following theorem shows that the point of maximum is a consistent estimator for the change point in rescaled time in the at-most-one-change situation.

Corollary C.4. *Let the assumptions of Theorem 3.2 hold and additionally* $\sqrt{T}\mathcal{E}_3(\mathbf{\Delta}_d, \mathbf{p}_d) \to \infty$ *a.s. Under the alternative of one abrupt change, i.e.* $g(x) = 1_{\{x>\vartheta\}}$ *for some* $0 < \vartheta < 1$*, the estimator*

$$\widehat{\vartheta}_T = \left\lfloor \frac{\arg\, max_k U^2_{d,T}(k/T)}{T} \right\rfloor$$

is consistent for the change point in rescaled time, i.e.

$$P\left(\left|\widehat{\vartheta}_T - \vartheta\right| \geqslant \epsilon \,|\, \mathbf{p}_d\right) \to 0 \qquad a.s.$$

An analogous statement holds, if the arg max *of* $w^2(k/T)U^2_{d,T}(k/T)$ *is used instead and* $w^2(x)\left((x-\vartheta)_+ - x(1-\vartheta)\right)^2$ *has a unique maximum at* ϑ*, which is the case for many standard weight functions such as* $w(t) = (t(1-t))^{-\beta}$ *for some* $0 \leqslant \beta < 1/2$*.*

C.4. The oracle in the case of non-invertibility

Let us now have a look at the situation if Σ is not invertible hence the above oracle does not exist. To this end, let us consider Case $\mathcal{C}.2$ above – other non-invertible dependent situations can essentially be viewed in a very similar fashion, but become a combination of the two scenarios below.

Case $\mathcal{C}.2$ (Fully dependent Components)**.** In this case $\Sigma = \mathbf{\Phi}_d\mathbf{\Phi}_d'$ is a rank 1 matrix and not invertible. Consequently, the oracle as in Definition 3.1 does not exist. To understand the situation better, we have to distinguish two scenarios:

(i) If $\mathbf{\Phi}_d$ is not a multiple of $\mathbf{\Delta}_d$ we can transform the data into a noise-free sequence that only contains the signal by projecting onto a vector that is orthogonal to $\mathbf{\Phi}_d$ (cancelling the noise term) but not to $\mathbf{\Delta}_d$. All such projections are in principle equivalent as they yield the same signal except for a different scaling which is not important if there is no noise present. Consequently, all such transformations could be called oracle projections.

(ii) On the other hand if $\mathbf{\Delta}_d$ is a multiple of $\mathbf{\Phi}_d$, then any projection cancelling the noise will also cancel the signal. Projections that are orthogonal to $\mathbf{\Phi}_d$ hence by definition also to $\mathbf{\Delta}_d$ will lead to a constant deterministic sequence hence to a degenerate situation. All other projections lead to the same (non-degenerate) time series except for multiplicative constants and different means (under which the proposed change point statistics are invariant by definition) so all of them could be called oracles.

The following interpretation also explains the above mathematical findings: In this situation, all components are obtained from one common factor $\{\eta_t\}$ with different weights according to $\mathbf{\Phi}_d$ i.e. they move in sync with those weights. If a change is proportional to $\mathbf{\Phi}_d$ it could either be attributed to the noise coming from $\{\eta_t\}$ or from a change, so it will be difficult to detect as we are essentially back in a duplicated rank one situation and no additional information about the change can be obtained from the multivariate situation. However, if it is not proportional to $\mathbf{\Phi}$, then it is immediately clear (with probability one) that a change in mean must have occurred (as the underlying time series no longer moves in sync). This can be seen to some extent in Figure 3.3, where the different panels in the figure mimic the different scenarios as outlined above (with a large value of ϕ being close to the non-invertible situation).

C.5. Misscaled projections for the mixed case

In this section we derive some mathematical theory for the mixed case $\mathcal{C}.3$ under misspecification explaining the intuition and simulation results already given in Section 3.3.2.

It holds $\tau^2({}_p\mathbf{o}) = \sum_{j=1}^d s_j^2\delta_j^2 + \left(\sum_{j=1}^d \delta_j\Phi_j\right)^2$. If additionally $\mathbf{\Delta}_d = k\mathbf{\Phi}_d$, for some $k > 0$, we get the following high dimensional efficiency for the pre-oracle by (3.4)

$$\mathcal{E}_3(\mathbf{\Delta}_d, {}_p\mathbf{o}) = \frac{\|\mathbf{\Delta}_d\|}{\sqrt{\sum_{i=1}^d s_i^2\left(\frac{\delta_i}{\|\mathbf{\Delta}_d\|}\right)^2 + \|\mathbf{\Phi}_d\|^2}}.$$

The high dimensional efficiency for the unscaled random projection is given by (confer Theorem 3.5 and (3.6))

$$\frac{\|\mathbf{\Delta}_d\|}{\sqrt{\sum_{j=1}^d s_j^2 + \|\mathbf{\Phi}_d\|^2}}.$$

As soon as s_j, Φ_j are of the same order, i.e. $0 < c \leqslant s_j, \Phi_j \leqslant C < \infty$ for all j, the pre-oracle behaves as badly as the unscaled random projection. The same holds for the quasi-oracle under the same assumptions. Interestingly, however, in this particular situation, even the oracle is of the same order as the random projection if the s_j are of the same order, i.e. $0 < c \leqslant s_j < C < \infty$. More precisely we get (for a proof we refer to the Section D)

$$\mathcal{E}_3(\mathbf{\Delta}_d, \mathbf{o}) = \frac{\|\mathbf{\Delta}_d\|}{\sqrt{1 + \sum_{j=1}^d \frac{\Phi_j^2}{s_j^2}}} \sqrt{\frac{\sum_{j=1}^d \frac{\delta_j^2}{s_j^2}}{\sum_{j=1}^d \delta_j^2}}. \tag{C.8}$$

Figure 3.3 shows simulations which confirm the underlying theory in finite samples.

On the other hand, if $\mathbf{\Delta}_d$ is orthogonal to $\mathbf{\Phi}_d$, then the noise from $\mathbf{\Phi}_d$ cancels for the pre-oracle projection and we get the rate

$$\mathcal{E}_3(\mathbf{\Delta}_d, {}_p\mathbf{o}) = \frac{\|\mathbf{\Delta}_d\|}{\sqrt{\sum_{i=1}^d s_i^2 \left(\frac{\delta_i}{\|\mathbf{\Delta}_d\|}\right)^2}},$$

which is of the order $\|\mathbf{\Delta}_d\|^2$ if the s_j are all of the same order. Anything between those two cases is possible and depends on the angle between $\mathbf{\Delta}$ and $\mathbf{\Phi}_d$ (again see Figures 3.3 and 3.4 for finite sample simulations).

Appendix D: Proofs

Proof of Theorem 3.1 and Theorem C.1. We need to prove the following functional central limit theorem for the triangular array of projected random variables $Y_{t,d} = \sum_{j=1}^d p_j(d) e_{j,t}(d)$ given the (possibly random) projection $\mathbf{p}_d = (p_1(d), \ldots, p_d(d))'$:

$$\left\{ \frac{1}{\sqrt{T\tau^2(\mathbf{p}_d)}} \sum_{t=1}^{\lfloor Tx \rfloor} Y_{t,d} : 0 \leqslant x \leqslant 1 \,|\, \mathbf{p}_d \right\} \xrightarrow{D[0,1]} \{W(x) : 0 \leqslant x \leqslant 1\} \qquad a.s., \tag{D.1}$$

where $\{W(\cdot)\}$ denotes a standard Wiener process.

The proof for tightness is analogous to the one given in Theorem 16.1 of Billingsley (1968) as it only depends on the independence across time (which also holds conditionally given $\mathbf{p}_d$ due to the independence of $\mathbf{p}_d$ and $\{\mathbf{e}_t(d)\}$). Similarly, the proof for the convergence of the finite dimensional distributions follows the proof of Theorem 10.1 in Billingsley (1968), where we need to use the Lindeberg-Levy-version of the univariate central limit theorem for triangular arrays. More precisely, we need to prove the Lindeberg condition given by

$$\mathrm{E}\left(\frac{Y_{1,d}^2}{\tau^2(\mathbf{p}_d)} 1_{\{Y_{1,d}/\tau(\mathbf{p}_d) \geqslant \epsilon\sqrt{T}\}} \,|\, \mathbf{p}_d \right) \to 0 \quad a.s.$$

for any $\epsilon > 0$. The following Lyapunov-type condition implies the above Lindeberg condition:

$$\mathrm{E}\left(\left| \frac{Y_{1,d}}{\tau(\mathbf{p}_d)} \right|^\nu \,|\, \mathbf{p}_d \right) = \mathrm{E}\left(\left| \frac{\mathbf{p}_d' \mathbf{e}_1(d)}{\tau(\mathbf{p}_d)} \right|^\nu \,|\, \mathbf{p}_d \right) = o(T^{\nu/2-1}) \qquad a.s., \tag{D.2}$$

where $\nu > 2$ as given in the theorem. Let

$$\tilde{\mathbf{p}}_d = \frac{\mathbf{p}_d}{\sqrt{\mathbf{p}_d' \operatorname{cov} \mathbf{e}_1(d) \mathbf{p}_d}},$$

then the above Lyapunov condition is equal to

$$\mathrm{E}\left(|\tilde{\mathbf{p}}_d' \mathbf{e}_1(d)|^\nu \,|\, \mathbf{p}_d \right) = o(T^{\nu/2-1}) \qquad a.s.$$

In the situation of a) $\operatorname{cov} \mathbf{e}_1(d) = \sum_{j\geqslant 1} \mathbf{a}_j(d)\mathbf{a}_j'(d)$ and we get by the Rosenthal inequality (confer e.g. Lin and Bai (23233010, 9.7c))

$$\begin{aligned}
&\mathrm{E}\left(\left|\sum_{j=m}^{n} \tilde{\mathbf{p}}_d'\mathbf{a}_j(d)\eta_{j,1}(d)\right|^{\nu} \mid \mathbf{p}_d\right) \\
&\leqslant O(1)\sum_{j=m}^{n} |\tilde{\mathbf{p}}_d'\mathbf{a}_j(d)|^{\nu}\ \mathrm{E}\,|\eta_{j,1}(d)|^{\nu} + O(1)\left(\sum_{j=m}^{n} (\tilde{\mathbf{p}}_d'\mathbf{a}_j(d))^2\ \operatorname{var}\eta_{j,1}(d)\right)^{\nu/2},
\end{aligned}$$

where the right-hand side is bounded for any m, n with a bound that does not depend on T or d and converges to zero for $m, n \to \infty$ as $\mathrm{E}\,|\eta_j(d)|^{\nu} \leqslant C$ hence $\operatorname{var}\eta_j(d) \leqslant 1 + C$ and by definition of $\tilde{\mathbf{p}}_d$ it holds $\sum_{j=m}^{n} |\tilde{\mathbf{p}}_d'\mathbf{a}_j(d)|^2 \leqslant \tilde{\mathbf{p}}_d' \operatorname{cov}\mathbf{e}_1(d)\,\tilde{\mathbf{p}}_d \leqslant 1$, hence also $|\tilde{\mathbf{p}}_d'\mathbf{a}_j(d)|^{\nu} \leqslant |\tilde{\mathbf{p}}_d'\mathbf{a}_j(d)|^2$ and $\sum_{j=m}^{n} |\tilde{\mathbf{p}}_d'\mathbf{a}_j(d)|^{\nu} \leqslant 1$.

Consequently, the infinite series exists in an L^{ν}-sense with the following uniform (in T and d) moment bound

$$\mathrm{E}\left(|\tilde{\mathbf{p}}_d'\mathbf{e}_1(d)|^{\nu} \mid \mathbf{p}_d\right) = O(1) = o(T^{\nu/2-1}) \qquad a.s. \tag{D.3}$$

To prove the Lyapunov-condition under the assumptions of b) we use the Jenssen-inequality which yields

$$\begin{aligned}
&\mathrm{E}\left(|\tilde{\mathbf{p}}_d'\mathbf{e}_1(d)|^{\nu} \mid \mathbf{p}_d\right) = \|\tilde{\mathbf{p}}_d\|_1^{\nu}\ \mathrm{E}\left(\left(\sum_{i=1}^{d} \frac{|\tilde{p}_{i,d}|}{\|\tilde{\mathbf{p}}_d\|_1} |e_{i,1}(d)|\right)^{\nu} \mid \mathbf{p}_d\right) \\
&\leqslant \|\tilde{\mathbf{p}}_d\|_1^{\nu} \sum_{i=1}^{d} \frac{|\tilde{p}_{i,d}|}{\|\tilde{\mathbf{p}}_d\|_1}\ \mathrm{E}\,|e_{i,1}(d)|^{\nu} \leqslant C\left(\frac{\|\mathbf{p}_d\|_1}{\sqrt{\mathbf{p}_d' \operatorname{cov}(\mathbf{e}_1(d))\mathbf{p}_d'}}\right)^{\nu} \\
&= o(T^{\nu/2-1}) \qquad a.s.
\end{aligned} \tag{D.4}$$

□

Proof of Lemma C.2. With the notation of the proof of Theorem 3.1 both estimators (as functions of $\mathbf{p}_d$) fulfill ($j = 1, 2$)

$$\frac{\widehat{\tau}_{j,d,T}^2(\mathbf{p}_d)}{\tau^2(\mathbf{p}_d)} = \widehat{\tau}_{j,d,T}^2(\tilde{\mathbf{p}}_d).$$

First by the independence across time we get by the van Bahr-Esseen inequality (confer e.g. Lin and Bai (23233010, 9.3 and 9.4)) for some constant $C > 0$, which may differ from line to line,

$$\begin{aligned}
&\mathrm{E}_{\mathbf{p}_d}\left|\sum_{j=a+1}^{b}\left((\tilde{\mathbf{p}}_d'\mathbf{e}_j(d))^2 - 1\right)\right|^{\nu/2} \leqslant C(b-a)^{\max(1,\nu/4)}\ \mathrm{E}_{\mathbf{p}_d}\left|(\tilde{\mathbf{p}}_d'\mathbf{e}_1(d))^2 - 1\right|^{\nu/2} \\
&\leqslant C\,(b-a)^{\max(1,\nu/4)}\ \max\left(1, \mathrm{E}_{\mathbf{p}_d}\,|\tilde{\mathbf{p}}_d'\mathbf{e}_1(d)|^{\nu}\right)
\end{aligned}$$

$$\leqslant \begin{cases} C(b-a)^{\max(1,\nu/4)} & a.s., \quad \text{in a)}, \\ C(b-a)^{\max(1,\nu/4)} \max\left(1, \left(\frac{\|\mathbf{p}_d\|_1}{\sqrt{\mathbf{p}_d' \operatorname{cov} \mathbf{e}_1(d) \mathbf{p}_d}}\right)^{\nu}\right), & \text{in b)}, \end{cases} \tag{D.5}$$

by (D.3) resp. (D.4), where $\mathrm{E}_{\mathbf{p}_d}$ denotes the conditional expectation given $\mathbf{p}_d$. An application of the Markov-inequality now yields for any $\epsilon > 0$

$$\begin{aligned} & P\left(\frac{1}{T}\left|\sum_{j=1}^{T}\left((\tilde{\mathbf{p}}_d' \mathbf{e}_j(d))^2 - 1\right)\right| \geqslant \epsilon \,\middle|\, \mathbf{p}_d\right) \\ & \leqslant \begin{cases} \frac{C}{\epsilon^{\nu/2}} T^{-\nu/2+\max(1,\nu/4)} & a.s., \quad \text{in a)}, \\ \frac{C}{\epsilon^{\nu/2}} T^{-\nu/2+\max(1,\nu/4)} o(T^{\nu/2-\nu/\min(\nu,4)}) & a.s., \quad \text{in b)}, \end{cases} \\ & \to 0 \qquad a.s. \end{aligned}$$

Similar arguments yield

$$P\left(\frac{1}{T}\left|\sum_{j=1}^{T}\tilde{\mathbf{p}}_d' \mathbf{e}_j(d)\right| \geqslant \epsilon \,\middle|\, \mathbf{p}_d\right) \to 0 \qquad a.s.$$

proving a) and b) for $\widehat{\tau}^2_{1,d,T}(\mathbf{p}_d)$.

From (D.5) it follows by Theorem B.1 resp. B.4 in Kirch (2006)

$$\begin{aligned} & \mathrm{E}_{\mathbf{p}_d} \max_{1\leqslant k\leqslant T}\left|\sum_{j=1}^{k}\left((\tilde{\mathbf{p}}_d' \mathbf{e}_j(d))^2 - 1\right)\right|^{\nu/2} \\ & \leqslant \begin{cases} CT^{\max(1,\nu/4)}(\log T)^{\frac{(4-\nu)_+\nu}{2(4-\nu)}} & a.s., \quad \text{in a)}, \\ CT^{\max(1,\nu/4)}(\log T)^{\frac{(4-\nu)_+\nu}{2(4-\nu)}} \max\left(1, \left(\frac{\|\mathbf{p}_d\|_1}{\sqrt{\mathbf{p}_d' \operatorname{cov} \mathbf{e}_1(d) \mathbf{p}_d}}\right)^{\nu}\right), & \text{in b)}, \end{cases} \\ & \to 0 \qquad a.s. \end{aligned}$$

An application of the Markov inequality now yields for any $\epsilon > 0$

$$P\left(\max_{1\leqslant k\leqslant T}\frac{1}{T}\left|\sum_{j=1}^{k}\left(((\tilde{\mathbf{p}}_d' \mathbf{e}_j(d))^2 - 1\right)\right| \geqslant \epsilon \,\middle|\, \mathbf{p}_d\right) \to 0 \qquad a.s.$$

By the independence across time it holds

$$\left\{\sum_{j=k+1}^{T}\left((\tilde{\mathbf{p}}_d' \mathbf{e}_j(d))^2 - 1\right) : 1 \leqslant k \leqslant T\right\} \stackrel{\mathcal{L}}{=} \left\{\sum_{j=1}^{T-k}\left((\tilde{\mathbf{p}}_d' \mathbf{e}_j(d))^2 - 1\right) : 1 \leqslant k \leqslant T\right\},$$

which implies

$$P\left(\max_{1\leqslant k\leqslant T}\frac{1}{T}\left|\sum_{j=k+1}^{T}\left((\tilde{\mathbf{p}}_d' \mathbf{e}_j(d))^2 - 1\right)\right| \geqslant \epsilon \,\middle|\, \mathbf{p}_d\right) \to 0 \qquad a.s.$$

Similar assertions can be obtained along the same lines for $\max_{1\leqslant k\leqslant T}\frac{1}{T}\left|\sum_{j=1}^{k}\tilde{\mathbf{p}}_d'\mathbf{e}_j(d)\right|$ as well as $\max_{1\leqslant k\leqslant T}\frac{1}{T}\left|\sum_{j=k+1}^{T}\tilde{\mathbf{p}}_d'\mathbf{e}_j(d)\right|$, which imply the assertion for $\widehat{\tau}^2_{2,d,T}(\mathbf{p}_d)$. □

Proof of Corollary C.3. By an application of the continuous mapping theorem and Theorem 3.1 we get the assertions for the truncated maxima resp. the sums over $[\tau T,(1-\tau)T]$ for any $\tau>0$ towards equivalently truncated limit distributions. Because we assume independence across time (with existing second moments) the Hájek-Rényi inequality yields for all $\epsilon>0$

$$P\left(\max_{1\leqslant k\leqslant \tau T} w(k/T)\left|\sum_{t=1}^{k}\tilde{\mathbf{p}}_d'\mathbf{e}_t(d)\right|\geqslant\epsilon\,\middle|\,\mathbf{p}_d\right)\to 0\quad a.s.$$
$$P\left(\max_{(1-\tau)T\leqslant k\leqslant} w(k/T)\left|\sum_{t=k+1}^{T}\tilde{\mathbf{p}}_d'\mathbf{e}_t(d)\right|\geqslant\epsilon\,\middle|\,\mathbf{p}_d\right)\to 0\quad a.s.$$

as $\tau\to 0$ uniformly in T, where the notation of the proof of Theorem 3.1 has been used. This in addition to an equivalent argument for the limit process shows that the truncation is asymptotically negligible proving the desired results. □

Proof of Theorem 3.2. We consider the situation where $\sqrt{T}\,\mathcal{E}_3(\boldsymbol{\Delta}_d,\mathbf{p}_d)$ converges a.s. Under alternatives it holds

$$\frac{U_{d,T}(x)}{\tau(\mathbf{p}_d)}=\frac{U_{d,T}(x;\mathbf{e})}{\tau(\mathbf{p}_d)}+\operatorname{sgn}(\boldsymbol{\Delta}_d'\mathbf{p}_d)\sqrt{T}\,\mathcal{E}_3(\boldsymbol{\Delta}_d,\mathbf{p}_d)\left(\frac{1}{T}\sum_{i=1}^{\lfloor Tx\rfloor}g(i/T)-\frac{\lfloor Tx\rfloor}{T^2}\sum_{j=1}^{T}g(j/T)\right),$$

where $U_{d,T}(x;\mathbf{e})$ is the corresponding functional of the error process. By Theorem 3.1 it holds

$$\left\{\frac{U_{d,T}(x;\mathbf{e})}{\tau(\mathbf{p}_d)}:0\leqslant x\leqslant 1\,|\,\mathbf{p}_d\right\}\xrightarrow{D[0,1]}\{B(x):0\leqslant x\leqslant 1\}\qquad a.s.$$

Furthermore, by the Riemann-integrability of $g(\cdot)$ it follows

$$\sup_{0\leqslant x\leqslant 1}\left|\frac{1}{T}\sum_{i=1}^{\lfloor Tx\rfloor}g(i/T)-\frac{\lfloor Tx\rfloor}{T^2}\sum_{j=1}^{T}g(j/T)-\left(\int_0^x g(t)\,dt-x\int_0^1 g(t)\,dt\right)\right|\to 0.$$

For any $\tau>0$

$$\max_{\tau\leqslant k/T\leqslant 1-\tau}w^2(k/T)\frac{U_{d,T}^2(k/T)}{\tau^2(\mathbf{p}_d)}=T\,\mathcal{E}_3^2(\boldsymbol{\Delta}_d,\mathbf{p}_d)\left(\sup_{\tau\leqslant x\leqslant 1-\tau}w^2(x)\left(\int_0^x g(t)\,dt-x\int_0^1 g(t)\,dt\right)^2+o_{P_{\mathbf{p}_d}}(1)\right)$$

almost surely, where $P_{\mathbf{p}_d}$ denotes the conditional probability given $\mathbf{p}_d$. Because by assumption $\sup_{\tau\leqslant x\leqslant 1-\tau} w^2(x)\left(\int_0^x g(t)\,dt - x\int_0^1 g(t)\,dt\right)^2 > 0$ for some $\tau > 0$, so that the above term becomes unbounded asymptotically. This gives the assertion for the max statistics, similar arguments give the assertion for the sum statistic. □

Proof of Corollary C.4. Similarly to the proof of Theorem 3.2 it follows (where the uniformity at 0 and 1 follows by the assumptions on the rate of divergence for $w(\cdot)$ at 0 or 1)

$$\sup_{0<x<1} w^2(x)\left|\frac{U_{d,T}^2(x)}{\tau^2(\mathbf{p}_d)T\,\mathcal{E}_3^2(\mathbf{\Delta}_d,\mathbf{p}_d)} - ((x-\vartheta)_+ - x(1-\vartheta))^2\right| = o_{P_{\mathbf{p}_d}}(1) \qquad a.s.,$$

which implies the assertion by standard arguments on noting that

$$\widehat{\vartheta}_T = \arg\max_{0\leqslant x\leqslant 1} w^2(x)\,\frac{U_{d,T}^2(x)}{\tau^2(\mathbf{p}_d)T\,\mathcal{E}_3^2(\mathbf{\Delta}_d,\mathbf{p}_d)},$$

$$\vartheta = \arg\max_{0\leqslant x\leqslant 1} w^2(x)\,((x-\vartheta)_+ - x(1-\vartheta))^2\,.$$ □

Proof of Proposition 3.3. The assertion follows from

$$\tau^2(\mathbf{p}_d) = \mathbf{p}_d'\Sigma\mathbf{p}_d = \|\Sigma^{1/2}\mathbf{p}_d\|^2,$$
$$|\langle\mathbf{\Delta}_d,\mathbf{p}_d\rangle| = (\Sigma^{-1/2}\mathbf{\Delta}_d)'(\Sigma^{1/2}\mathbf{p}_d) = \|\Sigma^{-1/2}\mathbf{\Delta}_d\|\|\Sigma^{1/2}\mathbf{p}_d\|\cos(\alpha_{\Sigma^{-1/2}\mathbf{\Delta}_d,\Sigma^{1/2}\mathbf{p}_d}).$$

□

Proof of Theorem 3.4. Let $\boldsymbol{X}_d = (X_1,\ldots,X_d)'$ be $\mathrm{N}(0,I_d)$, then by Marsaglia (1972) it holds $\mathbf{r}_d \overset{\mathcal{L}}{=} (X_1,\ldots,X_d)'/\|(X_1,\ldots,X_d)'\|$ and it follows by (3.4)

$$\mathcal{E}_3^2(\mathbf{\Delta}_d,\Sigma^{-1/2}\mathbf{r}_d)\,\frac{d}{\|\Sigma^{-1/2}\mathbf{\Delta}_d\|^2} \overset{\mathcal{L}}{=} \frac{\left|\frac{\boldsymbol{X}_d'\Sigma^{-1/2}\,\mathbf{\Delta}_d}{\|\Sigma^{-1/2}\mathbf{\Delta}_d\|}\right|^2}{\frac{\boldsymbol{X}_d'\boldsymbol{X}_d}{\mathrm{E}\,\boldsymbol{X}_d'\boldsymbol{X}_d}}$$

Since the numerator has a χ_1^2 distribution (not depending on d), there exist for any $\epsilon > 0$ constants $0 < c_1 < C_1 < \infty$ such that

$$\sup_{d\geqslant 1} P\left(c_1 \leqslant \left|\frac{\boldsymbol{X}_d'\Sigma^{-1/2}\,\mathbf{\Delta}_d}{\|\Sigma^{-1/2}\mathbf{\Delta}_d\|}\right|^2 \leqslant C_1\right) \geqslant 1-\epsilon.$$

Furthermore, the denominator has a χ_d^2-distribution divided by its expectation, consequently an application of the Markov-inequality yields for any $\epsilon > 0$ the existence of $0 < C_2 < \infty$ such that

$$\sup_{d\geqslant 1} P\left(\frac{\boldsymbol{X}_d'\boldsymbol{X}_d}{\mathrm{E}\,\boldsymbol{X}_d'\boldsymbol{X}_d} \geqslant C_2\right) \leqslant \epsilon.$$

By integration by parts we get $\mathrm{E}\left(\boldsymbol{X}_d'\boldsymbol{X}_d\right)^{-1} \leqslant 2/d$ for $d \geqslant 3$ so that another application of the Markov-inequality yields that for any $\epsilon > 0$ there exists $c_2 > 0$ such that

$$\limsup_{d\to\infty} P\left(\frac{\boldsymbol{X}_d'\boldsymbol{X}_d}{\mathrm{E}\,\boldsymbol{X}_d'\boldsymbol{X}_d} \leqslant c_2\right) \leqslant \epsilon,$$

completing the proof of the theorem by standard arguments. □

Proof of Theorem 3.5. Let $\boldsymbol{X}_d = (X_1, \ldots, X_d)'$ be $\mathrm{N}(0, I_d)$, then as in the proof of Theorem 3.4 it holds

$$\mathcal{E}_3^2(\boldsymbol{\Delta}, \mathbf{M}^{-1/2}\mathbf{r}_d)\,\frac{\mathrm{tr}(\mathbf{M}^{-1/2}\Sigma\mathbf{M}^{-1/2})}{\|\mathbf{M}^{-1/2}\boldsymbol{\Delta}_d\|^2} \overset{\mathcal{L}}{=} \frac{\left|\frac{\boldsymbol{X}_d'\mathbf{M}^{-1/2}\boldsymbol{\Delta}_d}{\|\mathbf{M}^{-1/2}\boldsymbol{\Delta}_d\|}\right|^2}{\frac{\boldsymbol{X}_d'\mathbf{M}^{-1/2}\Sigma\mathbf{M}^{-1/2}\boldsymbol{X}_d}{\mathrm{tr}(\mathbf{M}^{-1/2}\Sigma\mathbf{M}^{-1/2})}}.$$

The proof of the lower bound is analogous to the proof of Theorem 3.4 by noting that $(A = \mathbf{M}^{-1/2}\Sigma\mathbf{M}^{-1/2})$

$$\mathrm{E}\,\mathbf{X}'A\mathbf{X} = \mathrm{E}\sum_{i,j=1}^{d} a_{i,j}X_iX_j = \sum_{i,j=1}^{d} a_{i,j}\delta_{i,j} = \sum_{i=1}^{d} a_{i,i} = \mathrm{tr}(A).$$

For the proof of the upper bound, first note that by a spectral decomposition it holds

$$\frac{\mathbf{X}'\mathbf{M}^{-1/2}\Sigma\mathbf{M}^{-1/2}\mathbf{X}}{\mathrm{tr}(\mathbf{M}^{-1/2}\Sigma\mathbf{M}^{-1/2})} \overset{\mathcal{L}}{=} \sum_{j=1}^{d}\alpha_jX_j^2, \qquad \text{for some } 0 < \alpha_d \leqslant \ldots \leqslant \alpha_1, \quad \sum_{j=1}^{d}\alpha_j = 1.$$

From this we get on the one hand by the Markov inequality

$$P\left(\sum_{j=1}^{d}\alpha_jX_j^2 \leqslant c\right) \leqslant P(\alpha_1X_1^2 \leqslant c) \leqslant \left(\frac{c}{\alpha_1}\right)^{1/4}\mathrm{E}(|X_1^2|^{-1/4}),$$

where $\mathrm{E}(|X_1^2|^{-1/4}) = \Gamma(1/4)/(2^{1/4}\sqrt{\pi})$ exists (as can be seen using the density for a χ_1^2-distribution). On the other hand it holds for any $c \leqslant 1/2$ by another application of the Markov inequality

$$P\left(\sum_{j=1}^{d}\alpha_jX_j^2 \leqslant c\right) \leqslant P\left(\left|\sum_{j=1}^{d}\alpha_jX_j^2 - 1\right| \geqslant 1/2\right) \leqslant 8\sum_{i=1}^{d}\alpha_i^2 \leqslant 8\alpha_1.$$

By chosing $c = \min(1/2, (\mathrm{E}(|X_1^2|^{-1/4}))^{-4}/8\,\epsilon^5)$ we finally get

$$\begin{aligned}&\sup_{0<\alpha_d\leqslant\ldots\leqslant\alpha_1,\sum_{i=1}^{d}\alpha_i=1} P\left(\sum_{j=1}^{d}\alpha_jX_j^2 \leqslant c\right)\\ &\leqslant \sup_{0<\alpha_d\leqslant\ldots\leqslant\alpha_1,\sum_{i=1}^{d}\alpha_i=1} \min\left(\epsilon\left(\frac{\epsilon}{8\alpha_1}\right)^{1/4}, 8\alpha_1\right) \leqslant \epsilon,\end{aligned}$$

completing the proof. □

Proof of Theorem 3.6. By the Cauchy-Schwarz inequality

$$\tau^2(\mathbf{M}^{-1}\boldsymbol{\Delta}_d) = \boldsymbol{\Delta}_d'\mathbf{M}^{-1}\sum_{j\geqslant 1}\mathbf{a}_j\mathbf{a}_j'\mathbf{M}^{-1}\boldsymbol{\Delta}_d = \sum_{j\geqslant 1}(\mathbf{a}_j'\mathbf{M}^{-1}\boldsymbol{\Delta}_d)^2$$

$$\leqslant \sum_{j\geqslant 1}\mathbf{a}_j'\mathbf{M}^{-1}\mathbf{a}_j\ \boldsymbol{\Delta}_d'\mathbf{M}^{-1}\boldsymbol{\Delta}_d = \operatorname{tr}\left(\mathbf{M}^{-1/2}\sum_{j\geqslant 1}\mathbf{a}_j\mathbf{a}_j'\mathbf{M}^{-1/2}\right)\boldsymbol{\Delta}_d'\mathbf{M}^{-1}\boldsymbol{\Delta}_d,$$

which implies the assertion by (3.4). □

Proof of Proposition 3.7. Assertion a) follows from

$$|\langle\boldsymbol{\Delta}_d,\ {}_p\mathbf{o}\rangle|^2 = \left(\sum_{i=1}^d \frac{\delta_{i,T}^2}{\sigma_i^2}\,\sigma_i^2\right)^2 \geqslant c^2\left(\sum_{i=1}^d \frac{\delta_{i,T}^2}{\sigma_i^2}\right)^2 = c^2\,|\langle\boldsymbol{\Delta}_d,\ {}_q\mathbf{o}\rangle|^2,$$

$$\tau^2({}_p\mathbf{o}) = {}_p\mathbf{o}'\,\Sigma\,{}_p\mathbf{o} = \sum_{i=1}^d \frac{\delta_{i,T}^2}{\sigma_i^2}\,\sigma_i^4 \leqslant C^2\,|\langle\boldsymbol{\Delta}_d,\ {}_q\mathbf{o}\rangle|\,.$$

Concerning b) first note that by the Cauchy-Schwarz inequality with $\Lambda = \operatorname{diag}(\sigma_1^2,\ldots,\sigma_d^2)$

$$\tau^2({}_q\mathbf{o}) = \sum_{j\geqslant 1}(\boldsymbol{\Delta}_d'\Lambda^{-1}\mathbf{a}_j)^2 \leqslant \boldsymbol{\Delta}_d'\Lambda^{-2}\boldsymbol{\Delta}_d\sum_{j\geqslant 1}\mathbf{a}_j'\mathbf{a}_j \leqslant \frac{\boldsymbol{\Delta}_d'\boldsymbol{\Delta}_d}{c^2}\operatorname{tr}(\Sigma).$$

This implies assertion b) by (3.4) on noting that

$$|\boldsymbol{\Delta}_d'\Lambda^{-1}\boldsymbol{\Delta}_d|^2 \geqslant \frac{|\boldsymbol{\Delta}_d'\boldsymbol{\Delta}_d|^2}{C^2}.$$

□

Proof of Equation (C.8). By Proposition 3.3 it holds for $\boldsymbol{\Delta}_d = k\,\boldsymbol{\Phi}_d$

$$\mathcal{E}_3^2(\boldsymbol{\Delta}_d,\mathbf{o}) = \|\Sigma^{-1/2}\boldsymbol{\Delta}_d\|^2 = \boldsymbol{\Delta}_d'(\boldsymbol{D}+\boldsymbol{\Phi}_d\boldsymbol{\Phi}_d')^{-1}\boldsymbol{\Delta}_d,$$

where $\boldsymbol{D} = \operatorname{diag}(s_1^2,\ldots,s_d^2)'$. Hence

$$\begin{aligned}
&\boldsymbol{\Delta}_d'(\boldsymbol{D}+\boldsymbol{\Phi}_d\boldsymbol{\Phi}_d')^{-1}\boldsymbol{\Delta}_d\\
&= (\boldsymbol{D}^{-1/2}\boldsymbol{\Delta}_d)'\left(I_d + (\boldsymbol{D}^{-1/2}\boldsymbol{\Phi}_d)(\boldsymbol{D}^{-1/2}\boldsymbol{\Phi}_d)'\right)^{-1}\boldsymbol{D}^{-1/2}\boldsymbol{\Delta}_d\\
&= \frac{(\boldsymbol{D}^{-1/2}\boldsymbol{\Delta}_d)'\boldsymbol{D}^{-1/2}\boldsymbol{\Delta}_d}{1+\boldsymbol{D}^{-1/2}\boldsymbol{\Phi}_d'\boldsymbol{D}^{-1/2}\boldsymbol{\Phi}_d},
\end{aligned}$$

where the last line follows from the fact that $\boldsymbol{D}^{-1/2}\boldsymbol{\Delta}_d = k\boldsymbol{D}^{-1/2}\boldsymbol{\Phi}_d$ is an eigenvector of $I_d + (\boldsymbol{D}^{-1/2}\boldsymbol{\Phi}_d)(\boldsymbol{D}^{-1/2}\boldsymbol{\Phi}_d)'$ with eigenvalue $1+(\boldsymbol{D}^{-1/2}\boldsymbol{\Phi}_d)'\boldsymbol{D}^{-1/2}\boldsymbol{\Phi}_d$ hence also for the inverse of the matrix with inverse eigenvalue. □

Proof of Theorem 2.1. Similarly as in the proof of Theorem 3.2 it holds

$$Z_{T,i}(x) = Z_{T,i}(x;\mathbf{e}) + \delta_{i,T}\sqrt{T}\left(\frac{1}{T}\sum_{j=1}^{\lfloor Tx\rfloor} g(j/T) + \frac{\lfloor Tx\rfloor}{T^2}\sum_{j=1}^{T} g(j/T)\right),$$

where $Z_{T,i}(x;\mathbf{e})$ is the corresponding functional for the error sequence (rather than the actual observations). From this it follows

$$V_{d,T}(x) = V_{d,T}(x;\mathbf{e}) + T\,\mathcal{E}_1^2(\mathbf{\Delta}_d)\left(\int_0^x g(t)\,dt - x\int_0^1 g(t)\,dt + o(1)\right) + R_T(x),$$

where $R_T(x)$ is the mixed term given by

$$R_T(x) = \frac{2\sqrt{T}}{\sqrt{d}}\sum_{i=1}^{d}\frac{\delta_{i,T}}{\sigma_i^2}Z_{T,i}(x;\mathbf{e})\left(\int_0^x g(t)\,dt - x\int_0^1 g(t)\,dt + o(1)\right)$$

which by an application of the Hájek -Rényi inequality (across time) yields

$$P\left(\sup_{0\leqslant x\leqslant 1}|R_T(x)| \geqslant c\right) = O\left(1\right)\frac{1}{c^2}T\frac{1}{d}\sum_{i=1}^{d}\frac{\delta_i^2}{\sigma_i^2} = O_P(1)\frac{1}{c^2\sqrt{d}}\,T\,\mathcal{E}_1(\mathbf{\Delta}_d).$$

From this the assertion follows by an application of Theorem B.1. □

Proof of Lemma 2.2. The proof follows closely the proof of (28)–(30) in Horváth and Hušková (2012) but where we scale diagonally with the true variances. We will give a short sketch for the sake of completeness. The key is the following decomposition

$$\begin{aligned}
&V_{d,T}(x)\\
&= \frac{1}{\sqrt{d}}\sum_{i=1}^{d}\left(\frac{s_i^2}{s_i^2+\Phi_i^2}\frac{1}{T}\left(\sum_{t=1}^{\lfloor Tx\rfloor}\eta_{i,t}(d) - \frac{\lfloor Tx\rfloor}{T}\sum_{t=1}^{T}\eta_{i,T}(d)\right)^2 - \frac{\lfloor Tx\rfloor(T-\lfloor Tx\rfloor)}{T^2}\right)\\
&+ \frac{2}{\sqrt{d}}\left(\sum_{i=1}^{d}\frac{\Phi_i s_i}{s_i^2+\Phi_i^2}\frac{1}{\sqrt{T}}\left(\sum_{t=1}^{\lfloor Tx\rfloor}\eta_{i,t}(d) - \frac{\lfloor Tx\rfloor}{T}\sum_{t=1}^{T}\eta_{i,T}(d)\right)\right)\\
&\quad\cdot\frac{1}{\sqrt{T}}\left(\sum_{t=1}^{\lfloor Tx\rfloor}\eta_{d+1,t}(d) - \frac{\lfloor Tx\rfloor}{T}\sum_{t=1}^{T}\eta_{d+1,t}(d)\right)\\
&+ \frac{1}{T}\left(\sum_{t=1}^{\lfloor Tx\rfloor}\eta_{d+1,t}(d) - \frac{\lfloor Tx\rfloor}{T}\sum_{t=1}^{T}\eta_{d+1,t}(d)\right)^2\frac{1}{\sqrt{d}}A_d.
\end{aligned}$$

The first term converges to the limit given in a). To see this, note that the proof of the Lyapunov condition in Horváth and Hušková (2012) following equation

(39) still holds because $s_i^2/(s_i^2+\Phi_i^2)$ is uniformly bounded from above by assumption (showing that the numerator is bounded) while again by assumption

$$\frac{1}{d}\sum_{i=1}^{d}\frac{s_i^4}{(s_i^2+\phi_i^2)^2} \geqslant D > 0,$$

showing that the denominator is bounded. Similarly, the proof of tightness in Horváth and Hušková (2012) (equations (43) and following) remains valid. The asymptotic variance remains the same under a) and b) because by assumption

$$\left|\frac{1}{d}\sum_{i=1}^{d}\frac{s_i^4}{(s_i^2+\Phi_i^2)^2}-1\right| \leqslant \frac{3}{d}A_d \to 0.$$

The middle term in the above decomposition is bounded by an application of the Hájek -Rényi inequality

$$P\left(\sup_{0<x<1}\frac{1}{\sqrt{d}}\left|\sum_{i=1}^{d}\frac{\Phi_i s_i}{s_i^2+\Phi_i^2}\frac{1}{\sqrt{T}}\left(\sum_{t=1}^{\lfloor Tx\rfloor}\eta_{i,t}(d)-\frac{\lfloor Tx\rfloor}{T}\sum_{t=1}^{T}\eta_{i,T}(d)\right)\right| \geqslant D\right)$$
$$= O(1)\,\frac{1}{d}\sum_{j=1}^{d}\frac{\phi_i^2 s_i^2}{(s_i^2+\phi_i^2)^2} = O(1)\,\frac{1}{d}A_d,$$

which converges to 0 for a) and b) – for c) we multiply the original statistic by $\sqrt{d}/A_d$, which means this term is multiplied with d/A_d^2 leaving us with $1/A_d$ which converges to 0 if $A_d/\sqrt{d}\to\infty$.
Similarly, we can bound $\frac{1}{\sqrt{T}}\left(\sum_{t=1}^{\lfloor Tx\rfloor}\eta_{d+1,t}(d)-\frac{\lfloor Tx\rfloor}{T}\sum_{t=1}^{T}\eta_{d+1,t}(d)\right)$, showing that the middle term is asymptotically negligible. The assertions now follow by an application of the functional central limit theorem for $\frac{1}{T}\left(\sum_{t=1}^{\lfloor Tx\rfloor}\eta_{d+1,t}(d)-\frac{\lfloor Tx\rfloor}{T}\sum_{t=1}^{T}\eta_{d+1,t}(d)\right)^2$. □

Proof of Theorem 2.3. The proof is analogous to the one of Theorem 2.1 on noting that $\mathcal{E}_2^2(\mathbf{\Delta}_d)=\frac{\sqrt{d}}{A_d}\mathcal{E}_1^2(\mathbf{\Delta}_d)$ and $\sigma_i^2=s_i^2+\Phi_i^2$ by using Lemma 2.2 c) above. Concerning the remainder term $\widetilde{R}_T(x)$ note that $e_{i,t}=s_i\eta_{i,t}+\Phi_i\eta_{d+1,t}$, so that the remainder term can be split into two terms. The first term can be dealt with analogously to the proof of Theorem 2.1 and is of order $O_P\left(\sqrt{\frac{1}{A_d}T\mathcal{E}_2(\mathbf{\Delta}_d)}\right)$, while for the second summand we get by an application of the Cauchy-Schwarz-inequality

$$\sup_{0\leqslant x\leqslant 1}\left|\frac{1}{A_d}\sum_{i=1}^{d}\frac{\delta_i\phi_i}{\sigma_i^2}\left(\sum_{t=1}^{\lfloor Tx\rfloor}\eta_{d+1,t}-\frac{\lfloor Tx\rfloor}{T}\sum_{t=1}^{T}\eta_{d+1,t}\right)\right| = O_P(\sqrt{T})\sqrt{\frac{\sum_{i=1}^{d}\frac{\delta_i^2}{\sigma_i^2}}{A_d}}$$
$$= O\left(\sqrt{T\,\mathcal{E}_2^2(\mathbf{\Delta}_d)}\right).$$ □

Proof of Corollary 3.8. By an application of the Cauchy-Schwarz inequality it holds

$$\boldsymbol{\Delta}_d' \Lambda_d^{-1} \Sigma \Lambda_d^{-1} \boldsymbol{\Delta}_d = \sum_{i=1}^{d} \delta_{i,T}^2 \frac{s_i^2}{(s_i^2+\Phi_i^2)^2} + \left(\sum_{i=1}^{d} \frac{\delta_{i,T}\Phi_i}{s_i^2+\Phi_i^2} \right)^2$$
$$\leqslant \sum_{i=1}^{d} \frac{\delta_{i,T}^2}{\sigma_i^2} \left(1 + \sum_{i=1}^{d} \frac{\Phi_i^2}{\sigma_i^2} \right) = \boldsymbol{\Delta}_d' \Lambda_d^{-1} \boldsymbol{\Delta}_d \, (1 + A_d),$$

which implies assertion a) on noting that

$$\mathcal{E}_3^2(\boldsymbol{\Delta}_d, {}_q\mathbf{o}) = \frac{(\boldsymbol{\Delta}_d' \Lambda_d^{-1} \boldsymbol{\Delta}_d)^2}{\boldsymbol{\Delta}_d' \Lambda_d^{-1} \Sigma \Lambda_d^{-1} \boldsymbol{\Delta}_d}.$$

b) This follows immediately from Theorem 3.5 since by $0 < c \leqslant s_j^2 \leqslant C < \infty$ as well as as $\Phi_i^2 \leqslant C$, it follows that

$$\|\boldsymbol{\Delta}_d\|^2 \sim \boldsymbol{\Delta}_d' \operatorname{diag} \left(\frac{1}{s_1^2+\Phi_1^2}, \ldots, \frac{1}{s_d^2+\Phi_d^2} \right) \boldsymbol{\Delta}_d. \qquad \square$$

Acknowledgments

The first author was supported by the Engineering and Physical Sciences Research Council (UK) grants : EP/K021672/2 & EP/N031938/1. Some of this work was done while the second author was at KIT where her position was financed by the Stifterverband für die Deutsche Wissenschaft by funds of the Claussen-Simon-trust. Furthermore, this work was supported by the Ministry of Science, Research and Arts, Baden-Württemberg, Germany. We would also like to thank Alexandra Carpentier for pointing out the connections to minimax optimality to us. Finally, the authors would like to thank the Isaac Newton Institute for Mathematical Sciences, Cambridge, for support and hospitality during the programme 'Inference for Change-Point and Related Processes', where part of the work on this paper was undertaken.

References

J. A. D. Aston and C. Kirch. Detecting and estimating changes in dependent functional data. *Journal of Multivariate Analysis*, 109:204–220, 2012a. MR2922864

J. A. D. Aston and C. Kirch. Evaluating stationarity via change-point alternatives with applications to fMRI data. *Annals of Applied Statistics*, 6:1906–1948, 2012b. MR3058688

A. Aue and L. Horváth. Structural breaks in time series. *Journal of Time Series Analysis*, 34:1–16, 2013. MR3008012

A. Aue, R. Gabrys, L. Horváth, and P. Kokoszka. Estimation of a change-point in the mean function of functional data. *Journal of Multivariate Analysis*, 100:2254–2269, 2009a. MR2560367

A. Aue, S. Hörmann, L. Horváth, and M. Reimherr. Break detection in the covariance structure of multivariate time series models. *Annals of Statistics*, 37:4046–4087, 2009b. MR2572452

J. Bai. Common Breaks in Means and Variances for Panel Data. *Journal of Econometrics*, 157:78–92, 2010. MR2652280

Y. Baraud et al. Non-asymptotic minimax rates of testing in signal detection. *Bernoulli*, 8(5):577–606, 2002. MR1935648

I. Berkes, R. Gabrys, L. Horváth, and P. Kokoszka. Detecting changes in the mean of functional observations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71:927–946, 2009. MR2750251

P. J. Bickel and E. Levina. Regularized estimation of large covariance matrices. *Annals of Statistics*, 36:199–227, 2008. MR2387969

P. Billingsley. *Convergence of probability measures*. John Wiley & Sons, 1968. MR0233396

T. Cai and W. Liu. A direct estimation approach to sparse linear discriminant analysis. *Journal of the American Statistical Association*, 106(496):1566–1577, 2011. MR2896857

H. P. Chan and G. Walther. Optimal detection of multi-sample aligned sparse signals. *Annals of Statistics*, 43(5):1865–1895, 2015. MR3375870

J. Chan, L. Horváth, and M. Hušková. Darling–Erdős limit results for change–point detection in panel data. *Journal of Statistical Planning and Inference*, 2012. MR3011306

H. Cho. Change-point detection in panel data via double cusum statistic. *Electronic Journal of Statistics*, in press, 2015. MR3522667

H. Cho and P. Fryzlewicz. Multiple-change-point detection for high dimensional time series via sparsified binary segmentation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77(2):475–507, 2015. MR3310536

M. Csörgő and L. Horváth. *Limit Theorems in Change-Point Analysis*. Wiley, Chichester, 1997. MR2743035

A. Delaigle and P. Hall. Achieving near perfect classification for functional data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(2):267–286, 2012. MR2899863

I. Eckley, P. Fearnhead, and R. Killick. Analysis of changepoint models. In D. Barber, A. Cemgil, and S. Chiappa, editors, *Bayesian Time Series Models*, pages 215–238. Cambridge University Press, 2011. MR2894240

F. Enikeeva and Z. Harchaoui. High-dimensional change-point detection with sparse alternatives. *ArXiv e-prints*, Dec. 2013.

J. Fan, Y. Liao, and M. Mincheva. Large covariance estimation by thresholding principal orthogonal complements. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75:603–680, 2013. MR3091653

K. Frick, A. Munk, and H. Sieling. Multiscale change point inference. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76:495–580, 2014. MR3210728

W. A. Fuller. *Introduction to statistical time series*. John Wiley & Sons, 1996. MR1365746

N. J. Higham. *Accuracy and stability of numerical algorithms*. Siam, 2002. MR1927606

S. Hörmann and P. Kokoszka. Weakly dependent functional data. *Annals of Statistics*, 38:1845–1884, 2010. MR2662361

L. Horváth and M. Hušková. Change-point detection in panel data. *Journal of Time Series Analysis*, 33:631–648, 2012. MR2944843

L. Horváth and G. Rice. Extensions of some classical methods in change point analysis. *TEST*, 23:219–255, 2014. MR3210268

L. Horváth, P. Kokoszka, and J. Steinebach. Testing for Changes in Multivariate Dependent Observations with an Application to Temperature Changes. *Journal of Multivariate Analysis*, 68:96–119, 1999. MR1668911

Y. Ingster and I. A. Suslina. *Nonparametric goodness-of-fit testing under Gaussian models*, volume 169. Springer Science & Business Media, 2012. MR1991446

M. Jirak. Uniform change point tests in high dimension. *Ann. Statist.*, 43(6):2451–2483, 12 2015. MR3405600

C. Kirch. *Resampling Methods for the Change Analysis of Dependent Data*. PhD thesis, University of Cologne, Cologne, 2006. http://kups.ub.uni-koeln.de/volltexte/2006/1795/.

C. Kirch and J. T. Kamgaing. On the use of estimating functions in monitoring time series for change points. *Journal of Statistical Planning and Inference*, 161:25–49, 2015. MR3316549

C. Kirch and J. T. Kamgaing. Detection of change points in discrete valued time series. *Handbook of discrete valued time series. In: Davis RA, Holan SA, Lund RB, Ravishanker N*, 2016. MR3699407

C. Kirch and J. Tadjuidje Kamgaing. Testing for parameter stability in nonlinear autoregressive models. *Journal of Time Series Analysis*, 33:365–385, 2012. MR2915090

C. Kirch, B. Mushal, and H. Ombao. Detection of changes in multivariate time series with applications to eeg data. *Journal of the American Statistical Association*, 110:1197–1216, 2015. MR3420695

E. L. Lehmann. *Elements of Large Sample Theory*. Springer Berlin Heidelberg, 1999. MR1663158

Z. Lin and Z. Bai. *Probability inequalities*. Springer, 2010. MR2789096

M. Lopes, L. Jacob, and M. J. Wainwright. A more powerful two-sample test in high dimensions using random projection. In *Advances in Neural Information Processing Systems*, pages 1206–1214, 2011.

G. Marsaglia. Choosing a point from the surface of a sphere. *Annals of Mathematical Statistics*, 43:645–646, 1972.

H. Ombao, R. Von Sachs, and W. Guo. SLEX analysis of multivariate nonstationary time series. *Journal of the American Statistical Association*, 100:519–531, 2005. MR2160556

E. S. Page. Continuous Inspection Schemes. *Biometrika*, 41:100–115, 1954. MR0088850

D. N. Politis. Higher-order accurate, positive semi-definite estimation of large-sample covariance and spectral density matrices. *Econometric Theory*, 27:703–744, 2011. MR2822363

M. Robbins, C. Gallagher, R. Lund, and A. Aue. Mean shift testing in correlated data. *Journal of Time Series Analysis*, 32:498–511, 2011. MR2835683

M. S. Srivastava and M. Du. A test for the mean vector with fewer observations than the dimension. *Journal of Multivariate Analysis*, 99(3):386–402, 2008. MR2396970

L. Torgovitski. Detecting changes in hilbert space data based on "repeated" and change-aligned principal components. *arXiv preprint: 1509.07409*, 2015.

L. Wang, B. Peng, and R. Li. A high-dimensional nonparametric multivariate test for mean vector. *Journal of the American Statistical Association*, 110(512):1658–1669, 2015. MR3449062

T. Wang and R. J. Samworth. High-dimensional changepoint estimation via sparse projection. *Journal of the Royal Statistical Society: Series B*, 80:57–83, 2018. MR3744712

H. Zou, T. Hastie, and R. Tibshirani. Sparse principal component analysis. *Journal of Computational and Graphical Statistics*, 15:265–286, 2006. MR2252527