Change points in high dimensional settings[*]

John A D Aston[†]   Claudia Kirch[‡]

September 5, 2014

Abstract

While there is considerable work on change point analysis in univariate time series, more and more data being collected comes from high dimensional multivariate settings. This paper investigates change point detection procedures using projections and develops asymptotic theory for how full panel (multivariate) tests compare with both oracle and random projections. This is done by considering an analogous concept to asymptotic relative efficiency termed high dimensional efficiency. This provides the rate at which the change can get smaller with dimension while still being detectable. The effect of misspecification of the covariance on the power of the tests is 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. It is shown that if information concerning the direction of change is available, then projecting in this direction is always advantageous over the use of a panel statistic, in terms of size and power, particularly when the covariance is misspecified. Even if the change is not known, the projection method achieves a better power as long as the difference between the true change and the direction of the projection is small. The features of the tests are quantified in theory and simulations indicate that these results are indicative of small sample behaviour.

Keywords: CUSUM; High Dimensional Efficiency; Model Misspecification; Panel Data; Projections

AMS Subject Classification 2000: 62M10;

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

[*] The first author was supported by the Engineering and Physical Sciences Research Council (UK) grant : EP/K021672/2. The position of the second author was financed by the Stifterverband für die Deutsche Wissenschaft by funds of the Claussen-Simon-trust and this work was supported by the Ministry of Science, Research and Arts, Baden-Württemberg, Germany. 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.

[†] Statistical Laboratory, DPMMS, University of Cambridge, Cambridge, CB3 9HD, UK; j.aston@statslab.cam.ac.uk

[‡] Karlsruhe Institute of Technology (KIT), Institute for Stochastics, Kaiserstr. 89, D – 76133 Karlsruhe, Germany; claudia.kirch@kit.edu

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 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., 2014+] functional [Berkes et al., 2009, Aue et al., 2009a, Hörmann and Kokoszka, 2010, Aston and Kirch, 2012a] and high dimensional settings [Bai, 2010, Horváth and Hušková, 2012, Chan et al., 2012, Cho and Fryzlewicz, 2014+] have recently been proposed.

Instead of looking at more and more complicated models, this paper uses a simple mean change setting to illustrate how the power is influenced in high dimensional settings. The results and techniques can subsequently be extended to more complex change point setting as well as different statistical frameworks, such as two sample tests. We make use of the following two key concepts: Firstly, we investigate a class of tests based on projections. Secondly, 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).

The aims of the paper are threefold: Firstly, we will investigate the asymptotic properties of tests based on projections as a plausible way to include prior information into the tests. Secondly, by using high dimensional efficiency, we consider several projection tests (including oracle and random projections as benchmarks) and compare them with the efficiency of existing tests that take the full covariance structure into account. Finally, as in all high dimensional settings, the dependency between the components of the series can typically neither be effectively estimated nor even uniquely determined (for example if the sample size is less than the multivariate dimension) unless restrictions on the covariance are enforced. By considering the effect of misspecification of the model covariance on the size as well as efficiency we can quantify the implications of this for different tests.

Highest efficiency can only be achieved under knowledge of the direction of the change. In fact, data practitioners, in many cases, explicitly have prior knowledge in which direction changes are likely to occur. It should be noted at this point, that changes in mean are equivalent to changes of direction in multivariate time series. In frequentist testing situations, practitioners' main interest is in test statistics which have power against a range of related alternatives while still controlling the size. For example, an economist may check the performance of several companies looking for changes caused by a recession. There will often be a general idea as to which sectors of the economy will gain or lose by the recession and therefore a good idea, at least qualitatively, as to what a change will approximately look like (downward resp. upward shift depending on which sector a particular company is in) if there is a change present. Similarly, in medical studies, it will often be known a-priori whether genes are likely to be coregulated causing changes to be in similar directions for groups of genes in genetic time series.

Incorporating this a-priori information about how the change affects the components by using corresponding projections leads to a considerable power improvement if the change is indeed in the expected direction. It is also important that, as in many cases the a-priori knowledge is qualitative, the test has higher power than standard tests not only for that particular direction but also for other directions close by. Additionally, these projections lead to tests where the size is better controlled if no change is present. However, if such prior information is not reliable, it is important to be able to quantify the loss over tests which do not assume such information. In addition, while the prior information itself might be reliable, inherent misspecification in other parts of the model, such as the covariance structure, will have a detrimental effect on detection, and it is of interest to quantify this as well.

The results in this paper will be benchmarked against taking the simple approach of using a random projection in a single direction to reduce the dimension of the data. Random projections are becoming increasingly popular in high dimensional statistics with applications in Linear Discriminant Analysis [Durrant and Kabán, 2010] and two sample testing [Lopes et al., 2011, Srivastava et al., 2014]. This is primarily based on the insight from the Johnson-Lindenstrauss lemma that an optimal projection in the sense that the distances are preserved for a given set of data is independent of the dimension of the data [Johnson and Lindenstrauss, 1984] and thus random projections can often be a useful way to perform a dimension reduction for the data [Baraniuk et al., 2008]. However, in our context, we will see that a random projection will not work as well as truly multivariate methods, let alone projections with prior knowledge, but can only serve as a lower benchmark.

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 Tadjuidje Kamgaing, 2014a, Kirch and Tajduidje Kamgaing, 2014b]. 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. Furthermore, the developed theory applies equally to one- and two-sample testing and can be seen as somewhat analogous to methods for multivariate adaptive design [Minas et al., 2014].

The change is given by $\mathbf{\Delta}_d = (\delta_{1,T}, \ldots, \delta_{d,T})^T$ 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 (AMOC): $g(u) = \begin{cases} 0 & 0 \leq u \leq \theta_1 \\ 1 & \theta_1 < u < \theta_2 \\ 0 & \theta_2 < u \leq 1 \end{cases}$

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]. It will be shown that statistics based on partial sums are robust (in the sense of still having non-zero power) to a wide variety of g.

The model in (1.1) is defined for univariate ($d = 1$), multivariate (d fixed) or panel data ($d \to \infty$). We will consider the latter two, as these are of most interest in the high dimensional setting. In particular, the panel data (also known as "small n large p" or "high dimensional low sample size") setting is able to capture very well the small sample properties in situations where d is comparable or even larger than T using asymptotic considerations. It is this asymptotic framework that really enables a thorough investigation of the properties of various tests, as the rates at which various vanishing alternatives can be detected give an indication into the detection ability of the tests. In particular we will consider if tests are of the same order, $a_T \sim b_T$, which means that two constants c, C (independent of the dimension) exist such that $c \leqslant \frac{a_T}{b_T} \leqslant C$ as the dimension increases. 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 use of projections for detecting changes is investigated, particularly in terms of their size and power. In addition, the effect of the misspecification of the covariance structure on the tests will be investigated. In Section 3, the projection based statistics will be compared with the panel based change point statistics already suggested in Horváth and Hušková [2012], both in terms of control of size and power properties, particular with relation to the (mis)specification of the dependence structure. Section 4 concludes with some discussion of the different statistics proposed, while Section 5 gives the proofs of the results in the paper. In addition, rather than a separate simulation section, simulations will be interspersed throughout the theory. This is because many of the finite sample simulations give considerable intuition into the resulting asymptotic properties which are derived. 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 2.1.

2 Change Points and Projections

2.1 Projections

In model (1.1), the change $\mathbf{\Delta}_d = (\delta_{1,T}, \dots, \delta_{d,T})^T$ is always a one-dimensional 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. In fact, under (1.1) it holds

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

where $\mathbf{X}_d(t) = (X_{1,t}, \dots, X_{d,T})^T$, $\boldsymbol{\mu} = (\mu_1, \dots, \mu_d)^T$ and $\mathbf{e}_t = (e_{1,t}, \dots, e_{d,t})^T$. 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 the better the larger $\langle \mathbf{\Delta}_d, \mathbf{p}_d \rangle$ and the smaller the variance of $\langle \mathbf{e}_t, \mathbf{p}_d \rangle$ is. Consequently, an optimal projection in terms of power depends on $\mathbf{\Delta}_d$ as well as $\Sigma = \operatorname{var} \mathbf{e}_1$. In applications, certain changes are either expected or of particular interest e.g. an economist looking at the performance of several companies expecting changes caused by a recession will have a good idea which companies will profit or lose. This knowledge can be used to increase the power in directions close to the search direction $\boldsymbol{p}_d$ while decreasing it for changes that are close to orthogonal to it. Using projections can furthermore robustify the size of the test under the null hypothesis with respect to misspecification and estimation error.

In order to qualify this informal statement, we will consider contiguous changes for several change point tests, where $\|\mathbf{\Delta}_d\| \to 0$ but with such a rate that the power of the corresponding test is strictly between the size and one. Unlike for classical asymptotic efficiency, the information about detection power with respect to the dimension is contained in the rates, which will subsequently be called high dimensional efficiency. This high dimensional efficiency allows us to quantify and compare the power gain obtained by projections in comparison to statistics that use the full multivariate information. These results are confirmed by simulations using relatively small sample sizes. Furthermore, the general message holds true far beyond the particular test statistic or even this particular model, namely that (appropriate) projections can help stabilize size and at the same time increase the power for change point tests.

In order to be able to prove asymptotic results for change point statistics based on projections 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 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 be allowed.

Assumption $\mathcal{A}$. 1. Let $\eta_{1,t}(d), \eta_{2,t}(d), \ldots$ independent with $\operatorname{E} \eta_{i,t}(d) = 0$, $\operatorname{var} \eta_{i,t}(d) = 1$ and $\operatorname{E} |\eta_{i,t}(d)|^{\nu} \leqslant C < \infty$ 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 $e_t(d) = (e_{1,t}(d), \ldots, e_{d,t}(d))^T$ and $\boldsymbol{a}_j(d) = (a_{1,j}(d), \ldots, a_{d,j}(d))^T$

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

The following three cases of different dependency structures are very helpful in understanding different effects that can occur and will be used as examples throughout the paper:

Case $\mathcal{C}$. 1 (Independent Components)**.** The components are independent, i.e. $\mathbf{a}_j = (0, \ldots, s_j, \ldots, 0)^T$ 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. $\mathbf{a}_1 = \mathbf{\Phi}_d = (\Phi_1, \ldots, \Phi_d)^T$, $\mathbf{a}_j = \mathbf{0}$ for $j \geqslant 2$. In this case,

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

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

Of course, many other dependency structures are possible, but these three cases give insight into the cases of no, complete and some dependency respectively. In particular, as the change is always one dimensional, taking a one dimensional form of dependency, as in cases $\mathcal{C}$.2 and $\mathcal{C}$.3, still allows somewhat general conclusions to be drawn.

2.2 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, i.e.

$$U_{d,T}(x) = \langle \mathbf{Z}_T(x), \mathbf{p}_d \rangle = \frac{1}{\sqrt{T}} \sum_{t=1}^{\lfloor Tx \rfloor} \left(\langle \mathbf{X}_d(t), \mathbf{p}_d \rangle - \frac{1}{T} \sum_{j=1}^{T} \langle \mathbf{X}_d(j), \mathbf{p}_d \rangle \right), \tag{2.1}$$

$$Z_{T,i}(x) = \frac{1}{T^{1/2}} \left(\sum_{t=1}^{\lfloor Tx \rfloor} X_{i,t} - \frac{\lfloor Tx \rfloor}{T} \sum_{t=1}^{T} X_{i,t} \right), \tag{2.2}$$

where $\mathbf{X}_d(t) = (X_{1,1}, \ldots, X_{d,T})^T$.

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

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

versus the alternative

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

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

$$\tau^2(\mathbf{p}_d) = \mathbf{p}_d^T \operatorname{var}(\mathbf{e}_1(d))\, \mathbf{p}_d. \tag{2.3}$$

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 (2.9) (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., 2014+].

For the epidemic change, typical test statistics are given by

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

or

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

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. This efficiency provides information as to how small changes can be in comparison to the dimension while still being detectable.

2.3 Asymptotic behavior of Change point tests based on projections

In this section, we derive the asymptotic behavior of change point tests based on projections under rather general assumptions. 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.

2.3.1 Null Asymptotics

In the 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 2.1. *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 $\mathbf{p}_d^T \operatorname{cov}(\mathbf{e}_1(d))\mathbf{p}_d \neq 0$ (almost surely), which means that the projected data is not degenerate with probability one.*

a) 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

$$\left\{ \frac{U_{d,T}(x)}{\tau(\mathbf{p}_d)} : 0 \leqslant x \leqslant 1 \,|\, \mathbf{p}_d \right\} \overset{D[0,1]}{\longrightarrow} \{B(x) : 0 \leqslant x \leqslant 1\} \qquad a.s., \tag{2.4}$$

where $B(\cdot)$ is a standard Brownian bridge.

b) For i.i.d. error sequences $\{\mathbf{e}_t(d) : t = 1, \ldots, d\}$, $\mathbf{e}_t(d) = (e_{1,t}(d), \ldots, e_{d,t}(d))^T$ 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*

$$\frac{\|\mathbf{p}_d\|_1^2}{\mathbf{p}_d^T \operatorname{cov}(\mathbf{e}_t)\mathbf{p}_d^T} = o(T^{1-2/\nu}) \quad a.s., \tag{2.5}$$

where $\|\mathbf{a}\|_1 = \sum_{j=1}^d |a_j|$, then (2.4) *holds.*

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

$$P\left(\left| \frac{\widehat{\tau}_{d,T}^2}{\tau^2(\mathbf{p}_d)} - 1 \right| > \epsilon \right) \to 0 \qquad a.s. \tag{2.6}$$

Assumption (2.5) 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 (2.5) is always bounded by $\sqrt{d}$ if $\mathbf{p}_d^T \operatorname{cov}(e)\mathbf{p}_d/\|p_d\|^2$ 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. In these cases projection methods hold the size asymptotically, no matter what the dependency structure between components is and without having to estimate this dependency structure.

This is in contrast to the multivariate statistic which suffers from considerable size distortions if this underlying covariance structure is estimated incorrectly. The estimation of the covariance structure is a difficult problem in higher dimensions in particular since an estimator for the inverse is needed with additional numerical problems arising. The problem becomes even harder if time series errors are present, in which case the long-run covariance rather than the covariance matrix needs to be estimated [Hörmann and

Kokoszka, 2010, Aston and Kirch, 2012b, Kirch et al., 2014+]. While the size of the projection procedure is unaffected by the underlying dependency across components, we will see in the next section that for optimal power we need not only to know the change $\mathbf{\Delta}_d$ but also the inverse of the covariance matrix. Nevertheless the power of projection procedures turns out to be more robust with respect to misspecification than a size-corrected panel statistic, that takes the full multivariate information into account.

The following lemma shows the consistency of two different estimators for $\tau(\mathbf{p}_d)$ under the null hypothesis. The second one is typically still consistent in the presence of one mean change which usually leads to a power improvement in 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 2.2. *Consider*

$$\widehat{\tau}^2_{1,d,T}(\mathbf{p}_d) = \frac{1}{T}\sum_{j=1}^{T}\left(\mathbf{p}_d^T\mathbf{e}_t(d) - \frac{1}{T}\sum_{i=1}^{T}\mathbf{p}_d^T\mathbf{e}_t(d)\right)^2 \tag{2.7}$$

as well as

$$\widehat{\tau}^2_{2,d,T}(\mathbf{p}_d) = \frac{1}{T}\left(\sum_{j=1}^{\widehat{k}_{d,T}}\left(\mathbf{p}_d^T\mathbf{e}_j(d) - \frac{1}{T}\sum_{i=1}^{\widehat{k}_{d,T}}\mathbf{p}_d^T\mathbf{e}_i(d)\right)^2 + \sum_{j=\widehat{k}_{d,T}+1}^{T}\left(\mathbf{p}_d^T\mathbf{e}_t(d) - \frac{1}{T}\sum_{i=\widehat{k}_{d,T}+1}^{T}\mathbf{p}_d^T\mathbf{e}_i(d)\right)^2\right), \tag{2.8}$$

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

a) *Under the assumptions of Theorem 2.1 a) both estimators* (2.7) *as well as* (2.8) *fulfill* (2.6).

b) *Under the assumptions of Theorem 2.1 b), then estimator* (2.7) *fulfills* (2.6) *under the assumption*

$$\frac{\|\mathbf{p}_d\|_1^2}{\mathbf{p}_d^T\operatorname{cov}(\mathbf{e}_t)\mathbf{p}_d^T} = o(T^{1-2/\min(\nu,4)}) \qquad a.s.,$$

while estimator (2.8) *fulfills it under the assumption*

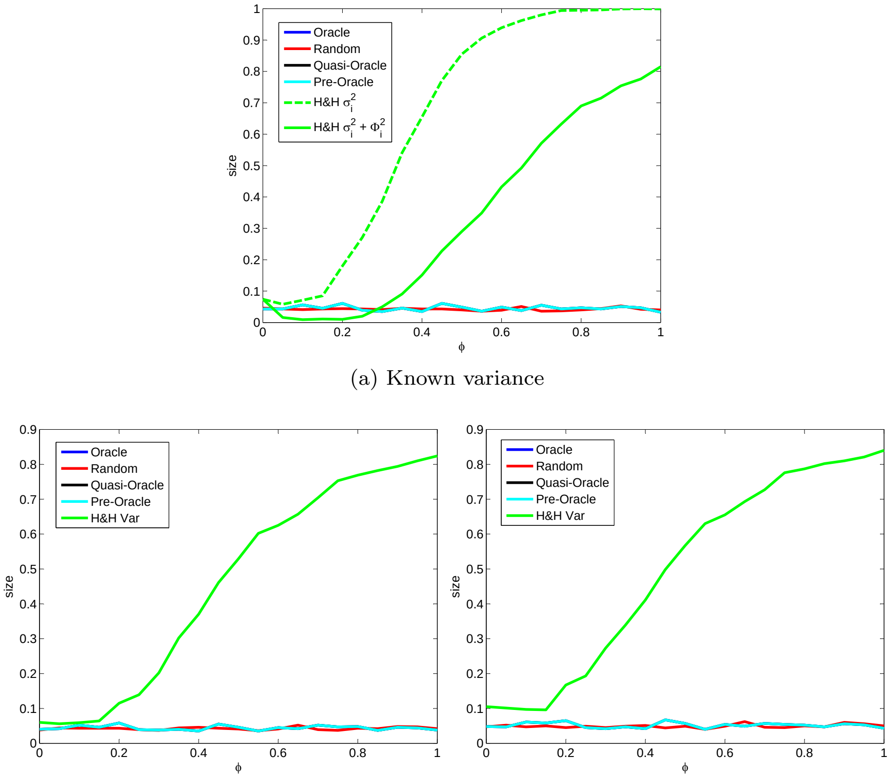
$$\frac{\|\mathbf{p}_d\|_1^2}{\mathbf{p}_d^T\operatorname{cov}(\mathbf{e}_t)\mathbf{p}_d^T} = o(T^{1-2/\min(\nu,4)}(\log T)^{-1}) \qquad a.s.,$$

The following theorem gives the null asymptotic for the simple CUSUM statistic for the at most one change, other statistics as given in Section 2.2 can be dealt with along the same lines.

Corollary 2.3. *Let the assumptions of Theorem 2.1 be fulfilled and $\widehat{\tau}(\mathbf{p}_d)$ fulfill* (2.6) *under the null hypothesis, then for all $x \in \mathbb{R}$ it holds under the null hypothesis*

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

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



(a) Known variance

(b) Estimated variance as in (2.7)

(c) Estimated variance as in (2.8)

Figure 2.1: 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 2.4 maintain the size of the tests. Those based on using the full information as described in Section 3 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).

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

$$\lim_{t\to 0} t^{\alpha} w(t) < \infty, \quad \lim_{t\to 1} (1-t)^{\alpha} w(t) \quad \textit{for some } 0 \leqslant \alpha < 1/2,$$
$$\sup_{\eta \leqslant t \leqslant 1-\eta} w(t) < \infty \qquad \textit{for all } 0 < \eta \leqslant \frac{1}{2}. \tag{2.9}$$

As can be seen in Figure 2.1, 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 and indeed the non-projection methods will be characterised in Section 2.4 below). The full tests, where size is not controlled, will be discussed in Section 3.

2.3.2 Asymptotic absolute high dimensional efficiency

As usual in statistics, large enough alternatives, i.e. large enough changes, will be detected by all statistics with the restriction that the change is not orthogonal to the projection vector for the projection method. In asymptotic theory this corresponds to fixed changes, where $\|\mathbf{\Delta}\| = c \neq 0$, for which the test has asymptotic power one.

To understand the small sample power of different statistics such asymptotics are therefore not suitable. Instead we consider the asymptotics for local or contiguous alternatives, meaning that we consider $\|\mathbf{\Delta}\| \to 0$ with such a rate that the test statistic has not power one asymptotically but is unbiased, i.e. the power is (strictly) larger than the size asymptotically. This notion is related to absolute relative efficiency: If the rate, with which changes can disappear is the same for all statistics of interest, and the limit distribution under the null is the same, then the additive shift in the limit distribution under those contiguous alternatives gives the absolute efficiency of the statistic (confer Noethers theorem for the case of a standard normal limit). In our setup, we obtain different contiguous rates for different statistics so that it is no longer the additive constant that is of interest but rather the rate with respect to d, that shows which statistic has the better power.

Theorem 2.4. *Denote*

$$\mathcal{E}_1^2(\mathbf{\Delta}_d, \mathbf{p}_d) := \frac{\|\mathbf{\Delta}_d\|^2 \|\mathbf{p}_d\|^2 \cos^2(\alpha_{\mathbf{\Delta}_d, \mathbf{p}_d})}{\tau^2(\mathbf{p}_d)} = \frac{|\langle \mathbf{\Delta}_d, \mathbf{p}_d \rangle|^2}{\tau^2(\mathbf{p}_d)}, \tag{2.10}$$

where $\tau^2(\mathbf{p}_d)$ is as in (2.3) and $\alpha_{\mathbf{u},\mathbf{v}}$ is the (smallest) angle between $\mathbf{u}$ and $\mathbf{v}$. Under the assumptions of Theorem 2.1 either a) or b) above on the errors respectively $\mathbf{p}_d$, it holds for the projection procedure:

a) If $\sqrt{T}\mathcal{E}_1(\mathbf{\Delta}_d, \mathbf{p}_d) \to \infty$ a.s., then

$$\left\{ \frac{U_{d,T}(x)}{\tau(\mathbf{p}_d)\, s_d \sqrt{T}\, \mathcal{E}_1(\mathbf{\Delta}_d, \mathbf{p}_d)} : 0 \leqslant x \leqslant 1 \,|\, \mathbf{p}_d \right\} \overset{D[0,1]}{\longrightarrow} \left\{ \int_0^x g(t)\, dt - x \int_0^1 g(t)\, dt : 0 \leqslant x \leqslant 1 \right\} \qquad a.s.,$$

where $s_d = sgn(\mathbf{\Delta}_d^T \mathbf{p}_d)$.

b) If $\sqrt{T}\mathcal{E}_1(\mathbf{\Delta}_d, \mathbf{p}_d) \to C_1 > 0$ a.s., then

$$\left\{ \frac{U_{d,T}(x)}{\tau(\mathbf{p}_d)} - s_d C_1 \left(\int_0^x g(t)\, dt - x \int_0^1 g(t)\, dt \right) : 0 \leqslant x \leqslant 1 \,|\, \mathbf{p}_d \right\} \overset{D[0,1]}{\longrightarrow} \{B(x) : 0 \leqslant x \leqslant 1\} \qquad a.s.,$$

where as in a) $s_d = sgn(\mathbf{\Delta}_d^T \mathbf{p}_d)$.

c) If $\sqrt{T}\mathcal{E}_1(\mathbf{\Delta}_d, \mathbf{p}_d) \to 0$ a.s., then

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

Corresponding assertions in a P-stochastic sense follow from the subsequence-principle if the assertion on the contiguous rate hold only in a P-stochastic sense.

We will call the rate $\mathcal{E}_1(\mathbf{\Delta}_d, \mathbf{p}_d)$ **high dimensional efficiency** as the above theorem shows that this is the rate with which changes can disappear such that appropriate tests have power strictly between the size and one. This can be seen via the term on the right-hand side of part a) of the theorem, which determines which type of statistic, e.g. maxima or sum type statistic with respective weights, has highest power for a given g. As this behavior is the same whether a projection or the full information is used, we concentrate on the high dimensional efficiency in this work. Similarly, one could consider local or contiguous changes, where not the size of the change $\|\mathbf{\Delta}_d\|$ disappears asymptotically but rather the duration of the change. We excluded this case by using the rescaled time version of the change by the function g. Particularly for multiple changes it makes sense to allow for the duration of the changes to get increasingly smaller asymptotically [Frick et al., 2014]. As the dependence on this type of change is the same for the projection, in both the multivariate as well as the panel statistic, as long as the same type of functional of the combined partial sum process is used, this does not give any additional insight in comparing the power of these statistics. However, some preliminary investigations suggest that while in the case of the second choice, using projections based on principle component analysis similar to Aston and Kirch [2012a] can be advantageous, this is not true for the setting discussed in this paper.

As an example, we state in a corollary that the weighted CUSUM-statistic has asymptotic power one for any non-constant g if the $T\mathcal{E}_1^2(\mathbf{\Delta}_d, \mathbf{p}_d) \to \infty$. If additionally there exists exactly one change, the corresponding change point estimator is consistent in rescaled time.

Corollary 2.5. *Let the assumptions of Theorem 2.4 a) hold.*

a) It holds for a weight function $w(\cdot)$ as in Corollary 2.3 and any $c > 0$

$$P\left(\max_{0\leqslant x\leqslant 1} w^2(k/T)U_{d,T}^2(k/T) > c \,|\, \mathbf{p}_d\right) \to 1 \quad a.s.,$$

if $w^2(x)\left(\int_0^x g(t)\,dt - x\int_0^1 g(t)\,dt\right)^2 \neq 0$. This shows in particular that for $w(x) > 0$ for $0 < x < 1$ any deviation from a stationary mean is detected by this statistic with asymptotic power one if $T\mathcal{E}_1^2(\mathbf{\Delta}, \mathbf{p}_d) \to \infty$.

b) Under the alternative of one abrupt change, i.e. $g(x) = 1_{\{x>\vartheta\}}$ for some $0 < \vartheta < 1$, the estimator

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

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 $w^2(k/T)U_{d,T}^2(k/T)$ 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 $0 \leqslant \beta < 1/2$.*

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 will also come from the situation where Σ is not invertible by considering case $\mathcal{C}.2$ above.

2.4 Oracle and random projections

In this section, we will further investigate the power gain obtained by projections – in particular, we will that the power depends only on the angle between the used projection and the change both properly scaled with the underlying covariance structure.

The highest power 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 $\Sigma^{-1/2}\mathbf{\Delta}_d$. In order to have a reasonable benchmark, we will compare this power to 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). 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

In Section 3 we will compare the power of the above projections with a procedure taking the full information into account. To this end we will use a panel data setting where $d \to \infty$ as $T \to \infty$ because then the full power information (including the information about d) is in the rates, whereas in the multivariate approach for d fixed it is harder to quantify as the information about d enters the asymptotic distribution in terms of different scales of the limit distribution. We will show that we lose an order $d^{1/4}$ in terms of high dimensional efficiency between the oracle and the full panel data statistic and another $d^{1/4}$ between the panel and the random projection.

2.4.1 Correctly scaled projections

The following proposition characterizes which projection yields an optimal high dimensional efficiency associated with the highest power.

Proposition 2.6. *If Σ is invertible, then*

$$\mathcal{E}_1(\mathbf{\Delta}, \mathbf{p}_d) = \|\Sigma^{-1/2}\mathbf{\Delta}_d\| \cos(\alpha_{\Sigma^{-1/2}\mathbf{\Delta}_d, \Sigma^{1/2}\mathbf{p}_d}). \tag{2.11}$$

Proposition 2.6 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 2.2).

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

Definition 2.1. *The projection $\mathbf{o} = \Sigma^{-1}\mathbf{\Delta}_d$ 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 2.6 the oracle choice leads to a high dimensional efficiency of $\mathcal{E}_1(\mathbf{\Delta}_d, \mathbf{o}) = \|\Sigma^{-1/2}\mathbf{\Delta}_d\|$.

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 $\Sigma^{-1/2}\mathbf{\Delta}$ and the (unit) projection vector, which is obviously maximized by a projection with $\Sigma^{-1/2}\mathbf{\Delta}/\|\Sigma^{-1/2}\mathbf{\Delta}\|$ which is equivalent to using $\mathbf{p}_d = \Sigma^{-1}\mathbf{\Delta}$ as a projection vector for the original non-standardized version.

So, if we know Σ and want to maximize 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 power 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

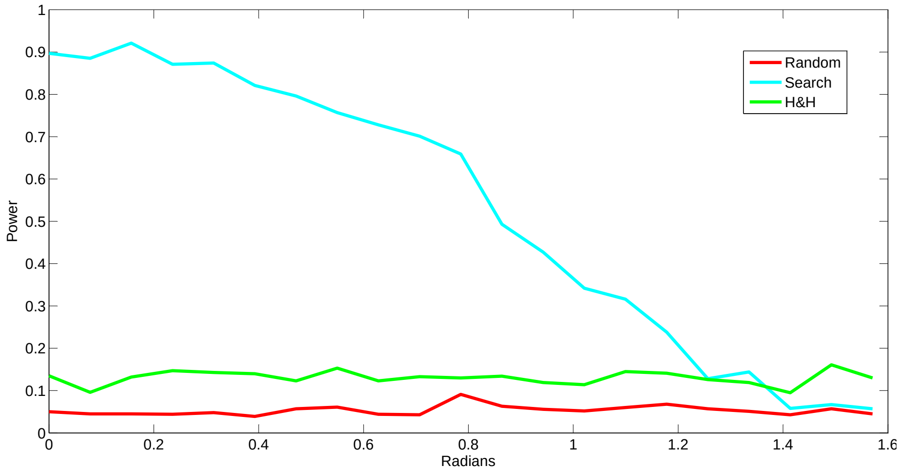


Figure 2.2: Power of tests as the angle between the search direction and the oracle increases. As can be seen, the search projection method decreases similarly to cosine of the angle, while the random projection and Horváth -Hušková tests as introduced in Section 3 are given for comparison. (Here $\Sigma_d = I_d$, $d = 200$, and $\mathbf{\Delta}_d = 0.05\ \mathbf{1_d}$, corresponding to Case $\mathcal{C}.1$).

$\mathbf{r}_d$ 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 $\mathbf{\Delta}_d$ is also transformed to $\Sigma^{-1/2}\mathbf{\Delta}_d$ 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 a full multivariate or panel statistic which will be investigated in detail in Section 3.

Such a random projection is the opposite of the oracle in the sense that absolutely no information about a possible change $\mathbf{\Delta}_d$ is used for the projection, while for the oracle the full and true information about $\mathbf{\Delta}_d$ is available and used.

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

Theorem 2.7. *Let the alternative hold, i.e. $\|\mathbf{\Delta}_d\| \neq 0$. 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 $c, C > 0$, such that*

$$P\left(c \leqslant \mathcal{E}_1^2(\mathbf{\Delta}_d, \mathbf{r}_{\Sigma,d})\,\frac{d}{\|\Sigma^{-1/2}\,\mathbf{\Delta}_d\|^2} \leqslant C\right) \geqslant 1-\epsilon.$$

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 $\mathbf{r}_d = (X_1, \ldots, X_d)^T / \|(X_1, \ldots, X_d)^T\|$ 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 2.6) it becomes apparent that we lose an order $\sqrt{d}$. In Section 3 we will see that the panel statistic taking the full multivariate information into account has a contiguous rate just between those two losing a power $d^{1/4}$ in comparison to the oracle but gaining $d^{1/4}$ in comparison to a scaled random projection. The finite sample nature of this can be clearly seen in Figure 2.3 where a change that can be detected for the oracle with constant power

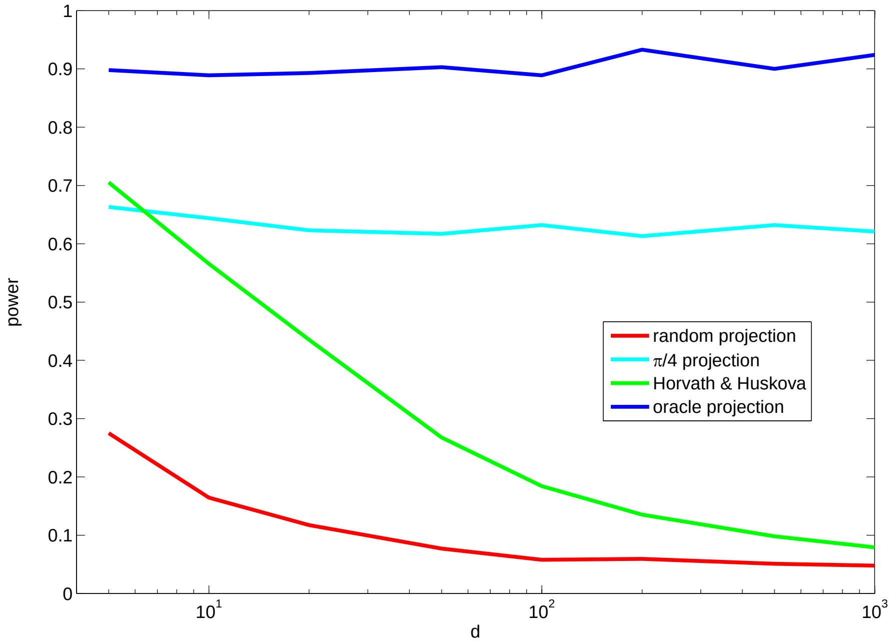


Figure 2.3: Power of the tests as d increases with a fixed sample size ($T = 100$). Here $\|\mathbf{\Delta}_d\| = \text{const.}$and $\Sigma_d = I_d$, i.e. $\|\Sigma^{-1/2}\mathbf{\Delta}_d\| = \text{const.}$, corresponding to Case $\mathcal{C}.1$. This gives roughly constant power for fixed angle projection tests (as $\|\mathbf{\Delta}_d\|$ is constant), while results in decreasing power for both the panel statistic test and random projections as predicted by theory.

as d increases rapidly loses power for the panel statistic as introduced in Section 3 as well as for the random projection.

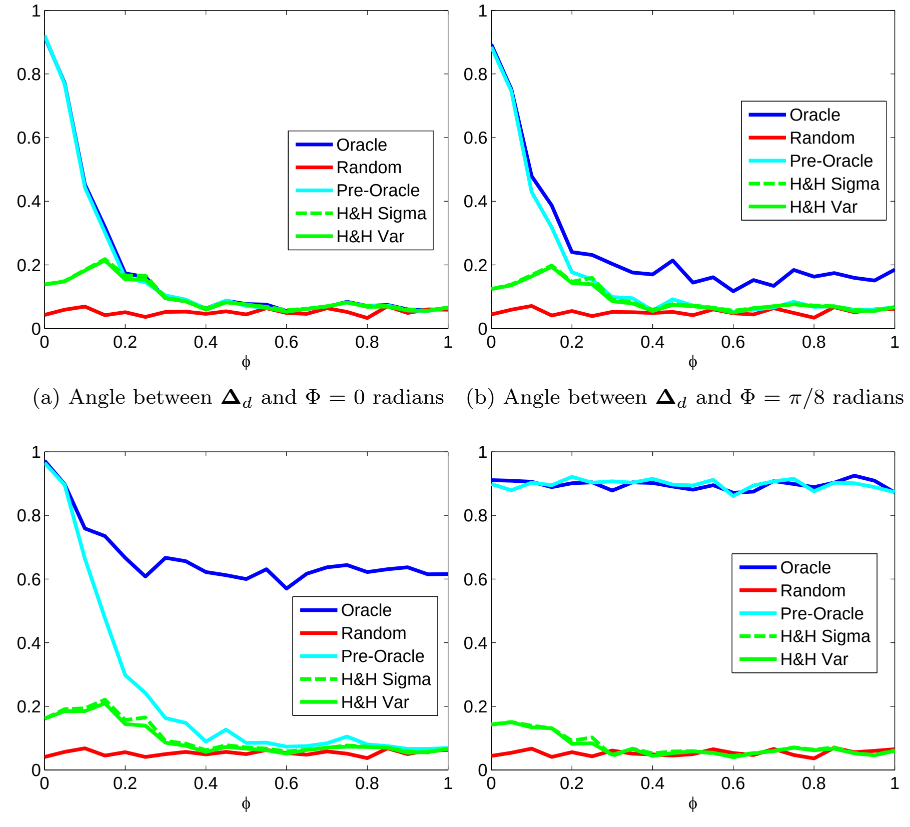
Let us now have a look at the situation if Σ is not invertible hence the above oracle does not exist. To this end, let us consider Case $\mathcal{C}.2$ above – other non-invertible dependent situations can essentially be viewed in a very similar fashion, but become a combination of the two scenarios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

Case $\mathcal{C}$. 2 (Fully dependent Components)**.** In this case $\Sigma = \mathbf{\Phi}_d\mathbf{\Phi}_d^T$ is a rank 1 matrix and not invertible. Consequently, the oracle as in Definition 2.1 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



(a) Angle between $\boldsymbol{\Delta}_d$ and $\Phi = 0$ radians (b) Angle between $\boldsymbol{\Delta}_d$ and $\Phi = \pi/8$ radians

(c) Angle between $\boldsymbol{\Delta}_d$ and $\Phi = \pi/4$ radians (d) Angle between $\boldsymbol{\Delta}_d$ and $\Phi = \pi/2$ radians

Figure 2.4: 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 2.4. 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 3 represent the 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 2.1. ($s_j = 1$, $\Phi_j = \phi$, $j = 1, \ldots, d$ with $d = 200$, $\|\boldsymbol{\Delta}_d\| = 0.05\sqrt{d}$, corresponding to Case $\mathcal{C}.3$), with ϕ as given on the x-axis.

weights according to $\boldsymbol{\Phi}_d$ i.e. they move in sync with those weights. If a change is proportional to $\boldsymbol{\Phi}_d$ 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 one-dimensional situation and no additional information about the change can be obtained from the multivariate situation. However, if it is not proportional to $\boldsymbol{\Phi}$, 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 2.4, 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).

2.4.2 Misscaled projections with respect to the covariance structure

The analysis in the previous section requires the knowledge or a precise estimate of the inverse 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 (2.8) as opposed to (2.7) to a multivariate situation as the estimation of a joint location already requires an initial weighting for the projection (or the multivariate statistic). Alternatively, component-wise 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 behavior of a **misscaled oracle** $\mathbf{o}_M = \mathbf{M}^{-1}\boldsymbol{\Delta}_d$ 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 model $\mathcal{A}.1$ is fulfilled.

The next theorem quantifies the high dimensional efficiency of the misscaled random projection by generalizing Theorem 2.7 to the misscaled situation.

Theorem 2.8. *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*

$$P\left(c \leqslant \mathcal{E}_1^2(\boldsymbol{\Delta}_d, \mathbf{r}_{\mathbf{M},d})\,\frac{tr\left(\mathbf{M}^{-1/2}\Sigma\mathbf{M}^{-1/2}\right)}{\|\mathbf{M}^{-1/2}\boldsymbol{\Delta}_d\|^2} \leqslant C\right) \geqslant 1-\epsilon,$$

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

$$\mathcal{E}_1^2(\boldsymbol{\Delta}_d, \mathbf{o}_M) \geqslant \frac{\|\mathbf{M}^{-1/2}\boldsymbol{\Delta}_d\|^2}{tr(\mathbf{M}^{-1/2}\Sigma\mathbf{M}^{-1/2})}$$

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 panel statistic discussed in Section 3 below.

Definition 2.2. *(i) The projection* ${}_p\mathbf{o} = \mathbf{\Delta}_d$ *is called* **pre-oracle**.

(ii) The projection ${}_q\mathbf{o} = \Lambda_d^{-1}\mathbf{\Delta}_d = (\delta_1/\sigma_1^2, \ldots, \delta_d/\sigma_d^2)^T$, $\Lambda_d = diag(\sigma_1^2, \ldots, \sigma_d^2)$ *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 are of the 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 quasi-oracle is even in the misspecified situation always better than an unscaled random projection.

Proposition 2.10. *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*

$$\frac{c^2}{C^2}\mathcal{E}_1^2(\mathbf{\Delta}_d, {}_q\mathbf{o}) \leqslant \mathcal{E}_1^2(\mathbf{\Delta}_d, {}_p\mathbf{o}) \leqslant \mathcal{E}_1^2(\mathbf{\Delta}, {}_q\mathbf{o}) = \|\Sigma^{-1/2}\mathbf{\Delta}_d\|^2.$$

b) Under Assumption $\mathcal{A}.1$*, it holds*

$$\mathcal{E}_1^2(\mathbf{\Delta}_d, {}_q\mathbf{o}) \geqslant \frac{c^2}{C^2}\,\frac{\|\mathbf{\Delta}_d\|^2}{tr(\Sigma)}.$$

We are now able to turn to our standard examples:

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 2.10, Theorem 2.7 and Theorem 2.8 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 next case shows that high dimensional efficiency of misscaled oracles can indeed become as bad as a random projection:

Case $\mathcal{C}$. 2 (Fully dependend components)**.** As already noted we have to distinguish two cases:

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

(ii) If $\mathbf{\Delta}_d$ is a multiple of $\mathbf{\Phi}_d$, 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 $\mathbf{\Phi}_d$ with probability 1.

The following case is essentially a mixture between the above two cases and a typical situation of how dependence between components can be introduced. Letting the weights of the common factor increase we get closer and closer to the fully dependent case $\mathcal{C}.2$. For this reason, in addition to the following, we also illustrate the behavior with finite sample simulations (see Figures 2.4 and 2.5).

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

$$\operatorname{tr}(\Sigma_3) = \sum_{j=1}^{d} s_j^2 + \sum_{j=1}^{d} \Phi_j^2. \tag{2.12}$$

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 competing procedures as in Section 3 below).

More precisely, in $\mathcal{C}.3$ 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 (2.10)

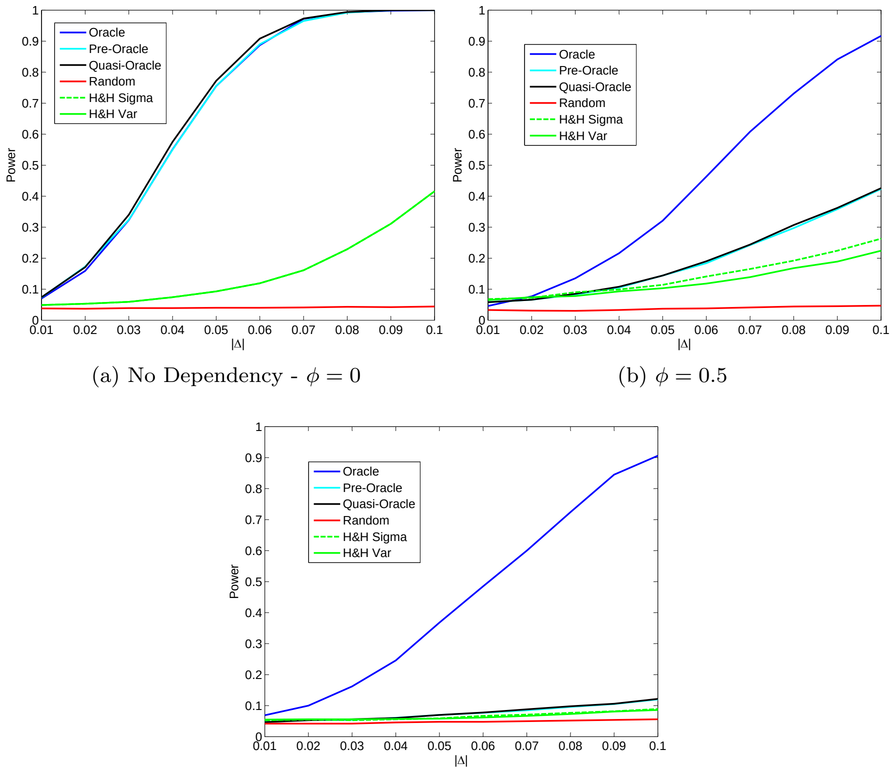
$$\mathcal{E}_1(\mathbf{\Delta}_d, {}_p\mathbf{o}) = \frac{\|\mathbf{\Delta}_d\|}{\sqrt{\sum_{i=1}^d s_i^2 \left(\frac{\delta_i}{\|\mathbf{\Delta}_d\|}\right)^2 + \|\mathbf{\Phi}_d\|^2}}.$$

The high dimensional efficiency for the unscaled random projection is given by (confer Theorem 2.8 and (2.12))

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

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 5)

$$\mathcal{E}_1(\mathbf{\Delta}_d, \mathbf{o}) = \frac{\|\mathbf{\Delta}_d\|}{\sqrt{1 + \sum_{j=1}^d \frac{\Phi_j^2}{s_j^2}}} \sqrt{\frac{\sum_{j=1}^d \frac{\delta_j^2}{s_j^2}}{\sum_{j=1}^d \delta_j^2}}. \tag{2.13}$$



(a) No Dependency - $\phi = 0$

(b) $\phi = 0.5$

(c) $\phi = 1$

Figure 2.5: 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$(\Phi, \mathbf{\Delta}_d)$=$\pi/4$, corresponding to Case $\mathcal{C}.3$, and size of change as given on the x-axis (multiplied by $\sqrt{d}$).

Figure 2.4 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

$$\mathcal{E}_1(\mathbf{\Delta}_d,\, {}_p\mathbf{o}) = \frac{\|\mathbf{\Delta}_d\|}{\sqrt{\sum_{i=1}^d s_i^2 \left(\frac{\delta_i}{\|\mathbf{\Delta}_d\|}\right)^2}},$$

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 2.4 and 2.5 for finite sample simulations).

The following interpretation also explains the above mathematical findings: 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.

We will see in Section 3 that the same statement holds true if we use a panel data (or multivariate) statistic (which can also be seen in Figures 2.4 and 2.5). As a matter of fact, the high dimensional efficiency of the misspecified panel statistic (i.e. where the statistic but not the critical values are constructed under the wrong assumption of independence between components) will be of the same order as a random projection for any choice $\mathbf{\Phi}_d$ with $\mathbf{\Phi}_d^T\mathbf{\Phi}_d \sim d$, irrespective of the direction of any change that might be present.

3 Power comparisons with change point tests for panel data

In this section, we will compare the power of the above projection tests with corresponding CUSUM tests that take the full multivariate information into account. First statistics of this type were developed for the multivariate setting with d fixed [Horváth et al., 1999]. The 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

$$V_d^M(x) = \mathbf{Z}_T(x)^T \boldsymbol{A} \mathbf{Z}_T(x)^T, \tag{3.1}$$

where $\mathbf{Z}_T(x) = (Z_{T,1}(x), \ldots, Z_{T,d}(x))^T$ is defined as in (2.2). 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. [2014+]. The choice $A = \Sigma^{-1}$ 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. For the panel data settings on the other hand, where $d \to \infty$, all the information about d is contained only in the rates rather than the limit distribution as in the previous sections. This makes the results interpretable in terms of the high dimensional efficiency. The panel null limit distribution differs from the one obtained for the projections but they are at least on the same scale, and not dependent on d nor the covariance structure Σ. 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).

We will now introduce the statistic for detecting changes in the mean introduced by Horváth and Hušková [2012]. Unlike in the above theory for projections, it is necessary to assume independence between components. Because the proofs are based on a central limit theorem across components, they cannot be generalized to uncorrelated (but dependent) data. 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.

We are interested in a comparison of the high dimensional efficiency for correctly specified covariance, i.e. $\boldsymbol{A} = \Sigma^{-1}$, in addition to a comparison in the misspecified case, $\boldsymbol{A} = \boldsymbol{M}^{-1}$. The latter has already been discussed by Horváth and Hušková [2012] to some extent. To be precise, a common factor is introduced as in $\mathcal{C}.3$ 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.

3.1 Asymptotic behavior for panel change point tests for independent panels

The above multivariate 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

$$V_{d,T}(x) = \frac{1}{\sqrt{d}} \sum_{i=1}^{d} \left(\frac{1}{\sigma_i^2} Z_{T,i}^2(x) - \frac{\lfloor Tx \rfloor (T - \lfloor Tx \rfloor)}{T^2} \right), \tag{3.2}$$

where $Z_{T,i}$ is as in (2.2) and $\sigma_i^2 = \operatorname{var} e_{i,1}$.

The following theorem gives a central limit theorem for this partial sum process (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 3.1. *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* $\limsup_{d\to\infty} \frac{1}{d} \sum_{i=1}^{d} \operatorname{E} |e_{i,t}|^{\nu} < \infty$ *for some* $\nu > 4$. *Furthermore, let* $\frac{d}{T^2} \to 0$. *Then, it holds under the null hypothesis of no change*

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

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

The following theorem derives the high dimensional efficiency in this setting.

Theorem 3.2. *Consider*

$$\mathcal{E}_2^2(\boldsymbol{\Delta}_d) = \frac{1}{\sqrt{d}} \, \|\Sigma^{-1/2} \boldsymbol{\Delta}_d\|^2.$$

Let the assumptions of Theorem 3.1 on the errors be fulfilled, which implies in particular that $\Sigma = diag(\sigma_1^2, \ldots, \sigma_d^2)$, *then the following assertions hold* *a) If $\sqrt{T}\,\mathcal{E}_2(\mathbf{\Delta}_d) \to \infty$, then*

$$\left\{\frac{V_{d,T}(x)}{T\,D^2_{d,T}(\mathbf{\Delta})} : 0 \leqslant x \leqslant 1 \,|\, \mathbf{p}_d\right\} \overset{D[0,1]}{\longrightarrow} \left\{\left(\int_0^x g(t)\,dt - x\int_0^1 g(t)\,dt\right)^2 : 0 \leqslant x \leqslant 1\right\}$$

b) If $\sqrt{T}\,\mathcal{E}_2(\mathbf{\Delta}_d) \to D_1$, then

$$V_{d,T}(x) \overset{D[0,1]}{\longrightarrow} \sqrt{2}(1-x)^2 W\left(\frac{x^2}{(1-x)^2}\right) + \left(\int_0^x g(t)\,dt - x\int_0^1 g(t)\,dt\right)^2 D_1^2.$$

c) If $\sqrt{T}\,\mathcal{E}_2(\mathbf{\Delta}_d) \to 0$, then

$$\{V_{d,T}(x) : 0 \leqslant x \leqslant 1 \,|\, \mathbf{p}_d\} \overset{D[0,1]}{\longrightarrow} \sqrt{2}(1-x)^2 W\left(\frac{x^2}{(1-x)^2}\right) \qquad a.s.$$

Equivalent assertions to Corollary 2.5 can be obtained analogously.

Comparing this high dimensional efficiency with the ones given in Theorem 2.4, Proposition 2.6 as well as Theorem 2.7, we note that the high dimensional efficiency of the full multivariate statistic is an order $d^{1/4}$ worse than for the oracle but a $d^{1/4}$ better than the scaled random projection (also see Figure 2.3). By Theorem 2.4 we also get an impression on how wrong our assumption on $\mathbf{\Delta}_d$ can be and still get a better efficiency than with the full multivariate information.We can see the finite sample nature of this phenomena in Figure 2.2.

3.2 Asymptotic behavior for panel 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.

Theorem 3.3. *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 (3.2), 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.

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

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

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

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

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

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

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

Because A_d in the above theorem is not feasible for estimation, this result cannot be used to construct test statistics with asymptotically correct size. On the other hand, it indicates that using the limit distribution from the previous section to derive critical values will result in asymptotically wrong sizes if a stronger contamination by a common factor is present. The simulations in Figure 2.1 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 2.1(a)), or whether a change is accounted for or not in the estimation (Figure 2.1(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 3.2, which is analogous to the proof of Theorem 3.2. Therefore, we will now concentrate on situation c) in the below proposition, which is the case where the noise coming from the common factor does not disappear asymptotically.

Theorem 3.4. *Consider the contiguous alternative sequence*

$$\mathcal{E}_3^2(\mathbf{\Delta}_d) = \frac{1}{A_d}\, \mathbf{\Delta}_d^T \; diag\left(\frac{1}{s_1^2+\Phi_1^2}, \dots, \frac{1}{s_d^2+\Phi_d^2}\right) \mathbf{\Delta}_d.$$

Let the assumptions of Theorem 3.3 on the errors be fulfilled and $A_d/\sqrt{d} \to \infty$, then the following assertions hold

a) If $\sqrt{T}\, \mathcal{E}_3(\mathbf{\Delta}_d) \to \infty$, then

$$\left\{\frac{V_{d,T}(x)}{T\,\mathcal{E}_3(\mathbf{\Delta})} : 0 \leqslant x \leqslant 1 \,|\, \mathbf{p}_d\right\} \overset{D[0,1]}{\longrightarrow} \left\{\left(\int_0^x g(t)\,dt - x\int_0^1 g(t)\,dt\right)^2 : 0 \leqslant x \leqslant 1\right\}$$

b) If $\sqrt{T}\, \mathcal{E}_3(\mathbf{\Delta}_d) \to D_2$, then

$$V_{d,T}(x) - D_2\,\tilde{R}_T(x) \overset{D[0,1]}{\longrightarrow} B^2(x) - x(1-x) + \left(\int_0^x g(t)\,dt - x\int_0^1 g(t)\,dt\right)^2 D_2^2,$$

where $\sup_{0\leqslant x\leqslant 1} |\tilde{R}_T(x)| = O_P(1)$.

c) If $\sqrt{T}\, \mathcal{E}_3(\mathbf{\Delta}_d) \to 0$, then

$$\{V_{d,T}(x) : 0 \leqslant x \leqslant 1 \,|\, \mathbf{p}_d\} \overset{D[0,1]}{\longrightarrow} B^2(x) - x(1-x) \qquad a.s.$$

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

Corollary 3.5. *Let the assumptions of Theorem 3.3 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 $\Sigma = diag(\sigma_1^2, \ldots, \sigma_d^2) + \mathbf{\Phi}\mathbf{\Phi}^T$, $\Lambda_d = diag(\sigma_1^2, \ldots, \sigma_d^2)$, *it holds*

$$\mathcal{E}_1^2(\mathbf{\Delta}_d, {}_q\mathbf{o}) \geqslant \frac{\mathbf{\Delta}_d^T \Lambda_d^{-1} \mathbf{\Delta}_d}{1 + A_d},$$

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

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.*

$$\mathcal{E}_3^2(\mathbf{\Delta}_d) = \frac{\mathbf{\Delta}_d^T \Lambda_d^{-1} \mathbf{\Delta}_d}{d} \left(A_d + o(1)\right).$$

In particular, for $A_d/d \to A > 0$ the misscaled panel statistic is always as bad as the random projection, this only holds for the misscaled (quasi-) projection if $\mathbf{\Delta}_d \sim \mathbf{\Phi}$. This effect can be clearly see in Figures 2.4 and 2.5, where in all cases H&H Sigma refers to the panel statistic using known variance, and H&H Var uses an estimated variance.

4 Conclusions

The primary aims of this paper were to introduce projection based statistics into the analysis of change points in high dimensions and compare and contrast these with the panel based statistics that are currently available. In summary, the following two assertions were proven: First, a suitable projection will substantially increase the power of detection but at the cost of a loss in power if the change is at a large angle away from the projection vector. Second, projections are more robust compared to the panel based statistic with respect to misspecification in the covariance structure both in terms of size and power.

One of the main tools for the comparison of the different tests in this paper is the use of relative efficiency rates, particularly the concept of high dimensional efficiency. This essentially allows a comparison of the magnitude of changes that can be detected asymptotically as the number of dimensions increases. All the tests in the paper are benchmarked against random projections. Because the space covered by far away angles increases rapidly with the dimension, the power of these becomes very poor in higher dimensions rendering random projections useless in practice for detecting change points. The panel statistic [Bai, 2010, Horváth and Hušková, 2012] test works well in situations where the panels are independent across dimension, and there is little to no information about the direction of the change. 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. Misspecification of the covariance structure can be problematic for all tests. However, if the direction of the likely change is known, then it is always preferable to use the corresponding projection (scaled with the assumed covariance structure), rather than either the panel statistic or a random projection, regardless of whether the covariance is misspecified or not.

This 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. Further many multiple change point procedures use binary segmentation or related methods to find the multiple change points, so much of the work here would apply equally in suitably defined sub intervals which are then assumed to contain at most one change. In addition, all the results here have been assessed with respect to choosing a single projection for the test which is optimal if the direction of the change is known. However, in some situations only qualitative information is known or several change scenarios are of interest. Then, it could be very beneficial to determine how best to combine this information into testing procedures based on several projections, where a standard subspace approach may not be ideal as the information about the likely direction of changes is lost. Finally, while the framework in this paper concentrates on tests with a given size, as soon as a-priori information is considered, then it is natural to ask whether related Bayesian approaches are of use, and indeed quantifying not only the a-priori direction of change, but also its uncertainty, prior to conducting the test is a natural line of further research.

5 Proofs

Proof of Theorem 2.1. 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 $\mathbf{p}_d = (p_1(d), \ldots, p_d(d))^T$:

$$\left\{ \frac{1}{\sqrt{T\tau^2(\mathbf{p}_d)}} \sum_{t=1}^{\lfloor Tx \rfloor} Y_{t,d} : 0 \leqslant x \leqslant 1 \,|\, \mathbf{p}_d \right\} \overset{D[0,1]}{\longrightarrow} \{W(x) : 0 \leqslant x \leqslant 1\} \qquad a.s., \tag{5.1}$$

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:

$$\mathrm{E}\left(\left| \frac{Y_{1,d}}{\tau(\mathbf{p}_d)} \right|^{\nu} \,|\, \mathbf{p}_d \right) = \mathrm{E}\left(\left| \frac{\mathbf{p}_d^T \mathbf{e}_1(d)}{\tau(\mathbf{p}_d)} \right|^{\nu} \,|\, \mathbf{p}_d \right) = o(T^{\nu/2-1}) \qquad a.s., \tag{5.2}$$

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

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

then the above Lyapunov condition is equal to

$$\mathrm{E}\left(\left| \tilde{\mathbf{p}}_d^T \mathbf{e}_1(d) \right|^{\nu} \,|\, \mathbf{p}_d \right) = o(T^{\nu/2-1}) \qquad a.s.$$

In the situation of a) $\operatorname{cov}\mathbf{e}_1(d) = \sum_{j\geqslant 1}\mathbf{a}_j(d)\mathbf{a}_j^T(d)$ and we get by the Rosenthal inequality (confer e.g. Lin and Bai [2010, 9.7c])

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

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 $\sum_{j=m}^{n}|\tilde{\mathbf{p}}_d^T\mathbf{a}_j(d)|^2\leqslant\tilde{\mathbf{p}}_d^T\operatorname{cov}\mathbf{e}_1(d)\,\tilde{\mathbf{p}}_d\leqslant 1$, hence also $|\tilde{\mathbf{p}}_d^T\mathbf{a}_j(d)|^{\nu}\leqslant|\tilde{\mathbf{p}}_d^T\mathbf{a}_j(d)|^2$ and $\sum_{j=m}^{n}|\tilde{\mathbf{p}}_d^T\mathbf{a}_j(d)|^{\nu}\leqslant 1$.

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

$$\mathrm{E}\left(\left|\tilde{\mathbf{p}}_d^T\mathbf{e}_1(d)\right|^{\nu}\mid\mathbf{p}_d\right) = O(1) = o(T^{\nu/2-1})\qquad a.s. \tag{5.3}$$

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

$$\begin{aligned}
&\mathrm{E}\left(\left|\tilde{\mathbf{p}}_d^T\mathbf{e}_1(d)\right|^{\nu}\mid\mathbf{p}_d\right) = \|\tilde{\mathbf{p}}_d\|_1^{\nu}\ \mathrm{E}\left(\left(\sum_{i=1}^{d}\frac{|\tilde{p}_{i,d}|}{\|\tilde{\mathbf{p}}_d\|_1}|e_{i,1}(d)|\right)^{\nu}\mid\mathbf{p}_d\right)\\
&\leqslant\|\tilde{\mathbf{p}}_d\|_1^{\nu}\sum_{i=1}^{d}\frac{|\tilde{p}_{i,d}|}{\|\tilde{\mathbf{p}}_d\|_1}\ \mathrm{E}\,|e_{i,1}(d)|^{\nu}\leqslant C\left(\frac{\|\mathbf{p}_d\|_1}{\sqrt{\mathbf{p}_d^T\operatorname{cov}(\mathbf{e}_1(d))\mathbf{p}_d^T}}\right)^{\nu} = o(T^{\nu/2-1})\qquad a.s.
\end{aligned} \tag{5.4}$$

■

Proof of Lemma 2.2. With the notation of the proof of Theorem 2.1 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 [2010, 9.3 and 9.4]) for some constant $C>0$, which may differ from line to line,

$$\begin{aligned}
&\mathrm{E}_{\mathbf{p}_d}\left|\sum_{j=a+1}^{b}\left(\left(\tilde{\mathbf{p}}_d^T\mathbf{e}_j(d)\right)^2-1\right)\right|^{\nu/2}\leqslant C(b-a)^{\max(1,\nu/4)}\ \mathrm{E}_{\mathbf{p}_d}\left|\left(\tilde{\mathbf{p}}_d^T\mathbf{e}_1(d)\right)^2-1\right|^{\nu/2}\\
&\leqslant C\,(b-a)^{\max(1,\nu/4)}\ \max\left(1,\mathrm{E}_{\mathbf{p}_d}\left|\tilde{\mathbf{p}}_d^T\mathbf{e}_1(d)\right|^{\nu}\right)\\
&\leqslant\begin{cases} C(b-a)^{\max(1,\nu/4)} & a.s., \qquad\text{in a)},\\ C(b-a)^{\max(1,\nu/4)}\max\left(1,\left(\frac{\|\mathbf{p}_d\|_1}{\sqrt{\mathbf{p}_d^T\operatorname{cov}\mathbf{e}_1(d)\mathbf{p}_d}}\right)^{\nu}\right), & \text{in b)},\end{cases}
\end{aligned} \tag{5.5}$$

by (5.3) resp. (5.4), where $\mathrm{E}_{\mathbf{p}_d}$ denotes the conditional expectation given $\mathbf{p}_d$. An application of the Markov-inequality now yields for any $\epsilon > 0$

$$
\begin{aligned}
&P\left(\frac{1}{T}\left|\sum_{j=1}^{T}\left(\left(\tilde{\mathbf{p}}_d^T\mathbf{e}_j(d)\right)^2-1\right)\right| \geqslant \epsilon \,\Big|\, \mathbf{p}_d\right)\\
&\leqslant \begin{cases} \frac{C}{\epsilon^{\nu/2}}T^{-\nu/2+\max(1,\nu/4)} \qquad a.s., & \text{in a)},\\ \frac{C}{\epsilon^{\nu/2}}T^{-\nu/2+\max(1,\nu/4)}o(T^{\nu/2-\nu/\min(\nu,4)}) \qquad a.s., & \text{in b)},\end{cases}\\
&\to 0 \qquad a.s.
\end{aligned}
$$

Similar arguments yield

$$
P\left(\frac{1}{T}\left|\sum_{j=1}^{T}\tilde{\mathbf{p}}_d^T\mathbf{e}_j(d)\right| \geqslant \epsilon \,\Big|\, \mathbf{p}_d\right) \to 0 \qquad a.s.
$$

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

From (5.5) it follows by Theorem B.1 resp. B.4 in Kirch [2006]

$$
\begin{aligned}
&\mathrm{E}_{\mathbf{p}_d}\max_{1\leqslant k\leqslant T}\left|\sum_{j=1}^{k}\left(\left(\tilde{\mathbf{p}}_d^T\mathbf{e}_j(d)\right)^2-1\right)\right|^{\nu/2}\\
&\leqslant \begin{cases} CT^{\max(1,\nu/4)}(\log T)^{\frac{(4-\nu)_+\nu}{2(4-\nu)}} \qquad a.s., & \text{in a)},\\ CT^{\max(1,\nu/4)}(\log T)^{\frac{(4-\nu)_+\nu}{2(4-\nu)}}\max\left(1,\left(\frac{\|\mathbf{p}_d\|_1}{\sqrt{\mathbf{p}_d^T\operatorname{cov}\mathbf{e}_1(d)\mathbf{p}_d}}\right)^{\nu}\right), & \text{in b)},\end{cases}\\
&\to 0 \qquad a.s.
\end{aligned}
$$

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

$$
P\left(\max_{1\leqslant k\leqslant T}\frac{1}{T}\left|\sum_{j=1}^{k}\left(((\tilde{\mathbf{p}}_d^T\mathbf{e}_j(d)\right)^2-1\right)\right| \geqslant \epsilon \,\Big|\, \mathbf{p}_d\right) \to 0 \qquad a.s.
$$

By the independence across time it holds

$$
\left\{\sum_{j=k+1}^{T}\left(\left(\tilde{\mathbf{p}}_d^T\mathbf{e}_j(d)\right)^2-1\right) : 1\leqslant k\leqslant T\right\} \overset{\mathcal{L}}{=} \left\{\sum_{j=1}^{T-k}\left(\left(\tilde{\mathbf{p}}_d^T\mathbf{e}_j(d)\right)^2-1\right) : 1\leqslant k\leqslant T\right\},
$$

which implies

$$
P\left(\max_{1\leqslant k\leqslant T}\frac{1}{T}\left|\sum_{j=k+1}^{T}\left(\left(\tilde{\mathbf{p}}_d^T\mathbf{e}_j(d)\right)^2-1\right)\right| \geqslant \epsilon \,\Big|\, \mathbf{p}_d\right) \to 0 \qquad a.s.
$$

Similar assertions can be obtained along the same lines for $\max_{1\leqslant k\leqslant T}\frac{1}{T}\left|\sum_{j=1}^{k}\tilde{\mathbf{p}}_d^T\mathbf{e}_j(d)\right|$ as well as $\max_{1\leqslant k\leqslant T}\frac{1}{T}\left|\sum_{j=k+1}^{T}\tilde{\mathbf{p}}_d^T\mathbf{e}_j(d)\right|$, which imply the assertion for $\widehat{\tau}^2_{2,d,T}(\mathbf{p}_d)$. ■

Proof of Corollary 2.3. By an application of the continuous mapping theorem and Theorem 2.1 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 $\epsilon > 0$

$$
\begin{aligned}
&P\left(\max_{1\leqslant k\leqslant \tau T} w(k/T)\left|\sum_{t=1}^{k}\tilde{\mathbf{p}}_d^T\mathbf{e}_t(d)\right| \geqslant \epsilon \,\Big|\, \mathbf{p}_d\right) \to 0 \quad a.s.\\
&P\left(\max_{(1-\tau)T\leqslant k\leqslant} w(k/T)\left|\sum_{t=k+1}^{T}\tilde{\mathbf{p}}_d^T\mathbf{e}_t(d)\right| \geqslant \epsilon \,\Big|\, \mathbf{p}_d\right) \to 0 \quad a.s.
\end{aligned}
$$

as $\tau \to 0$ uniformly in T, where the notation of the proof of Theorem 2.1 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 2.4. Under alternatives it holds

$$\frac{U_{d,T}(x)}{\tau(\mathbf{p}_d)} = \frac{U_{d,T}(x;\mathbf{e})}{\tau(\mathbf{p}_d)} + \operatorname{sgn}(\boldsymbol{\Delta}_d^T\mathbf{p}_d)\,\sqrt{T}\,\mathcal{E}_1(\boldsymbol{\Delta}_d,\mathbf{p}_d)\left(\frac{1}{T}\sum_{i=1}^{\lfloor Tx\rfloor} g(i/T) - \frac{\lfloor Tx\rfloor}{T^2}\sum_{j=1}^{T} g(j/T)\right),$$

where $U_{d,T}(x;\mathbf{e})$ is the corresponding functional of the error process. By Theorem 2.1 it holds

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

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

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

Putting everything together yields the assertions of the theorem. ■

Proof of Corollary 2.5. An application of Theorem 2.4 a) yields for any $\tau > 0$

$$\begin{aligned}&\max_{\tau\leqslant k/T\leqslant 1-\tau} w^2(k/T)\frac{U_{d,T}^2(k/T)}{\tau^2(\mathbf{p}_d)}\\ &= T\,\mathcal{E}_1^2(\boldsymbol{\Delta}_d,\mathbf{p}_d)\left(\sup_{\tau\leqslant x\leqslant 1-\tau} w^2(x)\left(\int_0^x g(t)\,dt - x\int_0^1 g(t)\,dt\right)^2 + o_{P_{\mathbf{p}_d}}(1)\right) \qquad a.s.,\end{aligned}$$

where $P_{\mathbf{p}_d}$ denotes the conditional probability given $\mathbf{p}_d$. This implies assertion a), 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. In the situation of b) it follows similarly (where the uniformity at 0 and 1 follows by the assumptions on the rate of divergence for $w(\cdot)$ at 0 or 1)

$$\sup_{0<x<1} w^2(x)\left|\frac{U_{d,T}^2(x)}{\tau^2(\mathbf{p}_d)T\,\mathcal{E}_1^2(\boldsymbol{\Delta}_d,\mathbf{p}_d)} - ((x-\vartheta)_+ - x(1-\vartheta))^2\right| = o_{P_{\mathbf{p}_d}}(1) \qquad a.s.,$$

which implies assertion b) by standard arguments on noting that

$$\widehat{\vartheta}_T = \arg\max_{0\leqslant x\leqslant 1} w^2(x)\,\frac{U_{d,T}^2(x)}{\tau^2(\mathbf{p}_d)T\,\mathcal{E}_1^2(\boldsymbol{\Delta}_d,\mathbf{p}_d)}, \quad \vartheta = \arg\max_{0\leqslant x\leqslant 1} w^2(x)\,((x-\vartheta)_+ - x(1-\vartheta))^2.$$

■

Proof of Proposition 2.6. The assertion follows from

$$\begin{aligned}&\tau^2(\mathbf{p}_d) = \mathbf{p}_d^T\Sigma\mathbf{p}_d = \|\Sigma^{1/2}\mathbf{p}_d\|^2,\\ &|\langle\boldsymbol{\Delta}_d,\mathbf{p}_d\rangle| = (\Sigma^{-1/2}\boldsymbol{\Delta}_d)^T(\Sigma^{1/2}\mathbf{p}_d) = \|\Sigma^{-1/2}\boldsymbol{\Delta}_d\|\,\|\Sigma^{1/2}\mathbf{p}_d\|\,\cos(\alpha_{\Sigma^{-1/2}\boldsymbol{\Delta}_d,\Sigma^{1/2}\mathbf{p}_d}).\end{aligned}$$

■

Proof of Theorem 2.7. Let $\boldsymbol{X}_d = (X_1,\ldots,X_d)^T$ be N$(0,I_d)$, then by Marsaglia [1972] it holds $\mathbf{r}_d \overset{\mathcal{L}}{=} (X_1,\ldots,X_d)^T/\|(X_1,\ldots,X_d)^T\|$ and it follows by (2.10)

$$\mathcal{E}_1^2(\boldsymbol{\Delta}_d,\Sigma^{-1/2}\mathbf{r}_d)\,\frac{d}{\|\Sigma^{-1/2}\boldsymbol{\Delta}_d\|^2} \overset{\mathcal{L}}{=} \frac{\left|\frac{\boldsymbol{X}_d^T\Sigma^{-1/2}\,\boldsymbol{\Delta}_d}{\|\Sigma^{-1/2}\boldsymbol{\Delta}_d\|}\right|^2}{\frac{\boldsymbol{X}_d^T\boldsymbol{X}_d}{\mathrm{E}\,\boldsymbol{X}_d^T\boldsymbol{X}_d}}$$

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

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

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

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

By integration by parts we get $\mathrm{E}\left(\boldsymbol{X}_d^T \boldsymbol{X}_d\right)^{-1} \leqslant 2/d$ 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

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

completing the proof of the theorem by standard arguments. ■

Proof of Theorem 2.8. Let $\boldsymbol{X}_d = (X_1, \ldots, X_d)^T$ be N(0,I_d), then as in the proof of Theorem 2.7 it holds

$$\mathcal{E}_1^2(\boldsymbol{\Delta}, \mathbf{M}^{-1/2}\mathbf{r}_d)\, \frac{\mathrm{tr}(\mathbf{M}^{-1/2}\Sigma\mathbf{M}^{-1/2})}{\|\mathbf{M}^{-1/2}\boldsymbol{\Delta}_d\|^2} \overset{\mathcal{L}}{=} \frac{\left| \frac{\boldsymbol{X}_d^T \mathbf{M}^{-1/2}\, \boldsymbol{\Delta}_d}{\|\mathbf{M}^{-1/2}\boldsymbol{\Delta}_d\|} \right|^2}{\frac{\boldsymbol{X}_d^T \mathbf{M}^{-1/2}\Sigma\mathbf{M}^{-1/2}\boldsymbol{X}_d}{\mathrm{tr}(\mathbf{M}^{-1/2}\Sigma\mathbf{M}^{-1/2})}}.$$

The proof of the lower bound is analogous to the proof of Theorem 2.7 by noting that ($A = \mathbf{M}^{-1/2}\Sigma\mathbf{M}^{-1/2}$)

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

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

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

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

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

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

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

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

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

completing the proof. ■

Proof of Theorem 2.9. By the Cauchy-Schwarz inequality

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

which implies the assertion by (2.10). ■

Proof of Proposition 2.10. Assertion a) follows from

$$\begin{aligned}|\langle\boldsymbol{\Delta}_d,\ {}_p\mathbf{o}\rangle|^2 &= \left(\sum_{i=1}^d\frac{\delta_{i,T}^2}{\sigma_i^2}\,\sigma_i^2\right)^2 \geqslant c^2\left(\sum_{i=1}^d\frac{\delta_{i,T}^2}{\sigma_i^2}\right)^2 = c^2\,|\langle\boldsymbol{\Delta}_d,\ {}_q\mathbf{o}\rangle|^2\,,\\ \tau^2({}_p\mathbf{o}) &= {}_p\mathbf{o}^T\,\Sigma\,{}_p\mathbf{o} = \sum_{i=1}^d\frac{\delta_{i,T}^2}{\sigma_i^2}\,\sigma_i^4 \leqslant C^2\,|\langle\boldsymbol{\Delta}_d,\ {}_q\mathbf{o}\rangle|\,.\end{aligned}$$

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

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

This implies assertion b) by (2.10) on noting that

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

■

Proof of Equation 2.13. By Proposition 2.6 it holds for $\boldsymbol{\Delta}_d = k\,\boldsymbol{\Phi}_d$

$$\mathcal{E}_1^2(\boldsymbol{\Delta}_d,\mathbf{o}) = \|\Sigma^{-1/2}\boldsymbol{\Delta}_d\|^2 = \boldsymbol{\Delta}_d^T(\boldsymbol{D}+\boldsymbol{\Phi}_d\boldsymbol{\Phi}_d^T)^{-1}\boldsymbol{\Delta}_d,$$

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

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

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 $I_d + (\boldsymbol{D}^{-1/2}\boldsymbol{\Phi}_d)(\boldsymbol{D}^{-1/2}\boldsymbol{\Phi}_d)^T$ with eigenvalue $1 + (\boldsymbol{D}^{-1/2}\boldsymbol{\Phi}_d)^T\boldsymbol{D}^{-1/2}\boldsymbol{\Phi}_d$ hence also for the inverse of the matrix with inverse eigenvalue. ■

Proof of Theorem 3.2. Similarly as in the proof of Theorem 2.4 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

$$V_{d,T}(x) = V_{d,T}(x;\mathbf{e}) + T\,\mathcal{E}_2^2(\mathbf{\Delta}_d)\left(\int_0^x g(t)\,dt - x\int_0^1 g(t)\,dt + o(1)\right) + R_T(x),$$

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

$$P\left(\sup_{0\leqslant x\leqslant 1}|R_T(x)| \geqslant c\right) = O\left(1\right)\frac{1}{c^2}T\frac{1}{d}\sum_{i=1}^d\frac{\delta_i^2}{\sigma_i^2} = O_P(1)\frac{1}{c^2\sqrt{d}}T\,\mathcal{E}_2(\mathbf{\Delta}_d).$$

From this the assertions follow by an application of Theorem 3.1. ■

Proof of Theorem 3.3. 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

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

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

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

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 3.4. The proof is analogous to the one of Theorem 3.2 on noting that $\mathcal{E}_3^2(\mathbf{\Delta}_d) = \frac{\sqrt{d}}{A_d}\mathcal{E}_2^2(\mathbf{\Delta}_d)$ and $\sigma_i^2 = s_i^2 + \Phi_i^2$ by using Theorem 3.3 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 3.2 and is of order $O_P\left(\sqrt{\frac{1}{A_d}T\mathcal{E}_3(\mathbf{\Delta}_d)}\right)$, while for the second summand we get by an application of the Cauchy-Schwarz-inequality

$$\sup_{0\leqslant x\leqslant 1}\left|\frac{1}{A_d}\sum_{i=1}^{d}\frac{\delta_i\phi_i}{\sigma_i^2}\left(\sum_{t=1}^{\lfloor Tx\rfloor}\eta_{d+1,t} - \frac{\lfloor Tx\rfloor}{T}\sum_{t=1}^{T}\eta_{d+1,t}\right)\right| = O_P(\sqrt{T})\sqrt{\frac{\sum_{i=1}^{d}\frac{\delta_i^2}{\sigma_i^2}}{A_d}}$$
$$= O\left(\sqrt{T\,\mathcal{E}_3^2(\mathbf{\Delta}_d)}\right).$$

■

Proof of Corollary 3.5. By an application of the Cauchy-Schwarz inequality it holds

$$\mathbf{\Delta}_d^T\Lambda_d^{-1}\Sigma\Lambda_d^{-1}\mathbf{\Delta}_d = \sum_{i=1}^{d}\delta_{i,T}^2\frac{s_i^2}{(s_i^2+\Phi_i^2)^2} + \left(\sum_{i=1}^{d}\frac{\delta_{i,T}\Phi_i}{s_i^2+\Phi_i^2}\right)^2$$
$$\leqslant \sum_{i=1}^{d}\frac{\delta_{i,T}^2}{\sigma_i^2}\left(1+\sum_{i=1}^{d}\frac{\Phi_i^2}{\sigma_i^2}\right) = \mathbf{\Delta}_d^T\Lambda_d^{-1}\mathbf{\Delta}_d\,(1+A_d),$$

which implies assertion a) on noting that

$$\mathcal{E}_1^2(\mathbf{\Delta}_d,\, {}_q\mathbf{o}) = \frac{(\mathbf{\Delta}_d^T\Lambda_d^{-1}\mathbf{\Delta}_d)^2}{\mathbf{\Delta}_d^T\Lambda_d^{-1}\Sigma\Lambda_d^{-1}\mathbf{\Delta}_d}.$$

b) This follows immediately from Theorem 2.8 since by $0 < c \leqslant s_j^2 \leqslant C < \infty$ as well as as $\Phi_i^2 \leqslant C$, it follows that

$$\|\mathbf{\Delta}_d\|^2 \sim \mathbf{\Delta}_d^T\,\mathrm{diag}\left(\frac{1}{s_1^2+\Phi_1^2},\ldots,\frac{1}{s_d^2+\Phi_d^2}\right)\mathbf{\Delta}_d.$$

■

References

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

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.

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

References

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.

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.

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

R. Baraniuk, M. Davenport, R. DeVore, and M. Wakin. A Simple Proof of the Restricted Isometry Property for Random Matrices. *Constructive Approximation*, 28: 253–263, 2008.

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.

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

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

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.

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)*, in press, 2014+.

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

R. J. Durrant and A. Kabán. Compressed Fisher linear discriminant analysis: Classification of randomly projected data. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1119–1128. ACM, 2010.

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.

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.

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.

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

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

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

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

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.

W. B. Johnson and J. Lindenstrauss. Extensions of lipschitz mappings into a hilbert space. *Contemporary mathematics*, 26:189–206, 1984.

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

C. Kirch and J. Tadjuidje Kamgaing. Monitoring time series based on estimating functions. Technical report, Technische Universität Kaiserslautern, Fachbereich Mathematik, 2014a.

C. Kirch and J. Tajduidje 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*, 2014b.

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*, in press, 2014+.

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

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

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.

G. Minas, J. A. D. Aston, and N. Stallard. Adaptive Multivariate Global Testing. *Journal of the American Statistical Association*, 109:613–623, 2014.

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.

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

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

R. Srivastava, P. Li, and D. Ruppert. RAPTT: An Exact Two-Sample Test in High Dimensions Using Random Projections. *ArXiv e-prints*, 1405.1792, 2014.

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