

MINIMAX RATES IN SPARSE, HIGH-DIMENSIONAL CHANGEPOINT DETECTION

BY HAORYANG LIU*, CHAO GAO* AND RICHARD J. SAMWORTH†

*Department of Statistics
George Herbert Jones Laboratory
5747 S. Ellis Avenue
Chicago, IL 60637
The University of Chicago
hyliu@uchicago.edu*

*chaogao@uchicago.edu
URL: <https://stat.uchicago.edu/~chaogao>*

*Statistical Laboratory
Wilberforce Road
Cambridge, CB3 0WB
United Kingdom
r.samworth@statslab.cam.ac.uk
URL: <http://www.statslab.cam.ac.uk/~rjs57>*

We study the detection of a sparse change in a high-dimensional mean vector as a minimax testing problem. Our first main contribution is to derive the exact minimax testing rate across all parameter regimes for n independent, p -variate Gaussian observations. This rate exhibits a phase transition when the sparsity level is of order $\sqrt{p \log \log(8n)}$ and has a very delicate dependence on the sample size: in a certain sparsity regime it involves a triple iterated logarithmic factor in n . Further, in a dense asymptotic regime, we identify the sharp leading constant, while in the corresponding sparse asymptotic regime, this constant is determined to within a factor of $\sqrt{2}$. Extensions that cover spatial and temporal dependence, primarily in the dense case, are also provided.

1. Introduction. The problem of changepoint detection has a long history (e.g. [Page, 1955](#)), but has undergone a remarkable renaissance over the last 5–10 years. This has been driven in part because these days sensors and other devices collect and store data on unprecedented scales, often at high frequency, which has placed a greater emphasis on the running time of changepoint detection algorithms ([Killick, Fearnhead and Eckley, 2012](#); [Frick, Munk and Sieling, 2014](#)). But it is also because nowadays these data streams are often monitored simultaneously as a multidimensional process, with a changepoint in a subset of the coordinates representing an event of interest. Examples include distributed denial of service attacks as detected by changes in traffic at certain internet routers ([Peng et al., 2004](#)) and changes in a subset of blood oxygen level dependent contrast in a subset of voxels in fMRI studies ([Aston and Kirch, 2012](#)). Away from time series contexts, the problem is also of interest, for instance in the detection of chromosomal copy number abnormality ([Zhang et al., 2010](#); [Wang and Samworth, 2018](#)). Key to the success of changepoint detection methods in such settings is the ability to borrow strength across the different coordinates, in order to be able to detect much smaller changes than would be possible through observation of any single coordinate in isolation.

*Research supported in part by NSF grant DMS-1712957 and NSF CAREER Award DMS-1847590

†Research supported by EPSRC Fellowship EP/P031447/1 and EPSRC Programme Grant EP/N031938/1.
MSC 2010 subject classifications: 62C20.

Keywords and phrases: minimax detection boundary, iterated logarithm, time series.

We initially consider a simple model where, for some $n \geq 2$, we observe a $p \times n$ matrix X that can be written as

$$(1) \quad X = \theta + E,$$

where $\theta \in \mathbb{R}^{p \times n}$ is deterministic and the entries of E are independent $N(0, 1)$ random variables. We wish to test the null hypothesis that the columns of θ are constant against the alternative that there exists a time $t_0 \in \{1, \dots, n-1\}$ at which these mean vectors change, in at most s out of the p coordinates. The difficulty of this problem is governed by a signal strength parameter ρ^2 that measures the squared Euclidean norm of the difference between the mean vectors, rescaled by $\frac{t_0(n-t_0)}{n}$; this latter quantity can be interpreted as an effective sample size. The goal is to identify the minimax testing rate in ρ^2 as a function of the problem parameters p , n and s , and we denote this by $\rho^*(p, n, s)^2$; this is the signal strength at which we can find a test making the sum of the Type I and Type II error probabilities arbitrarily small by choosing ρ^2 to be an appropriately large multiple of $\rho^*(p, n, s)^2$ (where the multiple is not allowed to depend on p , n and s), and at which any test has error probability sum arbitrarily close to 1 for a suitably small multiple of $\rho^*(p, n, s)^2$.

Our first main contribution, in Theorem 1, is to reveal a particularly subtle form of the exact minimax testing rate in the above problem, namely

$$\rho^*(p, n, s)^2 \asymp \begin{cases} \sqrt{p \log \log(8n)} & \text{if } s \geq \sqrt{p \log \log(8n)}, \\ s \log \left(\frac{ep \log \log(8n)}{s^2} \right) \vee \log \log(8n) & \text{if } s < \sqrt{p \log \log(8n)}. \end{cases}$$

This result provides a significant generalization of two known special cases in the literature, namely $\rho^*(1, n, 1)^2$ and $\rho^*(p, 2, s)^2$; see Section 2.1 for further discussion. Although our initial optimal testing procedure depends on the sparsity level s , which would often be unknown in practice, we show in Theorem 4 that it is possible to construct an adaptive test that achieves exactly the same rate (but is a little more complicated to describe).

The theorem described above is a finite-sample result, but does not provide information at the level of constants. By contrast, in Section 2.4, we study both dense and sparse asymptotic regimes, and identify the optimal constants exactly, in the former case, and to within a factor of $\sqrt{2}$ in the latter case. In combination with Theorem 1, then, we are able to provide really quite a precise picture of the minimax testing rate in this problem.

Sections 3 and 4 concern extensions of our results to more general data generating mechanisms that allow for spatial and temporal dependence respectively. In Section 3, we allow for cross-sectional dependence across the coordinates through a non-diagonal covariance matrix Σ for the (Gaussian) columns of E . We identify the sharp minimax testing rate when $s = p$, though the optimal procedure depends on three functionals of Σ , namely its trace, as well as its Frobenius and operator norms. Estimation of these quantities is confounded by the potential presence of the changepoint, but we are able to propose a robust method that retains the same guarantee under a couple of additional conditions. As an example, we consider covariance matrices that are a convex combination of the identity matrix and a matrix of ones; thus, each pair of distinct coordinates has the same (non-negative) covariance. Interestingly, we find here that this covariance structure can make the problem either harder or easier, depending on the sparsity level of the changepoint. In Section 4, we also focus on the case $s = p$ and allow dependence across the columns of E (which are still assumed to be jointly Gaussian), controlled through a bound B on the sum of the contributions of the operator norms of the off-diagonal blocks of the $np \times np$ covariance matrix. Again, interesting phase transition phenomena in the testing rate occur here, depending on the relative magnitudes of the parameters B , p and n .

Most prior work on multivariate changepoint detection has proceeded without a sparsity condition and in an asymptotic regime with n growing to infinity with the dimension fixed,

including [Basseville and Nikiforov \(1993\)](#), [Csörgő and Horváth \(1997\)](#), [Ombao et al. \(2005\)](#), [Aue et al. \(2009\)](#), [Kirch et al. \(2015\)](#), [Zhang et al. \(2010\)](#) and [Horváth and Hušková \(2012\)](#). [Bai \(2010\)](#) studied the least squares estimator of a change in mean for high-dimensional panel data. [Jirak \(2015\)](#), [Cho and Fryzlewicz \(2015\)](#), [Cho \(2016\)](#) and [Wang and Samworth \(2018\)](#) have all proposed CUSUM-based methods for the estimation of the location of a sparse, high-dimensional changepoint. [Aston and Kirch \(2018\)](#) introduce a notion of efficiency that quantifies the detection power of different statistics in high-dimensional settings. [Enikeeva and Harchaoui \(2019\)](#) study the sparse changepoint detection problem in an asymptotic regime in which $p \rightarrow \infty$, and at the same time $s \rightarrow \infty$ with $s/p \rightarrow \infty$ and the sample size not too large; we compare their results with ours in [Section 2.3](#). Further related work on high-dimensional changepoint problems include the detection of changes in covariance (e.g. [Aue et al., 2009](#); [Cribben and Yu, 2017](#); [Wang et al., 2017](#)) and in sparse dynamic networks ([Wang et al., 2018a](#)). We emphasize that in this work we focus entirely on the offline version of the changepoint testing problem, where the entire data stream is observed prior to the statistician attempting to determine whether or not a change in mean has occurred. For recent work on the corresponding online problem, where the data are observed sequentially and one wishes to declare a change as soon as possible after it has occurred, see, e.g., [Xie and Siegmund \(2013\)](#) and [Chen et al. \(2020\)](#).

Proofs of our results in [Sections 2 and 3](#) are given in [Section 5](#), while proofs of the results in [Section 4](#) and various auxiliary lemmas are provided in the supplementary material ([Liu, Gao and Samworth, 2020](#)). We close this section by introducing some notation that will be used throughout the paper. For $d \in \mathbb{N}$, we write $[d] := \{1, \dots, d\}$. Given $a, b \in \mathbb{R}$, we write $a \vee b := \max(a, b)$ and $a \wedge b := \min(a, b)$. We also write $a \lesssim b$ to mean that there exists a universal constant $C > 0$ such that $a \leq Cb$; moreover, $a \asymp b$ means $a \lesssim b$ and $b \lesssim a$. For a set S , we use $\mathbb{1}_S$ and $|S|$ to denote its indicator function and cardinality respectively. For a vector $v = (v_1, \dots, v_d)^T \in \mathbb{R}^d$, we define the norms $\|v\|_1 := \sum_{\ell=1}^d |v_\ell|$, $\|v\|^2 := \sum_{\ell=1}^d v_\ell^2$ and $\|v\|_\infty := \max_{\ell \in [d]} |v_\ell|$, and also define $\|v\|_0 := \sum_{\ell=1}^d \mathbb{1}_{\{v_\ell \neq 0\}}$. Given two vectors $u, v \in \mathbb{R}^d$ and a positive definite matrix $\Sigma \in \mathbb{R}^{d \times d}$, we define $\langle u, v \rangle_\Sigma^{-1} := u^T \Sigma^{-1} v$ and $\|v\|_{\Sigma^{-1}} := (v^T \Sigma^{-1} v)^{1/2}$ and omit the subscripts when $\Sigma = I_d$. More generally, the trace inner product of two matrices $A, B \in \mathbb{R}^{d_1 \times d_2}$ is defined as $\langle A, B \rangle := \sum_{\ell=1}^{d_1} \sum_{\ell'=1}^{d_2} A_{\ell\ell'} B_{\ell\ell'}$, while the Frobenius and operator norms of A are given by $\|A\|_F := \sqrt{\langle A, A \rangle}$ and $\|A\|_{\text{op}} := s_{\max}(A)$ respectively, where $s_{\max}(\cdot)$ denotes the largest singular value. The total variation distance between two probability measures P and Q on a measurable space $(\mathcal{X}, \mathcal{A})$ is defined as $\text{TV}(P, Q) := \sup_{A \in \mathcal{A}} |P(A) - Q(A)|$. Moreover, if P is absolutely continuous with respect to Q , then the Kullback–Leibler divergence is defined as $D(P\|Q) := \int_{\mathcal{X}} \log \frac{dP}{dQ} dP$, and the chi-squared divergence is defined as $\chi^2(P\|Q) := \int_{\mathcal{X}} \left(\frac{dP}{dQ} - 1\right)^2 dQ$. The notation \mathbb{P} and \mathbb{E} are generic probability and expectation operators whose distribution is determined from the context.

2. Main results. Recall that we consider the observation of a $p \times n$ matrix $X = \theta + E$, where $n \geq 2$, where θ is deterministic and where each entry of the error matrix E is an independent $N(0, 1)$ random variable. In other words, writing X_t and θ_t for the t th columns of X and θ respectively, we have $X_t \sim N_p(\theta_t, I_p)$. The goal of our paper is to test whether or not the sequence $\{\theta_t\}_{t \in [n]}$ has a changepoint. We define the parameter space of signals without a changepoint by

$$\Theta_0(p, n) := \left\{ \theta \in \mathbb{R}^{p \times n} : \theta_t = \mu \text{ for some } \mu \in \mathbb{R}^p \text{ and all } t \in [n] \right\}.$$

For $s \in [p]$ and $\rho > 0$, the space consisting of signals with a sparse structural change at time $t_0 \in [n-1]$ is defined by

$$\Theta^{(t_0)}(p, n, s, \rho) := \left\{ \theta = (\theta_1, \dots, \theta_n) \in \mathbb{R}^{p \times n} : \right.$$

$$\begin{aligned} \theta_t &= \mu_1 \text{ for some } \mu_1 \in \mathbb{R}^p \text{ for all } 1 \leq t \leq t_0, \\ \theta_t &= \mu_2 \text{ for some } \mu_2 \in \mathbb{R}^p \text{ for all } t_0 + 1 \leq t \leq n, \\ \|\mu_1 - \mu_2\|_0 &\leq s, \frac{t_0(n-t_0)}{n} \|\mu_1 - \mu_2\|^2 \geq \rho^2 \}. \end{aligned}$$

In the definition of $\Theta^{(t_0)}(p, n, s, \rho)$, the parameters p and n determine the size of the problem, while t_0 is the location of the changepoint. The quantities s and ρ parametrize the sparsity level and the magnitude of the structural change respectively. It is worth noting that $\|\mu_1 - \mu_2\|^2$ is normalized by the factor $\frac{t_0(n-t_0)}{n}$, which plays the role of the effective sample size of the problem. To understand this, consider the problem of testing the changepoint at location t_0 when $p = 1$. Then the natural test statistic is

$$\frac{1}{t_0} \sum_{t=1}^{t_0} X_t - \frac{1}{n-t_0} \sum_{t=t_0+1}^n X_t,$$

whose variance is $\frac{n}{t_0(n-t_0)}$. Hence the difficulty of changepoint detection problem depends on the location of the changepoint. Through the normalization factor $\frac{t_0(n-t_0)}{n}$, we can define a common signal strength parameter ρ across different possible changepoint locations. Taking a union over all such changepoint locations, the alternative hypothesis parameter space is given by

$$\Theta(p, n, s, \rho) := \bigcup_{t_0=1}^{n-1} \Theta^{(t_0)}(p, n, s, \rho).$$

We will address the problem of testing the two hypotheses

$$(2) \quad H_0 : \theta \in \Theta_0(p, n), \quad H_1 : \theta \in \Theta(p, n, s, \rho).$$

To this end, we let Ψ denote the class of possible test statistics, i.e. measurable functions $\psi : \mathbb{R}^{p \times n} \rightarrow [0, 1]$. We also define the minimax testing error by

$$\mathcal{R}(\rho) := \inf_{\psi \in \Psi} \left\{ \sup_{\theta \in \Theta_0(p, n)} \mathbb{E}_\theta \psi(X) + \sup_{\theta \in \Theta(p, n, s, \rho)} \mathbb{E}_\theta (1 - \psi(X)) \right\},$$

where we use \mathbb{P}_θ and \mathbb{E}_θ to denote probabilities and expectations under the data generating process (1). Our goal is to determine the order of the minimax rate of testing in this problem, as defined below.

DEFINITION 1. *We say $\rho^* = \rho^*(p, n, s)$ is the minimax rate of testing if the following two conditions are satisfied:*

1. *For any $\epsilon \in (0, 1)$, there exists $C_\epsilon > 0$, depending only on ϵ , such that $\mathcal{R}(C_\epsilon \rho^*) \leq \epsilon$ for any $C > C_\epsilon$.*
2. *For any $\epsilon \in (0, 1)$, there exists $c_\epsilon > 0$, depending only on ϵ , such that $\mathcal{R}(c_\epsilon \rho^*) \geq 1 - \epsilon$ for any $c \in (0, c_\epsilon)$.*

2.1. Special cases. Special cases of $\rho^*(p, n, s)$ are well understood in the literature. For instance, when $p = s = 1$, we recover the one-dimensional changepoint detection problem. [Arias-Castro et al. \(2011\)](#) showed that

$$(3) \quad \rho^*(1, n, 1)^2 \asymp \log \log(8n).$$

The rate (3) involves an iterated logarithmic factor, in contrast to a typical logarithmic factor in the minimax rate of sparse signal detection (e.g., Donoho and Jin, 2004; Arias-Castro et al., 2005; Berthet and Rigollet, 2013).

Another solved special case is when $n = 2$. In this setting, we observe $X_1 \sim N_p(\mu_1, I_p)$ and $X_2 \sim N_p(\mu_2, I_p)$, and the problem is to test whether or not $\mu_1 = \mu_2$. Since $X_1 - X_2$ is a sufficient statistic for $\mu_1 - \mu_2$, the problem can be further reduced to a sparse signal detection problem in a Gaussian sequence model. For this problem, Collier et al. (2017) established the minimax detection boundary

$$(4) \quad \rho^*(p, 2, s)^2 \asymp \begin{cases} \sqrt{p} & \text{if } s \geq \sqrt{p} \\ s \log\left(\frac{ep}{s^2}\right) & \text{if } s < \sqrt{p}. \end{cases}$$

It is interesting to notice the elbow effect in the rate (4). Above the sparsity level of \sqrt{p} , one obtains the parametric rate that can be achieved using the test that rejects H_0 if $\|X_1 - X_2\|_2^2 > 2p + c\sqrt{p}$ for an appropriate $c > 0$.

It is straightforward to extend both rates (3) and (4) to cases where either p or n is of a constant order. However, the general form of $\rho^*(p, n, s)$ is unknown in the statistical literature.

2.2. Minimax detection boundary. The main result of the paper is given by the following theorem.

THEOREM 1. *The minimax rate of the detection boundary of the problem (2) is given by*

$$(5) \quad \rho^*(p, n, s)^2 \asymp \begin{cases} \sqrt{p \log \log(8n)} & \text{if } s \geq \sqrt{p \log \log(8n)} \\ s \log\left(\frac{ep \log \log(8n)}{s^2}\right) \vee \log \log(8n) & \text{if } s < \sqrt{p \log \log(8n)}. \end{cases}$$

It is important to note that the minimax rate (5) is not a simple sum or multiplication of the rates (3) and (4) for constant p or n . The high-dimensional changepoint detection problem differs fundamentally from both its low-dimensional version and the sparse signal detection problem.

We observe that the minimax rate exhibits the two regimes in (5) only when $p \geq \log \log(8n)$, since if $p < \log \log(8n)$, then the condition $s \geq \sqrt{p \log \log(8n)}$ is empty, and (5) has just one regime. Compared with the rate (4), the phase transition boundary for the sparsity s becomes $\sqrt{p \log \log(8n)}$. In fact, the minimax rate (5) can be obtained by first replacing the p in (4) with $p \log \log(8n)$, and then adding the extra term (3).

The dependence of (5) on n is very delicate. Consider the range of sparsity where

$$\frac{\log \log(8n)}{\log(e \log \log(8n))} \vee \frac{\sqrt{p}}{(\log \log(8n))^C} \lesssim s \lesssim \sqrt{p \log \log(8n)},$$

for some universal constant $C > 0$. The rate (5) then becomes

$$\rho^*(p, n, s)^2 \asymp s \log(e \log \log(8n)).$$

That is, it grows with n at a $\log \log \log(\cdot)$ rate. To the best of our knowledge, such a triple iterated logarithmic rate has not been found in any other problem before in the statistical literature.

Last but not least, we remark that when p or n is a constant, the rate (5) recovers (3) and (4) as special cases.

2.2.1. *Upper Bound.* To derive the upper bound, we need to construct a testing procedure. We emphasize that the goal of hypothesis testing is to detect the existence of a changepoint; this is in contrast to the problem of changepoint estimation (Cho and Fryzlewicz, 2015; Wang and Samworth, 2018; Wang et al., 2018b), where the goal is to find the changepoint's location.

If we knew that the changepoint were between t and $n - t + 1$, it would be natural to define the Cumulative Sum (CUSUM)-type statistic

$$(6) \quad Y_t := \frac{(X_1 + \dots + X_t) - (X_{n-t+1} + \dots + X_n)}{\sqrt{2t}}.$$

Note that the definition of Y_t does not use the observations between $t + 1$ and $n - t$. This allows Y_t to detect any changepoint in this range, regardless of its location. The existence of a changepoint implies that $\mathbb{E}_\theta(Y_t) \neq 0$. Since the structural change only occurs in a sparse set of coordinates, we threshold the magnitude of each coordinate $Y_t(j)$ at level $a \geq 0$ to obtain

$$A_{t,a} := \sum_{j=1}^p \{Y_t(j)^2 - \nu_a\} \mathbb{1}_{\{|Y_t(j)| \geq a\}},$$

where $\nu_a := \mathbb{E}(Z^2 \mid |Z| \geq a)$ is the conditional second moment of $Z \sim N(0, 1)$, given that its magnitude is at least a . See Collier et al. (2017) for a similar strategy for the sparse signal detection problem. Note that $A_{t,0} = \sum_{j=1}^p \{Y_t(j)^2 - 1\}$ has a centered χ_p^2 distribution under H_0 .

Since the range of the potential changepoint locations is unknown, a natural first thought is to take a maximum of $A_{t,a}$ over $t \in [n/2]$. It turns out, however, that in high-dimensional settings it is very difficult to control the dependence between these different test statistics at the level of precision required to establish the minimax testing rate. A methodological contribution of this work, then, is the recognition that it suffices to compute a maximum of $A_{t,a}$ over a candidate set \mathcal{T} of locations, because if there exists a changepoint at time t_0 and $t_0/2 < \tilde{t} \leq t_0$ for some $\tilde{t} \in \mathcal{T}$, then $\|\mathbb{E}_\theta(Y_{\tilde{t}})\|$ and $\|\mathbb{E}_\theta(Y_{t_0})\|$ are of the same order of magnitude. This observation reflects a key difference between the changepoint testing and estimation problems. To this end, we define

$$\mathcal{T} := \left\{1, 2, 4, \dots, 2^{\lfloor \log_2(n/2) \rfloor}\right\},$$

so that $|\mathcal{T}| = 1 + \lfloor \log_2(n/2) \rfloor$. Then, for a given $r \geq 0$, the testing procedure we consider is given by

$$(7) \quad \psi \equiv \psi_{a,r}(X) := \mathbb{1}_{\{\max_{t \in \mathcal{T}} A_{t,a} > r\}}.$$

The theoretical performance of the test (7) is given by the following theorem. We use the notation $r^*(p, n, s)$ for the rate function on the right-hand side of (5).

PROPOSITION 2. *For any $\epsilon \in (0, 1)$, there exists $C > 0$, depending only on ϵ , such that the testing procedure (7) with $a^2 = 4 \log \left(\frac{ep \log \log(8n)}{s^2} \right) \mathbb{1}_{\{s < \sqrt{p \log \log(8n)}\}}$ and $r = Cr^*(p, n, s)$ satisfies*

$$\sup_{\theta \in \Theta_0(p, n)} \mathbb{E}_\theta \psi + \sup_{\theta \in \Theta(p, n, s, \rho)} \mathbb{E}_\theta (1 - \psi) \leq \epsilon,$$

as long as $\rho^2 \geq 32Cr^*(p, n, s)$.

Just as the minimax rate (5) has two regimes, the testing procedure (7) also uses two different strategies. In the dense regime $s \geq \sqrt{p \log \log(8n)}$, we have $a^2 = 0$ and thus (7) becomes simply $\psi = \mathbb{1}_{\{\max_{t \in \mathcal{T}} \|Y_t\|^2 - p > r\}}$. In the sparse regime $s < \sqrt{p \log \log(8n)}$, a thresholding rule is applied at level a , where $a^2 = 4 \log \left(\frac{ep \log \log(8n)}{s^2} \right)$. We discuss adaptivity to the sparsity level s in Section 2.3.

2.2.2. Lower Bound. We show that the testing procedure (7) is minimax optimal by stating a matching lower bound.

PROPOSITION 3. *For any $\epsilon \in (0, 1)$, there exists $c > 0$, depending only on ϵ , such that $\mathcal{R}(\rho) \geq 1 - \epsilon$ whenever $\rho^2 \leq cr^*(p, n, s)$.*

2.3. Adaptation to sparsity. The testing procedure (7) that achieves the minimax detection rate depends on knowledge of the sparsity s . In this section, we present an alternative procedure that is adaptive to s . The idea is to take supremum over a grid of sparsity levels. Recall the definition of the testing procedure $\psi_{a,r}$ in (7), and let us make the dependence on s explicit by writing

$$\psi^{(s)} := \psi_{a(s), r(s)},$$

where $a^2(s) := 4 \log \left(\frac{ep \log \log(8n)}{s^2} \right) \mathbb{1}_{\{s < \sqrt{p \log \log(8n)}\}}$ and $r(s) := Cr^*(p, n, s)$ as in Proposition 2. Then our adaptive test is defined by

$$\psi_{\text{adaptive}} := \max_{s \in \mathcal{S}} \psi^{(s)},$$

where

$$\mathcal{S} := \left\{ 1, 2, 4, \dots, 2^{\lceil \log_2(\sqrt{p \log \log(8n)}) \rceil - 1} \right\} \cup \{p\}.$$

The choice of this particular grid for \mathcal{S} is not essential (we could also take $\mathcal{S} := [p]$), but it reduces computation.

THEOREM 4. *For any $\epsilon \in (0, 1)$, there exists $C > 0$, depending only on ϵ , such that the testing procedure ψ_{adaptive} satisfies*

$$\sup_{\theta \in \Theta_0(p, n)} \mathbb{E}_\theta \psi_{\text{adaptive}} + \sup_{\theta \in \Theta(p, n, s, \rho)} \mathbb{E}_\theta (1 - \psi_{\text{adaptive}}) \leq \epsilon,$$

as long as $\rho^2 \geq 64Cr^*(p, n, s)$.

Theorem 4 shows that the minimax detection boundary (1) can be achieved adaptively without the knowledge of the sparsity level s . In the literature, changepoint detection with unknown sparsity was also investigated by [Enikeeva and Harchaoui \(2019\)](#). Their procedure has a vanishing testing error as long as

$$(8) \quad \rho^2 \gtrsim \min \left(\sqrt{p \log p} + \sqrt{p \log \log n}, s \log \frac{p}{s} \right),$$

under the additional assumptions that $p, s \rightarrow \infty$, $s/p \rightarrow 0$, and $\frac{\log n}{s \log(p/s)} \rightarrow 0$. Comparing (8) with the optimal rate (1), we see that [Enikeeva and Harchaoui \(2019\)](#) successfully identified the $\sqrt{p \log \log n}$ term in the dense regime and the $s \log p$ term in the sparse regime. However, we can also observe that the $\sqrt{p \log p}$ term is not necessary, and the rate (8) is in general not sharp without the assumption $\frac{\log n}{s \log(p/s)} \rightarrow 0$, especially when the sparsity level s is around $\sqrt{p \log \log n}$.

2.4. *Asymptotic constants.* A notable feature of our minimax detection boundary derived in Theorem 1 is that the rate is non-asymptotic, meaning that the result holds for arbitrary $n \geq 2$, $p \in \mathbb{N}$ and $s \in [p]$. On the other hand, if we are allowed to make a few asymptotic assumptions, we can give explicit constants for the lower and upper bounds. In this subsection, therefore, we let both the dimension p and the sparsity s be functions of n , and we consider asymptotics as $n \rightarrow \infty$.

THEOREM 5 (Dense regime). *Assume that $s^2/(p \log \log n) \rightarrow \infty$ as $n \rightarrow \infty$. Then, with*

$$\rho = \xi (p \log \log n)^{1/4},$$

we have $\mathcal{R}(\rho) \rightarrow 0$ when $\xi > \sqrt{2}$ and $\mathcal{R}(\rho) \rightarrow 1$ when $\xi < \sqrt{2}$.

THEOREM 6 (Sparse regime). *Assume that $s^2/p \rightarrow 0$ and $s/\log \log n \rightarrow \infty$ as $n \rightarrow \infty$. Then, with*

$$\rho = \xi \sqrt{s \log \left(\frac{p \log \log n}{s^2} \right)},$$

we have $\mathcal{R}(\rho) \rightarrow 0$ when $\xi > \sqrt{2}$ and $\mathcal{R}(\rho) \rightarrow 1$ when $\xi < 1$.

These two theorems characterize the asymptotic minimax upper and lower bounds of the changepoint detection problem under dense and sparse asymptotics respectively. While we are able to nail down the exact asymptotic constant in the dense regime, the optimal asymptotic constant in the sparse regime is a more involved problem (indeed, it appears to depend on more refined aspects of the asymptotic regime), and we therefore leave it as an open problem for future research.

3. Spatial dependence. In this section, we consider changepoint detection in settings with cross-sectional dependence in the p coordinates. To be specific, we now relax our previous assumption on the cross-sectional distribution by supposing only that $X_t \sim N_p(\theta_t, \Sigma)$ for some general positive definite covariance matrix $\Sigma \in \mathbb{R}^{p \times p}$; the goal remains to solve the testing problem (2). We retain the notation \mathbb{P}_θ and \mathbb{E}_θ for probabilities and expectations, with the dependence on Σ suppressed. Similar to Definition 1, we use the notation $\rho_\Sigma^*(p, n, s)$ for the minimax rate of testing in this problem with cross-sectional covariance Σ .

Our first result provides the minimax rate of the detection boundary in the dense case where $s = p$. This sets up a useful benchmark on the difficulty of the problem depending on the covariance structure.

THEOREM 7. *In the case $s = p$, the minimax rate of testing is given by*

$$(9) \quad \rho_\Sigma^*(p, n, p)^2 \asymp \|\Sigma\|_{\text{F}} \sqrt{\log \log(8n)} \vee \|\Sigma\|_{\text{op}} \log \log(8n).$$

For $\Sigma = I_p$, Theorem 7 yields $\rho_\Sigma^*(p, n, p)^2 \asymp \sqrt{p \log \log(8n)} \vee \log \log(8n)$, which recovers the result of Theorem 1 when $s = p$. The more general result for $s < p$ with a non-diagonal Σ is hard to obtain. This is because the proof of Theorem 7 relies on a diagonalization argument of the covariance matrix, which can affect the sparsity pattern of the change in the mean unless additional assumptions on Σ similar to Hall and Jin (2010) are imposed.

A test that achieves the optimal rate (9) is given by

$$(10) \quad \psi := \mathbb{1}_{\left\{ \max_{t \in \mathcal{T}} \|Y_t\|^2 - \text{Tr}(\Sigma) > C \left(\|\Sigma\|_{\text{F}} \sqrt{\log \log(8n)} \vee \|\Sigma\|_{\text{op}} \log \log(8n) \right) \right\}},$$

for an appropriate choice of $C > 0$. Though optimal, the procedure (10) relies on knowledge of Σ . In fact, one only needs to know $\text{Tr}(\Sigma)$, $\|\Sigma\|_{\text{F}}$ and $\|\Sigma\|_{\text{op}}$, rather than the entire covariance matrix Σ . To be even more specific, from a careful examination of the proof, we see that we only need to know $\text{Tr}(\Sigma)$ up to an additive error that is at most of the same order as the cut-off, whereas knowledge of the orders of $\|\Sigma\|_{\text{F}}$ and $\|\Sigma\|_{\text{op}}$, up to multiplication by universal constants, is enough.

We now discuss how to use X to estimate the three quantities $\text{Tr}(\Sigma)$, $\|\Sigma\|_{\text{F}}$ and $\|\Sigma\|_{\text{op}}$. The solution would be straightforward if we knew the location of the changepoint, but in more typical situations where the changepoint location is unknown, this becomes a robust covariance functional estimation problem. We assume that $n \geq 6$ and that $n/3$ is an integer, since a simple modification can be made if $n/3$ is not a integer. We can then divide $[n]$ into three consecutive blocks $\mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_3$, each of whose cardinalities is $n/3 \geq 2$. For $j \in [3]$, we compute the sample covariance matrix

$$\widehat{\Sigma}_{\mathcal{D}_j} := \frac{1}{|\mathcal{D}_j| - 1} \sum_{t \in \mathcal{D}_j} (X_t - \bar{X}_{\mathcal{D}_j})(X_t - \bar{X}_{\mathcal{D}_j})^T,$$

where $\bar{X}_{\mathcal{D}_j} := |\mathcal{D}_j|^{-1} \sum_{t \in \mathcal{D}_j} X_t$. We can then order these three estimators according to their trace, and Frobenius and operator norms, yielding

$$\begin{aligned} \text{Tr}(\widehat{\Sigma})^{(1)} &\leq \text{Tr}(\widehat{\Sigma})^{(2)} \leq \text{Tr}(\widehat{\Sigma})^{(3)}, \\ \|\widehat{\Sigma}\|_{\text{F}}^{(1)} &\leq \|\widehat{\Sigma}\|_{\text{F}}^{(2)} \leq \|\widehat{\Sigma}\|_{\text{F}}^{(3)}, \\ \|\widehat{\Sigma}\|_{\text{op}}^{(1)} &\leq \|\widehat{\Sigma}\|_{\text{op}}^{(2)} \leq \|\widehat{\Sigma}\|_{\text{op}}^{(3)}. \end{aligned}$$

The idea is that at least two of the three covariance matrix estimators $\widehat{\Sigma}_{\mathcal{D}_1}, \widehat{\Sigma}_{\mathcal{D}_2}, \widehat{\Sigma}_{\mathcal{D}_3}$ should be accurate, because there is at most one changepoint location. This motivates us to take the medians $\text{Tr}(\widehat{\Sigma})^{(2)}$, $\|\widehat{\Sigma}\|_{\text{F}}^{(2)}$ and $\|\widehat{\Sigma}\|_{\text{op}}^{(2)}$ with respect to the three functionals as our robust estimators. It is convenient to define $\Theta(p, n, s, 0) := \Theta_0(p, n) \cup (\cup_{\rho > 0} \Theta(p, n, s, \rho))$.

PROPOSITION 8. *Assume $p \leq cn$ for some $c > 0$, and fix an arbitrary positive definite $\Sigma \in \mathbb{R}^{p \times p}$ and $\theta \in \Theta(p, n, p, 0)$. Then given $\epsilon > 0$, there exists $C > 0$, depending only on c and ϵ , such that*

$$\begin{aligned} |\text{Tr}(\widehat{\Sigma})^{(2)} - \text{Tr}(\Sigma)| &\leq C \left(\frac{\sqrt{p} \|\Sigma\|_{\text{F}}}{\sqrt{n}} + \frac{p \|\Sigma\|_{\text{op}}}{n} \right), \\ \left| \|\widehat{\Sigma}\|_{\text{F}}^{(2)} - \|\Sigma\|_{\text{F}} \right| &\leq C \|\Sigma\|_{\text{op}} \sqrt{\frac{p^2}{n}}, \\ \left| \|\widehat{\Sigma}\|_{\text{op}}^{(2)} - \|\Sigma\|_{\text{op}} \right| &\leq C \|\Sigma\|_{\text{op}} \sqrt{\frac{p}{n}}, \end{aligned}$$

with \mathbb{P}_{θ} -probability at least $1 - \epsilon/4$.

With the help of Proposition 8, we can plug the estimators $\text{Tr}(\widehat{\Sigma})^{(2)}$, $\|\widehat{\Sigma}\|_{\text{F}}^{(2)}$ and $\|\widehat{\Sigma}\|_{\text{op}}^{(2)}$ into the procedure (10). This test is adaptive to the unknown covariance structure, and comes with the following performance guarantee.

COROLLARY 9. *Assume that $\sqrt{p} \|\Sigma\|_{\text{op}} \leq A \|\Sigma\|_{\text{F}}$ for some $A > 0$. Then given $\epsilon > 0$, there exist $c, C > 0$, depending only on A and ϵ , such that if $p \leq cn$, then the testing procedure*

$$\psi_{\text{Cov}} := \mathbb{1}_{\{\max_{t \in \mathcal{T}} \|Y_t\|^2 - \text{Tr}(\widehat{\Sigma})^{(2)} > C(\|\widehat{\Sigma}\|_{\text{F}}^{(2)} \sqrt{\log \log(8n)} \vee \|\widehat{\Sigma}\|_{\text{op}}^{(2)} \log \log(8n))\}}$$

satisfies

$$\sup_{\theta \in \Theta_0(p, n)} \mathbb{E}_\theta \psi_{\text{Cov}} + \sup_{\theta \in \Theta(p, n, p, \rho)} \mathbb{E}_\theta (1 - \psi_{\text{Cov}}) \leq \epsilon,$$

as long as $\rho^2 \geq 64C \left(\|\Sigma\|_{\text{F}} \sqrt{\log \log(8n)} \vee \|\Sigma\|_{\text{op}} \log \log(8n) \right)$.

REMARK 1. *The conditions $p \lesssim n$ and $\sqrt{p}\|\Sigma\|_{\text{op}} \lesssim \|\Sigma\|_{\text{F}}$ guarantee that $\|\widehat{\Sigma}\|_{\text{F}}^{(2)} - \|\Sigma\|_{\text{F}} \lesssim \|\Sigma\|_{\text{F}}$ and $\|\widehat{\Sigma}\|_{\text{op}}^{(2)} - \|\Sigma\|_{\text{op}} \lesssim \|\Sigma\|_{\text{op}}$ with high probability, by Proposition 8. Note that $\sqrt{p}\|\Sigma\|_{\text{op}} \lesssim \|\Sigma\|_{\text{F}}$ will be satisfied if all eigenvalues of Σ are of the same order. In fact, it is possible to weaken the condition $\sqrt{p}\|\Sigma\|_{\text{op}} \lesssim \|\Sigma\|_{\text{F}}$ using the notion of effective rank (Koltchinskii and Lounici, 2017); however, this greatly complicates the analysis, and we do not pursue this here. Alternatively, Corollary 9 also holds without the $\sqrt{p}\|\Sigma\|_{\text{op}} \leq A\|\Sigma\|_{\text{F}}$ condition but under the stronger dimensionality restriction $p^2 \leq cn$; this then allows for an arbitrary covariance matrix Σ .*

To better understand the influence of the covariance structure, consider, for $\gamma \in [0, 1)$, the covariance matrix

$$\Sigma(\gamma) := (1 - \gamma)I_p + \gamma \mathbf{1}_p \mathbf{1}_p^T,$$

which has diagonal entries 1 and off-diagonal entries γ . The parameter γ controls the pairwise spatial dependence; moreover, $\|\Sigma(\gamma)\|_{\text{F}}^2 = (1 - \gamma^2)p + p^2\gamma^2$ and $\|\Sigma(\gamma)\|_{\text{op}} = 1 + (p - 1)\gamma$. By Theorem 7, we have

$$(11) \quad \rho_{\Sigma(\gamma)}^*(p, n, p)^2 \asymp \sqrt{\{(1 - \gamma^2)p + p^2\gamma^2\} \log \log(8n)} \vee \{1 + (p - 1)\gamma\} \log \log(8n).$$

Thus the spatial dependence significantly increases the difficulty of the testing problem. In particular, if γ is of a constant order, then the minimax rate is $p \log \log(8n)$, which is much larger than the rate (9) for $\Sigma = I_p$.

However, the increased difficulty of testing in this example is just one part of the story. When we consider the sparsity factor s , the influence of the covariance structure can be the other way around. To illustrate this interesting phenomenon, we discuss a situation where s is small. Since $X_t \sim N_p(\theta_t, \Sigma(\gamma))$, we have that $Y_t \sim N_p(\Delta_t, \Sigma(\gamma))$ for $t < n/2$, where $\Delta_t := \frac{(\theta_1 + \dots + \theta_t) - (\theta_{n-t+1} + \dots + \theta_n)}{\sqrt{2t}}$. Hence, the distribution of Y_t can be expressed in terms of a factor model. That is,

$$(12) \quad Y_t(j) = \Delta_t(j) + \sqrt{\gamma}W_t + \sqrt{1 - \gamma}Z_{tj},$$

where $W_t, Z_{t1}, \dots, Z_{tp} \stackrel{\text{iid}}{\sim} N(0, 1)$. When there is no changepoint, we have $\Delta_t = 0$, so $Y_t(j)|W_t \stackrel{\text{iid}}{\sim} N(\sqrt{\gamma}W_t, 1 - \gamma)$ for all $j \in [p]$. When there is a changepoint between t and $n - t + 1$, we have $\|\Delta_t\|_0 \leq s$. In either case, then, we can estimate $\sqrt{\gamma}W_t$ by $\text{Median}(Y_t)$. This motivates the new statistic

$$(13) \quad \tilde{Y}_t := \frac{Y_t - \text{Median}(Y_t)\mathbf{1}_p}{\sqrt{1 - \gamma}}.$$

To construct a scalar summary of \tilde{Y}_t , we define the functions $f_a(x) := (x^2 - \nu_a)\mathbb{1}_{\{|x| \geq a\}}$ for $x \in \mathbb{R}$ and, for $C' \geq 0$, set

$$(14) \quad g_a(x) \equiv g_{a, C'}(x) := \inf \left\{ f_a(y) : |y - x| \leq C' \sqrt{\frac{\log \log(8n)}{p}} \right\}.$$

Note that $g_a(x) = f_a(x)$ when $C' = 0$. The use of a positive $C' > 0$ in (14) is to tolerate the error of $\text{Median}(Y_t)$ as an estimator of $\sqrt{\gamma}W_t$. The new testing procedure is then

$$(15) \quad \psi_{a,r,C'} := \mathbb{1}_{\{\max_{t \in \mathcal{T}} \sum_{j=1}^p g_a(\tilde{Y}_t(j)) > r\}}.$$

THEOREM 10. *Assume that $\gamma \in [0, 1)$ and $s \leq (p \log \log(8n))^{1/5}$. Then there exist universal constants $c, C' > 0$ such that if $\frac{\log \log(8n)}{p} \leq c$, then for any $\epsilon \in (0, 1)$, we can find $C > 0$ and $n_0 \in \mathbb{N}$, both depending only on ϵ , such that the testing procedure (15) with $a^2 = 4 \log \left(\frac{ep \log \log(8n)}{s^2} \right)$ and $r = C(1 - \gamma) \left(s \log \left(\frac{ep \log \log(8n)}{s^2} \right) \vee \log \log(8n) \right)$ satisfies*

$$\sup_{\theta \in \Theta_0(p,n)} \mathbb{E}_\theta \psi_{a,r,C'} + \sup_{\theta \in \Theta(p,n,s,\rho)} \mathbb{E}_\theta (1 - \psi_{a,r,C'}) \leq \epsilon,$$

for $n \geq n_0$, provided $\rho^2 \geq 32C(1 - \gamma) \{s \log(ep s^{-2} \log \log(8n)) \vee \log \log(8n)\}$.

Surprisingly, in the sparse regime, the spatial correlation helps changepoint detection, and the required signal strength for testing consistency decreases as γ increases. This is in stark contrast to (11) for the same covariance structure when $s = p$.

REMARK 2. *The testing procedure considered in Theorem 10 can be easily made adaptive to the unknown γ by taking advantage of Proposition 8. Since $\text{Tr}(\Sigma(\gamma)) = p + (p^2 - p)\gamma$, when $p \geq 2$ the estimator $\hat{\gamma} := \frac{\text{Tr}(\hat{\Sigma})^{(2)} - p}{p^2 - p}$ satisfies*

$$|\hat{\gamma} - \gamma| \lesssim \frac{\sqrt{(1 - \gamma^2)p + p^2 \gamma^2}}{p^{3/2} \sqrt{n}} + \frac{1 + (p - 1)\gamma}{pn},$$

with probability at least $1 - 2e^{-p}$. Then, the procedure with γ replaced by $\hat{\gamma}$ enjoys the same guarantee of Theorem 10 under mild extra conditions.

The next theorem shows that the rate achieved by Theorem 10 is minimax optimal.

THEOREM 11. *Assume that $\gamma \in [0, 1)$ and $s \leq \sqrt{p \log \log n}$. Then*

$$(16) \quad \rho_{\Sigma(\gamma)}^*(p, n, s)^2 \gtrsim (1 - \gamma) \left\{ s \log \left(\frac{ep \log \log(8n)}{s^2} \right) \vee \log \log(8n) \right\}.$$

To conclude this section, we remark that the dependence on γ of the minimax testing rate arises in part due to our choice of measuring departures from the null hypothesis in terms of a rescaled squared Euclidean distance. Other natural choices, such as a rescaled squared supremum norm distance (Jirak, 2015) may well lead to different phenomena.

4. Temporal dependence. In this section, we consider the situation where X_1, \dots, X_n form a multivariate time series. To be specific, in our model $X_t = \theta_t + E_t$ for $t \in [n]$, we now assume that the random vectors E_1, \dots, E_n are jointly Gaussian but not necessarily independent. The covariance structure of the error vectors can be parametrized by a covariance matrix $\Sigma \in \mathbb{R}^{pn \times pn}$, and for $B \geq 0$, we write $\Sigma \in \mathcal{C}(p, n, B)$ if:

1. $\text{Cov}(E_t) = I_p$ for all $t \in [n]$;
2. $\sum_{s \in [n] \setminus \{t\}} \|\text{Cov}(E_s, E_t)\|_{\text{op}} \leq B$ for all $t \in [n]$.

Thus the data generating process of X is completely determined by its mean matrix θ and covariance matrix $\Sigma \in \mathcal{C}(p, n, B)$, and we use the notion $\mathbb{P}_{\theta, \Sigma}$ and $\mathbb{E}_{\theta, \Sigma}$ for the corresponding probability and expectation. The case $B = 0$ reduces to the situation of observations at different time points being independent. Time series dependence in high-dimensional change-point problems has also been considered by [Wang and Samworth \(2018\)](#); their condition $\|\sum_{s=1}^n \text{Cov}(E_s, E_t)\|_{\text{op}} \leq B$ for all $t \in [n]$ is only slightly different from ours. We also mention here the work of [Horváth and Hušková \(2012\)](#), who study the asymptotic distributions of changepoint test statistics with dependent data, in a regime in which $p/\sqrt{n} \rightarrow 0$.

We focus on the case $s = p$ and do not consider the effect of sparsity. The minimax testing error is defined by

$$\mathcal{R}(\rho) := \inf_{\psi \in \Psi} \left\{ \sup_{\substack{\theta \in \Theta_0(p, n) \\ \Sigma \in \mathcal{C}(p, n, B)}} \mathbb{E}_{\theta, \Sigma} \psi + \sup_{\substack{\theta \in \Theta(p, n, p, \rho) \\ \Sigma \in \mathcal{C}(p, n, B)}} \mathbb{E}_{\theta, \Sigma} (1 - \psi) \right\}.$$

We also define the corresponding minimax rate of detection boundary $\rho_{\text{Temp}}^*(p, n, p, B)$ similarly to Definition 1. The testing procedure

$$(17) \quad \psi_{\text{Temp}} := \mathbb{1}_{\{\max_{t \in \mathcal{T}} \|Y_t\|^2 - p > r\}}$$

has the following property:

THEOREM 12. *For any $\epsilon \in (0, 1)$, there exists $C > 0$, depending only on ϵ , such that the test (17) with $r = C \left\{ Bp + (1 + B) \left(\sqrt{p \log \log(8n)} + \log \log(8n) \right) \right\}$ satisfies*

$$\sup_{\substack{\theta \in \Theta_0(p, n) \\ \Sigma \in \mathcal{C}(p, n, B)}} \mathbb{E}_{\theta, \Sigma} \psi_{\text{Temp}} + \sup_{\substack{\theta \in \Theta(p, n, p, \rho) \\ \Sigma \in \mathcal{C}(p, n, B)}} \mathbb{E}_{\theta, \Sigma} (1 - \psi_{\text{Temp}}) \leq \epsilon,$$

as long as $\rho^2 \geq 32C \left\{ Bp + (1 + B) \left(\sqrt{p \log \log(8n)} + \log \log(8n) \right) \right\}$.

Our final result provides the complementary lower bound.

THEOREM 13. *Assume that $B \leq D\sqrt{n/p}$ for some $D > 0$, and let*

$$(18) \quad \rho_{\text{Temp}}^{*2} \equiv \rho_{\text{Temp}}^*(p, n, p, B)^2 := Bp + (1 + B) \left\{ \sqrt{p \log \log(8n)} \vee \log \log(8n) \right\}.$$

Then given $\epsilon > 0$, there exist $c_{\epsilon, D} > 0$, depending only on ϵ and D , and $p_{\epsilon} \in \mathbb{N}$, depending only on ϵ , such that $\mathcal{R}(c\rho_{\text{Temp}}^) \geq 1 - \epsilon$ whenever $c \in (0, c_{\epsilon, D})$ and $p \geq p_{\epsilon}$.*

Together, Theorems 12 and 13 reveal the rate of the minimax detection boundary when $B \lesssim \sqrt{n/p}$. Observe that when $B = 0$, the rate (18) becomes $\sqrt{p \log \log(8n)} \vee \log \log(8n)$, which matches (5) when $s = p$. When $B > 0$, the rate (18) has an extra multiplicative factor $1 + B$ and an extra additive factor Bp , which are present for different reasons. Due to the dependence of the time series, one can think of $n/(1 + B)$ and $\rho^2/(1 + B)$ as being the effective sample size and signal strength respectively, instead of n and ρ^2 for the independent case, and this leads to the presence of the multiplicative factor $1 + B$. On the other hand, the additive term Bp arises from the fact that $\mathbb{E}_{\theta, \Sigma} \|Y_t\|^2 - p$ under the null hypothesis is not known completely due to the unknown covariance structure $\Sigma \in \mathcal{C}(p, n, B)$. In fact, in the construction of the lower bound, the relevant zero mean Gaussian distribution with unknown covariance can be approximated by a location mixture of Gaussians with known identity covariance. This allows us to relate the difficulties of the two problems. When $B = 0$, the class $\mathcal{C}(p, n, B)$ becomes a singleton, and we know that $\mathbb{E}_{\theta, \Sigma} \|Y_t\|^2 = p$ under the null, so this additional term disappears.

5. Proofs.

5.1. Proofs of results in Section 2.

PROOF OF PROPOSITION 2. In this proof, we seek to control Type I and Type II errors using tail probability bounds for chi-squared random variables and a version with truncated summands; these are given as Lemmas 1 and 5 respectively. Fixing $\epsilon \in (0, 1)$, set $C = C(\epsilon) := 50C_1/\epsilon$, where the universal constant $C_1 \geq 1$ is taken from Lemma 7. We first consider the case where $s \geq \sqrt{p \log \log(8n)}$. Then $a = 0$, so that $A_{t,a} = \sum_{j=1}^p Y_t(j)^2 - p$. Therefore, for any $\theta \in \Theta_0(p, n)$, we have $A_{t,a} \sim \chi_p^2 - p$. Then, by a union bound and Lemma 1, we obtain that with $x := \frac{C}{9} \log \log(8n)$,

$$\begin{aligned} \mathbb{E}_\theta \psi &= \mathbb{P}_\theta \left(\max_{t \in \mathcal{T}} A_{t,0} > C \sqrt{p \log \log(8n)} \right) \leq \mathbb{P}_\theta \left(\max_{t \in \mathcal{T}} A_{t,0} > 2\sqrt{px} + 2x \right) \\ (19) \qquad \qquad \qquad &\leq 2 \log(en) e^{-x} \leq \frac{\epsilon}{2}, \end{aligned}$$

where the final inequality holds because $C \geq 9 + 9 \log(4/\epsilon)$.

Now suppose that $\theta \in \Theta(p, n, s, \rho)$. For any $\theta \in \Theta(p, n, s, \rho)$, there exists some $t_0 \in [n-1]$ such that $X_1, \dots, X_{t_0} \stackrel{\text{iid}}{\sim} N_p(\mu_1, I_p)$ and $X_{t_0+1}, \dots, X_n \stackrel{\text{iid}}{\sim} N_p(\mu_2, I_p)$, where the vectors μ_1 and μ_2 satisfy $\frac{t_0(n-t_0)}{n} \|\mu_1 - \mu_2\|^2 \geq \rho^2$. Without loss of generality, we may assume that $t_0 \leq n/2$, since the case $t_0 > n/2$ can be handled by a symmetric argument. By the definition of \mathcal{T} , there exists a unique $\tilde{t} \in \mathcal{T}$ such that $t_0/2 < \tilde{t} \leq t_0$. Now $A_{\tilde{t},a} \sim \chi_{p,\delta^2}^2 - p$, where the non-centrality parameter δ^2 satisfies

$$\delta^2 = \frac{\tilde{t} \|\mu_1 - \mu_2\|^2}{2} \geq \frac{t_0 \|\mu_1 - \mu_2\|^2}{4} \geq \frac{t_0(n-t_0)}{4n} \|\mu_1 - \mu_2\|^2 \geq \frac{\rho^2}{4}.$$

Therefore, by Chebychev's inequality,

$$\begin{aligned} \mathbb{E}_\theta(1 - \psi) &\leq \mathbb{P}_\theta \left(\max_{t \in \mathcal{T}} \|Y_t\|^2 - p \leq \frac{\rho^2}{32} \right) \leq \mathbb{P}_\theta \left(\|Y_{\tilde{t}}\|^2 - p \leq \frac{\delta^2}{8} \right) \leq \frac{2(p + 2\delta^2)}{(7/8)^2 \delta^4} \\ (20) \qquad \qquad \qquad &\leq \frac{2(p + \rho^2/2)}{(7/32)^2 \rho^4} \leq \frac{2}{49C^2 \log \log(8n)} + \frac{32}{49C \sqrt{p \log \log(8n)}} \leq \frac{\epsilon}{2}, \end{aligned}$$

since $C \geq 49/(68\epsilon)$.

We now consider the case where $s < \sqrt{p \log \log(8n)}$, and first suppose that $\theta \in \Theta_0(p, n)$. By Lemma 5 and a union bound, we have

$$\begin{aligned} \mathbb{E}_\theta \psi &= \mathbb{P}_\theta \left(\max_{t \in \mathcal{T}} A_{t,a} > Cr^* \right) \leq \mathbb{P}_\theta \left(\max_{t \in \mathcal{T}} A_{t,a} > \sqrt{pe^{-a^2/2}x} + x \right) \\ (21) \qquad \qquad \qquad &\leq 2 \log(en) e^{-x} \leq \frac{\epsilon}{2}, \end{aligned}$$

where we still take $x = \frac{C}{9} \log \log(8n)$.

Finally, for $\theta \in \Theta(p, n, s, \rho)$, we define \tilde{t}, μ_1, μ_2 as in the dense case. Now

$$\max_{t \in \mathcal{T}} A_{t,a} \geq A_{\tilde{t},a} = \sum_{j=1}^p (Y_{\tilde{t}}(j)^2 - \nu_a) \mathbb{1}_{\{|Y_{\tilde{t}}(j)| \geq a\}},$$

where $Y_{\tilde{t}}(j) \sim N(\Delta_j, 1)$, with $\Delta_j := \sqrt{\frac{\tilde{t}}{2}} \{\mu_1(j) - \mu_2(j)\}$. By Lemma 6,

$$\mathbb{E}A_{\tilde{t},a} \geq \frac{1}{2} \sum_{j:|\Delta_j| \geq 8a} \Delta_j^2 = \frac{1}{2} \left(\sum_{j=1}^p \Delta_j^2 - \sum_{j:0 < |\Delta_j| < 8a} \Delta_j^2 \right) \geq \frac{1}{2} (\delta^2 - 64sa^2) \geq \frac{\delta^2}{4},$$

where the last inequality uses the fact that $4\delta^2 \geq \rho^2 \geq 8Csa^2$. Moreover, by Lemma 7, we have

$$\text{Var}(A_{\tilde{t},a}) = \sum_{j=1}^p \text{Var}\{(Y_{\tilde{t}}(j)^2 - \nu_a) \mathbb{1}_{\{|Y_{\tilde{t}}(j)| \geq a\}}\} \leq C_1 (pe^{-a^2/4} + sa^4 + \delta^2).$$

By Chebychev's inequality, we deduce that

$$\begin{aligned} \mathbb{E}_\theta(1 - \psi) &= \mathbb{P}_\theta \left(\max_{t \in \mathcal{T}} A_{t,a} \leq \frac{\rho^2}{32} \right) \leq \mathbb{P}_\theta \left(A_{\tilde{t},a} \leq \frac{\delta^2}{8} \right) \\ &\leq \frac{\text{Var}(A_{\tilde{t},a})}{(\mathbb{E}A_{\tilde{t},a} - \delta^2/8)^2} \leq \frac{C_1 (pe^{-a^2/4} + sa^4 + \delta^2)}{\delta^4/2^6} \\ (22) \quad &\leq \frac{C_1 pe^{-a^2/4} + C_1 sa^4 + C_1 \rho^2/4}{\rho^4/2^{10}} \leq \frac{C_1}{C^2} + \frac{16C_1}{C^2} + \frac{8C_1}{C} \leq \frac{\epsilon}{2}, \end{aligned}$$

as required. \square

The proof of Proposition 3 below is based on the lower bound technique that involves bounding the chi-squared divergence.

PROOF OF PROPOSITION 3. We only need to derive the lower bound for n that is sufficiently large. This is because when n is bounded, the minimax rate is reduced to the formula (4), and the derivation of the lower bound follows the same argument as in Collier et al. (2017). The strategy for our lower bound is to construct a suitable prior distribution on the alternative hypothesis parameter space and to bound the total variation distance between the null distribution and the mixture distribution induced by the prior. More precisely, by Lemmas 8 and 10, given $\eta > 0$, it suffices to find a probability measure ν with $\text{supp}(\nu) \subseteq \Theta(p, n, s, \rho)$ and a universal constant $c > 0$ such that

$$(23) \quad \mathbb{E}_{(\theta_1, \theta_2) \sim \nu \otimes \nu} \exp(\langle \theta_1, \theta_2 \rangle) \leq 1 + \eta,$$

whenever $\rho = c\rho^*$.

We first consider the case when $s \geq \sqrt{p \log \log(8n)}$. We define ν to be the distribution of $\theta = (\theta_{j\ell}) \in \Theta(p, n, s, \rho)$ with $\rho := \sqrt{s\beta}/\sqrt{2}$ for some $\beta = \beta(p, n, s)$ to be defined later, generated according to the following sampling process:

1. Uniformly sample a subset $S \subseteq [p]$ of cardinality s ;
2. Independently of S , generate $k \sim \text{Unif}\{0, 1, 2, \dots, \lfloor \log_2(n/2) \rfloor\}$;
3. Independently of (S, k) , sample $u = (u_1, \dots, u_p) \in \mathbb{R}^p$, where $u_1, \dots, u_p \stackrel{\text{iid}}{\sim} \text{Unif}(\{-1, 1\})$;
4. Given the triplet (S, k, u) sampled in the previous steps, define $\theta_{j\ell} := \frac{\beta}{\sqrt{2^k}} u_j$ for all $(j, \ell) \in S \times [2^k]$ and $\theta_{j\ell} := 0$ otherwise.

Since

$$\frac{2^k(n - 2^k)}{n} \frac{\beta^2}{2^k} \sum_{j \in S} u_j^2 = s\beta^2 \frac{n - 2^k}{n} \geq \frac{s\beta^2}{2},$$

we have $\text{supp}(\nu) \subseteq \Theta(p, n, s, \rho)$ with $\rho^2 = s\beta^2/2$. Suppose we independently sample triplets (S, k, u) and (T, l, v) from the first three steps and use these two triplets to construct θ_1 and θ_2 according to the fourth step. Then

$$\langle \theta_1, \theta_2 \rangle = (2^k \wedge 2^l) \frac{\beta^2}{\sqrt{2^{k+l}}} \sum_{j \in S \cap T} u_j v_j = \frac{\beta^2}{2^{|l-k|/2}} \sum_{j \in S \cap T} u_j v_j.$$

Thus

$$\mathbb{E}_{(\theta_1, \theta_2) \sim \nu \otimes \nu} \exp(\langle \theta_1, \theta_2 \rangle) = \mathbb{E} \exp\left(\frac{\beta^2}{2^{|l-k|/2}} \sum_{j \in S \cap T} u_j v_j\right),$$

where the expectation is over the joint distribution of (S, k, u, T, l, v) . But we also have that $u_j v_j \stackrel{\text{iid}}{\sim} \text{Unif}(\{-1, 1\})$, so

$$\begin{aligned} \mathbb{E}_{(\theta_1, \theta_2) \sim \nu \otimes \nu} \exp(\langle \theta_1, \theta_2 \rangle) &= \mathbb{E} \left\{ \mathbb{E} \left(\frac{1}{2} e^{\beta^2/2^{|l-k|/2}} + \frac{1}{2} e^{-\beta^2/2^{|l-k|/2}} \right) \right\}^{|S \cap T|} \\ &\leq \mathbb{E} \exp\left(|S \cap T| \frac{\beta^4}{2^{|l-k|+1}}\right), \end{aligned}$$

where the final inequality uses the fact that $(e^x + e^{-x})/2 \leq e^{x^2/2}$ for $x \in \mathbb{R}$ and Jensen's inequality. Note that $|S \cap T|$ is distributed according to the hypergeometric distribution¹ $\text{Hyp}(p, s, s)$. By the fact that the $\text{Hyp}(p, s, s)$ distribution is no larger, in the convex ordering sense, than the binomial distribution $\text{Bin}(s, s/p)$ (Hoeffding, 1963, Theorem 4), we have

$$\begin{aligned} \mathbb{E} \exp\left(|S \cap T| \frac{\beta^4}{2^{|l-k|+1}}\right) &\leq \left\{ \mathbb{E} \left(1 - \frac{s}{p} + \frac{s}{p} e^{\beta^4/2^{|l-k|+1}} \right) \right\}^s \\ (24) \quad &\leq \mathbb{E} \left\{ \left(1 + \frac{s}{2p} \frac{\beta^4}{2^{|l-k|}} e^{\beta^4/2^{|l-k|+1}} \right)^s \right\} =: \mathbb{E} L(l, k), \end{aligned}$$

say, where we have used $e^x - 1 \leq x e^x$ for all $x \geq 0$ and Jensen's inequality to derive the last inequality above. From now on, we set $\beta := \{c_1 p s^{-2} \log \log(8n)\}^{1/4}$, where $c_1 = c_1(\eta) \in (0, 1/4]$ will be chosen to be sufficiently small. The condition $s \geq \sqrt{p \log \log(8n)}$ ensures that $\beta \leq 1$. We first claim that

$$(25) \quad \mathbb{E}\{L(l, k) \mathbb{1}_{\{l=k\}}\} \leq \left\{ \left(1 + \frac{\eta}{4} \right) \mathbb{P}(l=k) \right\} \vee \frac{\eta}{4},$$

provided that $c_1 \leq \eta \log(1 + \frac{\eta}{4})/8$. To see this, first note that for $n \geq \exp(\exp(8/\eta))/8$, we have

$$\begin{aligned} \mathbb{E}\{L(l, k) \mathbb{1}_{\{l=k\}}\} &\leq \left(1 + \frac{c_1}{s} \log \log(8n) \right)^s \mathbb{P}(l=k) \leq \frac{\log^{1/4}(8n)}{1 + \lfloor \log_2(n/2) \rfloor} \\ &\leq \frac{\eta}{8} \log \log(8n) \frac{\log^{1/4}(8n)}{1 + \lfloor \log_2(n/2) \rfloor} \leq \frac{\eta}{4}. \end{aligned}$$

On the other hand, when $n < \exp(\exp(8/\eta))/8$, we have

$$\mathbb{E}\{L(l, k) \mathbb{1}_{\{l=k\}}\} \leq \log^{c_1}(8n) \mathbb{P}(l=k) \leq e^{8c_1/\eta} \mathbb{P}(l=k) \leq \left(1 + \frac{\eta}{4} \right) \mathbb{P}(l=k).$$

¹The $\text{Hyp}(p, s, r)$ distribution models the number of white balls drawn when sampling r balls without replacement from an urn containing p balls, s of which are white.

Moreover,

$$\begin{aligned}
& \mathbb{E}\left\{L(l, k) \mathbb{1}_{\{0 < |l-k| \leq (\eta/8) \log \log(8n)\}}\right\} \\
& \leq \left(1 + \frac{c_1}{s} \log \log(8n)\right)^s \mathbb{P}\left(0 < |l-k| \leq \frac{\eta}{8} \log \log(8n)\right) \\
(26) \quad & \leq \log^{1/4}(8n) \frac{\eta \log \log(8n)}{4(1 + \lfloor \log_2(n/2) \rfloor)} \leq \frac{\eta}{2}.
\end{aligned}$$

For the third term, we write $a_\eta := \sup_{n \geq 2} \frac{\log \log(8n)}{\log^{(\eta/8) \log(2)}(8n)}$. By reducing $\eta > 0$ and $c_1 = c_1(\eta)$ if necessary, we may assume that $c_1 a_\eta \leq \eta/8 \leq 1/2$, so that

$$\begin{aligned}
& \mathbb{E}\left\{L(l, k) \mathbb{1}_{\{|l-k| > (\eta/8) \log \log(8n)\}}\right\} \\
& \leq \left(1 + \frac{c_1 a_\eta}{s}\right)^s \mathbb{P}\{|l-k| > (\eta/8) \log \log(8n)\} \\
& \leq (1 + 2c_1 a_\eta) \mathbb{P}\{|l-k| > (\eta/8) \log \log(8n)\} \\
(27) \quad & \leq \left(1 + \frac{\eta}{4}\right) \mathbb{P}\{|l-k| > (\eta/8) \log \log(8n)\}.
\end{aligned}$$

From (25), (26) and (27), we conclude that

$$\mathbb{E}\{L(l, k)\} \leq 1 + \eta,$$

which establishes (23) in the case $s \geq \sqrt{p \log \log(8n)}$.

We now consider the case $s < \sqrt{p \log \log(8n)}$ and $s \log\left(\frac{ep \log \log(8n)}{s^2}\right) \geq \log \log(8n)$. The goal is to derive a lower bound with rate $s \log\left(\frac{ep \log \log(8n)}{s^2}\right)$. We use the same ν specified in the previous case except that in the third step, we set $u_j = 1$ for all $j \in S$. With this modification of ν , we have $\langle \theta_1, \theta_2 \rangle = |S \cap T| \frac{\beta^2}{2^{|l-k|/2}}$. Again, $|S \cap T|$ is distributed according to the hypergeometric distribution $\text{Hyp}(p, s, s)$, and

$$\begin{aligned}
(28) \quad & \mathbb{E}_{(\theta_1, \theta_2) \sim \nu \otimes \nu} \exp(\langle \theta_1, \theta_2 \rangle) = \mathbb{E} \exp\left(|S \cap T| \frac{\beta^2}{2^{|l-k|/2}}\right) \\
(29) \quad & \leq \left\{ \mathbb{E} \left(1 - \frac{s}{p} + \frac{s}{p} e^{\beta^2/2^{|l-k|/2}}\right) \right\}^s \leq \mathbb{E} \left\{ \left(1 + \frac{s}{p} e^{\beta^2/2^{|l-k|/2}}\right)^s \right\} =: \mathbb{E}R(l, k),
\end{aligned}$$

say. We take $\beta := \log^{1/2}\left(\frac{c_2 p \log \log(8n)}{s^2}\right)$, where $c_2 = c_2(\eta) \in (0, 1/4]$ will be chosen sufficiently small. Parallel to the bounds for $\mathbb{E}L(l, k)$, we will split into three terms. For the first term, we have

$$\begin{aligned}
& \mathbb{E}\{R(l, k) \mathbb{1}_{\{l=k\}}\} \leq \left(1 + \frac{c_2}{s} \log \log(8n)\right)^s \mathbb{P}(l=k) \\
& \leq \left\{ \left(1 + \frac{\eta}{4}\right) \mathbb{P}(l=k) \right\} \vee \frac{\eta}{4},
\end{aligned}$$

as before, as long as $c_2 \leq \eta \log\left(1 + \frac{\eta}{4}\right)/8$. For the second term,

$$\begin{aligned}
& \mathbb{E}\{R(l, k) \mathbb{1}_{\{0 < |l-k| \leq (\eta/8) \log \log(8n)\}}\} \\
& \leq \left(1 + \frac{s}{p} e^{\beta^2}\right)^s \mathbb{P}\left(0 < |l-k| \leq \frac{\eta}{8} \log \log(8n)\right)
\end{aligned}$$

$$\leq \left(1 + \frac{\log \log(8n)}{4s}\right)^s \frac{\eta \log \log(8n)}{4(1 + \lfloor \log_2(n/2) \rfloor)} \leq \frac{\eta}{2}.$$

For the third term, define $b_\eta := \sup_{n \geq 2} \exp\left(\frac{\log \log \log(8n)}{\log^{(\eta/16)} \log^2(8n)}\right)$. By reducing $c_2 = c_2(\eta)$ if necessary, we may assume that $c_2 \leq \log(1 + \eta/4)/b_\eta$. Then

$$\begin{aligned} & \mathbb{E}\{R(l, k) \mathbb{1}_{\{|l-k| > (\eta/8) \log \log(8n)\}}\} \\ & \leq \left\{1 + \frac{s}{p} \exp\left(\frac{\log(c_2 p/s^2) + \log \log \log(8n)}{\log^{(\eta/16)} \log^2(8n)}\right)\right\}^s \\ & \quad \times \mathbb{P}\{|l-k| > (\eta/8) \log \log(8n)\} \\ & \leq e^{c_2 b_\eta} \mathbb{P}\{|l-k| > (\eta/8) \log \log(8n)\} \\ & \leq \left(1 + \frac{\eta}{4}\right) \mathbb{P}\{|l-k| > (\eta/8) \log \log(8n)\}, \end{aligned}$$

which establishes (23) when $s < \sqrt{p \log \log(8n)}$ and $s \log\left(\frac{ep \log \log(8n)}{s^2}\right) \geq \log \log(8n)$.

The final case is $s < \sqrt{p \log \log(8n)}$ and $s \log\left(\frac{ep \log \log(8n)}{s^2}\right) < \log \log(8n)$. Notice that in our definition of the parameter space $\Theta^{(t_0)}(p, n, s, \rho)$, if we restrict μ_1 and μ_2 to agree in all coordinates except perhaps the first, then the testing problem is equivalent to testing between $\Theta_0(1, n)$ and $\Theta(1, n, 1, \rho)$. Therefore, the lower bound construction in Gao et al. (2019) applies directly here and we obtain the rate $\log \log(8n)$.

The result follows. \square

The proof of Theorem 4 uses several arguments from the proof of Proposition 2.

PROOF OF THEOREM 4. We first bound $\mathbb{E}\psi_{\text{adaptive}}$ for any $\theta \in \Theta_0(p, n)$, and let $\epsilon \in (0, 1)$. By a union bound, we have

$$(30) \quad \mathbb{E}\psi_{\text{adaptive}} \leq \sum_{s \in \mathcal{S}: s < \sqrt{p \log \log(8n)}} \mathbb{E}\psi^{(s)} + \mathbb{E}\psi^{(p)}.$$

By the same argument as that used in the proof of Proposition 2, we have $\mathbb{E}\psi^{(p)} \leq \epsilon/4$ as long as $C = C(\epsilon) > 0$ is chosen sufficiently large (in particular, it will need to be at least as large as the choice of C in the proof of Proposition 2).

For $s < \sqrt{p \log \log(8n)}$, we recall that $a^2 = a^2(s) = 4 \log\left(\frac{ep \log \log(8n)}{s^2}\right)$. As in (21), we have

$$\mathbb{E}\psi^{(s)} \leq 2 \log(en) e^{-x},$$

for any x such that

$$\sqrt{\frac{s^4}{e^2 p \log^2 \log(8n)}} x + x \leq C r^*(p, n, s).$$

The choice

$$x = \frac{C}{2} \left(\frac{p \log^2 \log(8n)}{s^2} \wedge r^*(p, n, s) \right)$$

satisfies this condition provided that $C \geq 2$. This choice of x gives the tail bound

$$\mathbb{E}_\theta \psi^{(s)} \leq 2 \log(en) \exp\left(-\frac{C p \log^2 \log(8n)}{2s^2}\right) + 2 \log(en) \exp\left(-\frac{C}{2} r^*(p, n, s)\right).$$

Moreover, provided we choose $C = C(\epsilon) > 0$ sufficiently large, we have

$$\begin{aligned} & 2 \log(en) \sum_{s \in \mathcal{S}: s < \sqrt{p \log \log(8n)}} \exp\left(-\frac{C p \log^2 \log(8n)}{2s^2}\right) \\ & \leq 2 \log(en) \sum_{k=0}^{\infty} \exp\left(-\frac{C}{2} 4^k \log \log(8n)\right) \leq \frac{\epsilon}{8}. \end{aligned}$$

Similarly,

$$\begin{aligned} & 2 \log(en) \sum_{s \in \mathcal{S}: s < \sqrt{p \log \log(8n)}} \exp\left(-\frac{C}{2} r^*(p, n, s)\right) \\ & \leq 2 \log(en) \sum_{s \in \mathcal{S}: s < \sqrt{p \log \log(8n)}} \exp\left\{-\frac{C}{4} s \log\left(\frac{ep \log \log(8n)}{s^2}\right) - \frac{C}{4} \log \log(8n)\right\} \\ & \leq \frac{\epsilon}{8} \sum_{s \in [p]: s < \sqrt{p \log \log(8n)}} \exp\left\{-\frac{C}{4} s \log\left(\frac{ep \log \log(8n)}{s^2}\right)\right\} \leq \frac{\epsilon}{8}. \end{aligned}$$

Therefore we have $\mathbb{E}_\theta \psi_{\text{adaptive}} \leq \epsilon/2$ by (30).

Finally, for $\theta \in \Theta(p, n, s, \rho)$, we bound $\mathbb{E}_\theta(1 - \psi_{\text{adaptive}})$. By the definition of \mathcal{S} , there exists a unique $\tilde{s} \in \mathcal{S}$ such that $s/2 < \tilde{s} \leq s$. Moreover, $\Theta(p, n, s, \rho) \subseteq \Theta(p, n, p \wedge 2\tilde{s}, \rho)$, so by the same argument used in the proof of Proposition 2, we have

$$\mathbb{E}_\theta(1 - \psi_{\text{adaptive}}) \leq \mathbb{E}_\theta(1 - \psi^{(\tilde{s})}) \leq \frac{\epsilon}{2},$$

as long as $\rho^2 \geq 32Cr^*(p, n, p \wedge 2\tilde{s})$. But $\tilde{s} \leq s$, so $r^*(p, n, p \wedge 2\tilde{s}) \leq 2r^*(p, n, s)$, which implies that $\rho^2 \geq 64Cr^*(p, n, s)$ is a sufficient condition for controlling the error under the alternative. \square

For Theorem 5 and Theorem 6, we will prove lower and upper bounds separately.

PROOF OF THEOREM 5 (LOWER BOUND). Since this proof is asymptotic, we assume in many places and without further comment (both here and in the upper bound proof that follows) that n is sufficiently large in developing our bounds. Let $f_{0n} := \phi$ be the density function of the standard normal distribution on $\mathbb{R}^{p \times n}$. Define $f_{1n}(x) := \int_{\text{supp}(\nu_n)} \phi(x - \theta) d\nu_n(\theta)$, where ν_n is the distribution of θ when θ is generated according to the following sampling process:

1. Uniformly sample a subset $S \subseteq [p]$ of cardinality s ;
2. Independently of S , generate $k \sim \text{Unif}(\mathcal{G}_n)$ where

$$\begin{aligned} \mathcal{G}_n := \{ & 0, \lfloor c \log \log \log n \rfloor, 2 \lfloor c \log \log \log n \rfloor, \dots, \\ & \lfloor c \log \log \log n \rfloor \lfloor \log_2(\sqrt{n}) / \lfloor c \log \log \log n \rfloor \rfloor \}, \end{aligned}$$

for some constant $c > 0$ to be chosen later;

3. Independently of (S, k) , sample $u = (u_1, \dots, u_p) \in \mathbb{R}^p$, where $u_1, \dots, u_p \stackrel{\text{iid}}{\sim} \text{Unif}(\{-1, 1\})$;
4. Given the triplet (S, k, u) sampled in the previous steps, define $\theta_{j\ell} := \frac{\beta}{\sqrt{2^k}} u_j$ for all $(j, \ell) \in S \times [2^k]$ and $\theta_{j\ell} := 0$ otherwise, where $\beta^2 = (2 - \epsilon) \sqrt{\frac{p \log \log n}{s^2}}$ for some small, constant $\epsilon > 0$.

Since

$$\frac{2^k(n - 2^k)}{n} \cdot \frac{\beta^2}{2^k} \sum_{j \in S} u_j^2 \geq \frac{n - \sqrt{n}}{n} (2 - \epsilon) \sqrt{p \log \log n},$$

we have $\text{supp}(\nu_n) \subseteq \Theta(p, n, s, \rho)$ with $\rho^2 = (2 - 2\epsilon) \sqrt{p \log \log n}$, which, following the argument in the proof of Lemma 8, yields that

$$\mathcal{R}(\rho) \geq \int_{\mathbb{R}^{p \times n}} f_{0n} \wedge f_{1n}.$$

Therefore, in order to show that $\mathcal{R}(\rho) \rightarrow 1$, it suffices to establish that $\frac{f_{1n}}{f_{0n}} \rightarrow 1$ in f_{0n} -probability as $n \rightarrow \infty$. By a truncated second moment argument, given as Lemma 9, it suffices to choose some measurable set $A_n \subseteq \mathbb{R}^{p \times n}$ and establish the following two conditions:

1. $\mathbb{E}_{X \sim f_{0n}} \left\{ \frac{f_{1n}(X)}{f_{0n}(X)} \mathbb{1}_{\{X \in A_n\}} \right\} \rightarrow 1$;
2. $\mathbb{E}_{X \sim f_{0n}} \left\{ \left(\frac{f_{1n}(X)}{f_{0n}(X)} \right)^2 \mathbb{1}_{\{X \in A_n\}} \right\} \rightarrow 1$.

By definition of f_{1n} and Lemma 10, the above two conditions are equivalent to:

1. $\mathbb{P}_{X \sim f_{1n}}(X \in A_n) \rightarrow 1$;
2. $\mathbb{E}_{(\theta_1, \theta_2) \sim \nu_n \otimes \nu_n} \left\{ \mathbb{P}_{X \sim N_{p \times n}(\theta_1 + \theta_2, I)}(X \in A_n) \exp(\langle \theta_1, \theta_2 \rangle) \right\} \rightarrow 1$.

Fix $\epsilon_1 > 0$ and define $\mathcal{I}_n := \{2^i : i \in \mathcal{G}_n\}$. Then we choose the truncation set to be

$$A_n := \left\{ X \in \mathbb{R}^{p \times n} : \max_{t \in \mathcal{I}_n} \|X_1 + \dots + X_t\|_2^2 / t \leq p + (2 + \epsilon_1) \sqrt{p \log \log n} \right\}.$$

Proof of Condition 1. To prove that $\mathbb{P}_{X \sim f_{1n}}(X \in A_n) \rightarrow 1$, it suffices to show that $\sup_{\theta \in \text{supp}(\nu_n)} \mathbb{P}_\theta(X \notin A_n) \rightarrow 0$. Assume that the true changepoint location in θ is t_0 . Then, $\|X_1 + \dots + X_t\|_2^2 / t$ follows a non-central chi-squared distribution with degrees of freedom p and non-centrality parameter

$$\frac{(2 - \epsilon) \sqrt{p \log \log n}}{(t/t_0) \vee (t_0/t)}.$$

We therefore divide the time grid into two parts $\mathcal{I}_n = \mathcal{I}_n^1 \cup \mathcal{I}_n^2$, where $\mathcal{I}_n^1 := \{t_0\}$ and $\mathcal{I}_n^2 := \mathcal{I}_n \setminus \{t_0\}$. Since $\lfloor c \log \log \log n \rfloor \rightarrow \infty$, the non-centrality parameter for $t \in \mathcal{I}_n^2$ is smaller than $\frac{\epsilon_1}{2} \sqrt{p \log \log n}$. Then

$$\begin{aligned} \mathbb{P}_\theta(X \notin A_n) &\leq \sum_{t \in \mathcal{I}_n^1 \cup \mathcal{I}_n^2} \mathbb{P}_\theta \left(\frac{\|X_1 + \dots + X_t\|_2^2}{t} > p + (2 + \epsilon_1) \sqrt{p \log \log n} \right) \\ &\leq \mathbb{P} \left(\chi_{p, (2 - \epsilon) \sqrt{p \log \log n}}^2 > p + (2 + \epsilon_1) \sqrt{p \log \log n} \right) \\ &\quad + (\log n) \mathbb{P} \left(\chi_{p, \epsilon_1 \sqrt{p \log \log n} / 2}^2 > p + (2 + \epsilon_1) \sqrt{p \log \log n} \right) \\ &\rightarrow 0, \end{aligned}$$

where the last line is by Lemma 2, and we have used the fact that $\frac{\log \log n}{\sqrt{p \log \log n}} \rightarrow 0$ as $n \rightarrow \infty$ by assumption. Moreover, the bound we just obtained is uniform over all $\theta \in \text{supp}(\nu_n)$, so $\mathbb{P}_{X \sim f_{1n}}(X \in A_n) \rightarrow 1$.

Proof of Condition 2. We independently sample (S, k, u) and (T, l, v) with distribution ν_n , and define θ_1 using (S, k, u) and θ_2 using (T, l, v) . Then

$$\begin{aligned} & \mathbb{E}_{(\theta_1, \theta_2) \sim \nu_n \otimes \nu_n} \left\{ \mathbb{P}_{X \sim N_{p \times n}(\theta_1 + \theta_2, I)}(X \in A_n) \exp(\langle \theta_1, \theta_2 \rangle) \right\} \\ & \leq \mathbb{E}_{k,l} \left[\mathbb{E}_{S,u,T,v} \left\{ \mathbb{P}_{X \sim N_{p \times n}(\theta_1 + \theta_2, I)}(X \in A_n) \exp(\langle \theta_1, \theta_2 \rangle) \right\} \mathbb{1}_{\{k=l\}} \right] \\ & \quad + \mathbb{E}_{k,l} \left[\mathbb{E}_{S,u,T,v} \left\{ \exp(\langle \theta_1, \theta_2 \rangle) \right\} \mathbb{1}_{\{k \neq l\}} \right]. \end{aligned}$$

By the same calculation as in the proof of Proposition 3, the second term above can be bounded as follows:

$$\begin{aligned} \mathbb{E}_{k,l} \left\{ \mathbb{E}_{S,u,T,v} \left(e^{\langle \theta_1, \theta_2 \rangle} \right) \mathbb{1}_{\{k \neq l\}} \right\} & \leq \mathbb{E}_{k,l} \left\{ \left(1 + \frac{s}{2p} \frac{\beta^4}{2^{|l-k|}} e^{\beta^4/2^{|l-k|+1}} \right)^s \mathbb{1}_{\{k \neq l\}} \right\} \\ & \leq \mathbb{E}_{k,l} \left\{ \exp \left(\frac{s^2}{2p} \frac{\beta^4}{2^{|l-k|}} e^{\beta^4/2^{|l-k|+1}} \right) \mathbb{1}_{\{k \neq l\}} \right\}. \end{aligned}$$

Notice that $\beta \rightarrow 0$ in our asymptotic regime and therefore $e^{\beta^4/2^{|l-k|+1}} \rightarrow 1$ uniformly for any $k, l \in \mathcal{G}_n$. Further notice that when $k, l \in \mathcal{G}_n$ satisfy $l \neq k$, we have $|l - k| \geq \lfloor c \log \log \log n \rfloor$ and therefore when c is sufficiently large,

$$\frac{s^2 \beta^4}{2p 2^{|l-k|}} \leq \frac{(2 - \epsilon)^2}{2} \frac{\log \log n}{2^{\lfloor c \log \log \log n \rfloor}} \rightarrow 0.$$

Thus we have shown that

$$\limsup_{n \rightarrow \infty} \mathbb{E}_{k,l} \left\{ \mathbb{E}_{S,u,T,v} \left(e^{\langle \theta_1, \theta_2 \rangle} \right) \mathbb{1}_{\{k \neq l\}} \right\} \leq 1.$$

Next, we need to show that

$$\mathbb{E}_{k,l} \left[\mathbb{E}_{S,u,T,v} \left\{ \mathbb{P}_{X \sim N_{p \times n}(\theta_1 + \theta_2, I)}(X \in A_n) \exp(\langle \theta_1, \theta_2 \rangle) \right\} \mathbb{1}_{\{k=l\}} \right] \rightarrow 0.$$

We let $\tilde{u} \in \mathbb{R}^p$ be the vector that equals to u on S and 0 otherwise and we define \tilde{v} in the same way for v . Then

$$\begin{aligned} & \mathbb{E}_{k,l} \left[\mathbb{E}_{S,u,T,v} \left\{ \mathbb{P}_{X \sim N_{p \times n}(\theta_1 + \theta_2, I)}(X \in A_n) e^{\langle \theta_1, \theta_2 \rangle} \right\} \mathbb{1}_{\{k=l\}} \right] \\ & = \mathbb{E}_{k,l} \left[\mathbb{E}_{S,u,T,v} \left\{ \mathbb{P}_{X \sim N_{p \times n}(\theta_1 + \theta_2, I)}(X \in A_n) e^{\langle \theta_1, \theta_2 \rangle} \mathbb{1}_{\{\|\tilde{u} + \tilde{v}\|^2 \leq (2 - \epsilon_2)s\}} \right\} \mathbb{1}_{\{k=l\}} \right] \\ (31) \quad & + \mathbb{E}_{k,l} \left[\mathbb{E}_{S,u,T,v} \left\{ \mathbb{P}_{X \sim N_{p \times n}(\theta_1 + \theta_2, I)}(X \in A_n) e^{\langle \theta_1, \theta_2 \rangle} \mathbb{1}_{\{\|\tilde{u} + \tilde{v}\|^2 > (2 - \epsilon_2)s\}} \right\} \mathbb{1}_{\{k=l\}} \right]. \end{aligned}$$

We first deal with the second term. When $k = l$, we let $t_0 = 2^k = 2^l$. When $\|\tilde{u} + \tilde{v}\|^2 > (2 - \epsilon_2)s$, the noncentrality parameter of $\|X_1 + \dots + X_{t_0}\|^2/t_0$ is

$$\beta^2 \|\tilde{u} + \tilde{v}\|^2 > (2 - \epsilon)(2 - \epsilon_2) \sqrt{p \log \log n}.$$

By Lemma 2, on the event that $k = l$ and $\|\tilde{u} + \tilde{v}\|^2 > (2 - \epsilon_2)s$, we have

$$\begin{aligned} \mathbb{P}_{X \sim N_{p \times n}(\theta_1 + \theta_2, I)}(X \in A_n) & \leq \mathbb{P} \left(\chi_{p, (2 - \epsilon)(2 - \epsilon_2) \sqrt{p \log \log n}}^2 \leq p + (2 + \epsilon_1) \sqrt{p \log \log n} \right) \\ & \leq \exp \left\{ -(1 + o(1)) \left(\frac{(2 - \epsilon)(2 - \epsilon_2) - (2 + \epsilon_1)}{2} \right)^2 \log \log n \right\} \\ & = \left(\frac{1}{\log n} \right)^{(1 + o(1)) \left(\frac{(2 - \epsilon)(2 - \epsilon_2) - (2 + \epsilon_1)}{2} \right)^2}. \end{aligned}$$

In addition, we also have

$$\mathbb{P}(k = l) = \frac{1}{\lceil \log_2(\sqrt{n}) / \lfloor c \log \log \log n \rfloor \rceil + 1}.$$

Therefore,

$$\begin{aligned} & \mathbb{E}_{k,l} \left[\mathbb{E}_{S,u,T,v} \left\{ \mathbb{P}_{X \sim N_{p \times n}(\theta_1 + \theta_2, I)} (X \in A_n) e^{(\theta_1, \theta_2)} \mathbb{1}_{\{\|\tilde{u} + \tilde{v}\|^2 > (2 - \epsilon_2)s\}} \right\} \mathbb{1}_{\{k=l\}} \right] \\ & \leq \left(\frac{1}{\log n} \right)^{(1+o(1)) \left(\frac{(2-\epsilon)(2-\epsilon_2) - (2+\epsilon_1)}{2} \right)^2} \exp \left(\frac{s^2 \beta^4}{2p} e^{\beta^4} \right) \mathbb{P}(k = l) \\ & = \frac{(\log n)^{(1+o(1)) \frac{(2-\epsilon)^2}{2}}}{(\log n)^{(1+o(1)) \left[\left(\frac{(2-\epsilon)(2-\epsilon_2) - (2+\epsilon_1)}{2} \right)^2 + 1 \right]}}. \end{aligned}$$

By choosing $\epsilon_1, \epsilon_2 > 0$ sufficiently small, we can ensure that the power of the $\log n$ factor in the denominator is arbitrarily close to $2 - 2\epsilon + \epsilon^2$, while the power of the $\log n$ factor in the numerator is arbitrarily close to $2 - 2\epsilon + \epsilon^2/2$. We deduce that

$$\mathbb{E}_{k,l} \left[\mathbb{E}_{S,u,T,v} \left\{ \mathbb{P}_{X \sim N_{p \times n}(\theta_1 + \theta_2, I)} (X \in A_n) e^{(\theta_1, \theta_2)} \mathbb{1}_{\{\|\tilde{u} + \tilde{v}\|^2 > (2 - \epsilon_2)s\}} \right\} \mathbb{1}_{\{k=l\}} \right] \rightarrow 0.$$

Finally, we analyse the first term on the right-hand side of (31). By the Cauchy–Schwarz inequality, we first obtain the bound

$$\begin{aligned} & \mathbb{E}_{k,l} \left[\mathbb{E}_{S,u,T,v} \left\{ \mathbb{P}_{X \sim N_{p \times n}(\theta_1 + \theta_2, I)} (X \in A_n) e^{(\theta_1, \theta_2)} \mathbb{1}_{\{\|\tilde{u} + \tilde{v}\|^2 \leq (2 - \epsilon_2)s\}} \right\} \mathbb{1}_{\{k=l\}} \right] \\ & \leq (\mathbb{E} e^{2(\theta_1, \theta_2)})^{1/2} \sqrt{\mathbb{P}(\|\tilde{u} + \tilde{v}\|^2 \leq (2 - \epsilon_2)s)}. \end{aligned}$$

Now, by construction,

$$\mathbb{E} e^{2(\theta_1, \theta_2)} \leq \mathbb{E} e^{\|\theta_1\|_{\mathbb{F}}^2 + \|\theta_2\|_{\mathbb{F}}^2} \leq e^{2s\beta^2} = e^{2(2-\epsilon)\sqrt{p \log \log n}}.$$

To bound $\mathbb{P}(\|\tilde{u} + \tilde{v}\|^2 \leq (2 - \epsilon_2)s)$, we condition on the event that $|S \cap T| = a \leq s$ and $|S \setminus T| = |T \setminus S| = s - a$. Denoting the distribution of a by r , we then have by Hoeffding's inequality that

$$\begin{aligned} \mathbb{P}(\|\tilde{u} + \tilde{v}\|^2 \leq (2 - \epsilon_2)s) &= \mathbb{E}_{a \sim r} \mathbb{P}(2(s - a) + 4 \cdot \text{Bin}(a, 1/2) \leq (2 - \epsilon_2)s \mid a) \\ &\leq \max_{a \leq s} \mathbb{P} \left(|\text{Bin}(a, 1/2) - a/2| \geq \frac{\epsilon_2}{4} s \mid a \right) \\ &\leq \max_{a \leq s} 2e^{-\frac{\epsilon_2^2 s^2}{8a}} = 2e^{-\epsilon_2^2 s/8}. \end{aligned}$$

Therefore,

$$\begin{aligned} & \mathbb{E}_{k,l} \left[\mathbb{E}_{S,u,T,v} \left\{ \mathbb{P}_{X \sim N_{p \times n}(\theta_1 + \theta_2, I)} (X \in A_n) e^{(\theta_1, \theta_2)} \mathbb{1}_{\{\|\tilde{u} + \tilde{v}\|^2 \leq (2 - \epsilon_2)s\}} \right\} \mathbb{1}_{\{k=l\}} \right] \\ & \leq 2e^{2(2-\epsilon)\sqrt{p \log \log n} - \epsilon_2^2 s/8} \rightarrow 0. \end{aligned}$$

This verifies Condition 2 and hence concludes our lower bound calculation. \square

PROOF OF THEOREM 5 (UPPER BOUND). For the upper bound, define

$$(32) \quad \tilde{Y}_t := \sqrt{\frac{t(n-t)}{n}} \left(\frac{1}{t} (X_1 + \dots + X_t) - \frac{1}{n-t} (X_{t+1} + \dots + X_n) \right).$$

Assume that $\rho^2 \geq 2\sqrt{(1+\epsilon)p \log \log n}$ for some constant $\epsilon \in (0, 1]$. For $\delta_1, \delta_2 > 0$, consider the test

$$\psi := \mathbb{1}_{\left\{ \max_{t \in \mathcal{T}_{\delta_2}} \|\tilde{Y}_t\|_2^2 - p \geq 2\sqrt{(1+\delta_1)p \log \log n} \right\}},$$

where $\mathcal{T}_{\delta_2} := \mathcal{T}_{\delta_2}^1 \cup \mathcal{T}_{\delta_2}^2$ with

$$\mathcal{T}_{\delta_2}^1 := \left\{ 1, \lfloor 1 + \delta_2 \rfloor, \lfloor (1 + \delta_2)^2 \rfloor, \dots, \lfloor (1 + \delta_2)^{\lfloor \log_{1+\delta_2}(n/2) \rfloor} \rfloor \right\},$$

and $\mathcal{T}_{\delta_2}^2 = \{n - t : t \in \mathcal{T}_{\delta_2}^1\}$. Under the null hypothesis, for a fixed t , we have $\|\tilde{Y}_t\|_2^2 \sim \chi_p^2$. Therefore, by Lemma 1 a union bound argument, we have

$$\begin{aligned} \sup_{\theta \in \Theta_0(p, n)} \mathbb{P}_\theta \left(\max_{t \in \mathcal{T}_{\delta_2}} \|\tilde{Y}_t\|_2^2 - p \geq 2\sqrt{(1+\delta_1)p \log \log n} \right) \\ \leq 2 \lfloor \log_{1+\delta_2}(n/2) \rfloor e^{-(1+o(1))(1+\delta_1) \log \log n} \rightarrow 0. \end{aligned}$$

Under the alternative hypothesis, for $\theta \in \Theta(p, n, s, \rho)$, assume that the true changepoint is at t_0 and that the mean vector before and after the changepoint is μ_1 and μ_2 , respectively. Assume $t_0 \leq n/2$ without loss of generality. Let $\tilde{t}_0 = \tilde{t}_0(\theta)$ be the closest point in \mathcal{T}_{δ_2} to t_0 such that $\tilde{t}_0 \leq t_0$. Then $\tilde{t}_0 \leq t_0 \leq (1 + \delta_2)\tilde{t}_0$. By the assumption that $\rho^2 \geq 2\sqrt{(1+\epsilon)p \log \log n}$, we have $\|\tilde{Y}_{\tilde{t}_0}\|_2^2 \sim \chi_{p, \lambda}^2$, where

$$\begin{aligned} \lambda = \frac{\tilde{t}_0(n - \tilde{t}_0)}{n} \left\| \frac{n - t_0}{n - \tilde{t}_0} (\mu_1 - \mu_2) \right\|^2 &\geq \frac{\tilde{t}_0(n - t_0)}{n} \|\mu_1 - \mu_2\|^2 \geq \frac{\tilde{t}_0}{t_0} \rho^2 \\ (33) \qquad \qquad \qquad &\geq \frac{\rho^2}{1 + \delta_2} \geq 2\sqrt{(1 + \epsilon/2)p \log \log n} \end{aligned}$$

if we choose $\delta_2 = \epsilon/8$. By Lemma 2, we therefore have

$$\begin{aligned} \sup_{\theta \in \Theta(p, n, s, \rho)} \mathbb{P}_\theta \left(\|\tilde{Y}_{\tilde{t}_0}\|_2^2 - p \leq 2\sqrt{(1+\delta_1)p \log \log n} \right) &\leq e^{-(1+o(1))(\sqrt{1+\epsilon/2} - \sqrt{1+\delta_1})^2 \log \log n} \\ &\rightarrow 0, \end{aligned}$$

as long as $\delta_1 \in (0, \epsilon/2)$. This completes the proof of the upper bound. \square

PROOF OF THEOREM 6 (LOWER BOUND). As in the proof of Theorem 5, we will allow n to be sufficiently large. Let $f_{0n} := \phi$ be the density function of the standard normal distribution on $\mathbb{R}^{p \times n}$. Define $f_{1n}(x) := \int_{\text{supp}(\nu_n)} \phi(x - \theta) d\nu_n(\theta)$, where ν_n is the distribution of θ when θ is generated according to the following sampling process:

1. Uniformly sample a subset $S \subseteq [p]$ of cardinality s ;
2. Independently of S , generate $k \sim \text{Unif}(\mathcal{G}_n)$ where

$$\begin{aligned} \mathcal{G}_n := \{0, \lfloor \log \log \log n \rfloor, 2 \lfloor \log \log \log n \rfloor, \dots, \\ \lfloor \log \log \log n \rfloor \lfloor \log_2(\sqrt{n}) / \lfloor \log \log \log n \rfloor \rfloor \}; \end{aligned}$$

3. Given (S, k) sampled in the previous steps, define $\theta_{j\ell} := \frac{\beta}{\sqrt{2^k}}$ for all $(j, \ell) \in S \times [2^k]$ and $\theta_{j\ell} := 0$ otherwise, where $\beta^2 = (1 - \epsilon) \log \left(\frac{p \log \log n}{s^2} \right)$ for some $\epsilon \in (0, 1)$.

Similar to our previous arguments, we let (S, k) and (T, l) be independent with distribution ν_n . Since $\frac{2^k(n-2^k)}{n} \cdot \frac{\beta^2}{2^k} |S| \geq \frac{n-\sqrt{n}}{n} (1-\epsilon) s \log \left(\frac{p \log \log n}{s^2} \right)$, we have $\text{supp}(\nu_n) \subseteq \Theta(p, n, s, \rho)$ with $\rho^2 = (1-2\epsilon) s \log \left(\frac{p \log \log n}{s^2} \right)$. Therefore, by Lemmas 8 and 10, it is sufficient to show

$$\limsup_{n \rightarrow \infty} \mathbb{E}_{(\theta_1, \theta_2) \sim \nu_n \otimes \nu_n} \exp(\langle \theta_1, \theta_2 \rangle) \leq 1.$$

By the same calculation that leads to (29), we have

$$\mathbb{E}_{(\theta_1, \theta_2) \sim \nu_n \otimes \nu_n} \left\{ \exp(\langle \theta_1, \theta_2 \rangle) \right\} \leq \mathbb{E}_{(l, k)} \left\{ \exp \left(\frac{s^2}{p} e^{\beta^2/2^{|l-k|/2}} \right) \right\}.$$

We split the right hand side above into two terms according to whether or not $l = k$. When $l \neq k$, we have $|l - k| \geq \lfloor \log \log \log n \rfloor$ due to the definition of ν_n . Therefore,

$$\mathbb{E}_{(l, k)} \left\{ \exp \left(\frac{s^2}{p} e^{\beta^2/2^{|l-k|/2}} \right) \mathbb{1}_{\{l \neq k\}} \right\} \leq \exp \left(\frac{s^2}{p} \left(\frac{p \log \log n}{s^2} \right)^{1/2 \lfloor \log \log \log n \rfloor / 2} \right) \rightarrow 1,$$

where we have used $s^2/p \rightarrow 0$ and $(\log \log n)^{1/\log \log n} \rightarrow 1$. When $l = k$, we have

$$\mathbb{E}_{(l, k)} \left\{ \exp \left(\frac{s^2}{p} e^{\beta^2/2^{|l-k|/2}} \right) \mathbb{1}_{\{l=k\}} \right\} \leq \frac{1}{\lfloor \log_2(\sqrt{n}) / \lfloor \log \log \log n \rfloor \rfloor} \exp \left(\frac{s^2}{p} e^{\beta^2} \right) \rightarrow 0,$$

according to the definition of β . Combining the two bounds, we have obtained the desired conclusion. \square

PROOF OF THEOREM 6 (UPPER BOUND). Consider the statistic

$$\tilde{A}_{t,a} = \sum_{j=1}^p (\tilde{Y}_t^2(j) - \nu_a) \mathbb{1}_{\{|\tilde{Y}_t(j)| \geq a\}},$$

where the definition of \tilde{Y}_t is given by (32). Recall the definition of \mathcal{T}_{δ_2} in the upper bound proof of Theorem 5. We then consider the testing procedure

$$\tilde{\psi} := \mathbb{1}_{\left\{ \max_{t \in \mathcal{T}_{\delta_2}} \tilde{A}_{t,a} \geq C^* \left(\sqrt{p e^{-a^2/2} 2 \log \log n} + 2 \log \log n \right) \right\}},$$

where $a = \sqrt{2 \log \left(\frac{p \log \log n}{s^2} \right)}$, and the constant $C^* > 0$ is taken from Lemma 5. Under the null hypothesis, i.e. for any $\theta \in \Theta_0(p, n)$, by Lemma 5 and a union bound argument, we have

$$\sup_{\theta \in \Theta_0(p, n)} \mathbb{E}_{\theta} \tilde{\psi} \leq 2 \left(\lfloor \log_{1+\delta_2}(n/2) \rfloor + 1 \right) \exp(-2 \log \log n) \rightarrow 0.$$

Next, we study $\mathbb{E}_{\theta}(1 - \tilde{\psi})$ under the alternative hypothesis. For $\theta \in \Theta(p, n, s, \rho)$ with ρ satisfying $\rho^2 \geq 2(1+\epsilon) s \log \left(\frac{p \log \log n}{s^2} \right)$, we denote the changepoint of θ by t_0 , and the mean vector before and after the changepoint by μ_1 and μ_2 , respectively. Assume that $t_0 \leq n/2$ without loss of generality. Let $\tilde{t}_0 = \tilde{t}_0(\theta)$ be the closest point in \mathcal{T}_{δ_2} to t_0 such that $\tilde{t}_0 \leq t_0$. Then we have $\tilde{t}_0 \leq t_0 \leq (1+\delta_1)\tilde{t}_0$. Write $\Delta_j := \mathbb{E}\{\tilde{Y}_{\tilde{t}_0}(j)\}$ as shorthand. By the assumption that $\rho^2 \geq 2(1+\epsilon) s \log \left(\frac{p \log \log n}{s^2} \right)$ and the same argument we have used in (33), we have

$$\sum_{j=1}^p \Delta_j^2 \geq \frac{\rho^2}{1+\delta_2} \geq 2(1+\epsilon/2) s \log \left(\frac{p \log \log n}{s^2} \right),$$

by choosing $\delta_2 = \epsilon/8$. By (S3) of Lemma 6, for any $\delta_1 > 0$, there exists $c_1^* = c_1^*(\delta_1) > 0$ such that

$$\begin{aligned} \mathbb{E}_\theta \tilde{A}_{t_0, a} &\geq c_1^* \sum_{j=1}^p \Delta_j^2 \mathbb{1}_{\{|\Delta_j| \geq (1+\delta_1)a\}} \\ &= c_1^* \left(\sum_{j=1}^p \Delta_j^2 - \sum_{j=1}^p \Delta_j^2 \mathbb{1}_{\{|\Delta_j| < (1+\delta_1)a\}} \right) \\ &\geq c_1^* \left\{ 2(1 + \epsilon/2)s \log \left(\frac{p \log \log n}{s^2} \right) - (1 + \delta_1)^2 s a^2 \right\} \\ &= \frac{c_1^* \epsilon}{2} s \log \left(\frac{p \log \log n}{s^2} \right), \end{aligned}$$

by choosing δ_1 such that $2(1 + \delta_1)^2 = 1 + \epsilon/2$. For the variance, due to Lemma 7, we have

$$\text{Var}_\theta(\tilde{A}_{t_0, a}) = O\left(pa^3 e^{-a^2/2} + sa^4\right).$$

Hence, following the same argument used in (22), we have

$$\begin{aligned} \sup_{\theta \in \Theta(p, n, s, \rho)} \mathbb{E}_\theta(1 - \tilde{\psi}) &\leq \sup_{\theta \in \Theta(p, n, s, \rho)} \mathbb{P}_\theta \left(\tilde{A}_{t_0, a} < C^* \left(\sqrt{pe^{-a^2/2} 2 \log \log n} + 2 \log \log n \right) \right) \\ &\rightarrow 0, \end{aligned}$$

which leads to the desired conclusion. \square

5.2. Proofs of results in Section 3.

PROOF OF THEOREM 7. For any $\theta \in \Theta(p, n, p, \rho)$, there exist $\mu_1, \mu_2 \in \mathbb{R}^p$ and $t \in [n]$, such that $X_1, \dots, X_t \stackrel{\text{iid}}{\sim} N_p(\mu_1, \Sigma)$ and $X_{t+1}, \dots, X_n \stackrel{\text{iid}}{\sim} N_p(\mu_2, \Sigma)$. The covariance matrix Σ admits the eigenvalue decomposition $\Sigma = U \Lambda U^T$ for some orthogonal $U \in \mathbb{R}^{p \times p}$ and $\Lambda = \text{diag}(\lambda) \in \mathbb{R}^{p \times p}$, where $\lambda := (\lambda_1, \dots, \lambda_p)^T$ and $\lambda_1 \geq \dots \geq \lambda_p > 0$. Then $U^T X_1, \dots, U^T X_t \stackrel{\text{iid}}{\sim} N_p(U^T \mu_1, \Lambda)$ and $U^T X_{t+1}, \dots, U^T X_n \stackrel{\text{iid}}{\sim} N_p(U^T \mu_2, \Lambda)$. Moreover, $\|U^T(\mu_1 - \mu_2)\| = \|\mu_1 - \mu_2\|$, so we can consider a diagonal Σ without loss of generality. From now on, we assume that $\Sigma = \Lambda$.

We first derive the upper bound. Consider the testing procedure

$$\psi = \mathbb{1}_{\{\max_{t \in \mathcal{T}} \|Y_t\|^2 - \text{Tr}(\Sigma) > r\}},$$

with $r = C \left(\sqrt{\|\Sigma\|_F^2 \log \log(8n)} + \|\Sigma\|_{\text{op}} \log \log(8n) \right)$ for some appropriate $C > 0$. Then the same argument in the proof of Proposition 2 together with Lemma 1 leads to the desired result.

We now derive the lower bound. We first seek to apply Lemmas 8 and 10 and given $\eta > 0$, find a probability measure ν with $\text{supp}(\nu) \subseteq \Theta(p, n, p, \rho)$ and a universal constant $c > 0$ such that

$$(34) \quad \mathbb{E}_{(\theta_1, \theta_2) \sim \nu \otimes \nu} \exp(\langle \theta_1, \theta_2 \rangle_{\Sigma^{-1}}) \leq 1 + \eta,$$

whenever $\rho = c\rho_\Sigma^*$. We define ν to be the distribution of $\theta = (\theta_{j\ell}) \in \Theta(p, n, p, \rho)$, sampled according to the following process:

1. Uniformly sample $k \in \{0, 1, 2, \dots, \lfloor \log_2(n/2) \rfloor\}$;

2. Independently of k , sample $u = (u_1, \dots, u_p)^T \in \mathbb{R}^p$ with independent coordinates, and with $u_j \sim \text{Unif}(\{-a_j, a_j\})$ for $j \in [p]$;
3. Given (k, u) sampled in the previous steps, define $\theta_{j\ell} := 2^{-k/2}u_j$ for all $(j, \ell) \in [p] \times [2^k]$ and $\theta_{j\ell} := 0$ otherwise.

If $\theta \sim \nu$, then $\theta \in \Theta(p, n, p, \rho)$ with $\rho^2 = \frac{1}{2} \sum_{j=1}^p a_j^2$. Suppose that we independently sample (k, u) and (l, v) from the first two steps and use these to construct θ_1 and θ_2 respectively according to the third step. Then, by direct calculation, we obtain

$$\langle \theta_1, \theta_2 \rangle_{\Sigma^{-1}} = (2^k \wedge 2^l) \frac{1}{\sqrt{2^{k+l}}} \sum_{j=1}^p \frac{u_j v_j}{\lambda_j} = \frac{1}{2^{|k-l|/2}} \sum_{j=1}^p \frac{u_j v_j}{\lambda_j}.$$

Observe that $u_j v_j \sim \text{Unif}(\{-a_j^2, a_j^2\})$, so

$$\begin{aligned} \mathbb{E}_{(\theta_1, \theta_2) \sim \nu \otimes \nu} \exp(\langle \theta_1, \theta_2 \rangle_{\Sigma^{-1}}) &= \mathbb{E} \exp\left(\frac{1}{2^{|k-l|/2}} \sum_{j=1}^p \frac{u_j v_j}{\lambda_j}\right) \\ &= \mathbb{E} \prod_{j=1}^p \left\{ \frac{1}{2} \exp\left(\frac{a_j^2}{2^{|k-l|/2} \lambda_j}\right) + \frac{1}{2} \exp\left(-\frac{a_j^2}{2^{|k-l|/2} \lambda_j}\right) \right\} \\ &\leq \mathbb{E} \exp\left(\frac{1}{2^{|k-l|}} \sum_{j=1}^p \frac{a_j^4}{\lambda_j^2}\right), \end{aligned}$$

where the last inequality above uses the fact that $(e^x + e^{-x})/2 \leq e^{x^2}$. We take $a_j^2 = \sqrt{c_1 \frac{\lambda_j^4 \log \log(8n)}{\|\lambda\|^2}}$ for some sufficiently small $c_1 = c_1(\eta) > 0$. Then it can be shown that

$$\mathbb{E} \exp\left(\frac{1}{2^{|k-l|}} \sum_{j=1}^p \frac{a_j^4}{\lambda_j^2}\right) = \mathbb{E} \exp\left(c_1 \frac{\log \log(8n)}{2^{|k-l|}}\right) \leq 1 + \eta$$

using very similar arguments to those employed in the proof of Proposition 2. We have therefore established (34), which implies the desired lower bound $\rho^2 = \frac{1}{2} \sum_{j=1}^p a_j^2 \asymp \sqrt{\|\Sigma\|_{\text{F}}^2 \log \log(8n)}$.

We also need to prove the lower bound $\|\Sigma\|_{\text{op}} \log \log(8n)$. Recall that we have assumed without loss of generality that Σ is diagonal with non-increasing diagonal elements. Then in our definition of the parameter space $\Theta^{(t_0)}(p, n, s, \rho)$, if we restrict μ_1 and μ_2 to agree in all coordinates except perhaps the first, then the testing problem is equivalent to testing between $\Theta_0(1, n)$ and $\Theta(1, n, 1, \rho)$ with variance $\lambda_1 = \|\Sigma\|_{\text{op}}$. Therefore, the lower bound construction in Gao et al. (2019) directly applies here and we obtain the desired rate $\|\Sigma\|_{\text{op}} \log \log(8n)$. \square

PROOF OF PROPOSITION 8. Suppose the index set \mathcal{D} does not include the changepoint. Then, by Lemma 3, we have that for every $x > 0$,

$$(35) \quad |\text{Tr}(\widehat{\Sigma}_{\mathcal{D}}) - \text{Tr}(\Sigma)| \leq 4 \left(\frac{\sqrt{x} \|\Sigma\|_{\text{F}}}{\sqrt{n}} + \frac{x \|\Sigma\|_{\text{op}}}{n} \right),$$

with probability at least $1 - 2e^{-x}$ (notice that substituting n for $n - 1$ means we multiply the right-hand side by at most 2). We will take $x = p \log(32/\epsilon)$, which guarantees that $e^{-x} \leq \epsilon/32$. Moreover, there exists a universal constant $\tilde{C} > 0$, such that for all $x \geq 1$

$$(36) \quad \|\widehat{\Sigma}_{\mathcal{D}} - \Sigma\|_{\text{op}} \leq \tilde{C} \|\Sigma\|_{\text{op}} \left(\sqrt{\frac{p}{n}} \vee \frac{p}{n} \vee \sqrt{\frac{x}{n}} \vee \frac{x}{n} \right),$$

with probability at least $1 - e^{-x}$ (Koltchinskii and Lounici, 2017, Theorem 1). Here we will take $x = p \log(16/\epsilon)$. From this we immediately have the error bounds for $\|\widehat{\Sigma}_{\mathcal{D}}\|_{\text{F}}$ and $\|\widehat{\Sigma}_{\mathcal{D}}\|_{\text{op}}$, because

$$\left| \|\widehat{\Sigma}_{\mathcal{D}}\|_{\text{op}} - \|\Sigma\|_{\text{op}} \right| \leq \|\widehat{\Sigma}_{\mathcal{D}} - \Sigma\|_{\text{op}},$$

and

$$\left| \|\widehat{\Sigma}_{\mathcal{D}}\|_{\text{F}} - \|\Sigma\|_{\text{F}} \right| \leq \|\widehat{\Sigma}_{\mathcal{D}} - \Sigma\|_{\text{F}} \leq \sqrt{p} \|\widehat{\Sigma}_{\mathcal{D}} - \Sigma\|_{\text{op}}.$$

Since there is only one changepoint, there exists an event of probability at least $1 - \epsilon/8$ on which at least two blocks among $\mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_3$ satisfy (35), and an event of probability at least $1 - \epsilon/8$ on which at least two blocks satisfy (36). The result therefore follows on taking $C = 4 \log(32/\epsilon) + \tilde{C}(c^{1/2} \vee 1) \log(16/\epsilon)$. \square

PROOF OF COROLLARY 9. Define a set of good events

$$G := \left\{ \left| \text{Tr}(\widehat{\Sigma})^{(2)} - \text{Tr}(\Sigma) \right| \leq (\|\Sigma\|_{\text{F}} + \|\Sigma\|_{\text{op}}) / 4, \right. \\ \left. \left| \|\widehat{\Sigma}\|_{\text{F}}^{(2)} - \|\Sigma\|_{\text{F}} \right| \leq \|\Sigma\|_{\text{F}} / 4, \quad \left| \|\widehat{\Sigma}\|_{\text{op}}^{(2)} - \|\Sigma\|_{\text{op}} \right| \leq \|\Sigma\|_{\text{op}} / 4 \right\}.$$

As a direct application of Proposition 8, given $\epsilon > 0$, there exists $c > 0$, depending only on A and ϵ , such that $\mathbb{P}_{\theta}(G^c) \leq \epsilon/4$ for any $\theta \in \Theta(p, n, p, 0)$. Hence, for $\theta \in \Theta_0(p, n)$, when $C \geq 1$, we have

$$\mathbb{E}_{\theta} \psi_{\text{Cov}} \leq \mathbb{P}_{\theta} \left(\left\{ \max_{t \in \mathcal{T}} \|Y_t\|^2 - \text{Tr}(\widehat{\Sigma})^{(2)} \right. \right. \\ \left. \left. > C \left(\|\widehat{\Sigma}\|_{\text{F}}^{(2)} \sqrt{\log \log(8n)} + \|\widehat{\Sigma}\|_{\text{op}}^{(2)} \log \log(8n) \right) \right\} \cap G \right) + \mathbb{P}_{\theta}(G^c) \\ \leq \mathbb{P}_{\theta} \left(\max_{t \in \mathcal{T}} \|Y_t\|^2 - \text{Tr}(\Sigma) > \frac{C}{2} \left(\|\Sigma\|_{\text{F}} \sqrt{\log \log(8n)} + \|\Sigma\|_{\text{op}} \log \log(8n) \right) \right) + \frac{\epsilon}{4}.$$

Therefore, by Theorem 7, we can choose $C = C(\epsilon) \geq 1$ large enough that the error under the null is at most $\epsilon/2$. A very similar argument also applies to $\mathbb{E}_{\theta}(1 - \psi_{\text{Cov}})$ for $\theta \in \Theta(p, n, p, \rho)$ with $\rho > 0$: when $\rho^2 \geq 64C \left(\|\Sigma\|_{\text{F}} \sqrt{\log \log(8n)} \vee \|\Sigma\|_{\text{op}} \log \log(8n) \right)$ and after increasing $C = C(\epsilon)$ if necessary, the error under the alternative is at most $\epsilon/2$, as required. \square

PROOF OF THEOREM 10. Recalling the representation of $Y_t(j)$ in (12), we define an oracle version of \tilde{Y}_t in (13) by

$$\tilde{Y}_t := \frac{Y_t - \sqrt{\gamma} W_t \mathbf{1}_p}{\sqrt{1 - \gamma}}.$$

Then

$$(37) \quad \|\tilde{Y}_t - \bar{Y}_t\|_{\infty} = \frac{|\text{Median}(Y_t) - \sqrt{\gamma} W_t|}{\sqrt{1 - \gamma}}.$$

By Lemma 11, there exist universal constants $C_1, C_2, C_3 > 0$ such that for any $\theta \in \Theta(p, n, s, 0)$, we have

$$\mathbb{P}_{\theta} \left\{ \frac{|\text{Median}(Y_t) - \sqrt{\gamma} W_t|}{\sqrt{1 - \gamma}} > C_1 \left(\frac{s}{p} + \sqrt{\frac{1+x}{p}} \right) \right\} \leq e^{-C_2 x},$$

as long as $C_3 \left(\frac{s}{p} + \sqrt{\frac{1+x}{p}} \right) \leq 1$. Using (37) and a union bound argument, there exists a universal constant $C_4 > 0$ such that

$$(38) \quad \max_{t \in \mathcal{T}} \|\tilde{Y}_t - \bar{Y}_t\|_\infty \leq C_4 \sqrt{\frac{\log \log(8n)}{p}},$$

with \mathbb{P}_θ -probability at least $1 - 1/\log^2(en)$, for any $\theta \in \Theta(p, n, s, 0)$, under the conditions $s \leq (p \log \log(8n))^{1/5}$ and $\frac{\log \log(8n)}{p} \leq c$. From now on, the event that (38) holds is denoted by G .

With the above preparations, we can analyze $\mathbb{E}_\theta \psi$ for any $\theta \in \Theta_0(p, n)$. Recalling the definition of $g_a(\cdot)$ in (14), we set C' in (14) to be C_4 in (38). Then, on the event G , we have $g_a(\tilde{Y}_t(j)) \leq f_a(\bar{Y}_t(j))$ for $j \in [p]$, and therefore given $\epsilon > 0$ we can choose $C = C(\epsilon) > 0$ in the definition of r and $n_0 = n_0(\epsilon) \in \mathbb{N}$ such that

$$\mathbb{E}_\theta \psi_{a,r,C'} \leq \mathbb{E}_\theta(\psi \mathbb{1}_G) + \mathbb{P}_\theta(G^c) \leq \mathbb{P}_\theta \left(\max_{t \in \mathcal{T}} \sum_{j=1}^p f_a(\bar{Y}_t(j)) > r \right) + \frac{1}{\log^2(en)} \leq \frac{\epsilon}{2},$$

for $n \geq n_0$, where the last inequality is by the same argument as in (21) in the proof of Proposition 2.

Now we analyze $\mathbb{E}_\theta(1 - \psi_{a,r,C'})$ for $\theta \in \Theta(p, n, s, \rho)$. Recall from the proof of Proposition 2 that given any $\theta \in \Theta(p, n, s, \rho)$, we may assume there exists $t_0 \leq n/2$ such that $\frac{t_0(n-t_0)}{n} \|\mu_1 - \mu_2\|^2 \geq \rho^2$; moreover, there exists a unique $\tilde{t} \in \mathcal{T}$ such that $t_0/2 < \tilde{t} \leq t_0$, and

$$\|\Delta_{\tilde{t}}\|^2 = \frac{\tilde{t} \|\mu_1 - \mu_2\|^2}{2} \geq \frac{t_0 \|\mu_1 - \mu_2\|^2}{4} \geq \frac{t_0(n-t_0)}{4n} \|\mu_1 - \mu_2\|^2 \geq \frac{\rho^2}{4}.$$

We introduce a function

$$h_a(x) := \inf \left\{ f_a(y) : |x - y| \leq \frac{a}{10} \right\} = \begin{cases} 0, & |x| \leq \frac{9}{10}a, \\ a^2 - \nu_a, & \frac{9}{10}a < |x| \leq \frac{11}{10}a, \\ (|x| - \frac{a}{10})^2 - \nu_a, & |x| > \frac{11}{10}a. \end{cases}$$

To gain some intuition, a plot of the functions $h_1(\cdot)$ and $f_1(\cdot)$ is shown in Figure 1. By reducing $c > 0$ if necessary, we may assume that $2C' \sqrt{\frac{\log \log(8n)}{p}} \leq \frac{a}{10}$, so we have on the event G that $g_a(\tilde{Y}_t(j)) \geq h_a(\bar{Y}_t(j))$ for $j \in [p]$. Thus

$$\begin{aligned} \mathbb{E}_\theta(1 - \psi_{a,r,C'}) &\leq \mathbb{E}_\theta \left\{ (1 - \psi_{a,r,C'}) \mathbb{1}_G \right\} + \mathbb{P}_\theta(G^c) \\ &\leq \mathbb{P}_\theta \left(\sum_{j=1}^p h_a(\bar{Y}_{\tilde{t}}(j)) \leq r \right) + \frac{1}{\log^2(en)}, \end{aligned}$$

and we now control the first term on the right-hand side. When $\Delta_{\tilde{t}}(j) = 0$, we have $\mathbb{E} h_a(\bar{Y}_{\tilde{t}}(j)) \leq \mathbb{E} f_a(\bar{Y}_{\tilde{t}}(j)) = 0$. Moreover, by Lemma 4,

$$\begin{aligned} & - \mathbb{E} h_a(\bar{Y}_{\tilde{t}}(j)) \\ &= 2(\nu_a - a^2) \{ \Phi(11a/10) - \Phi(9a/10) \} + 2 \int_{11a/10}^{\infty} \{ \nu_a - (x - a/10) \} \phi(x) dx \\ &\leq \frac{4a^3}{5} \phi(9a/10) + 6a^2 \{ 1 - \Phi(11a/10) \} \lesssim e^{-a^2/3}. \end{aligned}$$

Next, for $0 < |\Delta_{\tilde{t}}(j)| < 8(1 - \gamma)^{1/2}a$, we have $\mathbb{E} h_a(\bar{Y}_{\tilde{t}}(j)) \geq -(\nu_a - a^2) \geq -2a^2$, and by Lemma 6, we have $\mathbb{E} h_a(\bar{Y}_{\tilde{t}}(j)) \leq \mathbb{E} f_a(\bar{Y}_{\tilde{t}}(j)) \leq \frac{64\Delta_{\tilde{t}}^2(j)}{1-\gamma} + 1 \lesssim a^2$.

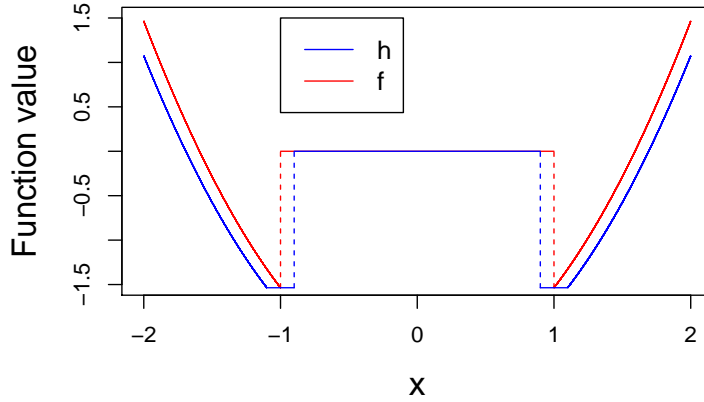


Fig 1: An illustration of the functions $f_a(\cdot)$ and $h_a(\cdot)$ for the special case $a = 1$.

Finally, we handle the case where $|\Delta_{\bar{t}}(j)| \geq 8(1 - \gamma)^{1/2}a$, and assume without loss of generality that $\Delta_{\bar{t}}(j) \geq 8(1 - \gamma)^{1/2}a$. Observe by Lemma 4 that for $x \geq 4a$, we have

$$(x - a/10)^2 - \nu_a \geq x^2 - \frac{ax}{5} - 3a^2 \geq x^2 - \frac{x^2}{20} - \frac{3x^2}{16} \geq 3x^2/4.$$

Hence

$$\begin{aligned} \mathbb{E}h_a(\bar{Y}_{\bar{t}}(j)) &\geq \frac{3}{4}\mathbb{E}(\bar{Y}_{\bar{t}}(j))^2\mathbb{1}_{\{\bar{Y}_{\bar{t}}(j) \geq 4a\}} - (\nu_a - a^2)\mathbb{P}(\bar{Y}_{\bar{t}}(j) < 4a) \\ &\geq \frac{3\Delta_{\bar{t}}^2(j)}{4(1 - \gamma)}\mathbb{P}(\bar{Y}_{\bar{t}}(j) \geq 4a) - 3a^2\mathbb{P}(\bar{Y}_{\bar{t}}(j) < 4a) \geq \frac{45\Delta_{\bar{t}}^2(j)}{128(1 - \gamma)}. \end{aligned}$$

Summarising then, we have

$$\mathbb{E}h_a(\bar{Y}_{\bar{t}}(j)) \begin{cases} \leq 0 \text{ and } \gtrsim -e^{-a^2/3} & \text{if } \Delta_{\bar{t}}(j) = 0, \\ \gtrsim -a^2 \text{ and } \lesssim a^2 & \text{if } 0 < |\Delta_{\bar{t}}(j)| < 8(1 - \gamma)^{1/2}a, \\ \geq \frac{45\Delta_{\bar{t}}^2(j)}{128(1 - \gamma)} & \text{if } |\Delta_{\bar{t}}(j)| \geq 8(1 - \gamma)^{1/2}a. \end{cases}$$

We now study $\text{Var} h_a(\bar{Y}_{\bar{t}}(j))$. When $\Delta_{\bar{t}}(j) = 0$, we have

$$\begin{aligned} \text{Var} h_a(\bar{Y}_{\bar{t}}(j)) &\leq \mathbb{E}h_a^2(\bar{Y}_{\bar{t}}(j)) \\ &\leq 2 \int_{9a/10}^{\infty} \{(\nu_a - a^2)^2 \vee ((x - a/10)^2 - \nu_a)^2\} \phi(x) dx \\ &\leq 2 \int_{9a/10}^{\infty} (\nu_a^2 \vee x^2) \phi(x) dx \lesssim e^{-a^2/4}. \end{aligned}$$

When $0 < |\Delta_{\bar{t}}(j)| < 2(1 - \gamma)^{1/2}a$, assuming that $\Delta_{\bar{t}}(j) > 0$ without loss of generality and writing $\theta := \Delta_{\bar{t}}(j)/(1 - \gamma)^{1/2}$ as shorthand, we have

$$\begin{aligned} \text{Var} h_a(\bar{Y}_{\bar{t}}(j)) &\leq \mathbb{E}h_a^2(\bar{Y}_{\bar{t}}(j)) \\ &\leq \left(\int_{-\infty}^{-\frac{9a}{10}} + \int_{\frac{9a}{10}}^{\theta+a} + \int_{\theta+a}^{\infty} \right) \{(\nu_a - a^2)^2 \vee ((|x| - a/10)^2 - \nu_a)^2\} \phi(x - \theta) dx \end{aligned}$$

$$\lesssim e^{-a^2/4} + a^4 + e^{-a^2/4} \lesssim a^4.$$

Finally, when $|\Delta_{\bar{t}}(j)| \geq 2(1-\gamma)^{1/2}a$, Let us define a random variable $L := \mathbb{1}_{\{\bar{Y}_{\bar{t}}(j) \geq 11a/10\}}$. Then assuming that $\Delta_{\bar{t}}(j) \geq 2(1-\gamma)^{1/2}a$ without loss of generality, we have that

$$\begin{aligned} & \text{Var } h_a(\bar{Y}_{\bar{t}}(j)) \\ &= \mathbb{E}\{\text{Var}(h_a(\bar{Y}_{\bar{t}}(j)) \mid L)\} + \text{Var}\{\mathbb{E}(h_a(\bar{Y}_{\bar{t}}(j)) \mid L)\} \\ &\leq \mathbb{P}(L=0)\mathbb{E}\{h_a^2(\bar{Y}_{\bar{t}}(j)) \mid L=0\} + \mathbb{P}(L=1)\text{Var}\{(\bar{Y}_{\bar{t}}(j) - a/10)^2 \mid L=1\} \\ &\quad + \mathbb{P}(L=0)\mathbb{P}(L=1) \\ &\quad \times \left\{ \left| \mathbb{E}\{h_a(\bar{Y}_{\bar{t}}(j)) \mid L=0\} \right| + \left| \mathbb{E}\{(\bar{Y}_{\bar{t}}(j) - a/10)^2 - \nu_a \mid L=1\} \right| \right\}^2. \end{aligned}$$

Now, similar to the proof of Lemma 7,

$$\begin{aligned} \left| \mathbb{E}\{h_a(\bar{Y}_{\bar{t}}(j)) \mid L=0\} \right| &\leq \nu_a + \mathbb{E}\{\bar{Y}_{\bar{t}}^2(j) \mid \bar{Y}_{\bar{t}}(j) < 11a/10\} \lesssim \frac{\Delta_{\bar{t}}^2(j)}{1-\gamma}, \\ \mathbb{E}\{h_a^2(\bar{Y}_{\bar{t}}(j)) \mid L=0\} &\leq 2\nu_a^2 + 2\mathbb{E}\{\bar{Y}_{\bar{t}}^4(j) \mid \bar{Y}_{\bar{t}}(j) < 11a/10\} \lesssim \frac{\Delta_{\bar{t}}^4(j)}{(1-\gamma)^2}, \\ \left| \mathbb{E}\{(\bar{Y}_{\bar{t}}(j) - a/10)^2 - \nu_a \mid L=1\} \right| &\leq \nu_a + \mathbb{E}\{\bar{Y}_{\bar{t}}^2(j) \mid \bar{Y}_{\bar{t}}(j) \geq 11a/10\} \lesssim \frac{\Delta_{\bar{t}}^2(j)}{1-\gamma}. \end{aligned}$$

But $\mathbb{P}(L=0) = \Phi(11a/10 - \Delta_{\bar{t}}(j)/(1-\gamma)^{1/2}) \leq \Phi\left(\frac{-9\Delta_{\bar{t}}(j)}{20(1-\gamma)^{1/2}}\right)$. Finally, we note that

$$\mathbb{P}(L=1)\text{Var}\{(\bar{Y}_{\bar{t}}(j) - a/10)^2 \mid L=1\} \leq \text{Var}\{(\bar{Y}_{\bar{t}}(j) - a/10)^2\} \lesssim \frac{\Delta_{\bar{t}}^2(j)}{1-\gamma}.$$

These observations allow us to deduce that

$$\text{Var}(h_a(\bar{Y}_{\bar{t}}(j))) \lesssim \begin{cases} e^{-a^2/4} & \text{if } \Delta_{\bar{t}}(j) = 0, \\ a^4 & \text{if } 0 < |\Delta_{\bar{t}}(j)| < 2(1-\gamma)^{1/2}a, \\ \frac{\Delta_{\bar{t}}(j)^2}{1-\gamma} & \text{if } |\Delta_{\bar{t}}(j)| \geq 2(1-\gamma)^{1/2}a. \end{cases}$$

The bound on the expectation then implies that

$$\sum_{j:\Delta_{\bar{t}}(j)=0} |\mathbb{E}h_a(\bar{Y}_{\bar{t}}(j))| \lesssim pe^{-a^2/3} \lesssim p \log \log(8n) \left(\frac{s^2}{p \log \log(8n)} \right)^{4/3} \leq s,$$

where we used the condition $s \leq (p \log \log(8n))^{1/5}$. We deduce similarly to the argument in the proof of Proposition 2 that

$$\sum_{j=1}^p \mathbb{E}h_a(\bar{Y}_{\bar{t}}(j)) \geq \frac{\|\Delta_{\bar{t}}\|^2}{4(1-\gamma)}$$

provided we choose $C = C(\epsilon) > 0$ sufficiently large in the definition of ρ . Moreover,

$$\sum_{j=1}^p \text{Var}(h_a(\bar{Y}_{\bar{t}}(j))) \lesssim pe^{-a^2/4} + sa^4 + \frac{\|\Delta_{\bar{t}}\|^2}{1-\gamma}.$$

By Chebychev's inequality, we deduce that

$$\begin{aligned} \mathbb{P}_\theta \left(\sum_{j=1}^p h_a(\bar{Y}_t(j)) \leq r \right) &\leq \mathbb{P}_\theta \left(\sum_{j=1}^p h_a(\bar{Y}_t(j)) \leq \frac{\|\Delta_{\bar{t}}\|^2}{8} \right) \\ &\lesssim \frac{pe^{-a^2/4} + sa^4 + \frac{\|\Delta_{\bar{t}}\|^2}{1-\gamma}}{\{\|\Delta_{\bar{t}}\|^2/(1-\gamma)\}^2} \lesssim \frac{pe^{-a^2/4} + sa^4 + \frac{\rho^2}{1-\gamma}}{\{\rho^2/(1-\gamma)\}^2}. \end{aligned}$$

Hence, by increasing $n_0 = n_0(\epsilon)$ and $C = C(\epsilon) > 0$ if necessary, we may conclude that $\mathbb{E}_\theta(1 - \psi_{a,r,C'}) \leq \epsilon/2$, as required. \square

PROOF OF THEOREM 11. The proof uses similar arguments to those in the proof of Proposition 3, but instead of establishing (23), we need to show that given $\eta > 0$, we can find a universal constant $c > 0$ such that $\mathbb{E}_{(\theta_1, \theta_2) \sim \nu \otimes \nu} \exp(\langle \theta_1, \theta_2 \rangle_{\Sigma(\gamma)^{-1}}) \leq 1 + \eta$ when $\rho \geq cr_{\Sigma(\gamma)}^*$, where $r_{\Sigma(\gamma)}^*$ denotes the right-hand side of (16). Since

$$(39) \quad \Sigma(\gamma)^{-1} = \kappa_1(\gamma)I_p - \kappa_2(\gamma)\mathbf{1}_p\mathbf{1}_p^T,$$

with $\kappa_1(\gamma) = \frac{1}{1-\gamma}$ and $\kappa_2(\gamma) = \frac{\gamma}{(1-\gamma)(1+(p-1)\gamma)}$, the calculation will be very similar, and essentially our argument replaces I_p in the proof of Proposition 3 by $\kappa_1(\gamma)I_p$.

First consider the case when $s \leq \sqrt{p \log \log(8n)}$ and $s \log \left(\frac{ep \log \log(8n)}{s^2} \right) \geq \log \log(8n)$. We define ν to be the distribution of θ , sampled according to the following process:

1. Uniformly sample a subset S of $[p]$ of cardinality s ;
2. Independently, sample k according to a uniform distribution on $\{0, 1, 2, \dots, \lfloor \log_2(n/2) \rfloor\}$;
3. Given (S, k) sampled in the previous steps, define $\theta_{j\ell} := \beta/2^{k/2}$ for all $(j, \ell) \in S \times [2^k]$ and $\theta_{j\ell} := 0$ otherwise, where $\beta > 0$.

Suppose that we generate θ_1 and θ_2 independently with distribution ν , where θ_1 is generated from (S, k) and θ_2 comes from (T, l) . By (39), we have $\langle \theta_1, \theta_2 \rangle_{\Sigma(\gamma)^{-1}} \leq \frac{\kappa_1(\gamma)\beta^2}{2^{|l-k|/2}} |S \cap T|$, and thus

$$\mathbb{E}_{(\theta_1, \theta_2) \sim \nu \otimes \nu} \exp(\langle \theta_1, \theta_2 \rangle_{\Sigma(\gamma)^{-1}}) \leq \mathbb{E} \exp \left(\frac{\kappa_1(\gamma)\beta^2}{2^{|l-k|/2}} |S \cap T| \right).$$

Note that we obtain the same formula as (28) except that the β^2 in (28) is replaced by $\kappa_1(\gamma)\beta^2$. This immediately implies that the same argument that bounds (28) can also be applied here and we obtain the lower bound with the desired rate $\kappa_1(\gamma)^{-1} s \log \left(\frac{ep \log \log(8n)}{s^2} \right)$.

Next we consider the case $s \leq \sqrt{p \log \log(8n)}$ and $s \log \left(\frac{ep \log \log(8n)}{s^2} \right) < \log \log(8n)$. The sampling process for $\theta \sim \nu$ is now:

1. Sample k from a uniform distribution on $\{0, 1, 2, \dots, \lfloor \log_2(n/2) \rfloor\}$;
2. Given k , define $\theta_{j\ell} := \beta/2^{k/2}$ for all $(j, \ell) \in [s] \times [2^k]$ and $\theta_{j\ell} := 0$ otherwise.

Similarly to before, $\langle \theta_1, \theta_2 \rangle_{\Sigma(\gamma)^{-1}} \leq \frac{\kappa_1(\gamma)\beta^2}{2^{|l-k|/2}} s$, and thus

$$\mathbb{E}_{(\theta_1, \theta_2) \sim \nu \otimes \nu} \exp(\langle \theta_1, \theta_2 \rangle_{\Sigma(\gamma)^{-1}}) \leq \mathbb{E} \exp \left(\frac{\kappa_1(\gamma)\beta^2}{2^{|l-k|/2}} s \right).$$

We can then set $\kappa_1(\gamma)\beta^2 s = c_1 \log \log(8n)$ for a sufficiently small $c_1 > 0$, and apply the same argument as in the proof of Gao et al. (2019, Proposition 4.2). The lower bound follows with rate $\rho^2 = s\beta^2 \asymp \kappa_1(\gamma)^{-1} \log \log(8n)$. \square

Acknowledgements. The authors thank the anonymous referees for helpful comments, and in particular for suggesting an improvement that led to the current version of Theorem 5.

SUPPLEMENTARY MATERIAL

Supplement: Supplement to “Minimax rates in sparse, high-dimensional change-point detection”
([link to be added](#)).

REFERENCES

- Arias-Castro, E., Candès, E. J. and Durand, A. (2011). Detection of an anomalous cluster in a network, *The Annals of Statistics* **39**(1): 278–304.
- Arias-Castro, E., Donoho, D. L. and Huo, X. (2005). Near-optimal detection of geometric objects by fast multiscale methods, *IEEE Transactions on Information Theory* **51**(7): 2402–2425.
- Aston, J. A. and Kirch, C. (2012). Evaluating stationarity via change-point alternatives with applications to fMRI data, *The Annals of Applied Statistics* **6**(4): 1906–1948.
- Aston, J. A. and Kirch, C. (2018). High dimensional efficiency with applications to change point tests, *Electronic Journal of Statistics* **12**(1): 1901–1947.
- Aue, A., Hörmann, S., Horváth, L. and Reimherr, M. (2009). Break detection in the covariance structure of multivariate time series models, *The Annals of Statistics* **37**(6B): 4046–4087.
- Bai, J. (2010). Common breaks in means and variances for panel data, *Journal of Econometrics* **157**(1): 78–92.
- Basseville, M. and Nikiforov, I. V. (1993). *Detection of Abrupt Changes: Theory and Application*, Vol. 104, Prentice Hall Englewood Cliffs.
- Berthet, Q. and Rigollet, P. (2013). Optimal detection of sparse principal components in high dimension, *The Annals of Statistics* **41**(4): 1780–1815.
- Chen, Y., Wang, T. and Samworth, R. J. (2020). High-dimensional, multiscale online changepoint detection, *arXiv preprint arXiv:2003.03668*.
- Cho, H. (2016). Change-point detection in panel data via double CUSUM statistic, *Electronic Journal of Statistics* **10**(2): 2000–2038.
- Cho, H. and Fryzlewicz, P. (2015). Multiple-change-point detection for high dimensional time series via sparsified binary segmentation, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **77**(2): 475–507.
- Collier, O., Comminges, L. and Tsybakov, A. B. (2017). Minimax estimation of linear and quadratic functionals on sparsity classes, *The Annals of Statistics* **45**(3): 923–958.
- Cribben, I. and Yu, Y. (2017). Estimating whole-brain dynamics by using spectral clustering, *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **66**(3): 607–627.
- Csörgő, M. and Horváth, L. (1997). *Limit Theorems in Change-Point Analysis*, Wiley, Chichester.
- Donoho, D. and Jin, J. (2004). Higher criticism for detecting sparse heterogeneous mixtures, *The Annals of Statistics* **32**(3): 962–994.
- Enikeeva, F. and Harchaoui, Z. (2019). High-dimensional change-point detection under sparse alternatives, *The Annals of Statistics* **47**(4): 2051–2079.
- Frick, K., Munk, A. and Sieling, H. (2014). Multiscale change point inference, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **76**(3): 495–580.
- Gao, C., Han, F. and Zhang, C.-H. (2019). On estimation of isotonic piecewise constant signals, *The Annals of Statistics, to appear*.
- Hall, P. and Jin, J. (2010). Innovated higher criticism for detecting sparse signals in correlated noise, *The Annals of Statistics* **38**(3): 1686–1732.
- Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables, *Journal of the American Statistical Association* **58**(301): 13–30.
- Horváth, L. and Hušková, M. (2012). Change-point detection in panel data, *Journal of Time Series Analysis* **33**(4): 631–648.
- Jirak, M. (2015). Uniform change point tests in high dimension, *The Annals of Statistics* **43**(6): 2451–2483.
- Killick, R., Fearnhead, P. and Eckley, I. A. (2012). Optimal detection of changepoints with a linear computational cost, *Journal of the American Statistical Association* **107**(500): 1590–1598.
- Kirch, C., Muhsal, B. and Ombao, H. (2015). Detection of changes in multivariate time series with application to eeg data, *Journal of the American Statistical Association* **110**(511): 1197–1216.
- Koltchinskii, V. and Lounici, K. (2017). Concentration inequalities and moment bounds for sample covariance operators, *Bernoulli* **23**(1): 110–133.

- Liu, H., Gao, C. and Samworth, R. J. (2020). Supplement to “Minimax rates in sparse, high-dimensional change-point detection”.
- Ombao, H., Von Sachs, R. and Guo, W. (2005). Slex analysis of multivariate nonstationary time series, *Journal of the American Statistical Association* **100**(470): 519–531.
- Page, E. (1955). A test for a change in a parameter occurring at an unknown point, *Biometrika* **42**(3-4): 523–527.
- Peng, T., Leckie, C. and Ramamohanarao, K. (2004). Proactively detecting distributed denial of service attacks using source IP address monitoring, in N. Mitrou, K. Kontovasilis, G. N. Rouskas, I. I. and L. Merakos (eds), *Networking 2004*, Springer, Berlin, pp. 771–782.
- Wang, D., Yu, Y. and Rinaldo, A. (2017). Optimal covariance change point localization in high dimension, *arXiv preprint arXiv:1712.09912* .
- Wang, D., Yu, Y. and Rinaldo, A. (2018a). Optimal change point detection and localization in sparse dynamic networks, *arXiv preprint arXiv:1809.09602* .
- Wang, D., Yu, Y. and Rinaldo, A. (2018b). Univariate mean change point detection: Penalization, CUSUM and optimality, *arXiv preprint arXiv:1810.09498* .
- Wang, T. and Samworth, R. J. (2018). High dimensional change point estimation via sparse projection, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **80**(1): 57–83.
- Xie, Y. and Siegmund, D. (2013). Sequential multi-sensor change-point detection, *The Annals of Statistics* **41**(2): 670–692.
- Zhang, N. R., Siegmund, D. O., Hanlee, J. and Li, J. Z. (2010). Detecting simultaneous changepoints in multiple sequences, *Biometrika* **97**(3): 631–645.