The Annals of Statistics
2020, Vol. 48, No. 3, 1815–1847
https://doi.org/10.1214/19-AOS1869

AN ADAPTABLE GENERALIZATION OF HOTELLING'S T^2 TEST IN HIGH DIMENSION

BY HAORAN LI[1,*], ALEXANDER AUE[1,**], DEBASHIS PAUL[1,†], JIE PENG[1,‡] AND PEI WANG[2]

[1]*Department of Statistics, University of California, Davis,* [*]*hrli@ucdavis.edu;* [**]*aaue@ucdavis.edu;* [†]*debpaul@ucdavis.edu;* [‡]*jiepeng@ucdavis.edu*

[2]*Icahn Institute, Department of Genetics and Genomic Sciences, School of Medicine at Mount Sinai,* pei.wang@mssm.edu

We propose a two-sample test for detecting the difference between mean vectors in a high-dimensional regime based on a ridge-regularized Hotelling's T^2. To choose the regularization parameter, a method is derived that aims at maximizing power within a class of local alternatives. We also propose a composite test that combines the optimal tests corresponding to a specific collection of local alternatives. Weak convergence of the stochastic process corresponding to the ridge-regularized Hotelling's T^2 is established and used to derive the cut-off values of the proposed test. Large sample properties are verified for a class of sub-Gaussian distributions. Through an extensive simulation study, the composite test is shown to compare favorably against a host of existing two-sample test procedures in a wide range of settings. The performance of the proposed test procedures is illustrated through an application to a breast cancer data set where the goal is to detect the pathways with different DNA copy number alterations across breast cancer subtypes.

1. Introduction. The focus of this paper is on the classical problem of testing for equality of means of two populations having an unknown but equal covariance matrix, when dimension is comparable to sample size. The standard solution to the two-sample testing problem is the well-known Hotelling's T^2 test (Anderson (1984), Muirhead (1982)). In spite of its central role in classical multivariate statistics, Hotelling's T^2 test has several limitations when dealing with data whose dimension p is comparable to, or larger than, the sum $n = n_1 + n_2$ of the two sample sizes n_1 and n_2. The test statistic is not defined for $p > n$ because of the singularity of the sample covariance matrix, but the test is also known to perform poorly in cases for which $p < n$ with p/n close to unity. For example, Bai and Saranadasa (1996) showed that the test is inconsistent in the asymptotic regime $p/n \to \gamma \in (0, 1)$.

Many approaches have been proposed in the literature to correct for the inconsistency of Hotelling's T^2 in high dimensions. One approach seeks to construct modified test statistics based on replacing the quadratic form involving the inverse sample covariance matrix with appropriate estimators of the squared distance between (rescaled) population means (Bai and Saranadasa (1996), Srivastava and Du (2008), Srivastava (2009), Dong et al. (2016), Chen and Qin (2010)). A different approach involves considering random projections of the data into a certain low-dimensional space and then using the Hotelling's T^2 statistics computed from the projected data (Lopes, Jacob and Wainwright (2011), Srivastava, Li and Ruppert (2016)).

Among other approaches to the problem under the "dense alternative" setting, Biswas and Ghosh (2014) considered nonparametric, graph-based two-sample tests and Chakraborty and Chaudhuri (2017) robust testing procedures. A different line of research involves assuming

Received May 2018; revised January 2019.
MSC2010 subject classifications. Primary 62J99; secondary 60B20.
Key words and phrases. Asymptotic property, covariance matrix, Hotelling's T^2 statistic, hypothesis testing, locally most powerful tests, random matrix theory.

certain forms of sparsity for the difference of mean vectors. Cai, Liu and Xia (2014) used this framework, in addition assuming that a "good" estimate of the precision matrix is available, and constructed tests based on the maximum componentwise mean difference of suitably transformed observations. Xu et al. (2016) proposed an adaptive two-sample test based on the class of ℓ_q-norms of the difference between sample means. Other recent contributions exploiting sparsity assumptions in high dimensions include Wang, Peng and Li (2015), Gregory et al. (2015), Chen, Li and Zhong (2014), Chang, Zhou and Zhou (2014) and Guo and Chen (2016). A different approach, exemplified by Gretton et al. (2012), formulates the test of equality of two populations in terms of a kernel-based discrepancy measure, with the kernel chosen adaptively from a collection of kernels. Despite similarities in terms of the use of power maximization as the principle behind selecting the regularization scheme, their work differs considerably from ours, in that here focus is on developing a data-driven procedure for selection of the regularization parameter for a regularized version of Hotelling's T^2 test for testing equality of the mean vectors for two populations of high-dimensional observations, as detailed in the next paragraph.

In this paper, we work under the scenario $p/n \to \gamma \in (0, \infty)$, assuming that the two sample sizes are asymptotically proportional. The proposed test statistic is built upon the *Regularized Hotelling's* T^2 *(RHT)* statistic introduced in Chen et al. (2011) for the one-sample case, but significantly extends its scope. The first major contribution of this work is to provide a Bayesian framework to analyze the power of the RHT, using a class of priors that captures the interaction between mean difference μ and population covariance Σ. This allows for the analytic study of power under local alternatives even without knowledge of Σ, in turn enabling the construction of a data-driven selection mechanism for the regularization parameter. Within this framework, it is also shown that the test of Bai and Saranadasa (1996) is the limit of a minimax RHT test. The second main contribution is the construction of a new composite test by combining the RHT statistics corresponding to a set of optimally chosen regularization parameters. This data-adaptive selection of λ allows the proposed test to have excellent power characteristics under various scenarios, such as different levels of decay of eigenvalues of Σ, and various types of structure of μ. We validate this property through extensive simulations involving a host of alternatives covering a wide range of mean and covariance structures. The proposed method has excellent empirical performance even when p is significantly larger than n. Because of these properties, and since the prefixes "robust" and "adaptive" are already part of the statistical nomenclature tied to specific contexts, the new composite testing procedure is termed "adaptable RHT," abbreviated as ARHT. We also establish the weak convergence of a normalized version of the stochastic process $(\mathrm{RHT}(\lambda) : \lambda \in C)$ to a Gaussian limit, where $C \subset \mathbb{R}_+$ is a compact interval. This result facilitates computation of the cut-off values for the ARHT test.

As a final key contribution, we establish the asymptotic behavior of the test by relaxing the assumption of Gaussianity to sub-Gaussianity. Establishing this result is nontrivial due to the lack of independence between sample mean and covariance matrix in non-Gaussian settings. Moreover, it is shown that a simple monotone transformation of the test statistic, or a χ^2 approximation, can significantly enhance the finite-sample behavior of the proposed tests.

The rest of the paper is organized as follows. Section 2 introduces the RHT statistic and studies a class of local alternatives. The adaptable RHT (ARHT) test statistic is considered in Section 3. Section 4 discusses finite-sample adjustments. Asymptotic analysis in the non-Gaussian case is given in Section 5. A simulation study is reported in Section 6 and an application to breast cancer data is described in Section 7. Section 8 has additional discussions. Proofs of the main theorems are presented in Section 9, and some auxiliary results are stated in the Appendix. Further technical details and additional simulation results are collected in the Supplementary Material (Li et al. (2019)). The tests provided in this paper have been implemented in the R package `ARHT`, which may be downloaded from the CRAN website.

2. Regularized Hotelling's T^2 test.

2.1. *Two-sample RHT.* This section introduces the two-sample regularized Hotelling's T^2 statistic. It is first assumed that $X_{ij} \sim \mathcal{N}(\mu_i, \Sigma)$, $j = 1, \dots, n_i$, $i = 1, 2$, are two independent samples with common $p \times p$ nonnegative population covariance $\Sigma \equiv \Sigma_p$. More general sub-Gaussian observations will be treated in Section 5. The matrix Σ can be estimated by its empirical counterpart, the "pooled" sample covariance matrix $S_n = (n-2)^{-1} \sum_{i=1}^{2} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)(X_{ij} - \bar{X}_i)^T$, where $n = n_1 + n_2$, $\bar{X}_i$ is the sample mean of the ith sample, and ^{T} is used to denote transposition of matrices and vectors. This framework has been assumed in much of the work on high-dimensional mean testing problems (Bai and Saranadasa (1996), Cai, Liu and Xia (2014)). The proposed test procedure is applicable even when the assumption of common population covariance is violated, although implications for the power characteristics of the test will be context-specific.

Due to the singularity of S_n when $p > n$, it is proposed to test $H_0\colon \mu_1 = \mu_2$ based on the family of ridge-regularized Hotelling's T^2 statistics

$$\mathrm{RHT}(\lambda) = \frac{n_1 n_2}{n_1 + n_2} (\bar{X}_1 - \bar{X}_2)^T (S_n + \lambda I_p)^{-1} (\bar{X}_1 - \bar{X}_2), \tag{1}$$

indexed by a tuning parameter $\lambda > 0$ controlling the regularization strength.

The limiting behavior of $\mathrm{RHT}(\lambda)$ is tied to the spectral properties of Σ. Let $\tau_{1,p} \geq \cdots \geq \tau_{p,p} \geq 0$ be the eigenvalues of Σ and $H_p(\tau) = p^{-1} \sum_{\ell=1}^{p} \mathbf{1}_{[\tau_{\ell,p}, \infty)}(\tau)$ its Empirical Spectral Distribution (ESD). The following assumptions are made.

C1 Σ_p is nonnegative definite and $\limsup_p \tau_{1,p} < \infty$;

C2 *High-dimensional setting*: $p, n \to \infty$ such that $n_1/n \to \kappa \in (0, 1)$, $\gamma_n = p/n \to \gamma \in (0, \infty)$ and $\sqrt{n} |p/n - \gamma| \to 0$;

C3 *Asymptotic stability of PSD*: $H_p(\tau)$ converges as $p \to \infty$ to a probability distribution function $H(\tau)$ at every point of continuity of H, and H is nondegenerate at 0. Moreover, $\sqrt{n} \| H_p - H \|_\infty \to 0$.

Since $\lambda > 0$ and in view of (1), it suffices in **C1** to require nonnegative definiteness of Σ_p rather than positive definiteness. The condition $\limsup_p \tau_{1,p} < \infty$ is necessary to obtain eigenvalue bounds. Condition **C2** ensures a well-balanced sampling design and defines the asymptotic regime in a way that dimensionality p and sample sizes n_1 and n_2 grow proportionately. Condition **C3** restricts the variability allowed in H_p as p increases, the $\sqrt{n}$-rate of convergence being a technical requirement needed to represent the asymptotic distribution of the normalized RHT statistics in terms of functionals of the *Population Spectral Distribution* (*PSD*) H.

Let I_p be the $p \times p$ identity matrix and, for $z \in \mathbb{C}$, denote by $R_n(z) = (S_n - z I_p)^{-1}$ and $m_{F_{n,p}}(z) = p^{-1} \operatorname{tr}\{R_n(z)\}$ the resolvent and *Stieltjes transform* of the ESD of S_n (see, e.g., Bai and Silverstein (2010) for more details). It is well known that $m_{F_{n,p}}(z)$ converges pointwise almost surely on $\mathbb{C}_+ = \{z = u + \imath v\colon v > 0\}$ to a nonrandom limiting distribution with Stieltjes transform $m_F(z)$ given as solution to the equation $m_F(z) = \int [\tau\{1 - \gamma - \gamma z m_F(z)\} - z]^{-1} \, dH(\tau)$. This convergence holds even when $z \in \mathbb{R}_-$ and m_F has a smooth extension to the negative reals. Following the same calculations as in Chen et al. (2011), under **C1**–**C3**, asymptotic mean and variance of the two-sample $\mathrm{RHT}(\lambda)$ under Gaussianity, are (up to multiplicative constants), given by

$$\Theta_1(\lambda, \gamma) = \frac{1 - \lambda m_F(-\lambda)}{1 - \gamma\{1 - \lambda m_F(-\lambda)\}}, \tag{2}$$

$$\Theta_2(\lambda, \gamma) = \frac{1 - \lambda m_F(-\lambda)}{[1 - \gamma\{1 - \lambda m_F(-\lambda)\}]^3} - \lambda \frac{\{m_F(-\lambda) - \lambda m_F'(-\lambda)\}}{[1 - \gamma\{1 - \lambda m_F(-\lambda)\}]^4}. \tag{3}$$

Moreover, the asymptotic normality of $\mathrm{RHT}(\lambda)$ can be established.

These expressions are derived by making use of the following key fact: for every fixed $\lambda > 0$, the random matrix $R_n(-\lambda) = (S_n + \lambda I)^{-1}$ has a *deterministic equivalent* (Bai and Silverstein (2010), Liu, Aue and Paul (2015), Paul and Aue (2014)) given by

$$D_p(-\lambda) = \left(\frac{1}{1+\gamma\Theta_1(\lambda,\gamma)}\Sigma_p + \lambda I_p\right)^{-1} \tag{4}$$

in the sense that, for symmetric matrices A bounded in operator norm,

$$\frac{1}{p}\operatorname{tr}\{R_n(-\lambda)A\} - \frac{1}{p}\operatorname{tr}\{D_p(-\lambda)A\} \to 0 \quad \text{with probability 1, as } n \to \infty. \tag{5}$$

These results hold more generally under the sub-Gaussian model described in Section 5.

Suppose $\Theta_j(\lambda,\gamma)$ is replaced with its empirical version $\hat{\Theta}_j(\lambda,\gamma_n)$ by substituting $m_F(-\lambda)$ with $m_{F_{n,p}}(-\lambda)$ and $m'_F(-\lambda)$ with $m'_{F_{n,p}}(-\lambda) = p^{-1}\operatorname{tr}\{R_n^2(-\lambda)\}$. Since $\hat{\Theta}_j(\lambda,\gamma_n)$ are $\sqrt{p}$-consistent estimators for $\Theta_j(\lambda,\gamma)$, $j = 1, 2$, the RHT test rejects the null hypothesis of equal means at asymptotic level $\alpha \in (0, 1)$ if

$$T_{n,p}(\lambda) = \sqrt{p}\frac{\{p^{-1}\operatorname{RHT}(\lambda) - \hat{\Theta}_1(\lambda,\gamma_n)\}}{\{2\hat{\Theta}_2(\lambda,\gamma_n)\}^{1/2}} > \xi_\alpha, \tag{6}$$

where ξ_α is the $1-\alpha$ quantile of the standard normal distribution $\mathcal{N}(0, 1)$.

REMARK 2.1. The test statistic by Bai and Saranadasa (1996) can be viewed as a limiting case of $\operatorname{RHT}(\lambda)$ as $\lambda \to \infty$. Specifically, observe that

$$T_{n,p}(\lambda) = \sqrt{p}\frac{p^{-1}\lambda\operatorname{RHT}(\lambda) - \lambda\hat{\Theta}_1(\lambda,\gamma_n)}{\{2\lambda^2\hat{\Theta}_2(\lambda,\gamma_n)\}^{1/2}},$$

while for any given observations X_{ij}, $j = 1, \dots, n_i$, $i = 1, 2$, as $\lambda \to \infty$,

$$\lambda(\bar{X}_1 - \bar{X}_2)^T(S_n + \lambda I_p)^{-1}(\bar{X}_1 - \bar{X}_2) \longrightarrow (\bar{X}_1 - \bar{X}_2)^T(\bar{X}_1 - \bar{X}_2),$$

$$\lambda\hat{\Theta}_1(\lambda,\gamma_n) \longrightarrow \frac{1}{p}\operatorname{tr}(S_n),$$

$$\lambda^2\hat{\Theta}_2(\lambda,\gamma_n) \longrightarrow \frac{1}{p}\operatorname{tr}(S_n^2) - \gamma_n\left[\frac{1}{p}\operatorname{tr}(S_n)\right]^2.$$

This implies that, as $\lambda \to \infty$, $\lambda\operatorname{RHT}(\lambda)$, $\lambda\hat{\Theta}_1(\lambda,\gamma_n)$, and $\lambda^2\hat{\Theta}_2(\lambda,\gamma_n)$ converge pointwise to the corresponding counterparts in the test of Bai and Saranadasa (1996) as given in display (4.5) of their paper, applying a rescaling of $\operatorname{RHT}(\lambda)$ to match their notation.

2.2. *Asymptotic power.* This subsection deals with the behavior of $\operatorname{RHT}(\lambda)$ under local alternatives, which is critical for the determination of an optimal regularization parameter λ. Defining $\mu = \mu_1 - \mu_2$, consider first a sequence of alternatives satisfying

$$\sqrt{n}\mu^T D_p(-\lambda)\mu \to q(\lambda,\gamma) \tag{7}$$

as $n \to \infty$ for some $q(\lambda,\gamma) > 0$, where $D_p(-\lambda)$ is the deterministic equivalent defined in (4). The following result determines the limit of the power function

$$\beta_n(\mu,\lambda) = \mathbb{P}_\mu\{T_{n,p}(\lambda) > \xi_\alpha\} \tag{8}$$

of the $\operatorname{RHT}(\lambda)$ test with asymptotic level α, where $\mathbb{P}_\mu$ denotes the distribution under μ.

THEOREM 2.1. *Suppose that* **C1**–**C3** *and* (7) *hold. Then, for any* $\lambda > 0$,

$$\beta_n(\mu, \lambda) \to \Phi\left(-\xi_\alpha + \kappa(1-\kappa)\frac{q(\lambda,\gamma)}{\{2\gamma\Theta_2(\lambda,\gamma)\}^{1/2}}\right) \quad (n \to \infty), \tag{9}$$

where Φ *denotes the standard normal CDF and* $\Theta_2(\lambda, \gamma)$ *is defined in* (3).

REMARK 2.2. (a) Let $\mathbf{E}_j$ denote the eigenprojection matrix associated with the jth largest eigenvalue $\tau_{j,p}$ of Σ_p. Suppose that there exists a sequence of functions $f_p : \mathbb{R}^+ \cup \{0\} \to \mathbb{R}^+ \cup \{0\}$ satisfying $f_p(\tau_{j,p}) = \sqrt{n} p \|\mathbf{E}_j \mu\|^2$, $j = 1, \ldots, p$, and a function f_∞ continuous on $\mathbb{R}^+ \cup \{0\}$ such that $\int |f_p(\tau) - f_\infty(\tau)| \, dH_p(\tau) \to 0$ as $p \to \infty$. (A sufficient condition for the latter is that $\|f_p - f_\infty\|_\infty \to 0$ as $p \to \infty$.) Then it follows from **C3** that (7) holds with

$$\begin{aligned} q(\lambda,\gamma) &= \{1 + \gamma\Theta_1(\lambda,\gamma)\} \int \frac{f_\infty(\tau)\, dH(\tau)}{\tau + \lambda\{1 + \gamma\Theta_1(\lambda,\gamma)\}} \\ &= \int \frac{f_\infty(\tau)\, dH(\tau)}{\tau\{1 - \gamma(1 - \lambda m_F(-\lambda))\} + \lambda}. \end{aligned} \tag{10}$$

The second line in (10) follows from the relationship $\{1 + \gamma\Theta_1(\lambda, \gamma)\}^{-1} = 1 - \gamma + \lambda\gamma m_F(-\lambda)$, for $\lambda > 0$.

(b) If $\Sigma_p = I_p$, then (7) is satisfied if $\sqrt{n}\|\mu\|^2 \to c^2 > 0$. In this case, $q(\lambda, \gamma) = c^2\Theta_1(\lambda, \gamma)$.

While deterministic local alternatives like (10) provide useful information, in the following, we focus on probabilistic alternatives that provide a convenient framework for incorporating structure. Focus is on the following class of priors for μ under the alternative hypothesis.

PA Assume that, under the alternative, $\mu = n^{-1/4}p^{-1/2}B\nu$ where B is a $p \times p$ matrix, and ν is random vector with independent coordinates such that $\mathbb{E}[\nu_i] = 0$, $\mathbb{E}[|\nu_i|^2] = 1$ and $\max_i \mathbb{E}[|\nu_i|^4] \le p^{c_\nu}$ for some $c_\nu \in (0, 1)$. Moreover, let $\mathbf{B} = BB^T$ with $\|\mathbf{B}\| \le C_1 < \infty$, and, as $n, p \to \infty$,

$$p^{-1} \operatorname{tr}\{D_p(-\lambda)\mathbf{B}\} \to q(\lambda, \gamma), \tag{11}$$

for some finite, positive constant $q(\lambda, \gamma)$.

REMARK 2.3. To better understand **PA**, first observe that μ has zero mean and covariance matrix $n^{-1/2}p^{-1}\mathbf{B}$. The factor $n^{-1/2}p^{-1}$ provides the scaling for the RHT test to have nontrivial local power. To check the meaning of (11), similar to the analysis in Remark 2.2, postulate the existence of functions $\tilde{f}_p$ satisfying $\tilde{f}_p(\tau_{j,p}) = \operatorname{tr}\{\mathbf{E}_j\mathbf{B}\}$ and $\int |\tilde{f}_p(\tau) - f_\infty(\tau)| \, dH_p(\tau) \to 0$ for some function f_∞ continuous on $\mathbb{R}^+ \cup \{0\}$. Then the limit in (11) exists and the corresponding $q(\lambda, \gamma)$ has the form given in (10). Thus, f_∞ can be viewed as a distribution of the total spectral mass of $\mathbf{B}$ (measured as $\operatorname{tr}\{\mathbf{B}\}$) across the eigensubspaces of Σ_p.

The framework **PA** encompasses both dense and sparse alternatives, as illustrated in the following special cases:

(I) *Dense alternative*: $\nu_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$.

(II) *Sparse alternative*: $\nu_i \overset{\text{i.i.d.}}{\sim} G_\eta$, for some $\eta \in (0, 1)$, where G_η is the discrete probability distribution which assigns mass $1 - p^{-\eta}$ on 0 and mass $(1/2)p^{-\eta}$ on the points $\pm p^{\eta/2}$.

If $B = I_p$ under (II), then μ is sparse, with the degree of sparsity determined by η.

THEOREM 2.2. *Suppose that* **C1–C3** *hold and that, under the alternative* $H_a : \mu \neq 0$, μ *has prior given by* **PA**. *Then, for any* $\lambda > 0$,

$$\beta_n(\mu, \lambda) - \Phi\left(-\xi_\alpha + \kappa(1-\kappa)\frac{p^{-1}\operatorname{tr}\{D_p(-\lambda)\mathbf{B}\}}{\{2\gamma\Theta_2(\lambda,\gamma)\}^{1/2}}\right) \to 0 \quad (n \to \infty), \tag{12}$$

where the convergence in (12) *holds in the* L^1-*sense. Note that the convergence in* (11) *is assumed to increase readability and interpretability. The asymptotic result of this theorem holds also if* $p^{-1}\operatorname{tr}\{D_p(-\lambda)\mathbf{B}\}$ *is replaced by* $q(\lambda, \gamma)$.

Theorem 2.2 notably shows that, even for alternatives that are sparse in the sense of (II), the proposed test has the same asymptotic power as for the dense alternatives (I), as long as the covariance structure is the same. The local power of the RHT test can be compared to a test based on maximizing coordinatewise t-statistics (as in Cai, Liu and Xia (2014)) under the sparse alternatives (II). For simplicity, let $\mathbf{B} = I_p$ and $\Sigma = I_p$. If $\eta \in (0, 1/2)$, then the size of each spike of the vector μ is of order $n^{-1/4}p^{-1/2+\eta/2} = o(n^{-1/2})$, while the maximum of the t-statistics is at least of the order $O_P(n^{-1/2})$ under the null hypothesis. This renders procedures based on maxima of t-statistics ineffective, while RHT still possesses nontrivial power. However, if $\eta > 1/2$, corresponding to a high degree of sparsity, tests based on maxima of t-statistics will outperform RHT. The RHT test shares this characteristic with the test of Chen and Qin (2010).

2.3. *Power under polynomial alternatives.* Computation of local power of the RHT test, as given in Theorem 2.2, involves the computation of $q(\lambda, \gamma)$ using (10), if **PA** holds and f_∞ is specified. However, this task remains challenging since the integral in (10) involves the unknown population spectral distribution H. In order to estimate $q(\lambda, \gamma)$, without having to estimate H (which is a difficult task in itself), it is convenient to have it in a closed form. Below, we formulate a scheme that allows us to compute $q(\lambda, \gamma)$ when f_∞ is a polynomial. The latter is true if **B** is a matrix polynomial in Σ. Since any arbitrary smooth function can be approximated by polynomials, this formulation is quite useful. Moreover, the choice of **B** as a matrix polynomial in Σ also allows for an easier interpretation of the structure of μ under the alternative.

It should be noted that the structural assumptions imply that the covariance of mean-difference μ diagonalizes in the eigenbasis of Σ, which is restrictive. However, this restriction enables us to make principled and data-adaptive choices of the regularization parameter λ. We specifically focus on the setting where **B** is quadratic in Σ, which elucidates many interesting phenomena in terms of the choice of optimal λ. Finally, the ARHT procedure described later is obtained by combining a small collection of simple probabilistic alternatives within this framework, and is seen to have robust performance characteristics.

Before proceeding further, we give a brief summary of how we utilize the expression for local power under this class of alternatives. First, in Section 2.4, they are utilized to compare the power characteristics of the RHT test with its "natural" competitors, namely, the Hotelling's T^2 test, the tests by Bai and Saranadasa (1996) and Chen and Qin (2010), and the random projection-based test by Lopes, Jacob and Wainwright (2011). In Section 2.5, we use them to devise a data-driven procedure for selecting the regularization parameter λ. In Section 2.6, we formulate and analyze a decision theoretic approach to selecting λ, an exercise which enhances our theoretical understanding of the RHT procedure in comparison with existing procedures. Finally, in Section 3, these expressions also enable us to propose the ARHT test by combining several optimally chosen regularization parameters.

By polynomial alternative, we refer to the following model: μ satisfies **PA** with $\mathbf{B} = \sum_{m=0}^{r} \pi_m \Sigma^m$, for pre-specified $\pi_0, \pi_1, \ldots, \pi_r$ such that $\mathbf{B}$ is positive semidefinite. Then

$$\operatorname{Var}(\mu) = \frac{1}{p\sqrt{n}} \sum_{m=0}^{r} \pi_m \Sigma^m. \tag{13}$$

We denote the prior $\mu \sim N(0, B)$ with $\mathbf{B}$ as in (13) by $\mathcal{P}_{\tilde{\pi}}$. Note that, in order for $\mathbf{B}$ to be positive semi-definite, it suffices that the real-valued polynomial $\sum_{m=0}^{r} \pi_m x^m$ is nonnegative on $[0, \|\Sigma\|]$. Unless $\Sigma = I_p$ or $\pi_0 = 1$, such a prior implies a certain distribution of the coefficients of μ in the spectral coordinate system. Specifically, larger values of π_m for higher powers m imply that μ has larger contribution from the leading eigenvectors of Σ.

Under model (13), (11) is satisfied and the limit equals

$$q(\lambda, \gamma) = \sum_{m=0}^{r} \pi_m \rho_m(-\lambda, \gamma), \tag{14}$$

with $\rho_m(-\lambda, \gamma)$ satisfying the recursive formula

$$\rho_{m+1}(-\lambda, \gamma) = \{1 + \gamma \Theta_1(\lambda, \gamma)\}\left\{\int \tau^m \, dH(\tau) - \lambda \rho_m(-\lambda, \gamma)\right\},$$

and $\rho_0(-\lambda, \gamma) = m_F(-\lambda)$. This formula, which can be deduced from Lemma 3 of Ledoit and Péché (2011) and the derivations given in the Supplementary Material, involves the population spectral moments $\int \tau^m \, dH(\tau)$. The latter can be estimated, since equations connecting the moments of H with the limits of the tracial moments $p^{-1} \operatorname{tr}\{S_n^m\}$, $m \geq 1$, are known (Bai, Chen and Yao (2010), Lemma 1).

2.4. *Power comparison.* We now use the probabilistic alternative framework determined by **PA** and (13) to analytically compare the power characteristics of the RHT procedure in comparison with some of the methods that are natural candidates in the sense of sharing the orthogonal invariance property enjoyed by RHT. As a first step, we derived the expressions for the power functions of these tests.

It can be checked that, under **C1**–**C3** and **PA**, together with sub-Gaussianity of the observations, the conditions imposed to derive asymptotics in Bai and Saranadasa (1996) and Chen and Qin (2010) are satisfied. Therefore, by making use of supporting results in their papers and the techniques used in this paper, the power β_{BS} of the test by Bai and Saranadasa (1996) (referred to as BS), and the power β_{CQ} of the test by Chen and Qin (2010) (referred to as CQ), can be shown to satisfy

$$\beta_{BS}(\mu) - \Phi\left(-\xi_\alpha + \kappa(1-\kappa) \frac{p^{-1} \operatorname{tr}(\mathbf{B})}{\{2\gamma \int \tau^2 \, dH(\tau)\}^{1/2}}\right) \xrightarrow{L_1} 0, \tag{15}$$

$$\beta_{CQ}(\mu) - \Phi\left(-\xi_\alpha + \kappa(1-\kappa) \frac{p^{-1} \operatorname{tr}(\mathbf{B})}{\{2\gamma \int \tau^2 \, dH(\tau)\}^{1/2}}\right) \xrightarrow{L_1} 0. \tag{16}$$

If we further assume that the observations are Gaussian, then we can also provide an expression for the asymptotic power of the random projection based test (referred to as RP) proposed by Lopes, Jacob and Wainwright (2011). Let $\beta_{RP}(\mu, P_k^T)$ be the power of the RP test, given a realization P_k of the rank-k random projection. The suggested k value is $p/2$. Then, it can be shown that for almost all sequences of projections P_k, as $n \to \infty$,

$$\beta_{RP}(\mu, P_k^T) - \Phi\left(-\xi_\alpha + \frac{\kappa(1-\kappa)}{\sqrt{2}} p^{-1} \operatorname{tr}\left[(P_k^T \mathbf{B} P_k)(P_k^T \Sigma_p P_k)^{-1}\right]\right) \xrightarrow{L_1} 0. \tag{17}$$

The asymptotic power $\beta_{HT}(\mu)$ of Hotelling's T^2 (referred to as HT) when $p/n \to \gamma$ is also derived in Bai and Saranadasa (1996). Making use of their results, under **PA**, we have, when $\gamma < 1$,

$$\beta_{HT}(\mu) - \Phi\left(-\xi_\alpha + \kappa(1-\kappa)\sqrt{\frac{1-\gamma}{2\gamma}} p^{-1}\operatorname{tr}(\Sigma_p^{-1}\mathbf{B})\right) \xrightarrow{L_1} 0. \tag{18}$$

Equations (12) and (14) together provide the corresponding expression for the local power of the RHT test under (13), that is, when $\mathbf{B} = \sum_{m=0}^{r} \pi_m \Sigma^m$.

At this point, a practical difficulty in analytically comparing power characteristics of these different methods presents itself. Notice that even though we have an analytical expression for the power of the RHT procedure, it still involves the functions $\rho_m(-\lambda, \gamma)$. These functions are available in closed form only if $\Sigma = I_p$. While $\Sigma = I_p$ is not the most compelling of cases, it is also a situation where the benefit of the ridge-type regularization is expected to be significantly reduced. Indeed, in such cases, choosing $\lambda = \infty$, which corresponds to replacing the normalizer $(S_n + \lambda I_p)$ by the identity matrix and, therefore, effectively reducing the RHT test to the BS test, at least intuitively appears to be the most reasonable option. On the other hand, the effect of appropriate normalization of the coordinates of $\bar{X}_1 - \bar{X}_2$ in the expression for the RHT statistic is expected to be much more significant when there is a degree of nondegeneracy in the spectral distribution of Σ. Keeping this in mind, and considering that a simple and interpretable model for the population spectral distribution H is useful for carrying out a meaningful comparison, we focus on the following example for H:

1. Assume a two-point mixture model for the spectral distribution H of Σ_p, namely,

$$a\delta_x + (1-a)\delta_y,$$

with $y = (1-ax)/(1-a)$, so that $p^{-1}\operatorname{tr}(\Sigma_p) = 1$. In other words, Σ_p has two distinct eigenvalues x and y with ratio a and $(1-a)$, respectively.

2. Suppose that the model **PA** together with (13) and assumptions **C1**–**C3** hold. In order to highlight the key features of the power characteristics of these tests, consider the three canonical settings for $\mathbf{B}$, namely, I_p, Σ and Σ^2.

In this model, the Marčenko–Pastur equation is cubic in $m_F(z)$ and given by

$$m_F(z) = \frac{a}{x\{1-\gamma-\gamma z m_F(z)\} - z} + \frac{1-a}{y\{1-\gamma-\gamma z m_F(z)\} - z}.$$

An lengthy yet explicit solution to $m_F(z)$ is available, but not displayed here. This solution in turn yields explicit solutions to $\Theta_1(\lambda, \gamma)$, $\Theta_2(\lambda, \gamma)$ and $q(\lambda, \gamma)$.

Figure 1 gives the power function $\beta_n(\mu, \lambda)$, approximated using the asymptotic results, against γ for different choices of a, x and $\mathbf{B}$ at selected values of λ. Specifically, we selected HT ($\lambda = 0$), $\lambda = 0.1$, $\lambda = 1$ and BS ($\lambda = \infty$). Additionally, for each γ the optimal λ_γ that maximizes $\beta_n(\mu, \lambda)$ was computed, whose power is then the best possible one among the ridge-regularized family of tests under probabilistic local alternatives. Other implementation details were $\alpha = 0.05$, $\kappa = 0.5$ and $\operatorname{Var}(\mu) = 10n^{-1/2}p^{-1}\mathbf{B}$ (chosen for visualization purposes).

Figures 2 and 3 display the power function $\beta_n(\mu, \lambda)$ against λ for $\gamma = 0.5, 0.9$ and 2. For comparison, the power of HT ($\gamma < 1$ only), BS/CQ and RP are given as horizontal lines.

Figure 2 and Figure 3 show clearly that when $\operatorname{Var}(\mu)$ is proportional to either I_p or Σ_p, the power gain of RHT with the optimal λ over either BS/CQ test or RP test is quite prominent when a relatively small fraction of eigenvalues is significantly bigger than the smaller eigenvalues. Figure 1 shows ridge-regularization can remarkably rescue the performance of HT

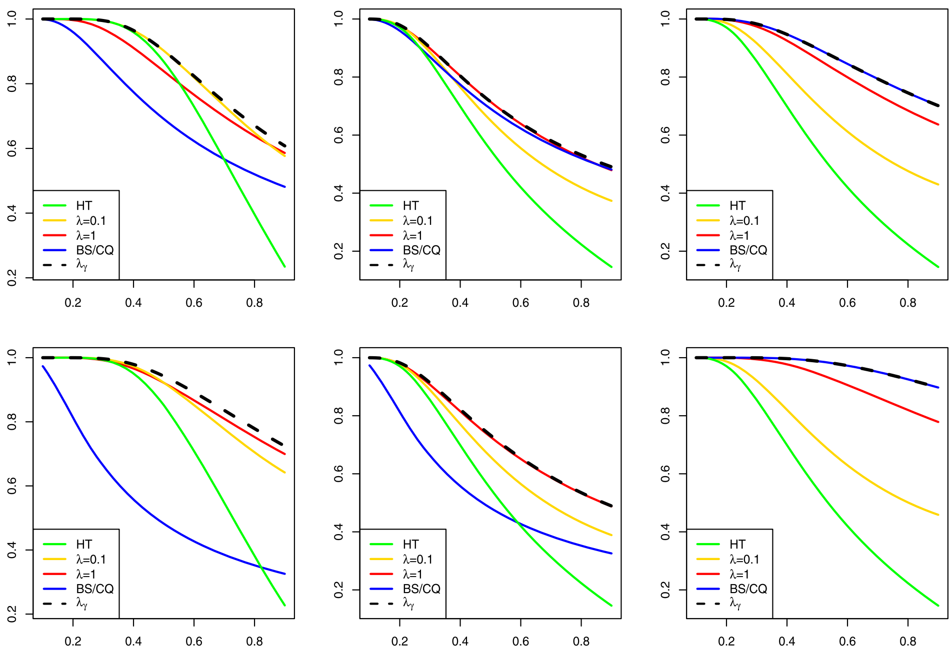


FIG. 1. *$\beta_n(\mu, \lambda)$ against γ. HT ($\lambda = 0$) (green), $\lambda = 0.1$ (yellow), $\lambda = 1$ (red), BS/CQ ($\lambda = \infty$) (blue), λ_γ (black, dashed). Columns (left to right)*: $\mathbf{B} = I_p, \Sigma_p, \Sigma_p^2$; *First Row*: $a = 0.5$ *and* $x = 0.4$; *Second Row*: $a = 0.9$ *and* $x = 0.6$.

when γ is close to 1. The rightmost panels of Figures 2 and 3, corresponding to the setting where $\text{Var}(\mu)$ is proportional to Σ^2, represent a case for which maximum power is attained for the largest λ. Here, the power of RHT is no better than the BS/CQ test under this class of alternatives and in all settings. In Section 2.6, we provide a mathematical result verifying this aspect that also forms the basis for a decision-theoretic framework to choose λ.

2.5. *Data-driven selection of* λ. Given a sequence of local probabilistic alternatives, the strategy is to choose λ by maximizing the "local power" function $\beta_n(\mu, \lambda)$. Theorems 2.1 and 2.2 suggest that λ should be chosen such that the ratio $Q(\lambda, \gamma) = q(\lambda, \gamma)\{\gamma\Theta_2(\lambda, \gamma)\}^{-1/2}$ is maximized, with $q(\lambda, \gamma)$ given by (11).

In the following, we recall two possible settings under **PA** where $q(\lambda, \gamma)$ can be computed explicitly: (i) Suppose that $\mathbf{B}$ is specified. In this case, $q(\lambda, \gamma)$ is estimated by $p^{-1}\operatorname{tr}((S_n + \lambda I_p)^{-1}\mathbf{B})$, the latter being a consistent estimator of the LHS of (11). (ii) Only the spectral mass distribution of $\mathbf{B}$ in the form of f_∞ (described in Remark 2.3) is specified. Then, as explained in Section 2.3, for polynomial f_∞, we obtain the expression (14) for $q(\lambda, \gamma)$, and this can be estimated consistently.

In order to effectively utilize the expression (14) for $q(\lambda, \gamma)$, we restrict to the case $r = 2$. There are several considerations that guide this choice of r. First, for $r = 2$, all quantities involved in estimating $q(\lambda, \gamma)$ can be computed explicitly without requiring knowledge of higher-order moments of the observations. Also, the corresponding estimating equations are more stable as they do not involve higher-order spectral moments. Second, the choice of $r = 2$ yields a significant, yet nontrivial, concentration of the prior covariance of μ (equivalently, $\mathbf{B}$) in the directions of the leading eigenvectors of Σ. Finally, the choice $r = 2$ allows for both convex and concave shapes for the spectral mass distribution f_∞ since the latter becomes a quadratic function.

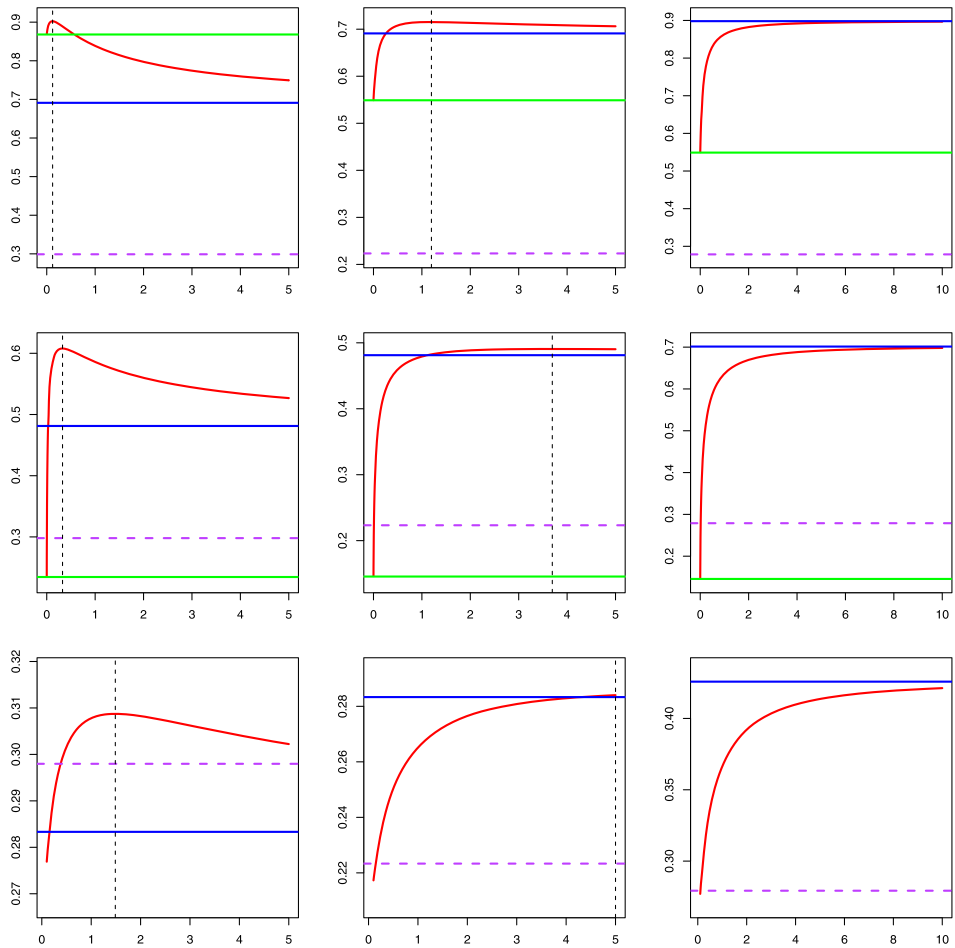


FIG. 2. $\beta_n(\mu, \lambda)$ *against* λ *when* $a = 0.5$, $x = 0.4$. *RHT* (*red*), *BS/CQ* ($\lambda = \infty$) (*blue*), *RP* (*dashed*, *purple*), *HT* ($\lambda = 0$) (*green*), (*locally*) *optimal lambda* (*vertical dash*). *Columns* (*left to right*): $\mathbf{B} = I_p$, Σ_p, Σ_p^2; *Rows* (*top to bottom*): $\gamma = 0.5, 0.9, 2$.

With $r = 2$, in order to estimate $q(\lambda, \gamma)$, it suffices to estimate

$$\begin{aligned} \rho_0(-\lambda, \gamma) &= m_F(-\lambda), \\ \rho_1(-\lambda, \gamma) &= \Theta_1(\lambda, \gamma), \\ \rho_2(-\lambda, \gamma) &= \{1 + \gamma \Theta_1(\lambda, \gamma)\}\{\phi_1 - \lambda \rho_1(-\lambda, \gamma)\}, \end{aligned} \tag{19}$$

where $\phi_1 = \int \tau \, dH(\tau)$. The latter can be estimated accurately by $\hat{\phi}_1 = p^{-1} \operatorname{tr}\{S_n\}$ (see Proposition A.2).

We state below the algorithm for data-driven selection of the regularization parameter λ.

ALGORITHM 2.1 (Empirical selection of λ). Perform the following steps:

1. Specify prior weights $\tilde{\pi} = (\pi_0, \pi_1, \pi_2)$;
2. For each λ, compute the estimates

$$\begin{aligned} \hat{\rho}_0(-\lambda, \gamma_n) &= m_{F_{n,p}}(-\lambda), \\ \hat{\rho}_1(-\lambda, \gamma_n) &= \hat{\Theta}_1(\lambda, \gamma_n), \\ \hat{\rho}_2(-\lambda, \gamma_n) &= \{1 + \gamma_n \hat{\Theta}_1(\lambda, \gamma_n)\}\{\hat{\phi}_1 - \lambda \hat{\rho}_1(-\lambda, \gamma_n)\}; \end{aligned}$$

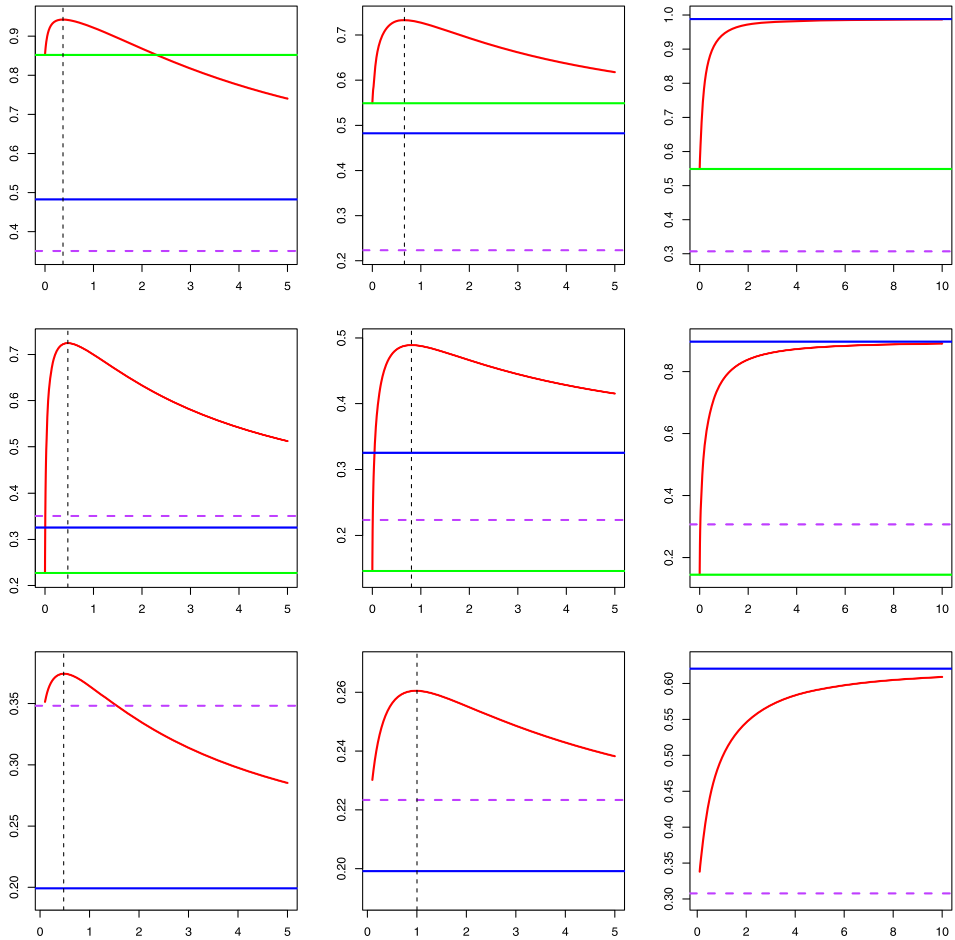


FIG. 3. *Same as Figure 2 but with $a = 0.9$, $x = 0.6$.*

3. For each λ, compute the estimate

$$\hat{Q}_n(\lambda, \gamma_n; \tilde{\pi}) = \sum_{m=0}^{2} \pi_m \hat{\rho}_m(-\lambda, \gamma_n)/\{\gamma_n \hat{\Theta}_2(\lambda, \gamma_n)\}^{1/2};$$

4. Select $\lambda_{\tilde{\pi}} \equiv \lambda_{\tilde{\pi},n} = \arg\max_\lambda \hat{Q}_n(\lambda, \gamma_n; \tilde{\pi})$ through a grid search.

Although in theory arbitrarily small positive λ are allowed in the test procedure, in practice, meaningful lower and upper bounds $\underline{\lambda}$ and $\overline{\lambda}$ are needed to ensure stability of the test statistic when $p \approx n$ or $p > n$. The recommended choices are $\underline{\lambda} = p^{-1}\operatorname{tr}\{S_n\}/100$ and $\overline{\lambda} = 20\|S_n\|$.

The behavior of the test with the data-driven tuning parameter is described in the next theorem.

THEOREM 2.3. *Let $[\underline{\lambda}, \overline{\lambda}]$ (with $\overline{\lambda} > \underline{\lambda} > 0$) be a nonempty interval. Let λ_∞ be any local maximizer of $Q(\lambda, \gamma; \tilde{\pi})$ on $[\underline{\lambda}, \overline{\lambda}]$. If conditions* **C1–C3** *are satisfied and if there is a $C > 0$ such that $\partial^2 Q(\lambda_\infty, \gamma; \tilde{\pi})/\partial\lambda^2 < -C$, then there exists a sequence $(\lambda_n : n \in \mathbb{N})$ of local maximizers of $(\hat{Q}_n(\lambda, \gamma_n; \tilde{\pi}) : n \in \mathbb{N})$, satisfying*

$$n^{1/4}|\lambda_n - \lambda_\infty| = O_p(1) \quad (n \to \infty). \tag{20}$$

Further, under the null hypothesis,

$$T_{n,p}(\lambda_n) = \frac{p^{1/2}\{p^{-1}\mathrm{RHT}(\lambda_n) - \hat{\Theta}_1(\lambda_n, \gamma_n)\}}{\{2\hat{\Theta}_2(\lambda_n, \gamma_n)\}^{1/2}} \Longrightarrow \mathcal{N}(0,1) \quad (n \to \infty), \tag{21}$$

where $\Longrightarrow$ *denotes convergence in distribution. The procedure is adaptive in the sense that the asymptotic power of the test based on* $T_{n,p}(\lambda_n)$ *is the same as that of* $T_{n,p}(\lambda_\infty)$ *under the sequence of priors specified by* $\tilde{\pi}$.

REMARK 2.4. In Theorem 2.3, if λ_∞ is a boundary point and $\partial Q(\lambda_\infty, \gamma; \tilde{\pi})/\partial\lambda \neq 0$, then the assumption on $\partial^2 Q(\lambda_\infty, \gamma; \tilde{\pi})/\partial\lambda^2$ can be dropped.

2.6. *Minimax selection of* λ. In Section 2.5, it is assumed that a specific prior $\tilde{\pi}$ is available. However, in practice, rather than a particular choice of $\tilde{\pi}$, we may have to consider a collection of such priors. In this subsection, a procedure for selecting the regularization parameter for the RHT test $T_{n,p}(\lambda)$ is presented that is based on the principle of minimaxity. Throughout this subsection, *minimax* refers to minimaxity within the class of all RHT tests.

Let $\mathcal{D} = \{T_{n,p}(\lambda) : \lambda \in [\underline{\lambda}, \overline{\lambda}]\}$, for $0 < \underline{\lambda} < \overline{\lambda} < \infty$ denote a class of normalized RHT test statistics. Also, let $\mathfrak{P}$ be a family of local priors for μ under the alternative. Notice that, for any $\alpha \in (0,1)$ the test $\delta_\alpha(\lambda) = \mathbf{1}(T_{n,p}(\lambda) > \xi_\alpha)$ has asymptotically level α. For any given prior $\mathcal{P}$ for μ under the alternative, define the *asymptotic Bayes risk* of the test $\delta_\alpha(\lambda)$ with respect to prior $\mathcal{P}$ as

$$R(\delta_\alpha(\lambda); \mathcal{P}) = \limsup_{n,p\to\infty}\bigl(1 - \mathbb{E}_{\mathcal{P}}[\beta_n(\mu, \lambda)]\bigr) = 1 - \liminf_{n,p\to\infty} \mathbb{E}_{\mathcal{P}}[\beta_n(\mu, \lambda)] \tag{22}$$

with $\beta_n(\mu, \lambda)$ as in (8). We say that $T_{n,p}(\lambda_*)$ is a *locally asymptotically minimax* (*LAM*) test within the class $\mathcal{D}$ and with respect to $\mathfrak{P}$, if for each $\alpha \in (0,1)$, the minimum value of $\sup_{\mathcal{P}\in\mathfrak{P}} R(\delta_\alpha(\lambda); \mathcal{P})$ over $\lambda \in [\underline{\lambda}, \overline{\lambda}]$ is attained at λ_*.

Consider now a family of priors $\mathfrak{P}_r(C)$ defined in the following way. For a constant $C > 0$, define

$$\Pi_r(C) = \left\{\tilde{\pi} = (\pi_0, \ldots, \pi_r) : \sum_{m=0}^{r} \pi_m x^m \geq 0 \text{ for } x \in [0, \infty), \sum_{m=0}^{r} \pi_m \phi_m = C\right\},$$

where $\phi_m = \int \tau^m \, dH(\tau)$. Let $\mathcal{P}_{\tilde{\pi}}$ denote the prior for μ satisfying **PA** and (13). Finally, let

$$\mathfrak{P}_r(C) = \{\mathcal{P}_{\tilde{\pi}} : \tilde{\pi} \in \Pi_r(C)\}.$$

The condition $\sum_{m=0}^{r} \pi_m x^m \geq 0$ for all $x \geq 0$ ensures that the matrix $\sum_{m=0}^{r} \pi_m \Sigma^m$ is nonnegative definite, while the condition $\sum_{m=0}^{r} \pi_m \phi_m = C$ means that as $p \to \infty$, $\sqrt{n}\,\mathrm{tr}\{\mathrm{Var}(\mu)\} \to C$. Observe that, for $\tilde{\pi} \in \Pi_r(C)$, the asymptotic Bayes risk $R(\delta_\alpha(\lambda); \mathcal{P}_{\tilde{\pi}})$ equals $1 - \Phi(-\xi_\alpha + \kappa(1-\kappa)Q(\lambda, \gamma; \tilde{\pi}))$ where $q(\lambda, \gamma) \equiv q(\lambda, \gamma; \tilde{\pi})$ is given by (14), implying that $\mathcal{P}_{\tilde{\pi}}$ actually constitutes an equivalence class of priors.

Restricting to $r = 2$, note that finding a LAM test within the class $\mathcal{D}$ and with respect to the family $\mathfrak{P}_2(C)$, means finding a $\lambda \in [\underline{\lambda}, \overline{\lambda}]$ that minimizes $\sup_{\tilde{\pi}\in\Pi_2(C)} R(\delta(\lambda); \mathcal{P}_{\tilde{\pi}})$. Without loss of generality, take $C = 1$ since the risk function is monotonically decreasing in $Q(\lambda, \gamma; \tilde{\pi})$, and the latter is a linear function of $\tilde{\pi}$. This leads to the following result.

PROPOSITION 2.1. *Under the conditions of Theorem 2.2, the LAM test within the class* $\mathcal{D}$, *with respect to the family* $\mathfrak{P}_2(C)$ *is* $T_{n,p}(\overline{\lambda})$.

The proof of this proposition is given in Section 9.6.

It can be verified that as $\lambda \to \infty$, the test statistic $\mathrm{RHT}(\lambda)$ converges pointwise to the corresponding test statistic by Bai and Saranadasa (1996), and the local asymptotic power of $\mathrm{RHT}(\lambda)$ under the class of alternatives $\mathfrak{P}_2(C)$ also converges to the corresponding power for the test by Bai and Saranadasa (1996). Thus, Proposition 2.1 shows that the test by Bai and Saranadasa (1996) is the limit of a locally asymptotically minimax test, namely the test $T_{n,p}(\bar{\lambda})$, as $\bar{\lambda} \to \infty$.

3. Adaptable RHT. Section 2.5 describes a data-driven procedure for selecting the optimal regularization parameter λ for prespecified prior weights $\tilde{\pi}$, whereas Section 2.6 derives an asymptotically minimax RHT test with respect to a class of priors. An extensive simulation analysis reveals that there is a considerable variation in the shape of the power function and the value of the corresponding Bayes rule, especially when the condition number of Σ is relatively large.

As an alternative to the minimax approach, which can be overly pessimistic, instead of considering a broad collection of priors, one might consider a convenient collection of priors that are representative of certain structural scenarios. Thus adopting a mildly conservative approach, define a new test statistic as the maximum of the RHT statistics corresponding to a set of regularization parameters that are optimal with respect to a specific collection of priors. Specifically, we propose the following test statistic, referred to as *Adaptable RHT* (*ARHT*):

$$\mathrm{ARHT}_{n,p}(\Pi) = \max_{\tilde{\pi} \in \Pi} T_{n,p}(\lambda_{\tilde{\pi}}), \tag{23}$$

where $T_{n,p}(\lambda)$ is defined in (6), $\lambda_{\tilde{\pi}}$ in Algorithm 2.1, and $\Pi = \{\tilde{\pi}_1, \dots, \tilde{\pi}_k\}$, $k \geq 1$, is a prespecified finite class of weights. A simple but effective choice of Π consists of the three *canonical weights* $\tilde{\pi} = (1, 0, 0)$, $(0, 1, 0)$ and $(0, 0, 1)$. We focus on this particular specification of Π, since a convex combination of these three weights cover a wide range of local alternatives, and this choice leads to very satisfactory empirical performance as is illustrated through simulations in Section 6. In particular, the ARHT procedure is shown to outperform the test by Bai and Saranadasa (1996) (the limiting LAM procedure) in most circumstances.

Determining the cut-off values of $\mathrm{ARHT}_{n,p}(\Pi)$ requires knowing the asymptotic distribution of the process $T_{n,p} = (T_{n,p}(\lambda) \colon \lambda \in [\underline{\lambda}, \bar{\lambda}])$ under the null hypothesis of equal means. From this, the case where $\Lambda = \{\lambda_{\tilde{\pi}_1}, \dots, \lambda_{\tilde{\pi}_k}\}$ is a collection of finitely many regularization parameters can be easily derived.

THEOREM 3.1. *If* **C1–C3** *are satisfied, then, under* H_0,

$$T_{n,p} \xrightarrow{d} Z \quad (n \to \infty),$$

where $\xrightarrow{d}$ *denotes weak convergence in the Skorohod space* $D[\underline{\lambda}, \bar{\lambda}]$ *and* $Z = (Z(\lambda) \colon \lambda \in [\underline{\lambda}, \bar{\lambda}])$ *a centered Gaussian process with covariance function*

$$\begin{aligned}\Gamma(\lambda, \lambda') &\\ &= \{1 + \gamma\Theta_1(\lambda, \gamma)\}\{1 + \gamma\Theta_1(\lambda', \gamma)\}\frac{\lambda'\Theta_1(\lambda', \gamma) - \lambda\Theta_1(\lambda, \gamma)}{(\lambda' - \lambda)\{\Theta_2(\lambda, \gamma)\Theta_2(\lambda', \gamma)\}^{1/2}},\end{aligned} \tag{24}$$

for $\lambda \neq \lambda'$, *and* $\Gamma(\lambda, \lambda) \equiv 1$. *In particular, for every* $k \geq 1$ *and every collection* $\Lambda = \{\lambda_1, \dots, \lambda_k\} \subset [\underline{\lambda}, \bar{\lambda}]$, *it holds that*

$$(T_{n,p}(\lambda_1), \dots, T_{n,p}(\lambda_k))^T \Longrightarrow N_k(0, \Gamma(\Lambda)) \quad (n \to \infty),$$

where the limit on the right-hand side is a k*-dimensional centered normal distribution with* $k \times k$ *covariance matrix* $\Gamma(\Lambda)$ *with entries* $\Gamma(\lambda_i, \lambda_j)$, $i, j = 1, \dots, k$.

Theorem 3.1 shows that $\mathrm{ARHT}_{n,p}(\Pi)$ has a nondegenerate limiting distribution under H_0. Theorem 3.1 can be used to determine the cut-off values of the test by deriving analytical formulae for the quantiles of the limiting distribution. Aiming to avoid complex calculations, a *parametric bootstrap* procedure is applied to approximate the cut-off values. Specifically, $\Gamma(\Lambda)$ is first estimated by $\hat{\Gamma}_n(\Lambda)$, and then bootstrap replicates are generated by simulating from $N_k(0, \hat{\Gamma}(\Lambda))$, thereby leading to an approximation of the null distribution of $\mathrm{ARHT}_{n,p}(\Pi)$. A natural candidate for the covariance estimator is

$$\begin{aligned}\hat{\Gamma}_n(\lambda, \lambda') &\\ &= \{1 + \gamma_n \hat{\Theta}_1(\lambda, \gamma_n)\}\{1 + \gamma_n \hat{\Theta}_1(\lambda', \gamma_n)\} \\ &\quad \times \frac{\lambda' \hat{\Theta}_1(\lambda', \gamma_n) - \lambda \hat{\Theta}_1(\lambda, \gamma_n)}{(\lambda' - \lambda)\{\hat{\Theta}_2(\lambda, \gamma_n)\hat{\Theta}_2(\lambda', \gamma_n)\}^{1/2}},\end{aligned} \tag{25}$$

for $\lambda \neq \lambda'$ and $\hat{\Gamma}_n(\lambda, \lambda) \equiv 1$.

REMARK 3.1. It should be noticed that $\hat{\Gamma}_n(\Lambda)$ defined through (25) may not be nonnegative definite even though it is symmetric. If such a case occurs, the resulting estimator can be projected to its closest nonnegative definite matrix simply by setting the negative eigenvalues to zero. This covariance matrix estimator is denoted by $\hat{\Gamma}_n^+(\Lambda)$ and is used for generating the bootstraps samples.

4. Calibration of type I error probability. Simulation studies reveal that the size of RHT tends to be slightly inflated. This is because a normal approximation is used to describe a quadratic form statistic, leading to skewed distributions in finite samples. Two remedies are proposed. The first is based on a power transformation of RHT, reducing skewness by calibrating higher-order terms in the test statistics. The second on choosing cut-off values of RHT based on quantiles of a normalized χ^2 distribution whose first two moments match those of RHT.

4.1. *Cube-root transformation.* In principle, any power transformation may be considered, but empirically, a near-symmetry of the null distribution is obtained by a cube-root transformation of the RHT statistic. Therefore, restricting to this case only, an application of the δ-method yields

$$\tilde{T}_{1/3}(\lambda) = \frac{p^{1/2}[\{p^{-1}\,\mathrm{RHT}(\lambda)\}^{1/3} - \hat{\Theta}_1^{1/3}(\lambda, \gamma_n)]}{(2^{1/2}/3)\hat{\Theta}_2^{1/2}(\lambda, \gamma_n)/\hat{\Theta}_1^{2/3}(\lambda, \gamma_n)} \Longrightarrow N(0, 1). \tag{26}$$

This gives rise to the cube-root transformed ARHT test statistic

$$\mathrm{ARHT}_{1/3}(\Pi) = \max_{\tilde{\pi} \in \Pi} \tilde{T}_{1/3}(\lambda_{\tilde{\pi}}).$$

A test based on $\mathrm{ARHT}_{1/3}(\Pi)$ for a finite set Π of weight vectors can be performed by making use of the covariance kernel Γ given in (24). $\mathrm{ARHT}_{1/3}$ is recommended for most practical applications since it nearly symmetrizes the null distribution of the test statistic even for moderate sample sizes. Algorithm 4.1 details the composite test procedure with the recommended $\mathrm{ARHT}_{1/3}$ statistic.

ALGORITHM 4.1 (Cube-root transformed ARHT).

1. Diagonalization: Compute the spectral decomposition of $S_n = P_n \Delta_n P_n^T$, apply the transformation $\bar{Y}_1 = P_n^T \bar{X}_1$, $\bar{Y}_1 = P_n^T \bar{X}_1$; and run the rest with $\bar{X}_1$, $\bar{X}_2$, S_n replaced by $\bar{Y}_1$, $\bar{Y}_2$ and Δ_n;

2. For each $\tilde{\pi}$ in Π, run Algorithm 2.1 and obtain $\Lambda = \{\lambda_{\tilde{\pi}} : \tilde{\pi} \in \Pi\}$;
3. Compute $\hat{\Gamma}_n^+(\Lambda)$;
4. Generate $\varepsilon_1, \ldots, \varepsilon_B$ with $\varepsilon_b = \max_{1 \le i \le k} Z_i^{(b)}$ with $Z^{(b)} \sim \mathcal{N}(0, \hat{\Gamma}_n^+(\Lambda))$;
5. Compute $\mathrm{ARHT}_{1/3}(\Pi)$;
6. Compute p-value as $B^{-1} \sum_{b=1}^{B} I\{\varepsilon_b > \mathrm{ARHT}_{1/3}(\Pi)\}$.

4.2. *χ^2-Approximation of cut-off values.* While the cube-root transformation is shown to be quite effective, a weighted chi-square approximation can also be used to calibrate the size of ARHT. This involves setting the cut-off values as quantiles of the maximum of a set of scaled χ^2 distributions, that is, random variables of the form $a\chi^2(\ell)$, where a is a normalizing constant and ℓ is the degree of freedom. For each pair (a, ℓ), the $a\chi^2(\ell)$ distribution is used to mimic the distribution of RHT in (1) for a given regularization parameter λ. The scale multipliers a and the degrees of freedom ℓ are selected so that the first two moments and the covariances of the χ^2 variables match with those of the corresponding RHT test. Details are given in the Supplementary Material. Unlike the cube-root transform of Section 4.1, this method only modifies cut-off values. Based on our simulations, both methods perform similar in terms of power curves.

5. Extension to sub-Gaussian distributions. The results presented thus far are now extended to a general class of sub-Gaussian distributions (see Chatterjee (2009)). The extension is achieved for the independent samples model

$$X_{ij} = \mu_i + \Sigma_p^{1/2} Z_{ij}, \quad j = 1, \ldots, n_i, i = 1, 2, \tag{27}$$

where $Z_{ij} = (z_{ij1}, \ldots, z_{ijp})^T$ are p-dimensional independent random vectors with i.i.d. entries satisfying $\mathbb{E}[z_{ijk}] = 0$, $\mathbb{E}[z_{ijk}^2] = 1$ and $\mathbb{E}[z_{ijk}^3] = 0$. To specify the distribution of z_{ijk}, introduce the following class of probability measures.

DEFINITION 5.1. For each $c_1, c_2 > 0$, let $\mathcal{L}(c_1, c_2)$ be the class of probability measures on the real line $\mathbb{R}$ that arises as laws of random variables $u(Z)$, where Z is a standard normal random variable and u is a twice continuously differentiable function such that, for all $x \in \mathbb{R}$,

$$|u'(x)| \le c_1 \quad \text{and} \quad |u''(x)| \le c_2. \tag{28}$$

Note that random variables in $\mathcal{L}(c_1, c_2)$ are sub-Gaussian and have continuous distribution, since u is a Lipschitz function with bounded Lipschitz constant. The first condition in (28) is used to control the magnitude of the variance of $u(Z)$, while the second condition is primarily for controlling the tail behavior of the statistic. This approach is particularly attractive as it only requires establishing appropriate upper bounds for the operator norms of the gradient and Hessian matrices of the statistic (with respect to the variables), and matching the first two asymptotic moments. However, the calculations in our setting are nontrivial since they require a detailed analysis of the resolvent of the sample covariance matrix.

THEOREM 5.1. *All previously stated results hold if the observations X_{ij} are as in (27) with the z_{ijk} satisfying Definition 5.1 together with $\mathbb{E}[z_{ijk}] = 0$, $\mathbb{E}[z_{ijk}^2] = 1$, $\mathbb{E}[z_{ijk}^3] = 0$, and Σ_p satisfying conditions* **C1–C3**.

Key to the proof of Theorem 5.1 is the consideration of a modified version of RHT, replacing S_n with the noncentered matrix $\widetilde{S}_n = n^{-1} \sum_{i=1}^{2} \sum_{j=1}^{n_i} X_{ij} X_{ij}^T$. Defining $U_{kl}(\lambda) = \bar{X}_k^T (\widetilde{S}_n + \lambda I_p)^{-1} \bar{X}_l$, $k, l = 1, 2$, the joint asymptotic normality of $(U_{11}(\lambda), U_{12}(\lambda), U_{22}(\lambda))$ can first be established. Then a suitable transformation of variables and an appropriate use of the δ-method prove the asymptotic normality of RHT(λ). The proof details for Theorem 5.1 are provided in the Supplementary Material. The derivation of the power function of the RHT test under local alternatives follows analogously.

Theorem 5.1 is expected to hold under even more general conditions than stated above. Indeed, in the one-sample testing problem, making use of the analytical framework adopted by Pan and Zhou (2011), asymptotic normality of RHT can be proved when Definition 5.1 is replaced by a bounded fourth moment assumption that is standard in spectral analysis of large covariance matrices. However, this derivation is rather technical and not readily extended to the two-sample setting due to certain structural differences between one- and two-sample settings under non-Gaussianity. Whether such generalizations are feasible in the present context is a topic for future research.

6. Simulations.

6.1. *Competing methods*. In this section, the proposed ARHT is compared by means of a simulation study to a host of popular competing methods, including the tests introduced by Bai and Saranadasa (1996) (BS), Chen and Qin (2010) (CQ), Lopes, Jacob and Wainwright (2011) (RP), and Cai, Liu and Xia (2014) (CLX.$\mathbf{\Omega}^{1/2}$ and CLX.$\mathbf{\Omega}$, corresponding to the two different transformation matrices $\Omega^{1/2}$ and $\Omega = \Sigma^{-1}$). In the following, ARHT, ARHT$_{1/3}$ and ARHT$_{\chi^2}$ denote the original, cubic-root transformed and χ^2-approximated ARHT procedure introduced in Sections 3, 4.1 and 4.2, respectively.

6.2. *Settings and results*. In the simulations, the observations X_{ij} are as in (27), while two different distributions for z_{ijk} are considered, namely the $N(0, 1)$ distribution and the t-distribution with four degrees of freedom, $t_{(4)}$, rescaled to unit variance. For the normal case, the sample sizes are chosen as $n_1 = n_2 = 50$. For the $t_{(4)}$ case, the sample sizes are chosen to be $n_1 = 30$ and $n_2 = 70$. The dimension p is 50, 200 or 1000, so that $\gamma = p/(n_1 + n_2) = 0.5$, 2 or 10. Results are here reported mainly for $p = 200$ and 1000, while the case $p = 50$ is reported in the Supplementary Material. The range of regularization parameters is chosen as $[\underline{\lambda}, \overline{\lambda}] = [0.01, 100]$, using a grid with progressively coarser spacings for determining the optimal $\lambda_n \equiv \lambda_{\pi,n}$.

The following three models for the covariance matrix $\Sigma = \Sigma_p$ are considered:

(i) The *identity matrix* (ID): Here, $\Sigma = I_p$;

(ii) The *sparse case* Σ_s: Here, $\Sigma = (p^{-1}\operatorname{tr}\{D\})^{-1}D$ with a diagonal matrix D whose eigenvalues are given by $\tau_j = 0.01 + (0.1 + j)^6$, $j = 1, \dots, p$;

(iii) The *dense case* Σ_d: Here, $\Sigma = P^T \Sigma_s P$ with a unitary matrix P randomly generated from the Haar measure and resampled for each different setting. Note that, for both Σ_s and Σ_d, the eigenvalues decay slowly to 0, so that no dominating leading eigenvalue exists.

Under the alternative, for each p, Σ and each replicate, the mean difference vector $\mu = \mu_1 - \mu_2$ is randomly generated from one of the four models: (1) $\mu \sim N(0, cI_p)$; (2) $\mu \sim N(0, c\Sigma)$; (3) $\mu \sim N(0, c\Sigma^2)$; and (4) μ is sparse with 5% randomly selected nonzero entries being either $-c$ or c with probability 1/2 each. The parameter c is used to control the signal size. The choices in (1)–(4), respectively, represent the cases where μ is uniform; is slightly tilted toward the eigenvectors corresponding to large eigenvalues of Σ; is heavily tilted toward the eigenvectors corresponding to large eigenvalues of Σ; and is sparse, respectively.

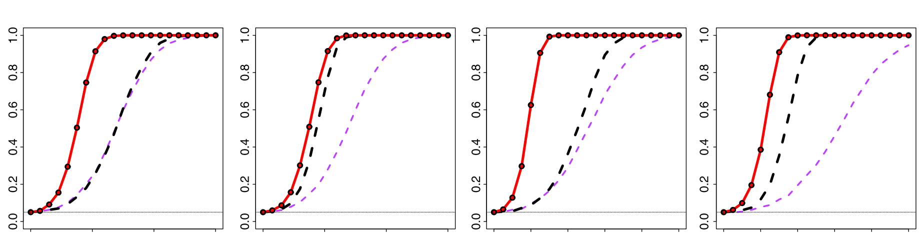
All tests are conducted at significance level $\alpha = 0.05$. There are two versions for each test: (a) utilizing (approximate) asymptotic cut-off values; and (b) utilizing the size-adjusted cut-off values based on the actual null distribution computed by simulations. Only results for the latter case are reported here; the former is in the Supplementary Material. Also, power graphs

TABLE 1
Empirical sizes of the various tests at the $\alpha = 0.05$ level

	Σ	p	ARHT	$\text{ARHT}_{1/3}$	ARHT_{χ^2}	BS	CQ	RP	$\text{CLX}.\Omega^{1/2}$	$\text{CLX}.\Omega$
$N(0,1)$	ID	50	0.0612	0.0447	0.0472	0.0609	0.0481	0.0520	0.0633	0.0637
$N(0,1)$	ID	200	0.0568	0.0473	0.0493	0.0561	0.0508	0.0490	0.0754	0.0757
$N(0,1)$	ID	1000	0.0539	0.0491	0.0510	0.0527	0.0517	0.0498	0.1004	0.1004
$N(0,1)$	Σ_d	50	0.0854	0.0489	0.0606	0.0695	0.0470	0.0485	0.0970	0.1101
$N(0,1)$	Σ_d	200	0.0917	0.0601	0.0705	0.0622	0.0486	0.0503	0.0833	0.0971
$N(0,1)$	Σ_d	1000	0.0626	0.0520	0.0347	0.0555	0.0484	0.0510	0.0991	0.0996
$N(0,1)$	Σ_s	50	0.0877	0.0492	0.0603	0.0688	0.0468	0.0508	0.0613	0.0615
$N(0,1)$	Σ_s	200	0.0938	0.0596	0.0707	0.0645	0.0487	0.0503	0.0773	0.0773
$N(0,1)$	Σ_s	1000	0.0642	0.0539	0.0347	0.0580	0.0510	0.0486	0.0991	0.0992
$t_{(4)}$	ID	50	0.0572	0.0395	0.0414	0.0516	0.0450	0.0477	0.0562	0.0563
$t_{(4)}$	ID	200	0.0541	0.0447	0.0456	0.0518	0.0505	0.0504	0.0611	0.0611
$t_{(4)}$	ID	1000	0.0502	0.0460	0.0443	0.0487	0.0527	0.0493	0.0735	0.0735
$t_{(4)}$	Σ_d	50	0.0836	0.0473	0.0582	0.0659	0.0468	0.0485	0.0815	0.0906
$t_{(4)}$	Σ_d	200	0.0912	0.0582	0.0692	0.0590	0.0484	0.0507	0.0759	0.0838
$t_{(4)}$	Σ_d	1000	0.0606	0.0503	0.0313	0.0541	0.0500	0.0494	0.0905	0.0906
$t_{(4)}$	Σ_s	50	0.0812	0.0451	0.0559	0.0634	0.0449	0.0481	0.0512	0.0512
$t_{(4)}$	Σ_s	200	0.0872	0.0551	0.0656	0.0565	0.0469	0.0474	0.0638	0.0638
$t_{(4)}$	Σ_s	1000	0.0584	0.0481	0.0246	0.0516	0.0502	0.0495	0.0730	0.0730

are given for the Gaussian case only, since power curves for the $t_{(4)}$ case are similar (see Supplementary Material). All empirical cut-off values, powers and sizes are calculated based on 10,000 replications. Empirical sizes for the various tests are shown in Table 1. Empirical power curves versus expected signal strength $(\sqrt{n}\mathbb{E}[\|\mu\|_2^2])^{1/2}$ are shown in Figures 4–7. Note that, in some of the settings, several of the power curves nearly overlap, creating an occlusion effect. For example, $\text{CLX}.\Omega^{1/2}$ is very similar to $\text{CLX}.\Omega$, therefore, only the latter is displayed. For the ease of illustration, power curves corresponding to the recommended $\text{ARHT}_{1/3}$ are plotted as the top layer.

6.3. *Summary of simulation results.* For each simulation configuration considered in this study, ARHT or its calibrated versions are as powerful as the procedure(s) with the best performance, except for the cases of sparse or uniform μ with sparse Σ and relatively large p (panels (a) and (d) of Figures 7 and 8). This serves as evidence for the robustness of ARHT procedures with respect to the structures of means under alternatives. The adaptable behavior also sets the proposed methodology apart from its competitors. The following observations are made based on the simulation outcomes:



(a) p=200, $\mu \sim N(0, cI)$ (b) p=200, sparse μ (c) p=1000, $\mu \sim N(0, cI)$ (d) p=1000, sparse μ

FIG. 4. *Size-adjusted empirical power with $X_{ij} \sim N(\cdot, \Sigma)$ and $\Sigma = \text{ID}$.* $\text{ARHT}_{1/3}$ (*solid, red*), χ^2 *approximation* (*circle*), *BS* (*solid, blue*), *CQ* (+), *RP* (*dashed, purple*) *and CLX*.Ω (*dashed*).

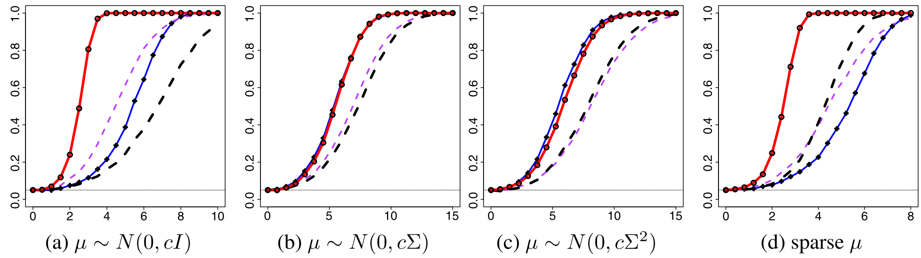


FIG. 5. *Size-adjusted empirical power with $X_{ij} \sim N(\cdot, \Sigma)$, $\Sigma = \Sigma_d$ and $p = 200$.* $\text{ARHT}_{1/3}$ (*solid, red*), χ^2 *approximation* (*circle*), *BS* (*solid, blue*), *CQ* (+), *RP* (*dashed, purple*) *and CLX*.Ω (*dashed*).

(1) When the dimension is high and there is no specific structure of μ and Σ that could be exploited, ARHT tends to outperform the other tests. Tilted alternatives are expected to be detrimental to the performance of both ARHT and RP. However, ARHT can be seen as only slightly less powerful than BS and CQ, which yield the best results for this case.

(2) In the case that Σ is equal to the identity matrix, the BS procedure is expected to give the best performance, since the test statistic is based on the true covariance matrix. Recalling that BS can be treated as $\text{RHT}(\infty)$, ARHT is shown to perform as well as BS in corresponding simulations (see Figure 4). This may be viewed as evidence of the effectiveness of the data-driven tuning parameter selection strategy detailed in Section 2.5.

(3) If both mean difference vector μ and covariance matrix Σ are sparse, the three CLX procedures are expected to perform the best. Specifically, the simulations reveal that the sparsity of μ alone does not guarantee superiority of CLX. This can be seen in the panel (d) of Figures 4–5. However, as evidenced in Figures 7 and 8, if Σ is sparse, then the performance of the CLX procedures is the best when μ is either uniform or sparse. The ARHT procedures are less sensitive to the structure imposed on the covariance matrix Σ than the CLX procedures, although they are less powerful in sparse settings.

The reason for the excellent performance of CLX for uniform μ (which is even better than for sparse μ) is that significant signals occur, with high probability due to uniform distribution of signal, at coordinates with very small variance due to their high signal-to-noise ratios. Consequently, l_∞-norm based methods, such as the CLX tests, are able to efficiently detect such signals. In contrast, all l_2-norm based methods, including ARHT, combine the signals over all coordinates and thus tend to miss such signals since the l_2 norm of μ is relatively small. When μ is sparse, such a phenomenon also happens but with smaller probability. When μ is tilted, on the other hand, this phenomenon is unlikely to occur. Therefore, what is at play is not only sparsity of μ, but also the matching of significant signals with small variances.

The results of this simulation study highlight the robustness or adaptivity of the proposed ARHT test to various different alternative scenarios and, therefore, demonstrate its potential usefulness for real world applications.

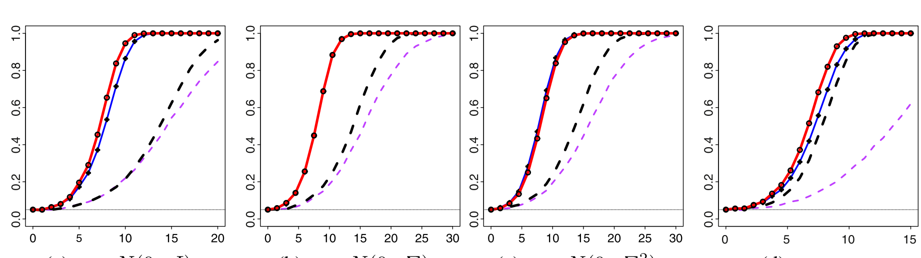
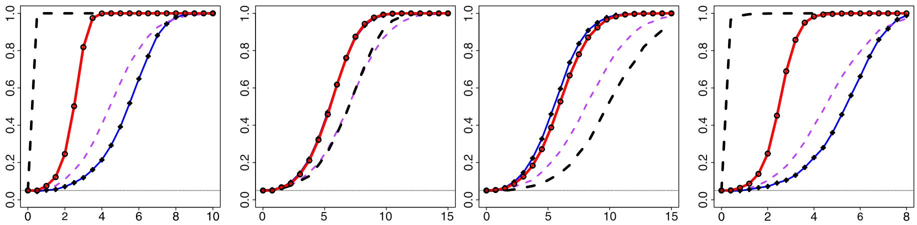


FIG. 6. *Same as in Figure 5 but with $p = 1000$.*



(a) $\mu \sim N(0, cI)$ (b) $\mu \sim N(0, c\Sigma)$ (c) $\mu \sim N(0, c\Sigma^2)$ (d) sparse μ

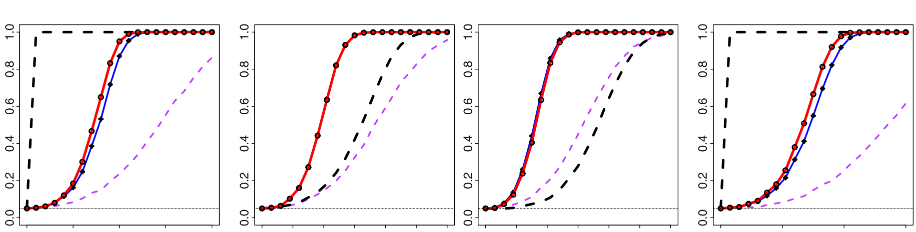
FIG. 7. *Size-adjusted empirical power with* $X_{ij} \sim N(\cdot, \Sigma)$, $\Sigma = \Sigma_s$ *and* $p = 200$. $\mathrm{ARHT}_{1/3}$ (*solid*, *red*), χ^2 *approximation* (*circle*), *BS* (*solid*, *blue*), *CQ* (+), *RP* (*dashed*, *purple*) *and CLX.*Ω (*dashed*).

7. Application. Breast cancer is one of the most common cancers with more than 1,300,000 cases and 450,000 deaths worldwide each year. Breast cancer is also a heterogeneous disease, consisting of several subtypes with distinct pathological and clinical characteristics. To better understand the disease mechanisms underlying different breast cancer subtypes, it is of great interest to characterize subtype-specific somatic *copy number alteration* (*CNA*) patterns, that have been shown to play critical roles in activating oncogenes and in inactivating tumor suppressors during the breast tumor development; see (Bergamaschi et al. (2006)). In this section, the proposed ARHT is applied to a TCGA (The Cancer Genome Atlas) breast cancer data set (Cancer Genome Atlas Network (2012)) to detect pathways showing distinct CNA patterns between different breast cancer subtypes.

Level-three segmented DNA copy number (CN) data of breast cancer tumor samples were obtained from the TCGA website. Focus is on a subset of 80 breast tumor samples, which are also subjected to deep protein-profiling by CPTAC (Clinical Proteomic Tumor Analysis Consortium) (Paulovich et al. (2010), Ellis et al. (2013), Mertins et al. (2016)). Thus findings from our analysis may lead to further investigations and knowledge generation through the corresponding protein profiles in the future. Specifically, among these 80 samples, 18, 29 and 33 samples belong to the Her2-enriched (Her2), Luminal A (Lum A) and Luminal B (Lum B) subtypes, respectively.

For the selected samples, first gene-level copy number estimates are derived based on the segmented CN profiles. Q-Q plots, provided in the Supplementary Material, suggest that the observations have heavier tails than normal distributions. To better illustrate the comparative performance of the proposed methods under high dimensions, consider the 36 largest KEGG pathways. The number of genes in these pathways ranging from 66 to 252, so that p/n varies between 0.75 and 3.5. For each pathway, interest is in testing whether genes in the pathway showed different copy number alterations between Lum (Lum A plus Lum B) vs. Her2, or Lum A vs. Lum B. These led to a total of 72 two-sample tests.

All testing methods discussed in the simulation studies were applied to this data set, except for ARHT_{χ^2}. The null distribution and the p-value for each method, were generated based



(a) $\mu \sim N(0, cI)$ (b) $\mu \sim N(0, c\Sigma)$ (c) $\mu \sim N(0, c\Sigma^2)$ (d) sparse μ

FIG. 8. *Same as in Figure 7 but with* $p = 1000$.

on 100,000 permutations, instead of applying the asymptotic theory, though the asymptotic and permutation-based cut-offs are similar for $\text{ARHT}_{1/3}$. Also, to control the familywise error rate, the p-values are further adjusted by FDR (Benjamini and Hochberg (1995)), and FDR-adjusted p-values below 0.01 indicate departure from null.

For the Lum vs. Her2 comparison, ARHT yielded the largest number of significant pathways followed by RP, while all other methods have similar behaviors with about half the detection rate of ARHT and RP. For the Lum A vs. Lum B comparison, the ARHT results are similar to those of BS and CQ, giving the largest number of significant pathways. On the other hand, in this case, RP only detected two while the three CLX methods did not detect any significant pathway.

One unique characteristic of Her2 subtype tumors is the amplification of gene ERBB2 and its neighboring genes in cytoband 17q12, including MED1, STARD3 and others. There are 7 pathways containing at least one of these genes. These pathways, whose annotations were colored in red in Figure 9, can serve as positive controls in the Her2 vs Lum comparison (Lamy et al. (2011)). Moreover, it has been shown that gene MAP3K1 and MAP2K4 have different CN loss activities in Lum A and Lum B tumors (Creighton (2012)). In addition, proliferation genes such as CCNB1, MKI67 and MYBL2 are more highly expressed in Lum B compared to Lum A, as shown in Tran and Bedard (2011). Thus, the pathways containing these genes can be viewed as positive controls in the Lum A vs. Lum B comparison analysis. As an illustrative reference, in Table 2, the performance of different procedures is summarized in terms of detecting the pathways known to have different CN alterations between subtypes, when FDR is controlled at 0.01. Interestingly, only the three ARHT procedures successfully detected all these pathways of positive controls, suggesting a superior power of ARHT procedures over the competitors. BS and CQ appeared to be the second best methods.

In summary, for this data, only ARHT consistently makes correct decisions on pathways known to be significant, while the other methods perform adequately for at most one of the comparisons—either Lum vs. Her2 or Lum A vs. Lum B. This provides further evidence in support of the power and robustness of ARHT.

8. Discussion. In this paper, a powerful and computationally tractable procedure for testing equality of mean vectors between two populations was presented that is based on a composite ridge-type regularization of Hotelling's T^2 statistics. Techniques from random matrix theory were used to derive the asymptotic null distribution under a regime where the dimension is comparable to the sample sizes. Extensive simulations were conducted to show that the proposed test has excellent power for a wide class of alternatives and is fairly robust to the structure of the covariance matrix as well as the distribution of the observations. Practical advantages of the proposed test were illustrated in the context of a breast cancer data analysis where the goal was to detect pathways with different DNA copy number alteration patterns between cancer subtypes.

There are several future research directions to pursue. On the technical side, aim could be on relaxing the distributional assumptions on the observations further, only requiring the existence of a certain number of moments. On the methodological front, aim could be on the extension of the framework to tests for mean difference under possibly unequal variances, and to deal with the MANOVA problem in high-dimensional settings. Another potentially interesting direction is to combine the proposed methodology with a variable screening strategy so that the test can be adapted to ultrahigh-dimensional settings.

9. Proofs of the main results. In this section, we provide the necessary technical support for the proposed methodology under the class of sub-Gaussian distributions $\mathcal{L}(c_1, c_2)$ introduced in Section 5. The technical details consist of the following four parts: (i) proof of

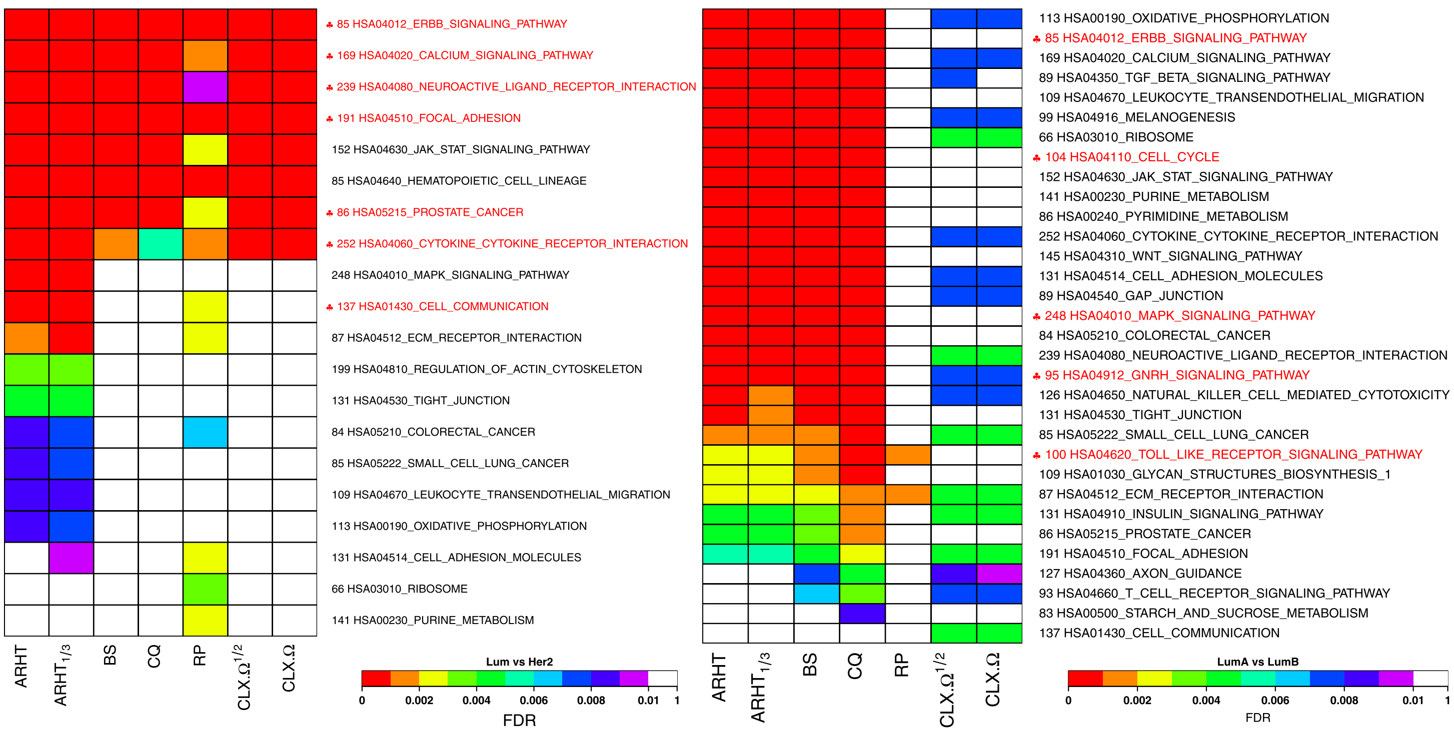


FIG. 9. *Lum vs. Her*2 (*left panel*) *and Lum A vs. Lum B* (*right panel*). *Row labels show pathway names and size* (p), *with those known to be significant highlighted in red.*

TABLE 2
Comparative performance on known significant pathways (at FDR level 0.01)

	Lum vs Her2	Lum A vs Lum B
ARHT	7/7	5/5
$\text{ARHT}_{1/3}$	7/7	5/5
BS	6/7	5/5
CQ	6/7	5/5
RP	7/7	1/5
$\text{CLX}.\Omega^{1/2}$	6/7	1/5
$\text{CLX}.\Omega$	6/7	1/5

asymptotic normality; (ii) proof of Theorem 2.1 and Theorem 2.2; (iii) proof of Theorem 2.3 and (iv) proof of Theorem 3.1.

The crucial difference between Gaussianity and non-Gaussianity is that in the Gaussian case, the sample covariance matrix S_n is independent of the sample means and can be written as sum of independent random elements. Indeed, under Gaussianity, $S_n = \sum_{i=1}^{n-2} \Sigma_p^{1/2} Y_i Y_i^T \Sigma_p^{1/2}$ with $Y_j \sim \mathcal{N}(0, (n-2)^{-1} I_p)$ is independent of the $\bar{X}_i$'s, with the latter normally distributed. However, in non-Gaussian settings, due to lack of independence between S_n and $\bar{X}_i$'s, their mutual correlation has to be disentangled carefully.

For this analysis, following common practice in random matrix theory, we use an uncentered version of the sample covariance, defined as

$$\tilde{S}_n = n^{-1} \sum_{i=1}^{2} \sum_{j=1}^{n} X_{ij} X_{ij}^T.$$

Note that

$$S_n = \frac{n}{n-2} \tilde{S}_n - \frac{n_1}{n-2} \bar{X}_1 \bar{X}_1^T - \frac{n_2}{n-2} \bar{X}_2 \bar{X}_2^T.$$

The statistic $(\bar{X}_1 - \bar{X}_2)^T (S_n + \lambda I_p)^{-1} (\bar{X}_1 - \bar{X}_2)$ changes nontrivially if S_n is replaced with $\tilde{S}_n$. It will be shown in the following proofs how to manipulate their difference. Recall the following definitions:

$$R_n(z) = (S_n - z I_p)^{-1}, \qquad \hat{\phi}_1 = p^{-1} \operatorname{tr}(S_n),$$

$$m_{F_{n,p}}(-\lambda) = p^{-1} \operatorname{tr}\{R_n(-\lambda)\},$$

$$\hat{\Theta}_1(\lambda, \gamma_n) = \frac{1 - \lambda m_{F_{n,p}}(-\lambda)}{1 - \gamma_n \{1 - \lambda m_{F_{n,p}}(-\lambda)\}},$$

$$\hat{\Theta}_2(\lambda, \gamma_n) = \frac{1 - \lambda m_{F_{n,p}}(-\lambda)}{[1 - \gamma_n \{1 - \lambda m_{F_{n,p}}(-\lambda)\}]^3} - \lambda \frac{\{m_{F_{n,p}}(-\lambda) - \lambda m'_{F_{n,p}}(-\lambda)\}}{[1 - \gamma_n \{1 - \lambda m_{F_{n,p}}(-\lambda)\}]^4}.$$

For the sake of brevity, S_n is replaced with $\tilde{S}_n$ in all these quantities and proofs are provided, even in the Gaussian case, using the thus modified versions. Because

$$|p^{-1} \operatorname{tr}(S_n) - p^{-1} \operatorname{tr}(\tilde{S}_n)| = O_p(1/p),$$

$$|p^{-1} \operatorname{tr}\{(S_n + \lambda I_p)^{-k}\} - p^{-1} \operatorname{tr}\{(\tilde{S}_n + \lambda I_p)^{-k}\}| \leq 2k\lambda^{-k} p^{-1},$$

all the derivations all results put forward in the rest of this section will also hold for the original quantities. The argument for the first relation is straightforward and the second argument is deduced from Proposition A.1. To lighten notation, $\hat{\phi}_1$, $R_n(z)$, $m_{F_{n,p}}(-\lambda)$, $\hat{\Theta}_1(\lambda, \gamma_n)$, $\hat{\Theta}_2(\lambda, \gamma_n)$, etc., are used to denote their counterparts after the replacement of S_n by $\tilde{S}_n$.

As mentioned above, the proposed statistic and other quadratic terms involving S_n will change significantly after the redefinition of S_n. Define

$$U_{ii'}(\lambda) = \bar{X}_i^T (\tilde{S}_n + \lambda I_p)^{-1} \bar{X}_{i'}, \quad i, i' = 1, 2. \tag{29}$$

The Woodbury matrix identity gives

$$\begin{aligned} &\frac{n}{n-2}\left(S_n + \frac{n}{n-2}\lambda I_p\right)^{-1} \\ &\quad = (\tilde{S}_n + \lambda I_p)^{-1} + (\tilde{S}_n + \lambda I_p)^{-1}(\bar{X}_1, \bar{X}_2)\mathbb{H}^{-1}\begin{pmatrix}\bar{X}_1^T \\ \bar{X}_2^T\end{pmatrix}(\tilde{S}_n + \lambda I_p)^{-1}, \end{aligned} \tag{30}$$

where

$$\mathbb{H} = \begin{pmatrix} nn_1^{-1} & 0 \\ 0 & nn_2^{-1} \end{pmatrix} - \begin{pmatrix} U_{11}(\lambda) & U_{12}(\lambda) \\ U_{21}(\lambda) & U_{22}(\lambda) \end{pmatrix}.$$

Therefore,

$$\begin{aligned} &\frac{n}{n-2}\mathrm{RHT}\left(\frac{n}{n-2}\lambda\right) \\ &\quad = \frac{n_1 n_2}{n_1 + n_2}\Bigg[(U_{11}(\lambda) + U_{22}(\lambda) - 2U_{12}(\lambda)) \\ &\qquad + \begin{pmatrix} U_{11}(\lambda) - U_{12}(\lambda) \\ U_{12}(\lambda) - U_{22}(\lambda) \end{pmatrix}^T \mathbb{H}^{-1} \begin{pmatrix} U_{11}(\lambda) - U_{12}(\lambda) \\ U_{12}(\lambda) - U_{22}(\lambda) \end{pmatrix}\Bigg]. \end{aligned} \tag{31}$$

9.1. *Proof of asymptotic normality under sub-Gaussianity.* It follows from (31) that $\mathrm{RHT}(n(n-2)^{-1}\lambda)$ can be expressed as a differentiable function of $U_{11}(\lambda)$, $U_{12}(\lambda)$ and $U_{22}(\lambda)$. Hence, the joint asymptotic normality of the latter implies the asymptotic normality of the former. Therefore, define an arbitrary linear combination,

$$\bar{R}(\lambda) = n^{1/2}[l_{11}U_{11}(\lambda) + l_{12}U_{12}(\lambda) + l_{22}U_{22}(\lambda)]$$

for any $l_{11}, l_{12}, l_{22} \in \mathbb{R}$. It suffices to show that $\bar{R}(\lambda)$ is asymptotically normal.

To this end, we use Theorem A.1. A key component of the proof is to establish the asymptotic orders of $\varrho_0(\bar{R})$, $\varrho_1(\bar{R})$ and $\varrho_2(\bar{R})$ and also $\mathrm{Var}(\bar{R})$. Since the gradient and Hessian of $\bar{R}(\lambda)$ are linear functions of those of $n^{1/2}U_{11}(\lambda)$, $n^{1/2}U_{12}(\lambda)$ and $n^{1/2}U_{22}(\lambda)$, it suffices to derive asymptotic orders of the functions ϱ_0, ϱ_1 and ϱ_2 with $n^{1/2}U_{11}$, $n^{1/2}U_{12}$ and $n^{1/2}U_{22}$ as arguments, then combining them through the Cauchy–Schwarz inequality. In the rest of the proof, only the asymptotic order of $\varrho_0(p^{1/2}U_{11})$, $\varrho_1(p^{1/2}U_{11})$ and $\varrho_2(p^{1/2}U_{11})$ is derived as similar arguments also work for U_{12} and U_{22}.

PROPOSITION 9.1. *Under the assumptions of Theorem* 5.1, $\varrho_0(\sqrt{n}U_{11}) = o(1)$.

PROPOSITION 9.2. *Under the assumptions of Theorem* 5.1, $\varrho_1(\sqrt{n}U_{11}) = o(n^{1/2})$.

PROPOSITION 9.3. *Under the assumptions of Theorem* 5.1, $\varrho_2(\sqrt{n}U_{11}) = O(n^{-1/2})$.

PROPOSITION 9.4. *Under the assumptions of Theorem 5.1,*

$$\mathbb{E}\bar{R}(\lambda) = \frac{(l_{11}/k + l_{22}/(1-\kappa))\gamma\Theta_1(\lambda,\gamma)}{1+\gamma\Theta_1(\lambda,\gamma)} + o(1),$$

$$\operatorname{Var}(\bar{R}(\lambda)) = \frac{[2l_{11}^2/\kappa^2 + l_{12}^2/(\kappa-\kappa^2) + 2l_{22}^2/(1-\kappa)^2]\gamma^2\Theta_2(\lambda,\gamma)}{(1+\gamma\Theta_1(\lambda,\gamma))^4} + o(1),$$

$$\operatorname{Cov}(\bar{R}(\lambda), \bar{R}(\lambda')) = \frac{[2l_{11}^2/\kappa^2 + l_{12}^2/(\kappa-\kappa^2) + 2l_{22}^2/(1-\kappa)^2]\gamma^2\Theta_3(\lambda,\lambda',\gamma)}{(1+\gamma\Theta_1(\lambda,\gamma))^2(1+\gamma\Theta_1(\lambda',\gamma))^2} + o(1),$$

where for $\lambda \neq \lambda'$,

$$\Theta_3(\lambda,\lambda',\gamma) = (1+\gamma\Theta_1(\lambda,\gamma))(1+\gamma\Theta_1(\lambda',\gamma))\frac{(\lambda'\Theta_1(\lambda',\gamma) - \lambda\Theta_1(\lambda,\gamma))}{(\lambda'-\lambda)}.$$

The proofs of these propositions are given in Section S.2. Since $\bar{R}$ has finite fourth moment, it follows immediately from Propositions 9.1 and 9.4 that

$$d_{\mathrm{TV}}(\bar{R}, U) \le 2\sqrt{5}(\operatorname{Var}(\bar{R}))^{-1}\{c_1c_2\varrho_0(\bar{R}) + c_1^3\varrho_1(\bar{R})\varrho_2(\bar{R})\} \to 0,$$

where U is a normal random variable with the same mean and variance as $\bar{R}$. The asymptotic normality of $\bar{R}$ now follows. From this, the asymptotic mean and variance of $\mathrm{RHT}(\lambda)$ follow from basic calculus, making use of the δ-method and the relation shown in (31). Details are omitted. Finally, we are able to conclude

$$\sqrt{p}\frac{\{p^{-1}\,\mathrm{RHT}(\lambda) - \Theta_1(\lambda,\gamma)\}}{\{2\Theta_2(\lambda,\gamma)\}^{1/2}} \Longrightarrow \mathcal{N}(0,1).$$

9.2. *Proof of Theorem* 2.1. Under the deterministic local alternative, we denote $Y_{ij} = X_{ij} - \mu_i$. Then

$$S_n = \frac{1}{n-2}\sum_{i=1}^{2}\sum_{j=1}^{n_i} Y_{ij}Y_{ij}^T - \frac{n_1}{n-2}\bar{Y}_1\bar{Y}_1^T - \frac{n_2}{n-2}\bar{Y}_2\bar{Y}_2^T.$$

Furthermore, redefine

$$\tilde{S}_n = \frac{1}{n}\sum_{i=1}^{2}\sum_{j=1}^{n_i} Y_{ij}Y_{ij}^T.$$

With $g_n = \kappa_n(1-\kappa_n)\{2\gamma_n\hat{\Theta}_2(\lambda,\gamma_n)\}^{-1/2}$, the statistic under the local alternative can be written as

$$\begin{aligned} T_{n,p}(\lambda) &= T_{n,p}^0(\lambda) + g_n n^{1/2}\mu^T(S_n + \lambda I_p)^{-1}\mu \\ &\quad - 2g_n n^{1/2}\mu^T(S_n+\lambda I_p)^{-1}\bar{Y}_1 + 2g_n n^{1/2}\mu^T(S_n+\lambda I_p)^{-1}\bar{Y}_2, \end{aligned} \tag{32}$$

where $T_{n,p}^0(\lambda)$ is the standardized statistic with $\{Y_{ij}\}$ as observations. We already proved $T_{n,p}^0(\lambda)$ converges to $\mathcal{N}(0,1)$ in distribution. To this end, it is enough to show that, under the stability condition (7),

$$n^{1/2}\mu^T(S_n+\lambda I_p)^{-1}\mu - q(\lambda,\gamma) = o_p(1),$$

$$n^{1/2}\mu^T(S_n+\lambda I_p)^{-1}\bar{Y}_i = o_p(1), \quad i = 1, 2.$$

Using the relation shown in (30), we can write

$$\begin{aligned}
&n^{1/2}\mu^T(S_n+\lambda I_p)^{-1}\mu\\
&\quad= n^{1/2}\mu^T(\tilde{S}_n+\lambda I_p)^{-1}\mu + n^{1/2}(U_{\mu,1},U_{\mu,2})\mathbb{H}^{-1}\begin{pmatrix}U_{\mu,1}\\ U_{\mu,2}\end{pmatrix},\\
&n^{1/2}\mu^T(S_n+\lambda I_p)^{-1}\bar{Y}_1\\
&\quad= n^{1/2}U_{\mu,1} + n^{1/2}(U_{\mu,1},U_{\mu,2})\mathbb{H}^{-1}\begin{pmatrix}U_{11}\\ U_{12}\end{pmatrix},\\
&n^{1/2}\mu^T(S_n+\lambda I_p)^{-1}\bar{Y}_2\\
&\quad= n^{1/2}U_{\mu,2} + n^{1/2}(U_{\mu,1},U_{\mu,2})\mathbb{H}^{-1}\begin{pmatrix}U_{12}\\ U_{22}\end{pmatrix},
\end{aligned}$$

where $U_{ii'}$ and $\mathbb{H}$ are defined in the same way as in (29) and (30), but with X_{ij} replaced by Y_{ij}, and $U_{\mu,i} = \mu^T(\tilde{S}_n+\lambda I_p)^{-1}\bar{Y}_i$, $i = 1, 2$.

Proposition 9.4 implies that U_{11}, U_{12}, U_{22} converge in probability to deterministic quantities and $\mathbb{H}$ converges in probability to a nonsingular matrix. Therefore, it suffices to show

$$n^{1/2}\mu^T(\tilde{S}_n+\lambda I_p)^{-1}\mu - q(\lambda,\gamma) = o_p(1), \tag{33}$$

$$n^{1/2}\mu^T(\tilde{S}_n+\lambda I_p)^{-1}\bar{Y}_i = o_p(1), \quad i = 1, 2. \tag{34}$$

Equation (33) is a special case of the limiting behavior of quadratic forms considered by El Karoui and Kösters (2011), and its proof follows along the material in Section 2 and Section 3 of their paper. The proof of (34) is given in Section S.3.5 of the Supplementary Material.

9.3. *Proof of Theorem 2.2.* We simply prove the result with $p^{-1}\operatorname{tr}\{D_p(-\lambda)\mathbf{B}\}$ replaced by $q(\lambda,\gamma)$. Under the prior distribution given by **PA**, decompose $T_{n,p}(\lambda)$ as

$$T_{n,p}(\lambda) = T^0_{n,p}(\lambda) + gq(\lambda,\gamma) + \sigma_n(\mu) + \sum_{i=1}^{2}\eta_n^{(i)}(Y) + \sum_{j=1}^{4}\delta_n^{(j)}(\mu,Y),$$

where, with $g = \kappa(1-\kappa)\{2\gamma\Theta_2(\lambda,\gamma)\}^{-1/2}$ and $g_n = \kappa(1-\kappa)\{2\gamma_n\hat{\Theta}_2(\lambda,\gamma_n)\}^{-1/2}$,

$$\begin{aligned}
\sigma_n(\mu) &= g\big[n^{1/2}\mu^T D(-\lambda)\mu - p^{-1}\operatorname{tr}\big(D(-\lambda)\mathbf{B}\big)\big],\\
\eta_n^{(1)}(Y) &= (g_n - g)q(\lambda,\gamma),\\
\eta_n^{(2)}(Y) &= g_n\big[p^{-1}\operatorname{tr}\big(D(-\lambda)\mathbf{B}\big) - q(\lambda,\gamma)\big],\\
\delta_n^{(1)}(\mu,Y) &= (g_n-g)\big[n^{1/2}\mu^T D(-\lambda)\mu - p^{-1}\operatorname{tr}\big(D(-\lambda)\mathbf{B}\big)\big],\\
\delta_n^{(2)}(\mu,Y) &= g_n\big[n^{1/2}\mu^T(S_n+\lambda I_p)^{-1}\mu - n^{1/2}\mu^T D(-\lambda)\mu\big],\\
\delta_n^{(3)}(\mu,Y) &= g_n n^{1/2}\mu^T(S_n+\lambda I_p)^{-1}\bar{Y}_1,\\
\delta_n^{(4)}(\mu,Y) &= g_n n^{1/2}\mu^T(S_n+\lambda I_p)^{-1}\bar{Y}_2.
\end{aligned}$$

Through this subsection, we use $\mathbb{P}_*$ to mean the prior probability measure of μ and use $\mathbb{P}_\mu$ to mean the probability of X_{ij} conditional on μ. The power under the alternative μ is then

$$\beta_n(\mu,\lambda) = \mathbb{P}_\mu\{T_{n,p}(\lambda) > \xi_\alpha\}.$$

To show (12), it suffices to show that for any $\epsilon > 0$ and any $\zeta > 0$, there exists a sufficiently large N, such that when $n > N$,

$$\mathbb{P}_*\big(\big|\beta_n(\mu,\lambda) - \Phi\big(-\xi_\alpha + gq(\lambda,\gamma)\big)\big| > \epsilon\big) < \zeta.$$

Due to Lemma 2.7 of Bai and Silverstein (1998) and the assumption $\mu = n^{-1/4}p^{-1/2}B\nu$,

$$n^{1/2}\mu^T D(-\lambda)\mu - p^{-1}\operatorname{tr}(D(-\lambda)BB^T) \xrightarrow{\mathbb{P}_*} 0.$$

Therefore, there exist a constant C_ϵ and a sufficiently large N_1 such that when $n > N_1$,

$$\mathbb{P}_*(K_\epsilon^{(1)}) \geq 1 - \zeta,$$

where

$$K_\epsilon^{(1)} = \{\mu : n^{1/2}\|\mu\|^2 \leq C_\epsilon\} \cap \{\mu : |\sigma_n(\mu)| \leq \epsilon\}.$$

Next, g_n is independent with μ and as introduced in Section 2.1,

$$g_n \xrightarrow{\mathbb{P}_\mu} g \quad \text{as } n, p \to \infty.$$

Therefore, when $\mu \in K_\epsilon^{(1)}$, as $n, p \to \infty$, with a tail bound not depending on μ,

$$\max_{i=1,2}|\eta_n^{(i)}(Y)| \xrightarrow{\mathbb{P}_\mu} 0 \quad \text{and} \quad |\delta_n^{(1)}(\mu, Y)| \xrightarrow{\mathbb{P}_\mu} 0.$$

As for $\delta_n^{(j)}(\mu, Y)$, $j = 2, 3, 4$, arguments analogous to those in Theorem 3.1 and Proposition 3.1 of El Karoui and Kösters (2011) show that, as $n, p \to \infty$,

$$n^{1/2}\mu^T(\tilde{S}_n + \lambda I_p)^{-1}\mu - n^{1/2}\mu^T D(-\lambda)\mu \xrightarrow{\mathbb{P}_\mu} 0,$$

with a tail bound only depending on $n^{1/2}\|\mu\|^2$. Moreover, the proof of Theorem 2.2 shows

$$n^{1/2}\mu^T(\tilde{S}_n + \lambda I_p)^{-1}\bar{Y}_i \xrightarrow{\mathbb{P}_\mu} 0, \quad i = 1, 2,$$

also with a tail bound only depending on $n^{1/2}\|\mu\|^2$ (see Section S.3.5 of the Supplementary Material). Together with the relation shown in (30), we conclude that on $\mu \in K_\epsilon^{(1)}$, with an uniform tail bound, as $n, p \to \infty$,

$$\max_{j=2,3,4}|\delta_n^{(j)}(\mu, Y)| \xrightarrow{\mathbb{P}_\mu} 0.$$

The analysis up to now implies that we can find a sufficiently large N_2 such that when $n > N_2$,

$$\mathbb{P}_\mu(K_\epsilon^{(2)}) > 1 - \epsilon,$$

for any $\mu \in K_\epsilon^{(1)}$, where

$$K_\epsilon^{(2)} = K_\epsilon^{(1)} \cap \Big\{Y_{ij} : \max_{i=1,2}|\eta_n^{(i)}(Y)| \leq \epsilon \text{ and } \max_{i=1,2,3,4}|\delta_n^{(i)}(\mu, Y)| \leq \epsilon\Big\}.$$

Since

$$\mathbb{P}_\mu(T_{n,p}(\lambda) > \xi_\alpha) = \mathbb{P}_\mu(\{T_{n,p}(\lambda) > \xi_\alpha\} \cap K_\epsilon^{(2)}) + \mathbb{P}_\mu(\{T_{n,p}(\lambda) > \xi_\alpha\} \cap \{K_\epsilon^{(2)}\}^c),$$

it follows that

$$\mathbb{P}_\mu(T_{n,p}(\lambda) > \xi_\alpha) \leq \epsilon + \mathbb{P}_\mu(T_{n,p}^0(\lambda) > \xi_\alpha - gq(\lambda, \gamma) - 7\epsilon),$$
$$\mathbb{P}_\mu(T_{n,p}(\lambda) > \xi_\alpha) \geq -\epsilon + \mathbb{P}_\mu(T_{n,p}^0(\lambda) > \xi_\alpha - gq(\lambda, \gamma) + 7\epsilon).$$

On the other hand, since $T_{n,p}^0(\lambda)$ is free of μ and converges in distribution to standard normal distribution, we can find a sufficiently large N_3 such that when $n > N_3$, for any $\mu \in K_\epsilon^{(1)}$,

$$P_\mu(T_{n,p}^0(\lambda) > \xi_\alpha - gq(\lambda, \gamma) - 7\epsilon) < \Phi(-\xi_\alpha + gq(\lambda, \gamma) - 7\epsilon) + \epsilon,$$
$$P_\mu(T_{n,p}^0(\lambda) > \xi_\alpha - gq(\lambda, \gamma) + 7\epsilon) > \Phi(-\xi_\alpha + gq(\lambda, \gamma) + 7\epsilon) - \epsilon.$$

In summary, on $\mu \in K_\epsilon^{(1)}$, when $n > \max_{i=1,2,3} N_i$,

$$\mathbb{P}_\mu\big(T_{n,p}(\lambda) > \xi_\alpha\big) \le 2\epsilon + \Phi\big(-\xi_\alpha + gq(\lambda,\gamma) - 7\epsilon\big),$$
$$\mathbb{P}_\mu\big(T_{n,p}(\lambda) > \xi_\alpha\big) \ge -2\epsilon + \Phi\big(-\xi_\alpha + gq(\lambda,\gamma) + 7\epsilon\big).$$

This completes the proof, since $\mathbb{P}_*(K_\epsilon^{(1)}) \ge 1 - \zeta$.

9.4. *Proof of Theorem* 2.3.

9.4.1. *Proof of* (20). To show the existence of a sequence of local maximizers of $\hat{Q}_n(\lambda, \gamma_n)$ as stated, it suffices to show that for any $\varepsilon \in (0, 1)$, there exists a constant $K > 0$, and an integer n_ε, such that, for $t = Kn^{-1/4}$,

$$\mathbb{P}\big\{\hat{Q}_n(\lambda_\infty \pm t, \gamma_n) - \hat{Q}_n(\lambda_\infty, \gamma_n) \le 0\big\} \ge \varepsilon$$

for all $n \ge n_\varepsilon$. If we use a stochastic term $\delta(t)$ to measure the difference between $\hat{Q}_n(\lambda, \gamma_n)$ and $Q(\lambda, \gamma)$ at $\lambda = \lambda_\infty \pm t$ and λ_∞, considering λ_∞ to be in the interior of $[\underline{\lambda}, \overline{\lambda}]$, a second-order Taylor expansion yields

$$\begin{aligned}\hat{Q}_n(\lambda_\infty \pm t, \gamma_n) - \hat{Q}_n(\lambda_\infty, \gamma_n) &= Q(\lambda_\infty \pm t, \gamma) - Q(\lambda_\infty, \gamma) + \delta(\pm t)\\ &= \frac{t^2}{2}\frac{\partial^2}{\partial\lambda^2}Q(\lambda_\infty, \gamma) + O\big(t^3\big) + \delta(\pm t).\end{aligned}$$

Since $O(t^3)$ is a smaller order term as $n \to \infty$ and $\partial^2 Q(\lambda_\infty, \gamma)/\partial\lambda^2 < 0$, it suffices to show that $n^{1/2}|\delta(\pm t)| = O_p(1)$ with an uniform tail bound in t. Again by Taylor expansion,

$$\begin{aligned}n^{1/2}\delta(\pm t) &= n^{1/2}t\bigg[\frac{\partial}{\partial\lambda}\hat{Q}_n(\lambda_\infty, \gamma_n) - \frac{\partial}{\partial\lambda}Q(\lambda_\infty, \gamma)\bigg]\\ &\quad + \frac{n^{1/2}t^2}{2}\bigg[\frac{\partial^2}{\partial\lambda^2}\hat{Q}_n(\lambda_\infty, \gamma_n) - \frac{\partial^2}{\partial\lambda^2}Q(\lambda_\infty, \gamma)\bigg]\\ &\quad + \frac{n^{1/2}t^3}{6}\frac{\partial^3}{\partial\lambda^3}\hat{Q}(\lambda_\infty + \alpha t, \gamma_n) - \frac{n^{1/2}t^3}{6}\frac{\partial^3}{\partial\lambda^3}Q(\lambda_\infty + \alpha t, \gamma)\end{aligned}$$

for some $\alpha \in [0, 1]$.

Now expressing $Q(\lambda, \gamma)$ $\hat{Q}_n(\lambda, \gamma)$ and their partial derivatives as continuous functions of $m_F(-\lambda_\infty)$, $m_F'(-\lambda_\infty)$, $m_F^{(3)}(-\lambda_\infty)$, $m_F^{(4)}(-\lambda_\infty)$, ϕ, γ, and their empirical counterparts, we use Proposition A.2–A.3 to deduce that

$$n^{1/4}\bigg|\frac{\partial}{\partial\lambda}\hat{Q}_n(\lambda_\infty, \gamma) - \frac{\partial}{\partial\lambda}Q(\lambda_\infty, \gamma)\bigg| \xrightarrow{P} 0,$$
$$\bigg|\frac{\partial^2}{\partial\lambda^2}\hat{Q}_n(\lambda_\infty, \gamma) - \frac{\partial^2}{\partial\lambda^2}Q(\lambda_\infty, \gamma)\bigg| \xrightarrow{P} 0,$$
$$\sup_{\lambda\in[\underline{\lambda},\overline{\lambda}]}\bigg|\frac{\partial^3}{\partial\lambda^3}\hat{Q}_n(\lambda, \gamma)\bigg| + \bigg|\frac{\partial^3}{\partial\lambda^3}Q(\lambda, \gamma)\bigg| = O_p(1),$$

which completes the proof. If λ_∞ is on the boundary and $\partial Q(\lambda_\infty, \gamma)/\partial\lambda < 0$, similar results follow from a first-order Taylor expansion.

9.4.2. *Proof of* (21). It remains to verify (21). To this end, note that it suffices to prove that

$$p^{1/2}\left|\frac{1}{p}\mathrm{RHT}(\lambda_n)-\hat{\Theta}_1(\lambda_n,\gamma_n)-\frac{1}{p}\mathrm{RHT}(\lambda_\infty)+\hat{\Theta}_1(\lambda_\infty,\gamma_n)\right|$$
$$\le p^{1/2}\left|\frac{1}{p}\frac{\partial}{\partial\lambda}\mathrm{RHT}(\lambda_\infty)-\frac{\partial}{\partial\lambda}\hat{\Theta}_1(\lambda_\infty,\gamma_n)\right||\lambda_n-\lambda_\infty|$$
$$+\frac{p^{1/2}}{2}\left|\frac{1}{p}\frac{\partial^2}{\partial\lambda^2}\mathrm{RHT}(\lambda_\infty)-\frac{\partial^2}{\partial\lambda^2}\hat{\Theta}_1(\lambda_\infty,\gamma_n)\right||\lambda_n-\lambda_\infty|^2$$
$$+\frac{p^{1/2}}{6}\left|\frac{1}{p}\frac{\partial^3}{\partial\lambda^3}\mathrm{RHT}(\lambda^*)-\frac{\partial^3}{\partial\lambda^3}\hat{\Theta}_1(\lambda^*,\gamma_n)\right||\lambda_n-\lambda_\infty|^3\xrightarrow{P}0,$$

where λ^* is in between λ_∞ and λ_n. So it is enough to show that

$$p^{1/4}\left|\frac{1}{p}\frac{\partial}{\partial\lambda}\mathrm{RHT}(\lambda_\infty)-\frac{\partial}{\partial\lambda}\hat{\Theta}_1(\lambda_\infty,\gamma_n)\right|\xrightarrow{P}0, \tag{35}$$

$$\left|\frac{1}{p}\frac{\partial^2}{\partial\lambda^2}\mathrm{RHT}(\lambda_\infty)-\frac{\partial^2}{\partial\lambda^2}\hat{\Theta}_1(\lambda_\infty,\gamma_n)\right|\xrightarrow{P}0, \tag{36}$$

$$\sup_{\lambda\in[\underline{\lambda},\overline{\lambda}]}\left|\frac{1}{p}\frac{\partial^3}{\partial\lambda^3}\mathrm{RHT}(\lambda)-\frac{\partial^3}{\partial\lambda^3}\hat{\Theta}_1(\lambda,\gamma_n)\right|=O_p(1). \tag{37}$$

Next,

$$\mathbb{E}\left|p^{-1}\frac{\partial^3}{\partial\lambda^3}\mathrm{RHT}(\lambda)\right|\le\frac{n_1n_2}{\underline{\lambda}^{-4}p(n_1+n_2)}\mathbb{E}\left|(\bar{X}_1-\bar{X}_2)^T(\bar{X}_1-\bar{X}_2)\right|=O(1)$$

for all $\lambda\in[\underline{\lambda},\overline{\lambda}]$. And Proposition A.3 shows the convergence of $\partial^3\hat{\Theta}_1(\lambda,\gamma_n)/\partial\lambda^3$ to $\partial^3\Theta_1(\lambda,\gamma)/\partial\lambda^3$ uniformly on $\lambda\in[\underline{\lambda},\overline{\lambda}]$, so that (37) holds.

For proving (35) and (36), note that Propositions A.4 and A.5 showed the convergence of $\partial\hat{\Theta}_1(\lambda,\gamma_n)/\partial\lambda$ to $-p^{-1}\operatorname{tr}[\{R_n(-\lambda)\}^2\Sigma_p]$, and the convergence of $\partial^2\hat{\Theta}_1(\lambda,\gamma_n)/\partial\lambda^2$ to $2p^{-1}\operatorname{tr}[\{R_n(-\lambda)\}^3\Sigma_p]$. So the proof will be complete if we can show

$$p^{1/4}\left|\frac{1}{p}\frac{\partial}{\partial\lambda}\mathrm{RHT}(\lambda_\infty)+\frac{1}{p}\operatorname{tr}\left[\{R_n(-\lambda_\infty)\}^2\Sigma_p\right]\right|\xrightarrow{P}0, \tag{38}$$

$$\left|\frac{1}{p}\frac{\partial^2}{\partial\lambda^3}\mathrm{RHT}(\lambda_\infty)-\frac{2}{p}\operatorname{tr}\left[\{R_n(-\lambda_\infty)\}^3\Sigma_p\right]\right|\xrightarrow{P}0. \tag{39}$$

We move the proofs of (38) and (39) to Sections S.3.6 and S.3.7 of the Supplementary Material, which are lengthy.

9.5. *Proof of Theorem* 3.1. To prove the process convergence stated in Theorem 3.1, we need to verify the convergence of finite-dimensional distributions and the tightness of the process.

(a) To show the distributional convergence of $\{\mathrm{RHT}(\lambda_1),\dots,\mathrm{RHT}(\lambda_k)\}$ for arbitrary integer k and fixed $\lambda_1,\dots,\lambda_k>0$, it suffices to show the joint normality of $\{U_{ii'}(\lambda_j),1\le i,i'\le 2,1\le j\le k\}$. Therefore, define an arbitrary linear combination

$$T_n=\sum_{i=1}^{2}\sum_{i'=1}^{2}\sum_{j=1}^{k}l_{ii'j}U_{ii'}(\lambda_j).$$

It suffices to show that T_n is asymptotically normal. We can derive asymptotic orders of the functions ϱ_0, ϱ_1 and ϱ_2 with each $U_{ii'}(\lambda_j)$ as arguments and combine them through the Cauchy–Schwarz inequality to get the asymptotic orders of ϱ_0, ϱ_1, ϱ_2 with T_n as the argument. The proof is essentially a repetition of the arguments in Section 9.1, and is hence omitted.

(b) To show tightness, note first that Proposition A.3 yields $\hat{\Theta}_2(\lambda, \gamma_n) \to_p \Theta_2(\lambda, \gamma)$ uniformly on $[\underline{\lambda}, \bar{\lambda}]$. This implies tightness of $(\hat{\Theta}_2(\lambda, \gamma_n)\colon \lambda \in [\underline{\lambda}, \bar{\lambda}])$. The sequence $n^{1/2}(p^{-1}\,\mathrm{RHT}(\lambda) - \hat{\Theta}_1(\lambda, \gamma_n))$ is shown to be tight in Pan and Zhou ((2011), Section 4) for observations with finite fourth moments but with $\Sigma = I_p$. Although their arguments are in a one-sample testing framework, they can easily be generalized to the two-sample testing case and for Σ satisfying **C1**–**C3**. Together with $\inf_{\lambda \in [\underline{\lambda}, \bar{\lambda}]} \Theta_2(\lambda, \gamma) > 0$, the convergence of the process follows.

(c) The covariance kernel can be computed via basic calculus, making use of Proposition 9.4 and the relation between $\bar{R}(\lambda)$ and $\mathrm{RHT}(\lambda)$ shown in (31).

9.6. *Proof of Proposition* 2.1. In order to find the minimax rule within $\mathcal{D}$, we first find $\tilde{\pi}_\lambda$ which minimizes $Q(\lambda, \gamma; \tilde{\pi})$ for $\tilde{\pi} \in \Pi_2(1)$, for every fixed λ. At this point, we make two important observations:

(i) $\Pi_2(1)$ is convex.

(ii) $(0, 0, 1/\phi_2)$ is an extremal point of the simplex $\Pi_2(1)$, while $\pi_0 \geq 0$ and $\pi_2 \geq 0$ for all $\tilde{\pi} = (\pi_0, \pi_1, \pi_2) \in \Pi_2(1)$.

Because of (i), and the fact that $Q(\lambda, \gamma; \tilde{\pi})$ is linear in $\tilde{\pi}$, the minimum occurs at the boundary of the set $\Pi_2(1)$.

The following proposition establishes that $\tilde{\pi}_\lambda = \phi_2^{-1}\mathbf{e}_2$, where $\mathbf{e}_2 = (0, 0, 1)$.

PROPOSITION 9.5. *For $j = 0, 1, \ldots,$ and $\phi_j = \int \tau^j\, dH(\tau)$,*

$$\phi_j^{-1}\rho_j(-\lambda, \gamma) \geq \phi_{j+1}^{-1}\rho_{j+1}(-\lambda, \gamma) \quad \textit{for all } \lambda > 0. \tag{40}$$

To verify the claim that $\tilde{\pi}_\lambda = \phi_2^{-1}\mathbf{e}_2$, observe that minimization of $Q(\lambda, \gamma; \tilde{\pi})$ is equivalent to minimization of $\sum_{j=0}^2 \pi_j \rho_j(-\lambda, \gamma)$ over $\tilde{\pi} \in \Pi_2(1)$. Using the fact that $\phi_0 = 1$, for any $\tilde{\pi} \in \Pi_2(1)$,

$$\begin{aligned}
&\sum_{j=0}^{2} \pi_j \rho_j(-\lambda, \gamma) - \phi_2^{-1}\rho_2(-\lambda, \gamma) \\
&\quad = \pi_0\big(\phi_0^{-1}\rho_0(-\lambda, \gamma) - \phi_1^{-1}\rho_1(-\lambda, \gamma)\big) \\
&\qquad + (1 - \phi_2\pi_2)\big(\phi_1^{-1}\rho_1(-\lambda, \gamma) - \phi_2^{-1}\rho_2(-\lambda, \gamma)\big),
\end{aligned}$$

which follows from substituting $\phi_1\pi_1 = 1 - \pi_0 - \phi_2\pi_2$. Now by (ii) and Proposition 9.5, the right-hand side is nonnegative, and equals zero only if $\tilde{\pi} = \phi_2^{-1}\mathbf{e}_2$, which verifies the claim.

The next step is therefore to find $\lambda \in [\underline{\lambda}, \overline{\lambda}]$ that maximizes $Q(\lambda, \gamma; \phi_2^{-1}\mathbf{e}_2) = \phi_2^{-1} Q(\lambda, \gamma; \mathbf{e}_2)$. Due to Proposition 9.6, stated below, the maximum occurs at $\lambda = \overline{\lambda}$, which shows that $T_{n,p}(\overline{\lambda})$ is LAM with respect the class $\mathfrak{P}_2(C)$ for any $C > 0$.

PROPOSITION 9.6. *The function $Q(\lambda, \gamma; \mathbf{e}_2)$ is nondecreasing on $[\underline{\lambda}, \infty)$ for any $\underline{\lambda} > 0$, where $\mathbf{e}_2 = (0, 0, 1)$.*

Proof of Propositions 9.5 and 9.6 are given in the Supplementary Material.

APPENDIX

Key propositions used in the proofs. In the following, c_1, c_2 and c_3 denote some universal positive constants, independent of λ. To lighten notation, some fixed parameters are ignored in the following expressions when it does not cause ambiguity; for example, weights $\widetilde{\pi}$ in $Q(\lambda, \gamma; \widetilde{\pi})$ may be dropped. The following propositions show the concentration of some quantities.

PROPOSITION A.1. *Suppose we have two matrices A and B with A symmetric and positive definite. For any vector Y and any integer $k \geq 1$,*

$$\left|\mathrm{tr}\{(A + YY^T)^{-k}B\} - \mathrm{tr}(A^{-k}B)\right| \leq \frac{k\|B\|}{\tau_A^k},$$

where τ_A is the smallest eigenvalue of A.

Recall that $\hat{\phi}_1 = p^{-1}\,\mathrm{tr}(S_n)$ and $\phi_1 = \int \tau\, dH(\tau)$.

PROPOSITION A.2. *If conditions **C1**–**C3** are satisfied, then for any $t > 0$,*

$$\mathbb{P}\{|\hat{\phi}_1 - \mathbb{E}\hat{\phi}_1| > t\} \leq c_1 \exp\{-\min(c_2 n t^2, c_3 n t)\}.$$

Moreover, $\sqrt{n}|\mathbb{E}\hat{\phi}_1 - \phi_1| \to 0$, as $n \to \infty$, since $\mathbb{E}\hat{\phi}_1 = \int \tau\, dH_p(\tau)$.

PROPOSITION A.3. *Define $m_{F_{n,p}}^{(k)}(-\lambda)$ to be the kth order derivative of $m_{F_{n,p}}(-\lambda)$ and $m_F^{(k)}(-\lambda)$ to be the kth order derivative of $m_F(-\lambda)$. If conditions **C1**–**C3** are satisfied, then for any $t > 0$, integer k and $\lambda \in [\underline{\lambda}, \overline{\lambda}]$,*

$$\mathbb{P}\left(\left|m_{F_{n,p}}^{(k)}(-\lambda) - \mathbb{E}m_{F_{n,p}}^{(k)}(-\lambda)\right| > t\right) \leq c_1 \exp(-c_2 n t^2).$$

Moreover,

$$n^{1/2}\left|\mathbb{E}m_{F_{n,p}}^{(k)}(-\lambda) - m_F^{(k)}(-\lambda)\right| \to 0.$$

It follows, as continuous and monotone functions in λ,

$$\sup_{\lambda \in [\underline{\lambda}, \overline{\lambda}]} \left|m_{F_{n,p}}^{(k)}(-\lambda) - m_F^{(k)}(-\lambda)\right| \xrightarrow{P} 0.$$

PROPOSITION A.4. *If conditions **C1**–**C3** are satisfied, then for any $\lambda \in [\underline{\lambda}, \overline{\lambda}]$,*

$$\frac{\partial}{\partial\lambda}\hat{\Theta}_1(\lambda, \gamma_n) = -\frac{1}{p}\,\mathrm{tr}\left[\{R_n(-\lambda)\}^2\Sigma_p\right] + o_p(n^{-1/4}).$$

PROPOSITION A.5. *If conditions **C1**–**C3** are satisfied, then for any $\lambda \in [\underline{\lambda}, \overline{\lambda}]$,*

$$\frac{\partial^2}{\partial\lambda^2}\hat{\Theta}_1(\lambda, \gamma_n) = \frac{2}{p}\,\mathrm{tr}\left[\{R_n(-\lambda)\}^3\Sigma_p\right] + o_p(1).$$

PROPOSITION A.6. *If conditions **C1**–**C3** are satisfied, then for any $\lambda, \lambda' \in [\underline{\lambda}, \overline{\lambda}]$, $\lambda \neq \lambda'$,*

$$\begin{aligned}
&\frac{1}{p}\,\mathrm{tr}\left[R_n(-\lambda)\Sigma_p R_n(-\lambda')\Sigma_p\right] \\
&\quad = \{1 + \gamma\Theta_1(\lambda, \gamma)\}\{1 + \gamma\Theta_1(\lambda', \gamma)\}\left\{\frac{\lambda'\Theta_1(\lambda', \gamma) - \lambda\Theta_1(\lambda, \gamma)}{\lambda' - \lambda}\right\} + o_p(1).
\end{aligned}$$

THEOREM A.1 (Theorem 2.2 of Chatterjee (2009)). *Let* $\mathbf{Z} = (z_1, \dots, z_n)$ *be a vector of independent random variables in* $\mathcal{L}(c_1, c_2)$ *for some finite* c_1, c_2. *Take any* $g \in C^2(\mathbb{R}^n)$ *and let* ∇g *and* $\nabla^2 g$ *denote the gradient and Hessian of* g. *Let*

$$\varrho_0(g) = \left(\mathbb{E}\sum_{i=1}^n \left|\frac{\partial g}{\partial z_i}(\mathbf{Z})\right|^4\right)^{1/2},$$

$$\varrho_1(g) = (\mathbb{E}\|\nabla g(\mathbf{Z})\|^4)^{1/4}, \qquad \varrho_2(g) = (\mathbb{E}\|\nabla^2 g(\mathbf{Z})\|^4)^{1/4},$$

where $\|\cdot\|$ *is the operator norm. Suppose* $W = g(\mathbf{Z})$ *has a finite fourth moment and let* $\sigma^2 = \mathrm{Var}(W)$. *Let* U *be a normal random variable having the same mean and variance as* W. *Then*

$$d_{\mathrm{TV}}(W, U) \le \sigma^{-2} 2\sqrt{5}\{c_1 c_2 \varrho_0(g) + c_1^3 \varrho_1(g)\varrho_2(g)\}, \tag{A.1}$$

where d_{TV} *is the total variation distance between two distributions.*

Acknowledgments. The authors thank the Editor (Professor Ed George), the Associate Editor and the referees for their insightful comments and suggestions that greatly improved the presentation of the material.

The first author was supported in part by NSF Grants DMS-1407530 and DMS-1811405.

The second author was supported in part by NSF Grants DMS-1305858 and DMS-1407530.

The third author was supported in part by NSF Grants DMS-1407530, DMS-1713120, DMS-1811405 and NIH Grant 1R01EB021707.

The fourth author was supported in part by NSF Grant DMS-1148643 and NIH Grant 1R01EB021707.

The fifth author was supported in part by NIH Grants R01GM108711 and U24 CA210993.

SUPPLEMENTARY MATERIAL

Supplement to "An adaptable generalization of Hotelling's T^2 test in high dimension" (DOI: 10.1214/19-AOS1869SUPP; .pdf). It contains additional simulation results and detailed proofs of the main theoretical results.

REFERENCES

ANDERSON, T. W. (1984). *An Introduction to Multivariate Statistical Analysis*, 2nd ed. *Wiley Publications in Statistics*. Wiley, New York. MR0091588

BAI, Z., CHEN, J. and YAO, J. (2010). On estimation of the population spectral distribution from a high-dimensional sample covariance matrix. *Aust. N. Z. J. Stat.* **52** 423–437. MR2791528 https://doi.org/10.1111/j.1467-842X.2010.00590.x

BAI, Z. and SARANADASA, H. (1996). Effect of high dimension: By an example of a two sample problem. *Statist. Sinica* **6** 311–329. MR1399305

BAI, Z. D. and SILVERSTEIN, J. W. (1998). No eigenvalues outside the support of the limiting spectral distribution of large-dimensional sample covariance matrices. *Ann. Probab.* **26** 316–345. MR1617051 https://doi.org/10.1214/aop/1022855421

BAI, Z. and SILVERSTEIN, J. W. (2010). *Spectral Analysis of Large Dimensional Random Matrices*, 2nd ed. *Springer Series in Statistics*. Springer, New York. MR2567175 https://doi.org/10.1007/978-1-4419-0661-8

BENJAMINI, Y. and HOCHBERG, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B* **57** 289–300. MR1325392

BERGAMASCHI, A., KIM, Y. H., WANG, P., SØRLIE, T., HERNANDEZ-BOUSSARD, T., LONNING, P. E., TIBSHIRANI, R., BØRRESEN-DALE, A. L. and POLLACK, J. R. (2006). Distinct patterns of DNA copy number alteration are associated with different clinicopathological features and gene-expression subtypes of breast cancer. *Genes Chromosomes Cancer* **45** 1033–40.

BISWAS, M. and GHOSH, A. K. (2014). A nonparametric two-sample test applicable to high dimensional data. *J. Multivariate Anal.* **123** 160–171. MR3130427 https://doi.org/10.1016/j.jmva.2013.09.004

CAI, T. T., LIU, W. and XIA, Y. (2014). Two-sample test of high dimensional means under dependence. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 349–372. MR3164870 https://doi.org/10.1111/rssb.12034

CHAKRABORTY, A. and CHAUDHURI, P. (2017). Tests for high-dimensional data based on means, spatial signs and spatial ranks. *Ann. Statist.* **45** 771–799. MR3650400 https://doi.org/10.1214/16-AOS1467

CHANG, J., ZHOU, W. and ZHOU, W. X. (2014). Simulation-based hypothesis testing of high dimensional means under covariance heterogeneity—An alternative road to high dimensional tests. Preprint. Available at arXiv:1406.1939.

CHATTERJEE, S. (2009). Fluctuations of eigenvalues and second order Poincaré inequalities. *Probab. Theory Related Fields* **143** 1–40. MR2449121 https://doi.org/10.1007/s00440-007-0118-6

CHEN, S. X., LI, J. and ZHONG, P. (2014). Two-sample tests for high dimensional means with thresholding and data transformation. Preprint. Available at arXiv:1410.2848.

CHEN, S. X. and QIN, Y.-L. (2010). A two-sample test for high-dimensional data with applications to gene-set testing. *Ann. Statist.* **38** 808–835. MR2604697 https://doi.org/10.1214/09-AOS716

CHEN, L. S., PAUL, D., PRENTICE, R. L. and WANG, P. (2011). A regularized Hotelling's T^2 test for pathway analysis in proteomic studies. *J. Amer. Statist. Assoc.* **106** 1345–1360. MR2896840 https://doi.org/10.1198/jasa.2011.ap10599

CREIGHTON, C. J. (2012). The molecular profile of luminal B breast cancer. *Biologics* **6** 289–297. https://doi.org/10.2147/BTT.S29923

DONG, K., PANG, H., TONG, T. and GENTON, M. G. (2016). Shrinkage-based diagonal Hotelling's tests for high-dimensional small sample size data. *J. Multivariate Anal.* **143** 127–142. MR3431423 https://doi.org/10.1016/j.jmva.2015.08.022

EL KAROUI, N. and KÖSTERS (2011). Geometric sensitivity of random matrix results: Consequences for shrinkage estimators of covariance and related statistical methods. Preprint. Available at arXiv:1105.1404.

ELLIS, M. J., GILLETTE, M., CARR, S. A., PAULOVICH, A. G., SMITH, R. D., RODLAND, K. K., TOWNSEND, R. R., KINSINGER, C., MESRI, M. et al. (2013). Connecting genomic alterations to cancer biology with proteomics: The NCI Clinical Proteomic Tumor Analysis Consortium. *Cancer Discov.* **3** 1108–1112.

GREGORY, K. B., CARROLL, R. J., BALADANDAYUTHAPANI, V. and LAHIRI, S. N. (2015). A two-sample test for equality of means in high dimension. *J. Amer. Statist. Assoc.* **110** 837–849. MR3367268 https://doi.org/10.1080/01621459.2014.934826

GRETTON, A., SRIPERUMBUDUR, B., SEJDINOVIC, D., STRATHMANN, H., BALAKRISHNAN, S., PONTIL, M. and FUKUMIZU, K. (2012). Optimal kernel choice for large-scale two-sample tests. *Adv. Neural Inf. Process. Syst.* **25** 1205–1213.

GUO, B. and CHEN, S. X. (2016). Tests for high dimensional generalized linear models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **78** 1079–1102. MR3557190 https://doi.org/10.1111/rssb.12152

LAMY, P.-J., FINA, F., BASCOUL-MOLLEVI, C., LABERENNE, A.-C., MARTIN, P.-M., OUAFIK, L. and JACOT, W. (2011). Quantification and clinical relevance of gene amplification at chromosome 17q12-q21 in human epidermal growth factor receptor 2-amplified breast cancers. *Breast Cancer Res.* **13** R15. https://doi.org/10.1186/bcr2824

LEDOIT, O. and PÉCHÉ, S. (2011). Eigenvectors of some large sample covariance matrix ensembles. *Probab. Theory Related Fields* **151** 233–264. MR2834718 https://doi.org/10.1007/s00440-010-0298-3

LI, H., AUE, A., PAUL, D., PENG, J. and WANG, P. (2020). Supplement to "An adaptable generalization of Hotelling's T^2 test in high dimension." https://doi.org/10.1214/19-AOS1869SUPP.

LIU, H., AUE, A. and PAUL, D. (2015). On the Marčenko–Pastur law for linear time series. *Ann. Statist.* **43** 675–712. MR3319140 https://doi.org/10.1214/14-AOS1294

LOPES, M. E., JACOB, L. and WAINWRIGHT, M. J. (2011). A more powerful two-sample test in high dimensions using random projection. *Adv. Neural Inf. Process. Syst.* 1206–1214.

MERTINS, P., MANI, D. R., RUGGLES, K. V., GILLETTE, M. A., CLAUSER, K. R., WANG, P., WANG, X., QIAO, J. W., CAO, S. et al. (2016). Proteogenomics connects somatic mutations to signalling in breast cancer. *Nature* **534** 55–62. https://doi.org/10.1038/nature18003

MUIRHEAD, R. J. (1982). *Aspects of Multivariate Statistical Theory. Wiley Series in Probability and Mathematical Statistics*. Wiley, New York. MR0652932

CANCER GENOME ATLAS NETWORK (2012). Comprehensive molecular portraits of human breast tumours. *Nature* **490** 61–70. https://doi.org/10.1038/nature11412

PAN, G. M. and ZHOU, W. (2011). Central limit theorem for Hotelling's T^2 statistic under large dimension. *Ann. Appl. Probab.* **21** 1860–1910. MR2884053 https://doi.org/10.1214/10-AAP742

PAUL, D. and AUE, A. (2014). Random matrix theory in statistics: A review. *J. Statist. Plann. Inference* **150** 1–29. MR3206718 https://doi.org/10.1016/j.jspi.2013.09.005

PAULOVICH, A. G., BILLHEIMER, D., HAM, A. J., VEGA-MONTOTO, L., RUDNICK, P. A., TABB, D. L., WANG, P. et al. (2010). Interlaboratory study characterizing a yeast performance standard for benchmarking LC-MS platform performance. *Mol. Cell. Proteomics* **9** 242–254.

SRIVASTAVA, M. S. (2009). A test for the mean vector with fewer observations than the dimension under non-normality. *J. Multivariate Anal.* **100** 518–532. MR2483435 https://doi.org/10.1016/j.jmva.2008.06.006

SRIVASTAVA, M. S. and DU, M. (2008). A test for the mean vector with fewer observations than the dimension. *J. Multivariate Anal.* **99** 386–402. MR2396970 https://doi.org/10.1016/j.jmva.2006.11.002

SRIVASTAVA, R., LI, P. and RUPPERT, D. (2016). RAPTT: An exact two-sample test in high dimensions using random projections. *J. Comput. Graph. Statist.* **25** 954–970. MR3533647 https://doi.org/10.1080/10618600.2015.1062771

TRAN, B. and BEDARD, P. L. (2011). Luminal-B breast cancer and novel therapeutic targets. *Breast Cancer Res.* **13** 221. https://doi.org/10.1186/bcr2904

WANG, L., PENG, B. and LI, R. (2015). A high-dimensional nonparametric multivariate test for mean vector. *J. Amer. Statist. Assoc.* **110** 1658–1669. MR3449062 https://doi.org/10.1080/01621459.2014.988215

XU, G., LIN, L., WEI, P. and PAN, W. (2016). An adaptive two-sample test for high-dimensional means. *Biometrika* **103** 609–624. MR3551787 https://doi.org/10.1093/biomet/asw029