Bayesian Analysis (2024) **19**, Number 3, pp. 869–893

Bayesian Optimal Two-Sample Tests for High-Dimensional Gaussian Populations[*]

Kyoungjae Lee[†], Kisung You[‡], and Lizhen Lin[§]

Abstract. We propose minimax optimal Bayesian two-sample tests for testing equality of high-dimensional mean vectors and covariance matrices between two populations. In many applications including genomics and medical imaging, it is natural to assume that only a few entries of two mean vectors or covariance matrices are different. Many existing tests that rely on aggregating the difference between empirical means or covariance matrices are not optimal or yield low power under such setups. Motivated by this, we develop Bayesian two-sample tests employing a divide-and-conquer idea, which is powerful especially when the differences between two populations are rare but large. The proposed two-sample tests manifest closed forms of Bayes factors and allow scalable computations even in high-dimensions. We prove that the proposed tests are consistent under relatively mild conditions compared to existing tests in the literature. Furthermore, the testable regions from the proposed tests turn out to be minimax optimal in terms of rates. Simulation studies show clear advantages of the proposed tests over other state-of-the-art methods in various scenarios. Our tests are also applied to the analysis of the gene expression data of two cancer data sets.

Keywords: Bayesian hypothesis test, Bayes factor consistency, high-dimensional covariance matrix, optimal high-dimensional tests.

MSC2020 subject classifications: Primary 62F15, 62F03; secondary 62H15.

1 Introduction

Consider two samples of observations from high-dimensional normal models

$$\begin{aligned} X_i \mid \mu_1, \Sigma_1 &\overset{i.i.d.}{\sim} N_p(\mu_1, \Sigma_1), \quad i = 1, \ldots, n_1, \\ Y_i \mid \mu_2, \Sigma_2 &\overset{i.i.d.}{\sim} N_p(\mu_2, \Sigma_2), \quad i = 1, \ldots, n_2, \end{aligned} \tag{1.1}$$

where $N_p(\mu, \Sigma)$ is the p-dimensional normal distribution with mean vector $\mu \in \mathbb{R}^p$ and covariance matrix $\Sigma \in \mathbb{R}^{p\times p}$, and the number of variables p can increase to infinity as the sample sizes (n_1 and n_2) grow. Given two samples of such observations, there is abundant interest in testing the equality of high-dimensional mean vectors or covariance

[*] This research was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2020R1A4A1018207). LL would like to acknowledge the generous support of NSF grants DMS CAREER 1654579 and DMS 2113642.

[†] Department of Statistics, Sungkyunkwan University, Seoul 03063, South Korea, leekjstat@gmail.com

[‡] Department of Internal Medicine, Yale University School of Medicine, New Haven, CT 06510, USA, kisung.you@yale.edu

[§] Department of Applied and Computational Mathematics and Statistics, The University of Notre Dame, Notre Dame, IN 46556, USA, lizhen.lin@nd.edu

https://doi.org/10.1214/23-BA1373

matrices with applications in medical imaging, genetics and biology (Tsai and Chen, 2009; Shen et al., 2011). Although there is an emerging literature on high-dimensional hypothesis testing, most of the literature has focused on proposing frequentist testing statistics with relatively little work on developing Bayesian hypothesis tests in particular for high-dimensional problems. Bayesian tests, which typically are based on Bayes factors with appropriate design of prior distributions for the model under the null and the alternative operate differently from their frequentist counterparts, and there is independent interest in developing Bayesian testing approaches. We add to the limited literature by developing powerful and scalable Bayesian high-dimensional tests for testing the equality of means and covariance matrices between two populations.

Our initial focus is on the two-sample mean test, where we assume $\Sigma_1 = \Sigma_2$ and test whether $\mu_1 = \mu_2$ in model (1.1). When $\mu_1 \neq \mu_2$, we call the nonzero elements in the mean difference vector $\mu_1 - \mu_2 \in \mathbb{R}^p$ the *signals*. It is well known that the power of a test depends on both the number and the magnitude of the signals. From a frequentist perspective, Bai and Saranadasa (1996) and Srivastava and Du (2008) proposed high-dimensional two-sample mean tests based on estimators of $\|A(\mu_1 - \mu_2)\|_2^2$ for some positive definite matrix $A \in \mathbb{R}^{p\times p}$, where $\|\cdot\|_2$ denotes the vector ℓ_2-norm. We call these tests ℓ_2-type tests because their test statistics involve the ℓ_2-norm. It is known that ℓ_2-type tests tend to have good power when there are many signals, i.e., when a large portion of $\mu_1 - \mu_2$ is nonzero. When there are many but small signals, ℓ_2-type tests tend to show better performance over other types of tests.

In many applications, however, it is more natural to assume rare signals, where only few entries of $\mu_1 - \mu_2 \in \mathbb{R}^p$ are nonzero. Under the presence of few relatively large signals, it is well known that maximum-type tests tend to outperform ℓ_2-type tests. Here, a maximum-type test refers to a class of tests whose test statistic involves the maximum-norm. Cai et al. (2014) proposed a consistent maximum-type test for high-dimensional two-sample mean test. They standardized the difference between sample mean vectors using an estimated precision matrix based on either the constrained ℓ_1-minimization for inverse matrix estimation (CLIME) (Cai et al., 2011) or the inverse of the adaptive thresholding estimator for a covariance matrix (Cai and Liu, 2011). Because their test statistics depend on an estimated precision matrix, practical performance of the tests could be impacted by performance of the estimated precision matrix.

Besides the aforementioned papers, many other interesting studies have been conducted for the two-sample testing setup. Gregory et al. (2015) proposed a two-sample mean test which bypasses the needs of the estimation of precision matrix and is robust to highly unequal covariance matrices between two populations. Xu et al. (2016) proposed an adaptive two-sample mean test that retains high power against a wide range of alternatives. Cao et al. (2018) developed a test for compositional data based on the centered log-ratio transformation. Recently, Wang et al. (2019) suggested a robust version of the maximum-type test for contaminated data.

Our second focus is the two-sample covariance test of whether $\Sigma_1 = \Sigma_2$ or not in model (1.1) under the assumption $\mu_1 = \mu_2 = 0$. In this case, we call the nonzero entries in $\Sigma_1 - \Sigma_2 \in \mathbb{R}^{p\times p}$ the signals. Some frequentist tests have also been suggested in the literature for two-sample covariance in high-dimensional settings. Schott (2007) and Li and Chen (2012) proposed to test equality of covariance matrices based on an estimator of $\|\Sigma_1 - \Sigma_2\|_F^2$, where $\|\cdot\|_F$ denotes the Frobenius norm defined at Section 2.1. Srivastava and Yanagihara (2010) suggested a test based on a consistent estimator of $\mathrm{tr}(\Sigma_1^2)/\{\mathrm{tr}(\Sigma_1)\}^2 - \mathrm{tr}(\Sigma_2^2)/\{\mathrm{tr}(\Sigma_2)\}^2$. These tests can be categorized as ℓ_2-type tests. A two-sample covariance test based on super-diagonals was proposed by He and Chen (2018) whose test turned out to be more powerful than other existing tests when Σ_1 and Σ_2 have bandable structures. However, the aforementioned tests target many signals, where most of components of $\Sigma_1 - \Sigma_2$ are nonzero. Thus, they might be less powerful under the rare signals setting, where only a few entries in $\Sigma_1 - \Sigma_2$ are nonzero. In such a situation, a maximum-type test might outperform ℓ_2-type tests. Cai et al. (2013) proposed a maximum-type test for two-sample covariance testing. Similar to two-sample mean test in Cai et al. (2014), Cai et al. (2013) standardized the difference between sample covariances and took the maximum over the standardized sample covariances. Recently, Zheng et al. (2017) combined the two tests in Li and Chen (2012) and Cai et al. (2013) by taking weighted average to handle both rare and many signals.

On the other hand, up to our knowledge, no theoretically supported Bayesian method has been proposed for high-dimensional two-sample tests, except a recent work of Zoh et al. (2018). They proposed a Bayesian test for high-dimensional two-sample mean test by reducing the dimension of data via random projections. They proved consistency of the proposed Bayesian test under the joint distribution of *data and prior*, where the true mean vector is a random variable from the prior distribution.

In this paper, we develop scalable Bayesian two-sample tests supported by theoretical guarantees. Since rare signals can be more realistic in many applications, our goal is to develop a consistent Bayesian test achieving good power when there are rare signals. To this end, we apply the maximum pairwise Bayes factor approach suggested by Lee et al. (2021), which is essentially a divide-and-conquer idea. Rather than comparing the whole mean vectors or covariance matrices at once, we divide them into smaller pieces. Although we employ the general idea of modularization by Lee et al. (2021), the former work however only focuses on one-sample testing of the structure of covariance matrices. Substantial new developments have been made in this work which differs in terms of problem setup, prior choice, theory development as well as computational approach.

The main contributions of this paper can be summarized as follows. The proposed Bayesian tests are scalable with simple implementations that can be readily used by practitioners. It accelerates the computation speed by circumventing computational issues such as inversion of a large matrix. Furthermore, up to our knowledge, these are the first results on Bayes factor consistency in high-dimensional two-sample testings. We prove that the proposed Bayesian tests are consistent under both null and alternative under mild conditions (Theorems 2.1 and 3.1), where the true parameter is a fixed unknown quantity, which differentiates our results from those in Zoh et al. (2018). Simulation studies show that the proposed tests have the desired property of being much more powerful than ℓ_2-type tests under rare signals settings. Besides the development of new Bayesian methods, our proposal also improves state-of-the-art methods theoretically and empirically. We show that the derived testable regions from the proposed tests are optimal in terms of convergence rates (Theorem 3.2), and the required conditions for achieving the theoretical results are much weaker than those used in existing literature. Furthermore, although there are existing frequentist maximum-type tests (Cai et al., 2013, 2014), the proposed tests in this paper outperform the contenders in various settings.

The rest of paper is organized as follows. Sections 2 and 3 present the proposed Bayesian two-sample tests for mean vectors and covariance matrices, respectively. In Section 4, the practical performance of the proposed methods is evaluated based on numerical study. Concluding remarks are given in Section 5, and proofs of the main results are included in the supplementary material (Lee et al., 2023).

2 Two-sample mean test

2.1 Notation

For any given constants a and b, we denote the maximum and minimum between the two by $a \vee b$ and $a \wedge b$. For a vector $x = (x_1, \ldots, x_p)^T$ and a positive integer q, we denote the vector ℓ_q-norm as $\|x\|_q = \big(\sum_{j=1}^p x_j^q\big)^{1/q}$. For any positive sequences a_n and b_n, $a_n \ll b_n$, or equivalently $a_n = o(b_n)$, means that $a_n/b_n \longrightarrow 0$ as $n \to \infty$. We denote $a_n = O(b_n)$ if there exists a constant $C > 0$ such that $a_n/b_n \le C$ for all large n, and $a_n \asymp b_n$ means that $a_n = O(b_n)$ and $b_n = O(a_n)$. For a given matrix $A \in \mathbb{R}^{p\times p}$, we denote the Frobenius norm $\|A\|_F = \big(\sum_{i=1}^p \sum_{j=1}^p a_{ij}^2\big)^{1/2}$, the matrix ℓ_1-norm $\|A\|_1 = \sup_{x\in\mathbb{R}^p, \|x\|_1=1} \|Ax\|_1$ and the matrix maximum norm $\|A\|_{\max} = \max_{1\le i\le j\le p} |a_{ij}|$. The maximum and minimum eigenvalues of a matrix A are denoted by $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$, respectively. For given positive numbers a and b, $IG(a, b)$ denotes the inverse-gamma distribution with shape parameter a and rate parameter b.

2.2 Maximum pairwise Bayes factor for two-sample mean test

Suppose that we observe the data from two populations

$$\begin{aligned} X_i \mid \mu_1, \Sigma &\overset{i.i.d.}{\sim} N_p(\mu_1, \Sigma), \quad i = 1, \ldots, n_1, \\ Y_i \mid \mu_2, \Sigma &\overset{i.i.d.}{\sim} N_p(\mu_2, \Sigma), \quad i = 1, \ldots, n_2, \end{aligned} \tag{2.1}$$

where $\mu_1, \mu_2 \in \mathbb{R}^p$ and Σ is a $p \times p$ covariance matrix. Let $\mathbf{X}_{n_1} = (X_1, \ldots, X_{n_1})^T \in \mathbb{R}^{n_1\times p}$ and $\mathbf{Y}_{n_2} = (Y_1, \ldots, Y_{n_2})^T \in \mathbb{R}^{n_2\times p}$ be the data matrices for each population. We are interested in the testing problem

$$H_0 : \mu_1 = \mu_2 \quad \text{versus} \quad H_1 : \mu_1 \ne \mu_2. \tag{2.2}$$

Bayesian hypothesis tests are typically based on Bayes factors. To construct a Bayes factor for two-sample mean test, marginal likelihoods should be calculated based on priors for each hypothesis. Using normal priors for mean vectors and the Jeffreys' prior for a covariance matrix, which corresponds to a default choice, the resulting Bayes factor can be calculated in a closed form when $1 < p < n - 2$. See Zoh et al. (2018) for the details. However, the Bayes factor under such priors involves the inverse of a pooled sample covariance matrix, which prevents one from using when $p \geq n-2$. Zoh et al. (2018) suggested projecting the data to a lower-dimensional subspace to reduce the dimensionality.

In this paper, we apply the maximum pairwise Bayes factor (mxPBF) approach suggested by Lee et al. (2021). Specifically, we compare two mean vectors by comparing them element-by-element. For a given integer $1 \leq j \leq p$, let $\tilde{X}_j = (X_{1j}, \ldots, X_{n_1 j})^T$ and $\tilde{Y}_j = (Y_{1j}, \ldots, Y_{n_2 j})^T$ be the jth columns of $\mathbf{X}_{n_1}$ and $\mathbf{Y}_{n_2}$, respectively. From model (2.1), we have the following marginal models

$$\begin{aligned} \tilde{X}_j \mid \mu_{1j}, \sigma_{jj} &\sim N_{n_1}(\mu_{1j} 1_{n_1}, \sigma_{jj} I_{n_1}), \\ \tilde{Y}_j \mid \mu_{2j}, \sigma_{jj} &\sim N_{n_2}(\mu_{2j} 1_{n_2}, \sigma_{jj} I_{n_2}), \end{aligned}$$

where $\mu_k = (\mu_{k1}, \ldots, \mu_{kp})^T$ for $k = 1, 2$, $\Sigma = (\sigma_{ij})$ and $1_q = (1, \ldots, 1)^T \in \mathbb{R}^q$. The hypothesis testing problem (2.2) can be reformulated as

$$H_{0j} : \mu_{1j} = \mu_{2j} \quad \text{versus} \quad H_{1j} : \mu_{1j} \neq \mu_{2j},$$

in the sense that H_0 is true if and only if H_{0j} is true for all $j = 1, \ldots, p$. Thus, we will first construct Bayesian tests for each testing problem H_{0j} versus H_{1j} and calculate *pairwise Bayes factors* (PBFs) based on $(\tilde{X}_j, \tilde{Y}_j)$ for $j = 1, \ldots, p$. For a given $1 \leq j \leq p$, we suggest the following prior $\pi_{0j}(\mu_j, \sigma_{jj})$ under H_{0j},

$$\begin{aligned} \mu_j \mid \sigma_{jj} &\sim N\Big(\bar{Z}_j, \frac{\sigma_{jj}}{n\gamma}\Big), \\ \pi(\sigma_{jj}) &\propto \sigma_{jj}^{-1}, \end{aligned}$$

where $\mu_j = \mu_{1j} = \mu_{2j}$, and the following prior $\pi_{1j}(\mu_{1j}, \mu_{2j}, \sigma_{jj})$ under H_{1j},

$$\begin{aligned} \mu_{1j} \mid \sigma_{jj} &\sim N\Big(\bar{X}_j, \frac{\sigma_{jj}}{n_1\gamma}\Big), \\ \mu_{2j} \mid \sigma_{jj} &\sim N\Big(\bar{Y}_j, \frac{\sigma_{jj}}{n_2\gamma}\Big), \\ \pi(\sigma_{jj}) &\propto \sigma_{jj}^{-1}, \end{aligned}$$

where $\tilde{Z}_j = (\tilde{X}_j^T, \tilde{Y}_j^T)^T = (Z_{1j}, \ldots, Z_{nj})^T$, $\bar{Z}_j = n^{-1}\sum_{i=1}^n Z_{ij}$, $\bar{X}_j = {n_1}^{-1}\sum_{i=1}^{n_1} X_{ij}$, $\bar{Y}_j = {n_2}^{-1}\sum_{i=1}^{n_2} Y_{ij}$, $n = n_1 + n_2$ and $\gamma = (n \vee p)^{-\alpha}$. Throughout this paper, we consider α as a fixed positive constant.

For any vector v, define the projection matrix $H_v = v(v^T v)^{-1} v^T$. Let $\widehat{\sigma}^2_{Z_j} = n^{-1}\tilde{Z}_j^T(I_n - H_{1_n})\tilde{Z}_j$, $\widehat{\sigma}^2_{X_j} = {n_1}^{-1}\tilde{X}_j^T(I_{n_1} - H_{1_{n_1}})\tilde{X}_j$ and $\widehat{\sigma}^2_{Y_j} = {n_2}^{-1}\tilde{Y}_j^T(I_{n_2} - H_{1_{n_2}})\tilde{Y}_j$. Then, the resulting log PBF is

$$\begin{aligned} \log B_{10}(\tilde{X}_j, \tilde{Y}_j) &:= \log \frac{p(\tilde{X}_j, \tilde{Y}_j \mid H_{1j})}{p(\tilde{X}_j, \tilde{Y}_j \mid H_{0j})} \\ &= \frac{1}{2}\log\Big(\frac{\gamma}{1+\gamma}\Big) + \frac{n}{2}\log\left(\frac{n\widehat{\sigma}^2_{Z_j}}{n_1\widehat{\sigma}^2_{X_j} + n_2\widehat{\sigma}^2_{Y_j}}\right), \end{aligned} \tag{2.3}$$

where

$$p(\tilde{X}_j, \tilde{Y}_j \mid H_{0j}) = \iint p(\tilde{X}_j \mid \mu_j, \sigma_{jj}, H_{0j}) p(\tilde{Y}_j \mid \mu_j, \sigma_{jj}, H_{0j}) \pi_{0j}(\mu_j, \sigma_{jj}) d\mu_j d\sigma_{jj},$$

$$p(\tilde{X}_j, \tilde{Y}_j \mid H_{1j}) = \iiint p(\tilde{X}_j \mid \mu_{1j}, \sigma_{jj}, H_{1j}) p(\tilde{Y}_j \mid \mu_{2j}, \sigma_{jj}, H_{1j}) \times \pi_{1j}(\mu_{1j}, \mu_{2j}, \sigma_{jj}) d\mu_{1j} d\mu_{2j} d\sigma_{jj}.$$

The derivation of (2.3) is given in the supplementary material. To aggregate PBFs for all $j = 1, \ldots, p$, we define the mxPBF as

$$B^{\mu}_{\max,10}(\mathbf{X}_{n_1}, \mathbf{Y}_{n_2}) \quad := \quad \max_{1 \le j \le p} B_{10}(\tilde{X}_j, \tilde{Y}_j). \tag{2.4}$$

Then one can conduct a Bayesian test based on the mxPBF by rejecting the null $H_0 : \mu_1 = \mu_2$ if the mxPBF is larger than a prespecified threshold. Note that the mxPBF $B^{\mu}_{\max,10}(\mathbf{X}_{n_1}, \mathbf{Y}_{n_2})$ supports $H_1 : \mu_1 \neq \mu_2$ if and only if there is at least one strong evidence in favor of $H_{1j} : \mu_{1j} \neq \mu_{2j}$.

We note here that the proposed mxPBF is quite a naive approach that completely ignores the dependence structure of the data and treats p variables as independent of each other. Thus, the proposed method does not require priors on the entries σ_{ij} for $1 \le i \neq j \le p$, which also simplifies computations compared to the approach of Zoh et al. (2018).

As pointed out by a referee, the prior used for the mxPBF depends on the data. A data-dependent prior uses the data twice, so in general, it might hurt sequential analyses and Bayesian updating. Nevertheless, empirical priors are somewhat routinely used in Bayesian analysis, and we have used the data-dependent priors to avoid introducing unnecessary conditions when proving theoretical results. In the subsequent section, we partially justify the above concerns by establishing the consistency of mxPBF.

2.3 Bayes factor consistency

A mxPBF is said to be consistent if it (i) converges to zero under H_0 and (ii) diverges to infinity under H_1 in probability. Let $\mu_{01} = (\mu_{01,j}) \in \mathbb{R}^p$ and $\mu_{02} = (\mu_{02,j}) \in \mathbb{R}^p$ be true mean vectors for each population, respectively, and $\Sigma_0 = (\sigma_{0,ij}) \in \mathbb{R}^{p \times p}$ be the true covariance matrix. Theorem 2.1 shows that, despite the ignorance of the dependence structure, the mxPBF is consistent under mild conditions.

Theorem 2.1. *Consider model (2.1) and the two-sample mean test $H_0 : \mu_1 = \mu_2$ versus $H_1 : \mu_1 \neq \mu_2$. Assume that $\log p \le n\epsilon_0$ and*

$$\alpha \quad > \quad \frac{2(1+\epsilon_0)}{1 - 3\sqrt{C_1 \epsilon_0}}, \tag{2.5}$$

for some constant $0 < \epsilon_0 < 1$ and any constant $C_1 > 1$ arbitrarily close to 1. Then, the mxPBF (2.3) is consistent under H_0: for some constant $c > 0$,

$$B^{\mu}_{\max,10}(\mathbf{X}_{n_1}, \mathbf{Y}_{n_2}) \quad = \quad O_p\{(n \vee p)^{-c}\} \quad \text{under } H_0.$$

When H_1 is true, assume that there is at least one of indices $1 \le j \le p$ satisfying

$$\frac{n_1 n_2(\mu_{01,j}-\mu_{02,j})^2}{n^2\sigma_{0,jj}} \geq \left[\sqrt{2C_1}+\sqrt{2C_1+\alpha C_1\{1+(1+8C_1)\epsilon_0\}}\right]^2 \frac{\log(n\vee p)}{n}. \quad (2.6)$$

Then, the mxPBF is also consistent under H_1: for some constant $c'>0$,

$$\left\{B^{\mu}_{\max,10}(\mathbf{X}_{n_1},\mathbf{Y}_{n_2})\right\}^{-1} \;=\; O_p\{(n\vee p)^{-c'}\} \quad \textit{under } H_1.$$

It is worthwhile to compare our findings to those of previous studies. As mentioned earlier, the test statistic of Cai et al. (2014) depends on an estimated precision matrix that some conditions for consistent estimation of the precision matrix are required. For example, it was assumed that Ω_0 has bounded eigenvalues and absolute correlations of X_i, Y_i, $\Omega_0 X_i$ and $\Omega_0 Y_i$ are bounded away from 1, where Ω_0 is the true precision matrix. Furthermore, Ω_0 is assumed to satisfy $\|\Omega_0\|_1^2 = o(\sqrt{n/(\log p)^3})$ or stronger sparsity assumption, which essentially means that a large amount of entries in Ω_0 is sufficiently small. They also assumed that $\mu_{01}-\mu_{02}$ has at most p^r nonzero entries, where $r\in[0,1/4)$.

On the other hand, theoretical results in Theorem 2.1 do not require any condition on the true precision matrix and allow the number of nonzero entries in $\mu_{01}-\mu_{02}$ to have the same order with p. Therefore, we suspect that the mxPBF would perform better than the maximum-type test in Cai et al. (2014) when these conditions are violated. Indeed, we find empirical evidences for this conjecture in our simulation study in Section 4.2.

Recently, Zoh et al. (2018) proposed a Bayesian two-sample mean test and proved consistency of the Bayes factor in high-dimensional settings. They used random projections to reduce the dimensionality of the data and assumed that the reduced dimension has the same order with $(n_1\wedge n_2)$. To conduct a Bayesian test, a single random projection matrix was considered, which can lead to different results depending on the generated projection matrix. Furthermore, they did not provide a condition on the lower bound of $\mu_{01}-\mu_{02}$ of $\mu_{01}-\mu_{02}$ like condition (2.6) to ensure consistency under the true alternative. They assumed that $\mu_1-\mu_2$ is a random vector under $H_1:\mu_1\neq\mu_2$ rather than considering a fixed true value $\mu_{01}-\mu_{02}$, which differentiates our results from those in Zoh et al. (2018).

We note here that, under regularity assumptions, condition (2.6) is rate-optimal in the following sense. When $\log(n\vee p)\asymp\log p$, $\|\Omega_0\|_1^2 = O(1)$ and $\sum_{j=1}^p I(\mu_{01,j}\neq\mu_{02,j}) = p^r$ for some $0<r<1/4$, Theorem 3 in Cai et al. (2014) implies that no consistent test exists that also has vanishing Type I error for the alternative class satisfying condition (2.6). Thus, condition (2.6) is minimax rate-optimal, and the proposed mxPBF-based test provides a minimax optimal testable region with respect to the maximum norm. Cai et al. (2014) assumed the condition $\max_j(\mu_{01,j}-\mu_{02,j})^2\geq C\log p/n$ for some constant $C>0$, which is similar to (2.6).

3 Two-sample covariance test

In this section, we propose Bayesian two-sample tests for testing the equity of high-dimensional covariance matrices and consider their theoretical properties in terms of Bayes factor consistency and optimality of the testing regions.

3.1 Maximum pairwise Bayes factor for two-sample covariance test

Suppose that we observe the data from two populations

$$\begin{aligned} X_i \mid \Sigma_1 &\overset{i.i.d.}{\sim} N_p(0, \Sigma_1), \quad i = 1, \ldots, n_1, \\ Y_i \mid \Sigma_2 &\overset{i.i.d.}{\sim} N_p(0, \Sigma_2), \quad i = 1, \ldots, n_2, \end{aligned} \tag{3.1}$$

where $\Sigma_1 = (\sigma_{1,ij})$ and $\Sigma_2 = (\sigma_{2,ij})$ are $p \times p$ covariance matrices. In this section, we consider the testing problem

$$H_0 : \Sigma_1 = \Sigma_2 \quad \text{versus} \quad H_1 : \Sigma_1 \neq \Sigma_2. \tag{3.2}$$

To apply the mxPBF approach, we need to divide the comparison of two covariance matrices into smaller problems. Among various options for that, we use the reparametrization trick used in Lee et al. (2021). Specifically, for a given pair (i, j) with $1 \le i \neq j \le p$, (3.1) induces the conditional distributions

$$\begin{aligned} \tilde{X}_i \mid \tilde{X}_j, a_{1,ij}, \tau_{1,ij} &\sim N_{n_1}\big(a_{1,ij}\tilde{X}_j, \tau_{1,ij} I_{n_1}\big), \\ \tilde{Y}_i \mid \tilde{Y}_j, a_{2,ij}, \tau_{2,ij} &\sim N_{n_2}\big(a_{2,ij}\tilde{Y}_j, \tau_{2,ij} I_{n_2}\big), \end{aligned} \tag{3.3}$$

where $a_{k,ij} = \sigma_{k,ij}/\sigma_{k,jj}$, $\tau_{k,ij} = \sigma_{k,ii}(1 - \rho_{k,ij}^2)$ and $\rho_{k,ij} = \sigma_{k,ij}/(\sigma_{k,ii}\sigma_{k,jj})^{1/2}$ for $k = 1, 2$. The hypothesis testing problem (3.2) can be reformulated as

$$H_{0,ij} : a_{1,ij} = a_{2,ij} \text{ and } \tau_{1,ij} = \tau_{2,ij} \quad \text{versus} \quad H_{1,ij} : \text{ not } H_{0,ij}, \tag{3.4}$$

in the sense that H_0 is true if and only if $H_{0,ij}$ is true for all pairs (i, j), $1 \le i \neq j \le p$.

To construct a Bayesian test for testing (3.4), we suggest the following prior distribution $\pi_{0,ij}(a_{ij}, \tau_{ij})$ under $H_{0,ij}$,

$$\begin{aligned} a_{ij} \mid \tau_{ij} &\sim N\Big(\widehat{a}_{ij}, \frac{\tau_{ij}}{\gamma \|\tilde{Z}_j\|_2^2}\Big), \\ \tau_{ij} &\sim IG(a_0, b_{0,ij}), \end{aligned}$$

where $a_{ij} = a_{1,ij} = a_{2,ij}$ and $\tau_{ij} = \tau_{1,ij} = \tau_{2,ij}$, and the prior $\pi_{1,ij}(a_{1,ij}, a_{2,ij}, \tau_{1,ij}, \tau_{2,ij})$ under $H_{1,ij}$,

$$\begin{aligned} a_{1,ij} \mid \tau_{1,ij} &\sim N\Big(\widehat{a}_{1,ij}, \frac{\tau_{1,ij}}{\gamma \|\tilde{X}_j\|_2^2}\Big), \quad a_{2,ij} \mid \tau_{2,ij} \sim N\Big(\widehat{a}_{2,ij}, \frac{\tau_{2,ij}}{\gamma \|\tilde{Y}_j\|_2^2}\Big), \\ \tau_{1,ij} &\sim IG(a_0, b_{01,ij}), \quad \tau_{2,ij} \sim IG(a_0, b_{02,ij}), \end{aligned}$$

where $a_0, b_{0,ij}, b_{01,ij}$ and $b_{02,ij}$ are positive constants, $\gamma = (n \vee p)^{-\alpha}$, $\widehat{a}_{ij} = \tilde{Z}_i^T \tilde{Z}_j / \|\tilde{Z}_j\|_2^2$, $\widehat{a}_{1,ij} = \tilde{X}_i^T \tilde{X}_j / \|\tilde{X}_j\|_2^2$ and $\widehat{a}_{2,ij} = \tilde{Y}_i^T \tilde{Y}_j / \|\tilde{Y}_j\|_2^2$. Let $\widehat{\tau}_{ij} = n^{-1} \tilde{Z}_i^T (I_n - H_{\tilde{Z}_j}) \tilde{Z}_i$, $\widehat{\tau}_{1,ij} = n_1^{-1} \tilde{X}_i^T (I_{n_1} - H_{\tilde{X}_j}) \tilde{X}_i$ and $\widehat{\tau}_{2,ij} = n_2^{-1} \tilde{Y}_i^T (I_{n_2} - H_{\tilde{Y}_j}) \tilde{Y}_i$. The resulting log PBF is given by

$$\begin{aligned} &\log B_{10}(\tilde{X}_i, \tilde{Y}_i, \tilde{X}_j, \tilde{Y}_j) \\ &:= \log \frac{p(\tilde{X}_i, \tilde{Y}_i \mid \tilde{X}_j, \tilde{Y}_j, H_{1,ij})}{p(\tilde{X}_i, \tilde{Y}_i \mid \tilde{X}_j, \tilde{Y}_j, H_{0,ij})} \end{aligned} \tag{3.5}$$

$$= \frac{1}{2}\log\Big(\frac{\gamma}{1+\gamma}\Big) + \log\Gamma\Big(\frac{n_1}{2}+a_0\Big) + \log\Gamma\Big(\frac{n_2}{2}+a_0\Big)$$
$$- \log\Gamma\Big(\frac{n}{2}+a_0\Big) + \log\Big(\frac{b_{01,ij}^{a_0} b_{02,ij}^{a_0}}{b_{0,ij}^{a_0}\Gamma(a_0)}\Big)$$
$$- \Big(\frac{n_1}{2}+a_0\Big)\log\big(b_{01,ij}+\frac{n_1}{2}\widehat{\tau}_{1,ij}\big) - \Big(\frac{n_2}{2}+a_0\Big)\log\big(b_{02,ij}+\frac{n_2}{2}\widehat{\tau}_{2,ij}\big) \qquad (3.6)$$
$$+ \Big(\frac{n}{2}+a_0\Big)\log\big(b_{0,ij}+\frac{n}{2}\widehat{\tau}_{ij}\big), \qquad (3.7)$$

where Γ is the Gamma function and

$$p(\tilde{X}_i, \tilde{Y}_i \mid \tilde{X}_j, \tilde{Y}_j, H_{0,ij})$$
$$= \iint p(\tilde{X}_i \mid \tilde{X}_j, a_{ij}, \tau_{ij}, H_{0,ij}) p(\tilde{Y}_i \mid \tilde{Y}_j, a_{ij}, \tau_{ij}, H_{0,ij}) \pi_{0,ij}(a_{ij}, \tau_{ij}) da_{ij} d\tau_{ij},$$
$$p(\tilde{X}_i, \tilde{Y}_i \mid \tilde{X}_j, \tilde{Y}_j, H_{1,ij})$$
$$= \iiiint p(\tilde{X}_i \mid \tilde{X}_j, a_{1,ij}, \tau_{1,ij}, H_{1,ij}) p(\tilde{Y}_i \mid \tilde{Y}_j, a_{2,ij}, \tau_{2,ij}, H_{1,ij})$$
$$\times\ \pi_{1,ij}(a_{1,ij}, a_{2,ij}, \tau_{1,ij}, \tau_{2,ij}) da_{1,ij} da_{2,ij} d\tau_{1,ij} d\tau_{2,ij}.$$

The derivation of (3.5) is given in the supplementary material. Then, the mxPBF for two-sample covariance test is given by

$$B^{\Sigma}_{\max,10}(\mathbf{X}_{n_1}, \mathbf{Y}_{n_2}) \quad := \quad \max_{i\neq j} B_{10}(\tilde{X}_i, \tilde{Y}_i, \tilde{X}_j, \tilde{Y}_j). \qquad (3.8)$$

Similar to the two-sample mean test, one can conduct a Bayesian test based on the mxPBF by rejecting the null $H_0 : \Sigma_1 = \Sigma_2$ if the mxPBF is larger than a prespecified threshold.

We note here that the mxPBF approach for two-sample covariance test essentially treats each pair, the ith and jth variables, as if they were independent of the other $p-2$ variables. Despite this naive treatment, Theorem 3.1 in the subsequent section shows that the mxPBF is consistent under mild conditions.

3.2 Bayes factor consistency

In this section, we show that the mxPBF in (3.8) is consistent for high-dimensional two-sample covariance test. We first introduce sufficient conditions that guarantee consistency of the mxPBF. The first condition, (A1), roughly means that $p = O(\exp(n^c))$ for some $0 < c < 1$.

(A1) $\epsilon_{0k} := \log(n \vee p)/n_k = o(1)$ for $k = 1, 2$.

When $H_0 : \Sigma_1 = \Sigma_2$ is true, we denote Σ_0 as the true covariance matrix. Furthermore, we define $a_{0,ij} = \sigma_{0,ij}/\sigma_{0,jj}, \tau_{0,ij} = \sigma_{0,ii}(1-\rho^2_{0,ij})$, and $R_0 = (\rho_{0,ij})$ is a correlation matrix. Condition (A2) is a sufficient condition for consistency under the null $H_0 : \Sigma_1 = \Sigma_2$.

(A2) $\min_{i\neq j}\tau_{0,ij} \gg \{\log(n\vee p)\}^{-1}$.

Condition (A2) is satisfied if $\min_{1\le i\le p}\sigma_{0,ii} > \epsilon$ and $\max_{i\neq j}\rho_{0,ij}^2 < 1-\epsilon$ for some small constant $\epsilon > 0$. However, in fact, condition (A2) allows more general cases where possibly $\sigma_{0,ii}\to 0$ and $\rho_{0,ij}^2 \to 1$ as $p\to\infty$ at certain rates.

When $H_1 : \Sigma_1 \neq \Sigma_2$ is true, we denote $\Sigma_{01} = (\sigma_{01,ij})$ and $\Sigma_{02} = (\sigma_{02,ij})$ as the true covariance matrices for each population. Furthermore, we define $a_{0k,ij} = \sigma_{0k,ij}/\sigma_{0k,jj}, \tau_{0k,ij} = \sigma_{0k,ii}(1-\rho_{0k,ij}^2)$ and $R_{0k} = (\rho_{0k,ij})$ is a correlation matrix for $k = 1, 2$. Under the alternative $H_1 : \Sigma_1 \neq \Sigma_2$, we assume that $(\Sigma_{01}, \Sigma_{02})$ satisfies the following condition (A3) or (A3*). Note that the constant α is the hyperparameter used in the priors $\pi_{0,ij}$ and $\pi_{1,ij}$.

(A3) There exists a pair (i, j) with $i \neq j$ such that

$$\{\log(n\vee p)\}^{-1} \ll \tau_{01,ij}\wedge\tau_{02,ij} \quad\le\quad \tau_{01,ij}\vee\tau_{02,ij} \ll (n\vee p),$$

satisfying either

$$\frac{\tau_{01,ij}}{\tau_{02,ij}} \quad > \quad \frac{1+C_{\rm bm}\sqrt{\epsilon_{01}}}{1-4\sqrt{C_1(\epsilon_{01}\vee\epsilon_{02})}},$$

or

$$\frac{\tau_{02,ij}}{\tau_{01,ij}} \quad > \quad \frac{1+C_{\rm bm}\sqrt{\epsilon_{02}}}{1-4\sqrt{C_1(\epsilon_{01}\vee\epsilon_{02})}},$$

for some constant $C_{\rm bm}^2 > 8(\alpha+1)$ and any constant $C_1 > 1$ arbitrarily close to 1.

(A3*) There exists a pair (i, j) with $i \neq j$ such that $\sigma_{01,ii}\vee\sigma_{02,ii} \ll (n\vee p)$ and

$$(a_{01,ij}-a_{02,ij})^2 \quad\ge\quad \frac{25}{2}C_1\sum_{k=1}^{2}\Big\{\frac{\tau_{0k,ij}\epsilon_{0k}}{\sigma_{0k,jj}(1-2\sqrt{C_1\epsilon_{0k}})}\Big\}, \tag{3.9}$$

$$(a_{01,ij}-a_{02,ij})^2 \quad\ge\quad \frac{10n}{n+2a_0}\sum_{k=1}^{2}\Big\{\frac{\epsilon_{0k}}{\sigma_{0k,jj}(1-2\sqrt{C_1\epsilon_{0k}})}\Big\} \tag{3.10}$$

$$\times\Big[\frac{b_{0,ij}}{\log(n\vee p)}+\Big\{\sum_{k=1}^{2}\sigma_{0k,ii}(1+4\sqrt{C_1\epsilon_{0k}})+\frac{2b_{0,ij}}{n}\Big\}C_{{\rm bm},a}\Big],$$

for some constant $C_{{\rm bm},a} > \alpha + a_0 + 1$ and any constant $C_1 > 1$ arbitrarily close to 1.

Conditions (A3) and (A3*) may seem complicated at first glance, but it can be transformed into simpler conditions. For given positive constants $\alpha, C_{\rm bm}$ and $C_{{\rm bm},a}$ such that $C_{\rm bm}^2 > 8(\alpha + 1)$ and $C_{{\rm bm},a} > \alpha + 1$, define a class of two covariance matrices

$$H_1(C_{\rm bm}, C_{{\rm bm},a}) \quad := \quad \Big\{(\Sigma_1,\Sigma_2) : (\Sigma_1,\Sigma_2) \text{ satisfies condition (A3) or (A3}^\star)\Big\}.$$

Conditions (A3) and (A3*) specify the minimum difference condition between Σ_{01} and Σ_{02} to consistently detect the alternative $H_1 : \Sigma_1 \neq \Sigma_2$ under the reparametrization using $\{a_{0k,ij}, \tau_{0k,ij} : k = 1, 2 \text{ and } 1 \le i \neq j \le p\}$. Suppose that

$$\begin{aligned} \max_{1\le k\le 2} \max_{1\le i\neq j\le p} \rho^2_{0k,ij} &\le 1 - c_0, \\ \{\log(n\vee p)\}^{-1} \ll \min_{1\le k\le 2} \min_{1\le i\le p} \sigma_{0k,ii} &\le \max_{1\le k\le 2} \max_{1\le i\le p} \sigma_{0k,ii} \ll (n\vee p), \end{aligned} \tag{3.11}$$

for some small constant $c_0 >$. If $\alpha > 1$, $n_1 \asymp n_2$ and

$$\begin{aligned} \widetilde{H}_1(C_\star, c_0) := \Big\{ (\Sigma_1, \Sigma_2) : \max_{1\le i\le j\le p} \frac{(\sigma_{1,ij} - \sigma_{1,ij})^2}{\sigma_{1,ii}\sigma_{1,jj} + \sigma_{2,ii}\sigma_{2,jj}} \ge C_\star \frac{\log(n\vee p)}{n}, \\ (\Sigma_1, \Sigma_2) \text{ satisfies conditions in (3.11) with } c_0 \Big\}, \end{aligned} \tag{3.12}$$

then $\widetilde{H}_1(C_\star, c_0) \subset H_1(C_{\rm bm}, C_{{\rm bm},a})$ for some large constant $C_\star > 0$ by Lemma 3.1 in the supplementary material. Condition (3.12) characterizes the difference between Σ_{01} and Σ_{02} using the squared maximum *standardized difference*. Hence, conditions (A3) and (A3*) can essentially be understood as the squared maximum standardized difference condition given at (3.12). Cai et al. (2013) also used a similar difference measure between Σ_{01} and Σ_{02}.

The following theorem shows consistency of the mxPBF (3.8). We note that the condition $\lim_{(n_1\wedge n_2)\to\infty} n_1/n = 1/2$ in Theorem 3.1 can be relaxed to $n_1 \asymp n_2$, although constants in conditions (A2), (A3) and (A3*) should be changed accordingly.

Theorem 3.1. *Consider model (3.1) and the two-sample covariance test $H_0 : \Sigma_1 = \Sigma_2$ versus $H_1 : \Sigma_1 \neq \Sigma_2$. Assume that $\lim_{(n_1\wedge n_2)\to\infty} n_1/n = 1/2$ and condition (A1) holds. Then, under H_0, if $\alpha > 12$ and condition (A2) holds, for some constant $c > 0$,*

$$B^{\Sigma}_{\max,10}(\mathbf{X}_{n_1}, \mathbf{Y}_{n_2}) \quad = \quad O_p\{(n\vee p)^{-c}\}.$$

Under H_1, if $(\Sigma_{01}, \Sigma_{02}) \in H_1(C_{\rm bm}, C_{{\rm bm},a})$, for some constant $c' > 0$,

$$\{B^{\Sigma}_{\max,10}(\mathbf{X}_{n_1}, \mathbf{Y}_{n_2})\}^{-1} \quad = \quad O_p\{(n\vee p)^{-c}\}.$$

Cai et al. (2013) considered a high-dimensional setting, $(\log p)^5 = o(n_k)$, while we assume a weaker condition, $\log p = o(n_k)$ for $k = 1, 2$ (condition (A1)). For given constants $C > 0$ and $0 < r < 1$, define $s_j(C) = \text{card}\{i : |\rho_{01,ij}| \ge (\log p)^{-1-C} \text{ or } |\rho_{02,ij}| \ge (\log p)^{-1-C}\}$ and $\Lambda(r) = \{i : |\rho_{01,ij}| > r \text{ or } |\rho_{02,ij}| > r \text{ for some } j \neq i\}$, where $\text{card}(A)$ means the cardinality of the set A. Cai et al. (2013) assumed that there exist $\Gamma \subset \{1, \ldots, p\}$, $C > 0$ and $0 < r < 1$ such that $\text{card}(\Gamma) = o(p)$, $\max_{j\neq\Gamma} s_j(C) = o(p^c)$ for some constant $c > 0$, and $\text{card}(\Lambda(r)) = o(p)$. These conditions essentially restrict the number of highly correlated variables. They are satisfied if $\lambda_{\max}(R_{01}) \vee \lambda_{\max}(R_{02}) \le C'$ for some constant $C' > 0$ and $\|R_{01}\|_{\max} \vee \|R_{02}\|_{\max} \le r < 1$. The power of their test tends to one if

$$\max_{1\le i\le j\le p} \frac{(\sigma_{01,ij} - \sigma_{01,ij})^2}{n_1^{-1}\theta_{01,ij} + n_2^{-1}\theta_{02,ij}}, \quad \ge \quad C \log p,$$

for $C \geq 4$, where $\theta_{01,ij} = \mathrm{Var}(X_{1i}X_{1j})$ and $\theta_{02,ij} = \mathrm{Var}(Y_{1i}Y_{1j})$. This condition is equivalent to condition (3.12) in terms of the rate. Thus, compared with those used in Cai et al. (2013), we obtain consistency of the mxPBF under weaker conditions for (n, p) and similar conditions for the true covariance matrices.

One of the interesting findings from Theorem 3.1 is that the mxPBF does not require any standardization step. Cai et al. (2013) mentioned that the standardization of the test statistic is necessary to deal with a wide range of variability and heteroscedasticity of sample covariances.

However, the mxPBF (3.8) still enjoys consistency for the similar parameter space without standardization. Although we did not mention earlier, a similar phenomenon is observed for the two-sample mean test: the proposed mxPBF (2.4) does not require a standardization step while having similar properties with a standardized test.

Another important finding is that condition (A3) (or (A3*)) is rate-optimal to guarantee consistency under $H_0 : \Sigma_1 = \Sigma_2$ as well as $H_1 : \Sigma_1 \neq \Sigma_2$. Theorem 3.2 shows that, for some small constants C_{bm} and $C_{\mathrm{bm},a} > 0$, no consistent test exists that also has vanishing Type I error for the alternative class $(\Sigma_{01}, \Sigma_{02}) \in H_1(C_{\mathrm{bm}}, C_{\mathrm{bm},a})$.

Theorem 3.2. *Let $\mathbb{E}_{\Sigma_{01},\Sigma_{02}}$ be the expectation corresponding to model (3.1) with $(\Sigma_{01}, \Sigma_{02})$. Suppose that $n_1 \asymp n_2$ and $p \geq n^c$ for some constant $c > 0$. Let $0 < \beta_0 < 1$ be a fixed constant. Then, there exists small constants C_1, C_{bm} and $C_{\mathrm{bm},a} > 0$ such that for all large n,*

$$\inf_{\phi \in \mathcal{T}} \sup_{(\Sigma_{01},\Sigma_{02}) \in H_1(C_{\mathrm{bm}}, C_{\mathrm{bm},a})} \mathbb{E}_{\Sigma_{01},\Sigma_{02}}(1-\phi) \quad \geq \quad \beta_0,$$

where $\mathcal{T}$ is the set of tests over the multivariate normal distributions such that $\mathbb{E}_0\phi \longrightarrow 0$ as $n \to \infty$ for any $\phi \in \mathcal{T}$, and $\mathbb{E}_0$ is the expectation corresponding to model (3.1) under $H_0 : \Sigma_1 = \Sigma_2$.

4 Numerical results

4.1 Choice of hyperparameters

The proposed mxPBFs for mean vectors and covariance matrices have hyperparameters that need to be determined. In this section, we provide some guidelines for their choice. The mxPBF (2.4) for mean vectors has a hyperparameter α, while the mxPBF (3.8) for covariance matrices have hyperparameters $a_0, b_{0,ij}, b_{01,ij}, b_{02,ij}$ and α. We first suggest, for the mxPBF (3.8), using $a_0 = b_{0,ij} = b_{01,ij} = b_{02,ij} = 0.01$ for all $1 \leq i \neq j \leq p$. Note that the above choice of hyperparameters does not affect the mxPBF (3.8) too much.

The choice of hyperparameter α in the mxPBFs (2.4) and (3.8) is more crucial to the performance. Although Theorems 2.1 and 3.1 provide theoretical conditions for α, they might be overly conservative in practice. Especially, Theorem 3.1 requires $\alpha > 12$ to guarantee consistency of the mxPBF (3.8) under the null. However, in simulation studies, we found this choice produced many false negatives leading to low sensitivity. Therefore, we suggest to choose α that controls an empirical false positive rate (FPR) at a prespecified level. Note that due to the nature of the mxPBF, it may be natural to use a method that controls FPR.

Here we describe a unified FPR-based method for choosing α, which can be used for both two-sample mean and covariance tests. Suppose we have observed samples from two populations, say $\mathbf{X}_{n_1}$ and $\mathbf{Y}_{n_2}$. Let $\hat{\mu}_{\text{pool}} = (n_1\hat{\mu}_1 + n_2\hat{\mu}_2)/n$ and $\hat{\Sigma}_{\text{pool}} = \{\sum_{i=1}^{n_1}(X_i - \hat{\mu}_1)(X_i - \hat{\mu}_1)^T + \sum_{i=1}^{n_2}(Y_i - \hat{\mu}_2)(Y_i - \hat{\mu}_2)^T\}/(n-2)$ be the pooled sample mean vector and covariance matrix, respectively, where $\hat{\mu}_1 = \sum_{i=1}^{n_1} X_i/n_1$ and $\hat{\mu}_2 = \sum_{i=1}^{n_2} Y_i/n_2$. When $\hat{\Sigma}_{\text{pool}}$ is not positive definite, we make it positive definite by adding $\{-\lambda_{\min}(\hat{\Sigma}_{\text{pool}}) + 0.1^3\}I_p$ to $\hat{\Sigma}_{\text{pool}}$. We generate a simulated dataset $\mathbf{X}_{\text{sim}} = (X_{1,\text{sim}}, \ldots, X_{n_1,\text{sim}})^T \in \mathbb{R}^{n_1 \times p}$ and $\mathbf{Y}_{\text{sim}} = (Y_{1,\text{sim}}, \ldots, Y_{n_2,\text{sim}})^T \in \mathbb{R}^{n_2 \times p}$, where $X_{i,\text{sim}}$s and $Y_{i,\text{sim}}$s are random samples from $N_p(\hat{\mu}_{\text{pool}}, \hat{\Sigma}_{\text{pool}})$. Note that under the null hypothesis, we assume $\mu_1 = \mu_2$ and $\Sigma_1 = \Sigma_2$. Thus, $(\mathbf{X}_{\text{sim}}, \mathbf{Y}_{\text{sim}})$ can be considered as a simulated dataset from the null hypothesis whose mean vector and covariance matrix are roughly close to those of the true data-generating distributions. For each simulated dataset, we calculate the mxPBF based on $(\mathbf{X}_{\text{sim}}, \mathbf{Y}_{\text{sim}})$ with a hyperparameter value α, say $B_{\max,10,\alpha}(\mathbf{X}_{\text{sim}}, \mathbf{Y}_{\text{sim}})$, and reject the null if the mxPBF is larger than 10. Note that if we reject the null, it corresponds to a false positive. By generating N simulated datasets $(\mathbf{X}_{\text{sim}}^{(s)}, \mathbf{Y}_{\text{sim}}^{(s)})_{s=1}^N$, we can calculate the following empirical FPR for each α,

$$\widehat{\text{FPR}}_\alpha = N^{-1}\sum_{s=1}^{N} I\Big(B_{\max,10,\alpha}(\mathbf{X}_{\text{sim}}^{(s)}, \mathbf{Y}_{\text{sim}}^{(s)}) > 10\Big). \tag{4.1}$$

To calculate the empirical FPR, we use a fixed threshold 10. Note that after calculating the mxPBF once for some α, the mxPBF for some $\alpha^* \neq \alpha$ can be easily calculated as $B_{\max,10,\alpha^*}(\mathbf{X}_{\text{sim}}, \mathbf{Y}_{\text{sim}}) = B_{\max,10,\alpha}(\mathbf{X}_{\text{sim}}, \mathbf{Y}_{\text{sim}}) - 0.5\log\{\gamma/(1+\gamma)\}) + 0.5\log\{\gamma^*/(1+\gamma^*)\}$, where $\gamma = (n \vee p)^{-\alpha}$ and $\gamma^* = (n \vee p)^{-\alpha^*}$. Then, among a grid of values $\alpha \in \{0.01, 0.02, \ldots, 15\}$, for example, we can select the minimum value of α, say $\hat{\alpha}$, that achieves a prespecified FPR level. It follows to calculate the value of the mxPBF using it, say $B_{\max,10,\hat{\alpha}}(\mathbf{X}_{n_1}, \mathbf{Y}_{n_2})$. For a given threshold C_{th}, we reject the null if $B_{\max,10,\hat{\alpha}}(\mathbf{X}_{n_1}, \mathbf{Y}_{n_2}) > C_{\text{th}}$.

Throughout numerical studies in Section 4, we control the empirical FPR at 0.05. In the supplementary material, we have described exact values of α's chosen by the FPR-based method in each setting and compared their performances with α's satisfying the theoretical conditions.

Regarding the empirical FPR, we admit that non-asymptotic control of the empirical FPR is not guaranteed by the current theoretical results. However, we can say that the empirical FPR converges to zero in probability as $\min(n_1, n_2) \to \infty$ and $N \to \infty$ as long as we only consider the range of α satisfying the conditions in Theorems 2.1 and 3.1. In the two-sample covariance test, for instance, we need to focus on the values of α larger than 12, and select the minimum value of α that achieves a prespecified FPR level.

As pointed out by a referee, the choice of threshold 10 in (4.1) seems somewhat arbitrary. This value of the threshold is chosen because it is a common default threshold for Bayes factors, and no rigorous theoretical support is involved in this choice. For a partial justification, we would like to mention that the choice of the threshold might not be very critical in practice at least in terms of specificity. Suppose we use the same threshold C_{th} for (4.1) and the mxPBF, $B_{\max,10,\hat{\alpha}}(\mathbf{X}_{n_1}, \mathbf{Y}_{n_2})$, to make the final decision, and control the empirical FPR at 0.05. Then if the empirical FPR is reasonably work well, we would have FPR values close to 0.05 (or equivalently, specificity values close to 0.95), regardless of the value of C_{th}. We acknowledge that this is just a rough justification. For more rigorous theoretical justification, we plan to develop a theoretically supported threshold selection as a future work.

4.2 Simulation study: two-sample mean test

In this section, we illustrate performance of the mxPBF for the two-sample mean test through simulation studies. We generate the data as follows: $X_1, \ldots, X_n \overset{i.i.d.}{\sim} N_p(\mu_1, \Sigma)$ and $Y_1, \ldots, Y_n \overset{i.i.d.}{\sim} N_p(\mu_2, \Sigma)$ with $n = 100$ and $p \in \{100, 300\}$. Under the null hypothesis, $H_0 : \mu_1 = \mu_2$, we set $\mu_1 = \mu_2 = 0 \in \mathbb{R}^p$. Under the alternative hypothesis, $H_1 : \mu_1 \neq \mu_2$, we set $\mu_1 = 0 \in \mathbb{R}^p$ and randomly choose n_0 entries in μ_2, say $\{\mu_{2,j} : 1 \leq j_1 < \cdots < j_{n_0} \leq p\}$, and set $\mu_{2,j} = \mu > 0$ for all $j = j_1, \ldots, j_{n_0}$ and $\mu_{2,j} = 0$ for the rest. Thus, n_0 and μ are the number and magnitude of signals in the alternative, respectively. Here, signals mean nonzero elements in $\mu_2 - \mu_1 \in \mathbb{R}^p$. In our simulation study, the following scenarios for alternatives are considered:

1. (H_{1R}: Rare signals) To demonstrate a situation where only a few signals exist, we set $n_0 = 5$ and consider various magnitudes of signals

$$\mu \in \{0.2, 0.25, 0.3, 0.35, 0.4, 0.5, 0.6, 0.8, 1.0, 1.5\}.$$

2. (H_{1M}: Many signals) To demonstrate a situation where a lot of signals exist, we set $n_0 = p/2$ and consider various magnitudes of signals

$$\mu \in \{0.025, 0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.4, 0.5, 0.6\}.$$

   Note that relatively smaller signals are used compared to "rare signals" setting, due to the larger number of signals.

Furthermore, we consider the following two settings for the true covariance matrix Σ:

1. (Sparse $\Omega = \Sigma^{-1}$) To demonstrate a situation where the true precision matrix is sparse, we randomly choose 1% of entries in $\Omega = (\omega_{ij})$ and set their value to $\omega_{ij} = 0.3$. The rest of entries in Ω are set to 0. When the resulting Ω is not positive definite, we make it positive definite by adding $\{-\lambda_{\min}(\Omega) + 0.1^3\} I_p$ to Ω. Finally, we set $\Sigma = \Omega^{-1}$.
2. (Dense $\Omega = \Sigma^{-1}$) To demonstrate a situation where the true precision matrix is dense, we randomly choose 40% of entries in Ω and set their value to $\omega_{ij} = 0.3$. The rest of the steps for constructing Σ is the same as above.

Note that the words “rare” and “many” describe how many nonzeros are in $\mu_2 - \mu_1 \in \mathbb{R}^p$, thus they are related to signals. On the other hand, the words “sparse” and “dense” describe how many nonzeros are in the common precision matrix Ω, regardless of the number of signals.

In each setting and hypothesis, 500 data sets are generated. Recall that the mxPBF for the two-sample mean test naively treats p variables as independent of each other.

For the proposed mxPBF for the two-sample mean test, the hyperparameter α is chosen by the FPR-based method described in Section 4.1. We reject the null hypothesis H_0 if the mxPBF is larger than some threshold $C_{\text{th}} > 0$. As contenders, we consider the tests proposed by Bai and Saranadasa (1996), Srivastava and Du (2008) and Cai et al. (2014), which will be simply denoted as BS, SD and CLX, respectively. Here, CLX means the two-sample mean test based on the CLIME, while CLX.AT refers to the two-sample mean test based on the inverse of the adaptive thresholding estimator with the tuning parameter $\delta = 2$ as a default choice. To choose the tuning parameter in CLIME, we used a cross-validation with `sugm` and `sugm.select` functions in the R package `flare`. Note that BS and SD are ℓ_2-type tests, while mxPBF, CLX and CLX.AT are maximum-type tests. It is expected that ℓ_2-type tests perform better (worse) than maximum-type tests in “many signals” (“rare signals”) setting.

To illustrate performance of each test, receiver operating characteristic (ROC) curves are drawn. The curves are obtained by adjusting thresholds and significance levels for the mxPBF and frequentist tests, respectively. Furthermore, we check the practical performance of mxPBF at a fixed threshold. As a default choice, threshold $C_{\text{th}} = 10$ is used. Note that $C_{\text{th}} = 10$ corresponds to “strong evidence” for the alternative hypothesis based on the criteria suggested by Jeffreys (1998) and Kass and Raftery (1995).

Figure 1 shows ROC curves based on 500 simulated data sets for each hypothesis, $H_0 : \mu_1 = \mu_2$ and $H_{1R} : \mu_1 \neq \mu_2$, with $p = 100$. Here, H_{1R} represents the “rare signals” scenario where $\mu_2 - \mu_1 \in \mathbb{R}^p$ has only five nonzero elements with size μ. The dots in Figure 1 show the results for the mxPBF with $C_{\text{th}} = 10$ and the FPR-based chosen α. In the rare signals scenario, the maximum-type tests overall slightly work better than the ℓ_2-type tests as expected. However, when the true precision matrix Ω is dense, CLX failed to infer because `sugm.select` R function for the cross-validation-based CLIME produced zero matrices. We suspect this is because the CLIME targets sparse precision matrices. For this reason, the results for CLX are shown only in sparse Ω setting. On the other hand, the mxPBF outperforms other tests in the dense Ω setting in term of the area under the curve (AUC). Furthermore, the mxPBF with $C_{\text{th}} = 10$ performs reasonably well. Especially in the dense Ω setting, when $\mu \geq 0.8$, its specificity and sensitivity are close to 1, while other tests suffer from low specificity or low sensitivity. This clearly shows the relative advantage of the mxPBF-based two-sample mean test over the existing tests.

Figure 2 shows ROC curves based on 500 simulated data sets for each hypothesis, $H_0 : \mu_1 = \mu_2$ and $H_{1M} : \mu_1 \neq \mu_2$, with $p = 100$. The dots in Figure 2 show the results for the mxPBF with $C_{\text{th}} = 10$ and the FPR-based chosen α. Here, H_{1M} represents the “many signals” scenario where $\mu_2 - \mu_1 \in \mathbb{R}^p$ has $p/2 = 50$ nonzero elements with

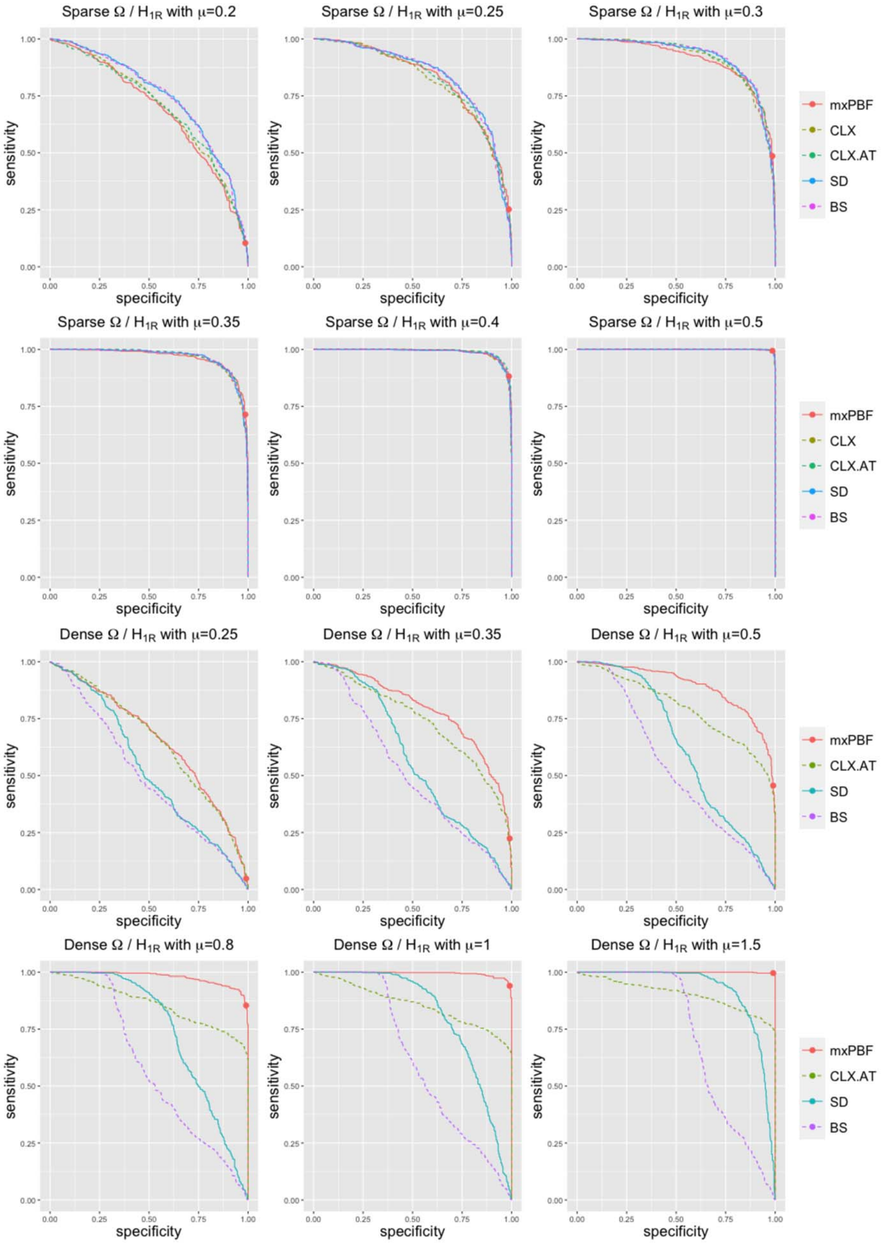

Figure 1: ROC curves for the two-sample mean tests based on 500 simulated data sets for each hypothesis, H_0 and H_{1R}, with $p = 100$. The mxPBF, SD and BS represent the test proposed in this paper, Srivastava and Du (2008) and Bai and Saranadasa (1996), respectively. The CLX and CLX.AT mean the tests proposed by Cai et al. (2014) based on the CLIME and the adaptive thresholding estimator, respectively. The dots show the results with $C_{\text{th}} = 10$ for the mxPBF.

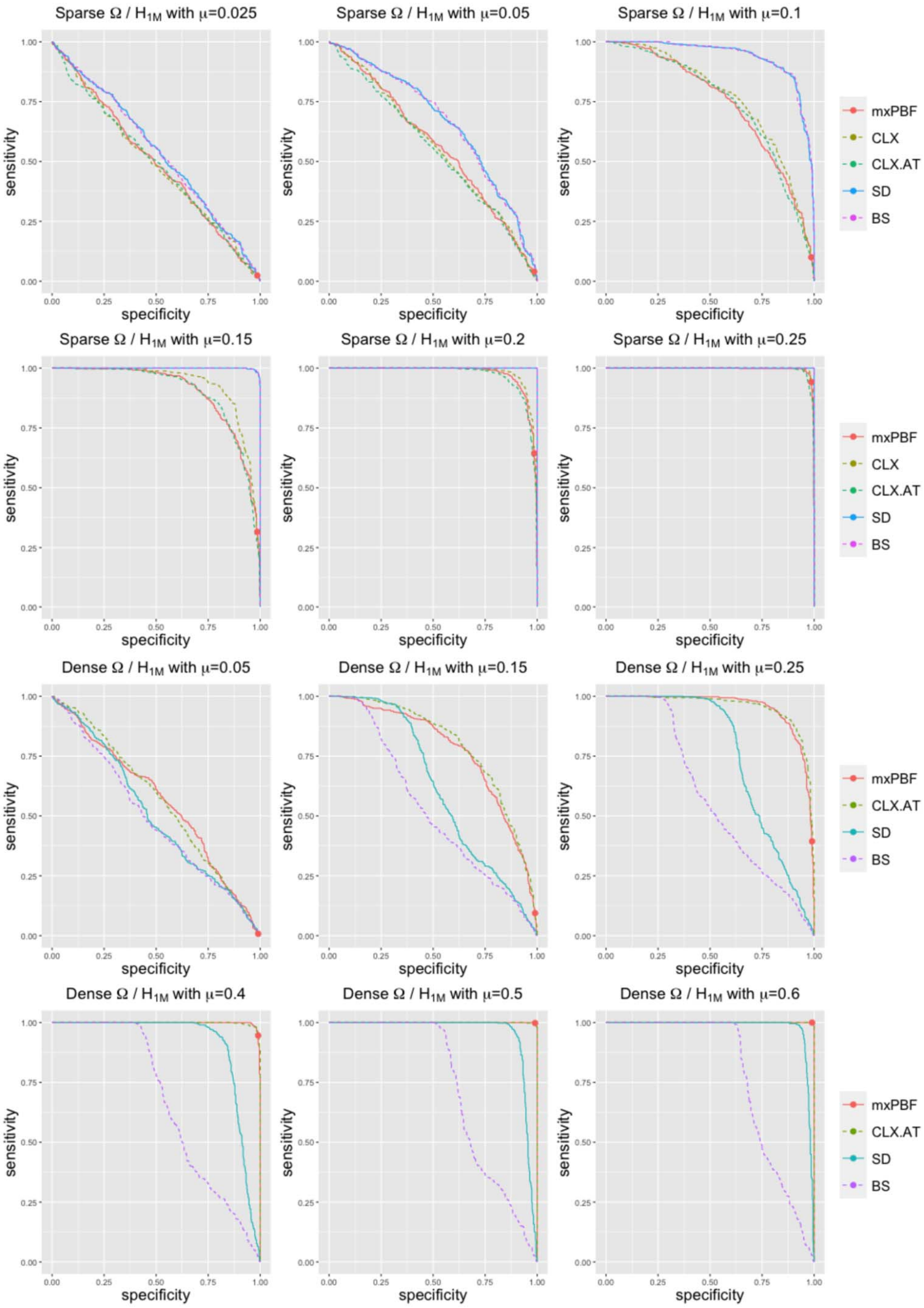

Figure 2: ROC curves for the two-sample mean tests based on 500 simulated data sets for each hypothesis, H_0 and H_{1M}, with $p = 100$.

size μ. When the true precision matrix Ω is sparse and $\mu \geq 0.1$, overall, the ℓ_2-type tests slightly work better than the maximum-type tests as expected. However, when the true precision matrix Ω is dense, somewhat surprisingly, the mxPBF and CLX.AT outperform the ℓ_2-type tests. This observation can be partially explained by theoretical properties of the ℓ_2-type tests: Bai and Saranadasa (1996) and Srivastava and Du (2008) showed that powers of their tests decrease as the Frobenius norm of the true covariance and correlation matrices increase, respectively. Indeed, in our simulations, we find that $\|\Sigma\|_F$ and $\|R\|_F$ are much larger in the dense Ω setting than in the sparse Ω setting. We further confirmed that, when H_{1M} is true, the ℓ_2-type tests tend to fail to reject H_0 even when the size of signals is large. Therefore, this observation suggests another advantage of the mxPBF that reasonable performance is maintained even when $\|\Sigma\|_F$ is large.

When $p = 300$, similar phenomena are observed, thus we omit it here for reasons of space. The results with $p = 300$ including ROC curves and descriptions are deferred to the Supplementary material.

4.3 Simulation study: two-sample covariance test

Now, we illustrate performance of the mxPBF for two-sample covariance test. We generate the data as follows: $X_1, \ldots, X_n \overset{i.i.d.}{\sim} N_p(0, \Sigma_1)$ and $Y_1, \ldots, Y_n \overset{i.i.d.}{\sim} N_p(0, \Sigma_2)$ with $n = 100$ and $p \in \{100, 300\}$. Under the null hypothesis, $H_0 : \Sigma_1 = \Sigma_2$, we set $\Sigma_1 = \Sigma_2 \equiv \Sigma \in \mathbb{R}^{p\times p}$. Under the alternative hypothesis, $H_1 : \Sigma_1 \neq \Sigma_2$, we set $\Sigma_1 \equiv \Sigma$ and $\Sigma_2 = \Sigma_1 + U$ for some matrix $U \in \mathbb{R}^{p\times p}$ containing signals. If Σ_1 or Σ_2 is not positive definite, we add a small diagonal matrix $\delta_1 I_p$ to them, where $\delta_1 = |\min\{\lambda_{\min}(\Sigma_1), \lambda_{\min}(\Sigma_2)\}| + 0.05$. Note that the matrix U determines the number and magnitude of signals in the alternative hypothesis. In our simulation study, the following two different scenarios for generating U are considered:

1. (H_{1R}: Rare signals) To demonstrate a situation where only few signals exist, we randomly select five entries in the lower triangular part of U and generate their values from $\mathrm{Unif}(0, \psi)$ with
$$\psi \in \{0.5, 0.8, 1.5, 3, 6, 15\}.$$
2. (H_{1M}: Many signals) To demonstrate a situation where a lot of signals exist, we generate $u = (u_1, \ldots, u_p)^T$ from $u_j \overset{i.i.d.}{\sim} \mathrm{Unif}(0, \psi)$ for
$$\psi \in \{0.2, 0.3, 0.5, 0.7, 1, 1.5\}.$$
For this "many signals" setting, we set $U = uu^T$ that leads to $p(p + 1)/2$ signals in U (except upper triangular part). Note that relatively smaller signals are used compared to "rare signals" setting, due to the larger number of signals.

Note that in the above, ψ is the magnitude of signals. Furthermore, we consider the following two settings for Σ_1:

1. (Sparse Σ_1) To demonstrate a situation where Σ_1 is sparse, we randomly choose 5% of entries in the lower triangular part of $\Delta_1 = (\delta_{1,jk})$ and set their value to $\delta_{1,jk} = 0.5$. The rest of entries in Δ_1 are set to 0. To make it positive definite, we set $\Delta = \Delta_1 + \delta I_p$, where $\delta = |\lambda_{\min}(\Delta_1)| + 0.05$. Finally, we set $\Sigma_1 = D^{1/2}\Delta D^{1/2}$, where $D = diag(d_j)$ and $d_j \overset{i.i.d.}{\sim}$ Unif$(0.5, 2.5)$. This setting corresponds to Model 3 in Cai et al. (2013).

2. (Dense Σ_1) To demonstrate a situation where Σ_1 is dense, we set $\Sigma_1 = O\Delta O$, where $O = diag(\omega_j)$, $\omega_j \overset{i.i.d.}{\sim}$ Unif$(1, 5)$, $\Delta = (\delta_{ij})$ and $\delta_{ij} = (-1)^{i+j}0.4^{|i-j|^{1/10}}$. This setting corresponds to Model 4 in Cai et al. (2013).

Similar to the previous two-sample mean case, the words "rare" and "many" describe how many nonzeros are in $\Sigma_2 - \Sigma_1 \in \mathbb{R}^{p\times p}$, thus they are related to signals. On the other hand, the words "sparse" and "dense" describe how many nonzeros are in the covariance matrix of the first group, Σ_1, regardless of the number of signals.

In each setting and hypothesis, we generate 500 simulated data. Recall that the mxPBF for two-sample covariance test naively treats each pair, the ith and jth variables, as independent of other $p - 2$ variables, thus the above sparse and dense Σ_1 settings violate the assumption of the mxPBF.

For the proposed mxPBF for two-sample covariance test, as described in Section 4.1, we use $a_0 = b_{0,ij} = b_{01,ij} = b_{02,ij} = 0.01$ for all $1 \le i \ne j \le p$, and the hyperparameter α is chosen by the FPR-based method.

For comparison, we consider the tests proposed by Schott (2007), Li and Chen (2012) and Cai et al. (2013), which will be denoted as Sch, LC and CLX, respectively. Note that Sch and LC are ℓ_2-type tests, while mxPBF and CLX are maximum-type tests. Because the unbiased version of the test in Li and Chen (2012) is computationally expensive, we use the biased version as suggested by Li and Chen (2012). Similar to the simulation study for two-sample mean test, ROC curves are drawn to demonstrate performance of each test.

Figure 3 shows ROC curves based on 500 simulated data sets for each hypothesis, $H_0 : \Sigma_1 = \Sigma_2$ and $H_{1R} : \Sigma_1 \ne \Sigma_2$, with $p = 100$. When Σ_1 is sparse and signals are moderate ($\psi \ge 0.8$), the maximum-type tests work better than the ℓ_2-type tests as expected. The performance of the ℓ_2-type tests are slowly improved as ψ gets larger. Similar phenomena are observed in the dense Σ_1 setting, but in this case, the ℓ_2-type tests do not work well even when there are large signals ($\psi = 15$). Overall, we find that the mxPBF shows better performance than CLX in terms of the AUC.

Figure 4 shows ROC curves based on 500 simulated data sets for each hypothesis, $H_0 : \Sigma_1 = \Sigma_2$ and $H_{1M} : \Sigma_1 \ne \Sigma_2$, with $p = 100$. As expected, the ℓ_2-type tests slightly work better than the maximum-type tests when Σ_1 is sparse and $\psi \ge 0.5$. Note that the performance of the maximum-type tests are also rapidly improved as the signal ψ gets larger. Somewhat surprisingly, when Σ_1 is dense and signals are moderate ($\psi \ge 1$), the maximum-type tests outperform the ℓ_2-type tests. We suspect that it is likely that the conditions for deriving the null distribution of Sch and LC are violated in the dense

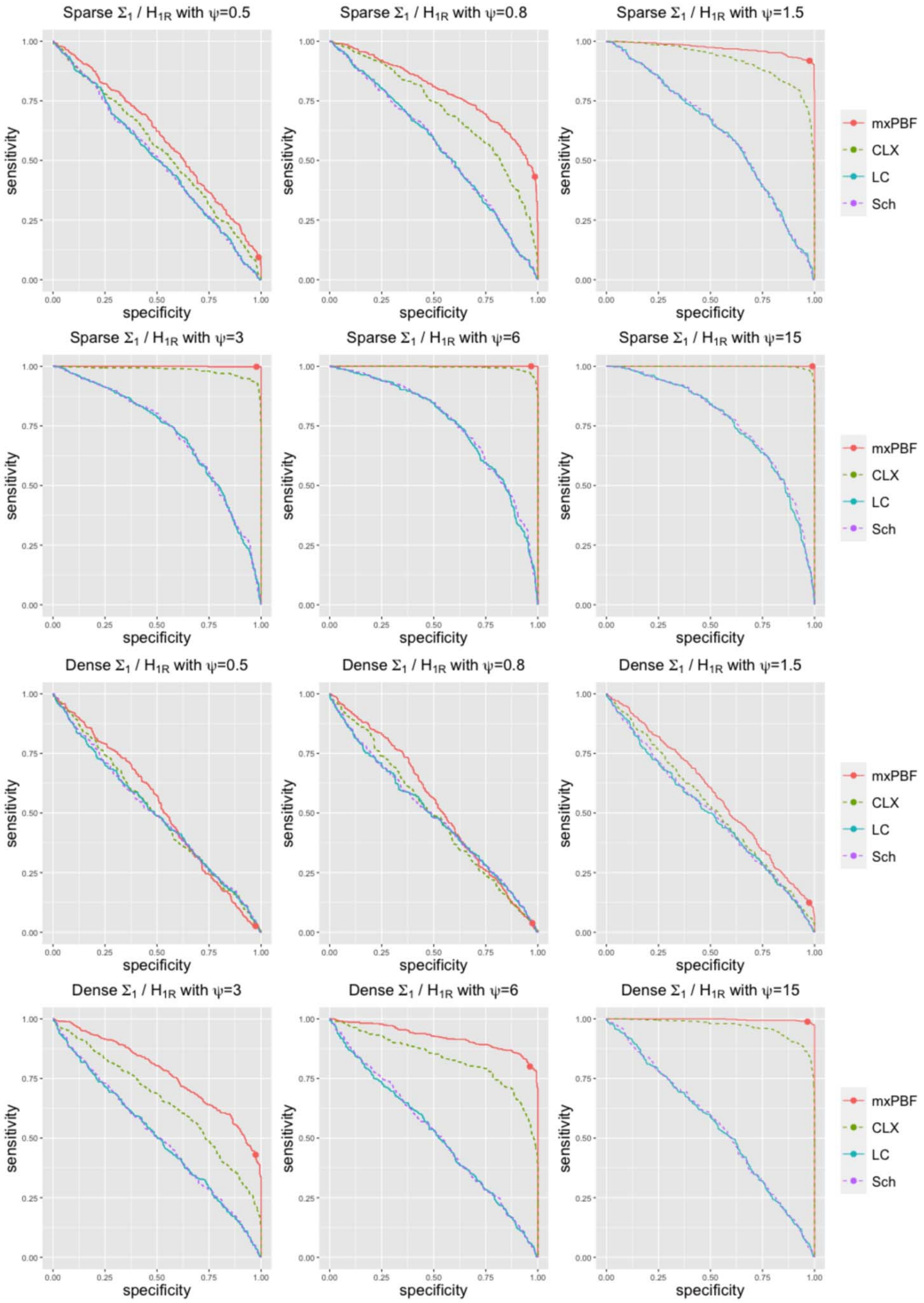

Figure 3: ROC curves for the two-sample covariance tests based on 500 simulated data sets for each hypothesis, H_0 and H_{1R}, with $p = 100$. The mxPBF, CLX, LC and Sch represent the test proposed in this paper, Cai et al. (2013), Li and Chen (2012) and Schott (2007), respectively. The dots show the results with $C_{\text{th}} = 10$ for the mxPBF.

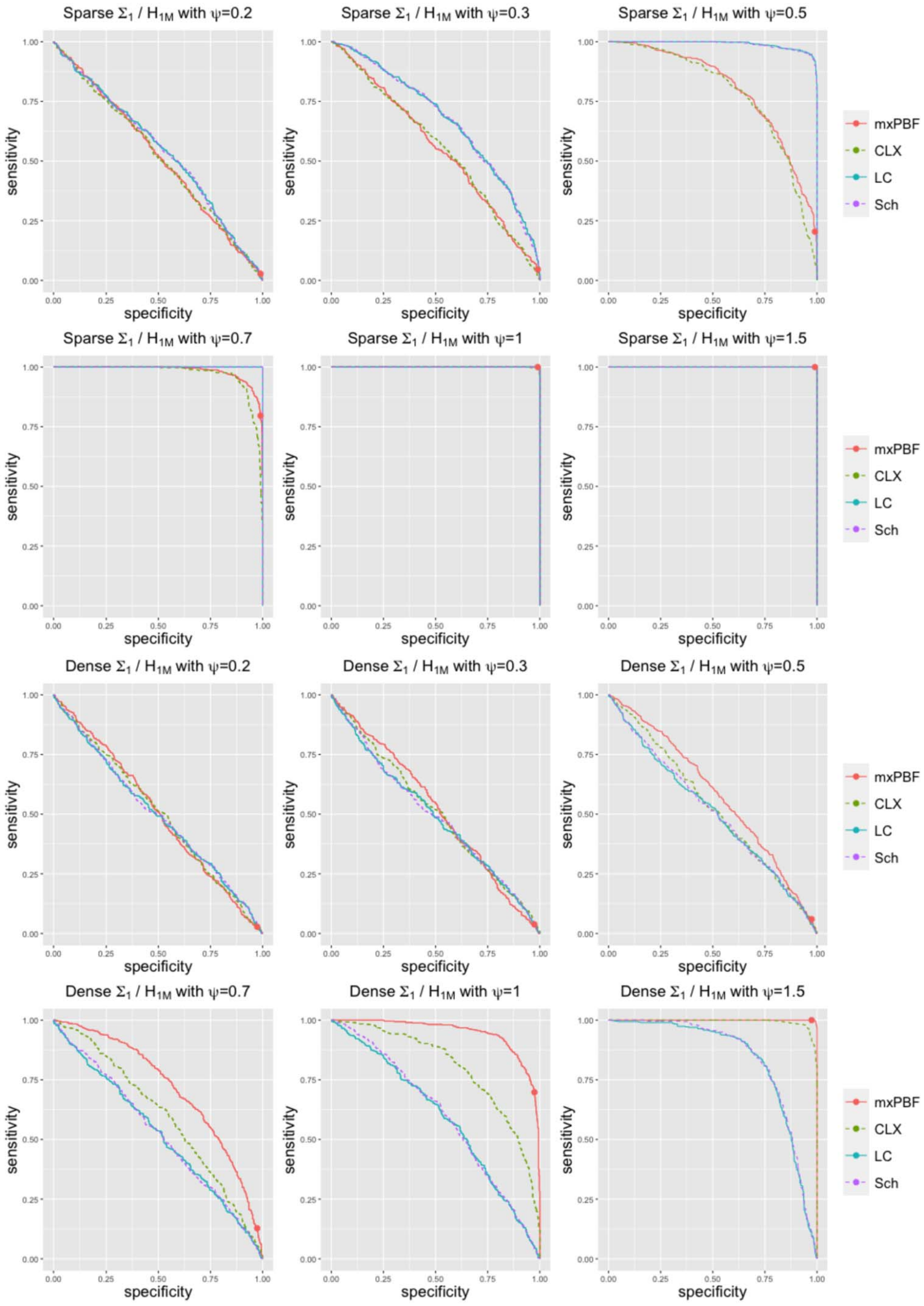

Figure 4: ROC curves for the two-sample covariance tests based on 500 simulated data sets for each hypothesis, H_0 and H_{1M}, with $p = 100$.

Σ_1 setting. Schott (2007) and Li and Chen (2012) assumed that $\lim_{p\to\infty} \mathrm{tr}(\Sigma^i)/p = \gamma_i \in (0,\infty)$ for $i = 1,\dots,8$ and $\mathrm{tr}(\Sigma^4) = o\{\mathrm{tr}(\Sigma^2)^2\}$, respectively, to derive the null distribution. In our settings, we find that $\mathrm{tr}(\Sigma^i)/p$ and $\mathrm{tr}(\Sigma^4)/\mathrm{tr}(\Sigma^2)^2$ are much larger in the dense Σ_1 setting than in the sparse Σ_1 setting. This partially supports our conjecture, although more rigorous investigation might be needed to determine the exact cause.

The dots in Figures 3 and 4 show the results for the mxPBF with $C_{\mathrm{th}} = 10$ and the FPR-based chosen α. The mxPBF with this default choice seems to work well if there is a reasonable amount of signals.

Lastly, we note that the experiment for $p = 300$ showed similar phenomena whose results including ROC curves and descriptions are deferred to the Supplementary material due to lack of space.

4.4 Real data analysis

In this section, we apply the proposed two-sample mean and covariance tests to two real datasets, small round blue cell tumors (SRBCT) dataset and prostate cancer dataset, respectively. For both datasets, the sample sizes are quite small compared to the number of variables. Thus, based on this numerical study, we would like to illustrate the practical performance of mxPBF-based tests in "small n large p" situations.

We first apply two-sample mean tests to the SRBCT dataset. The SRBCT dataset is available in the `R` package `plsgenomics`. This is a gene expression data having 83 samples with 2308 genes ($p = 2308$) from the microarray experiments in (Khan et al., 2001). Among 83 samples, we focus on 11 cases of Burkitt lymphoma (BL) ($n_1 = 11$) and 18 cases of neuroblastoma (NB) ($n_2 = 18$). Our main interest is to test equality of mean vectors of the gene expressions between BL and NB tumors. We apply the mxPBF, CLX.AT, SD and BS to test equality of mean vectors. Note that CLX.AT is used because the lack of prior information about the sparsity of the covariance matrix. For this dataset, the value of the mxPBF is greater than 10^8, and p-values of CLX.AT, SD and BS are less than 10^{-15}. Therefore, all the tests reject the null hypothesis, $H_0 : \mu_1 = \mu_2$, if we use the default choices, threshold $C_{\mathrm{th}} = 10$ for the mxPBF and significance level 0.05 for the frequentist tests.

The prostate cancer dataset is available in the `R` package `SIS`. This dataset contains 12600 gene expressions from 52 patients with prostate tumors ($n_1 = 52$) and 50 patients with normal prostate ($n_2 = 50$). As suggested by Cai et al. (2013), 5000 genes ($p = 5000$) with the largest absolute values of the t statistics are selected. Data were centered prior to analysis. In this dataset, we would like to test equality of covariance matrices of the gene expressions between tumor and normal samples. We apply the mxPBF, CLX, LC and Sch to test equality of covariance matrices. For this dataset, the value of the mxPBF is greater than 10^{32}, and p-values of CLX, LC and Sch are less than 0.0058, 10^{-15} and 10^{-15}, respectively. Therefore, all the tests reject the null hypothesis, $H_0 : \Sigma_1 = \Sigma_2$, if we use the default choices, threshold $C_{\mathrm{th}} = 10$ for the mxPBF and significance level 0.05 for the frequentist tests.

5 Discussion

In this paper, we propose a Bayesian two-sample mean test and a Bayesian two-sample covariance test in high-dimensional settings based on the idea of the maximum pairwise Bayes factor (Lee et al., 2021). These tests are not only computationally scalable but also enjoy Bayes factor consistency under relatively weak or similar conditions compared to existing tests. The proposed methods can be applied to change point detection for mean vectors or covariance matrices, which is indeed one of our ongoing works. Note that from the first data point, using only a subset of data within a certain window, a two-sample test can be sequentially conducted to detect change points. Due to consistency of the proposed mxPBF-based two-sample tests, it is expected that the resulting change point detection procedures can consistently detect and estimate change points.

As pointed out by a referee, the proposed tests can be applied when there are two independent datasets. Note that the proposed priors for the two-sample mean and covariance tests are data-dependent. Thus, if we have two independent datasets, the first dataset can be used to train the prior, while the second dataset can be used in the likelihood. In that case, the explicit forms of the resulting mxPBFs will be changed accordingly, but we can still proceed with the tests based on them. It might be worthwhile to investigate whether the consistency results in Sections 2 and 3 are still valid in this context, possibly with some additional assumptions.

Supplementary Material

Supplementary Material for “Bayesian Optimal Two-Sample Tests for High-Dimensional Gaussian Populations” (DOI: 10.1214/23-BA1373SUPP; .pdf).

References

Bai, Z. and Saranadasa, H. (1996). “Effect of high dimension: by an example of a two sample problem.” *Statistica Sinica*, 311–329. MR1399305. 870, 883, 884, 886

Cai, T. and Liu, W. (2011). “Adaptive thresholding for sparse covariance matrix estimation.” *Journal of the American Statistical Association*, 106(494): 672–684. MR2847949. doi: https://doi.org/10.1198/jasa.2011.tm10560. 870

Cai, T., Liu, W., and Luo, X. (2011). “A constrained ℓ_1 minimization approach to sparse precision matrix estimation.” *Journal of the American Statistical Association*, 106(494): 594–607. MR2847973. doi: https://doi.org/10.1198/jasa.2011.tm10155. 870

Cai, T., Liu, W., and Xia, Y. (2013). “Two-sample covariance matrix testing and support recovery in high-dimensional and sparse settings.” *Journal of the American Statistical Association*, 108(501): 265–277. MR3174618. doi: https://doi.org/10.1080/01621459.2012.758041. 871, 872, 879, 880, 887, 888, 890

Cai, T. T., Liu, W., and Xia, Y. (2014). “Two-sample test of high dimensional means under dependence.” *Journal of the Royal Statistical Society: Series B (Statistical*

Methodology), 76(2): 349–372. MR3164870. doi: https://doi.org/10.1111/rssb.12034. 870, 871, 872, 875, 883, 884

Cao, Y., Lin, W., and Li, H. (2018). "Two-sample tests of high-dimensional means for compositional data." *Biometrika*, 105(1): 115–132. MR3768869. doi: https://doi.org/10.1093/biomet/asx060. 870

Gregory, K. B., Carroll, R. J., Baladandayuthapani, V., and Lahiri, S. N. (2015). "A two-sample test for equality of means in high dimension." *Journal of the American Statistical Association*, 110(510): 837–849. MR3367268. doi: https://doi.org/10.1080/01621459.2014.934826. 870

He, J. and Chen, S. X. (2018). "High-dimensional two-sample covariance matrix testing via super-diagonals." *Statistica Sinica*, 28: 2671–2696. MR3839879. 871

Jeffreys, H. (1998). *The theory of probability*. OUP Oxford. MR1647885. 883

Kass, R. E. and Raftery, A. E. (1995). "Bayes factors." *Journal of the American Statistical Association*, 90(430): 773–795. MR3363402. doi: https://doi.org/10.1080/01621459.1995.10476572. 883

Khan, J., Wei, J. S., Ringner, M., Saal, L. H., Ladanyi, M., Westermann, F., Berthold, F., Schwab, M., Antonescu, C. R., Peterson, C., et al. (2001). "Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks." *Nature Medicine*, 7(6): 673–679. 890

Lee, K., Lin, L., and Dunson, D. (2021). "Maximum pairwise Bayes factors for covariance structure testing." *Electronic Journal of Statistics*, 15(2): 4384 – 4419. MR4312207. doi: https://doi.org/10.1214/21-ejs1900. 871, 873, 876, 891

Lee, K., You, K., and Lin, L. (2023). "Supplementary Material for "Bayesian Optimal Two-Sample Tests for High-Dimensional Gaussian Populations"." *Bayesian Analysis*. doi: https://doi.org/10.1214/23-BA1373SUPP. 872

Li, J. and Chen, S. X. (2012). "Two sample tests for high-dimensional covariance matrices." *The Annals of Statistics*, 40(2): 908–940. MR2985938. doi: https://doi.org/10.1214/12-AOS993. 871, 887, 888, 890

Schott, J. R. (2007). "A test for the equality of covariance matrices when the dimension is large relative to the sample sizes." *Computational Statistics & Data Analysis*, 51(12): 6535–6542. MR2408613. doi: https://doi.org/10.1016/j.csda.2007.03.004. 870, 887, 888, 890

Shen, Y., Lin, Z., and Zhu, J. (2011). "Shrinkage-based regularization tests for high-dimensional data with application to gene set analysis." *Computational Statistics & Data Analysis*, 55(7): 2221–2233. MR2786983. doi: https://doi.org/10.1016/j.csda.2010.12.013. 870

Srivastava, M. S. and Du, M. (2008). "A test for the mean vector with fewer observations than the dimension." *Journal of Multivariate Analysis*, 99(3): 386–402. MR2396970. doi: https://doi.org/10.1016/j.jmva.2006.11.002. 870, 883, 884, 886

Srivastava, M. S. and Yanagihara, H. (2010). "Testing the equality of several covariance matrices with fewer observations than the dimension." *Journal of Multivariate Analysis*, 101(6): 1319–1329. MR2609494. doi: https://doi.org/10.1016/j.jmva.2009.12.010. 871

Tsai, C.-A. and Chen, J. J. (2009). "Multivariate analysis of variance test for gene set analysis." *Bioinformatics*, 25(7): 897–903. 870

Wang, W., Lin, N., and Tang, X. (2019). "Robust two-sample test of high-dimensional mean vectors under dependence." *Journal of Multivariate Analysis*, 169: 312–329. MR3875602. doi: https://doi.org/10.1016/j.jmva.2018.09.013. 870

Xu, G., Lin, L., Wei, P., and Pan, W. (2016). "An adaptive two-sample test for high-dimensional means." *Biometrika*, 103(3): 609–624. MR3551787. doi: https://doi.org/10.1093/biomet/asw029. 870

Zheng, S., Lin, R., Guo, J., and Yin, G. (2017). "Testing homogeneity of high-dimensional covariance matrices." *Statistica Sinica*. Accepted. MR4285484. doi: https://doi.org/10.5705/ss.202017.0275. 871

Zoh, R. S., Sarkar, A., Carroll, R. J., and Mallick, B. K. (2018). "A Powerful Bayesian Test for Equality of Means in High Dimensions." *Journal of the American Statistical Association*, 113(524): 1733–1741. MR3902242. doi: https://doi.org/10.1080/01621459.2017.1371024. 871, 872, 873, 874, 875

Acknowledgments

We are very grateful to the Editor, Associate Editor and two reviewers for their valuable comments which have led to significant improvement in the earlier version of our paper.