A More Powerful Two-Sample Test in High Dimensions using Random Projection

Miles E. Lopes[1] **Laurent Jacob**[1] **Martin J. Wainwright**[1,2]
Departments of Statistics[1] and EECS[2]
University of California, Berkeley
Berkeley, CA 94720-3860
{mlopes,laurent,wainwrig}@stat.berkeley.edu

Abstract

We consider the hypothesis testing problem of detecting a shift between the means of two multivariate normal distributions in the high-dimensional setting, allowing for the data dimension p to exceed the sample size n. Our contribution is a new test statistic for the two-sample test of means that integrates a random projection with the classical Hotelling T^2 statistic. Working within a high-dimensional framework that allows $(p, n) \to \infty$, we first derive an asymptotic power function for our test, and then provide sufficient conditions for it to achieve greater power than other state-of-the-art tests. Using ROC curves generated from simulated data, we demonstrate superior performance against competing tests in the parameter regimes anticipated by our theoretical results. Lastly, we illustrate an advantage of our procedure with comparisons on a high-dimensional gene expression dataset involving the discrimination of different types of cancer.

1 Introduction

Two-sample hypothesis tests are concerned with the question of whether two samples of data are generated from the same distribution. Such tests are among the most widely used inference procedures in treatment-control studies in science and engineering [1]. Application domains such as molecular biology and fMRI have stimulated considerable interest in detecting shifts between distributions in the high-dimensional setting, where the two samples of data $\{X_1, \dots, X_{n_1}\}$ and $\{Y_1, \dots, Y_{n_2}\}$ are subsets of $\mathbb{R}^p$, and $n_1, n_2 \ll p$ [e.g., 2–5]. In transcriptomics, for instance, p gene expression measures on the order of hundreds or thousands may be used to investigate differences between two biological conditions, and it is often difficult to obtain sample sizes n_1 and n_2 larger than several dozen in each condition. In high-dimensional situations such as these, classical methods may be ineffective, or not applicable at all. Likewise, there has been growing interest in developing testing procedures that are better suited to deal with the effects of dimension [*e.g.*, 6–10].

A fundamental instance of the general two-sample problem is the *two-sample test of means* with Gaussian data. In this case, two independent sets of samples $\{X_1, \dots, X_{n_1}\}$ and $\{Y_1, \dots, Y_{n_2}\}$ are generated in an i.i.d. manner from p-dimensional multivariate normal distributions $N(\mu_1, \Sigma)$ and $N(\mu_2, \Sigma)$ respectively, where the mean vectors $\mu_1, \mu_2 \in \mathbb{R}^p$ and covariance matrix $\Sigma \succ 0$ are all fixed and unknown. The hypothesis testing problem of interest is

$$\mathbf{H}_0 : \mu_1 = \mu_2 \;\text{ versus }\; \mathbf{H}_1 : \mu_1 \neq \mu_2. \tag{1}$$

The most well-known test statistic for this problem is the Hotelling T^2 statistic, defined by

$$T^2 := \frac{n_1 n_2}{n_1 + n_2} (\bar{X} - \bar{Y})^\top \widehat{\Sigma}^{-1} (\bar{X} - \bar{Y}), \tag{2}$$

where $\bar{X} := \frac{1}{n_1}\sum_{j=1}^{n_1} X_j$ and $\bar{Y} := \frac{1}{n_2}\sum_{j=1}^{n_2} Y_j$ are the sample means, and $\widehat{\Sigma}$ is the pooled sample covariance matrix, given by $\widehat{\Sigma} := \frac{1}{n}\sum_{j=1}^{n_1}(X_j - \bar{X})(X_j - \bar{X})^\top + \frac{1}{n}\sum_{j=1}^{n_2}(Y_j - \bar{Y})(Y_j - \bar{Y})^\top$, with $n := n_1 + n_2 - 2$.

When $p > n$, the matrix $\widehat{\Sigma}$ is singular, and the Hotelling test is not well-defined. Even when $p \le n$, the Hotelling test is known to perform poorly if p is nearly as large as n. This behavior was demonstrated in a seminal paper of Bai and Saranadasa [6] (or BS for short), who studied the performance of the Hotelling test under $(p, n) \to \infty$ with $p/n \to 1 - \epsilon$, and showed that the asymptotic power of the test suffers for small values of $\epsilon > 0$. In subsequent years, a number of improvements on the Hotelling test in the high-dimensional setting have been proposed [e.g., 6–9].

In this paper, we propose a new test statistic for the two-sample test of means with multivariate normal data, applicable when $p \ge n/2$. We provide an explicit asymptotic power function for our test with $(p, n) \to \infty$, and show that under certain conditions, our test has greater asymptotic power than other state-of-the-art tests. These comparison results are valid with p/n tending to a positive constant or infinity. In addition to its advantage in terms of asymptotic power, our procedure specifies *exact* level-α critical values for multivariate normal data, whereas competing procedures offer only approximate level-α critical values. Furthermore, our experiments in Section 4 suggest that the critical values of our test may also be more robust than those of competing tests. Lastly, the computational cost of our procedure is modest in the $n < p$ setting, being of order $\mathcal{O}(n^2 p)$.

The remainder of this paper is organized as follows. In Section 2, we provide background on hypothesis testing and describe our testing procedure. Section 3 is devoted to a number of theoretical results about its performance. Theorem 1 in Section 3.1 provides an asymptotic power function, and Theorems 2 and 3 in Sections 3.3 and 3.4 give sufficient conditions for achieving greater power than state-of-the-art tests in the sense of asymptotic relative efficiency. In Section 4 we provide performance comparisons with ROC curves on synthetic data against a broader collection of methods, including some recent kernel-based and non-parametric approaches such as MMD [11], KFDA [12], and TreeRank [10]. Lastly, we study a high-dimensional gene expression dataset involving the discrimination of different cancer types, demonstrating that our test's false positive rate is reliable in practice. We refer the reader to the preprint [13] for proofs of our theoretical results.

Notation. Let $\delta := \mu_1 - \mu_2$ denote the *shift vector* between the distributions $N(\mu_1, \Sigma)$ and $N(\mu_2, \Sigma)$, and define the ordered pair of parameters $\theta := (\delta, \Sigma)$. Let $z_{1-\alpha}$ denote the $1 - \alpha$ quantile of the standard normal distribution, and let Φ be its cumulative distribution function. If A is a matrix in $\mathbb{R}^{p \times p}$, let $|\!|\!| A |\!|\!|_2$ denote its spectral norm (maximum singular value), and define the Frobenius norm $|\!|\!| A |\!|\!|_F := \sqrt{\sum_{i,j} A_{ij}^2}$. When all the eigenvalues of A are real, we denote them by $\lambda_{\min}(A) = \lambda_p(A) \le \cdots \le \lambda_1(A) = \lambda_{\max}(A)$. For a positive-definite covariance matrix Σ, let $D_\sigma := \operatorname{diag}(\Sigma)$, and define the associated correlation matrix $R := D_\sigma^{-1/2} \Sigma D_\sigma^{-1/2}$. We use the notation $f(n) \lesssim g(n)$ if there is some absolute constant c such that the inequality $f(n) \le c\, n$ holds for all large n. If both $f(n) \lesssim g(n)$ and $g(n) \lesssim f(n)$ hold, then we write $f(n) \asymp g(n)$. The notation $f(n) = o(g(n))$ means $f(n)/g(n) \to 0$ as $n \to \infty$.

2 Background and random projection method

For the remainder of the paper, we retain the set-up for the two-sample test of means (1) with Gaussian data, assuming throughout that $p \ge n/2$, and $n = n_1 + n_2 - 2$.

Review of hypothesis testing terminology. The primary focus of our results will be on the comparison of *power* between test statistics, and here we give precise meaning to this notion. When testing a null hypothesis $\mathbf{H}_0$ versus an alternative hypothesis $\mathbf{H}_1$, a procedure based on a test statistic T specifies a *critical value*, such that $\mathbf{H}_0$ is rejected if T exceeds that critical value, and $\mathbf{H}_0$ is accepted otherwise. The chosen critical value fixes a trade-off between the risk of rejecting $\mathbf{H}_0$ when $\mathbf{H}_0$ actually holds, and the risk of accepting $\mathbf{H}_0$ when $\mathbf{H}_1$ holds. The former error is referred to as a type I error and the latter as a type II error. A test is said to have level α if the probability of committing a type I error is at most α. Finally, at a given level α, the *power* of a test is the probability of rejecting $\mathbf{H}_0$ under $\mathbf{H}_1$, *i.e.*, 1 minus the probability of a type II error. When evaluating testing procedures at a given level α, we seek to identify the one with the greatest power.

Past work. The Hotelling T^2 statistic (2) discriminates between the hypotheses $\mathbf{H}_0$ and $\mathbf{H}_1$ by providing an estimate of the "statistical distance" separating the distributions $N(\mu_1, \Sigma)$ and $N(\mu_2, \Sigma)$. More specifically, the Hotelling statistic is essentially an estimate of the Kullback-Leibler (KL) divergence $D_{\text{KL}}\big(N(\mu_1, \Sigma)\|N(\mu_2, \Sigma)\big) = \frac{1}{2}\delta^\top\Sigma^{-1}\delta$, where $\delta := \mu_1 - \mu_2$. Due to the fact that the pooled sample covariance matrix $\widehat{\Sigma}$ in the definition of T^2 is not invertible when $p > n$, several recent procedures have offered substitutes for the Hotelling statistic in the high-dimensional setting: Bai and Saranadasa [6], Srivastava and Du [7, 8], Chen and Qin [9], hereafter BS, SD and CQ respectively. Up to now, the route toward circumventing this difficulty has been to form an estimate of Σ that is diagonal, and hence easily invertible. We shall see later that this limited use of covariance structure sacrifices power when the data exhibit non-trivial correlation. In this regard, our procedure is motivated by the idea that covariance structure may be used more effectively by testing with projected samples in a space of lower dimension.

Intuition for random projection. To provide some further intuition for our method, it is possible to consider the problem (1) in terms of a competition between the dimension p, and the statistical distance separating $\mathbf{H}_0$ and $\mathbf{H}_1$. On one hand, the accumulation of variance from a large number of variables makes it difficult to discriminate between the hypotheses, and thus, it is desirable to reduce the dimension of the data. On the other hand, most methods for reducing dimension will also bring $\mathbf{H}_0$ and $\mathbf{H}_1$ "closer together," making them harder to distinguish. Mindful of the fact that the Hotelling test measures the separation of $\mathbf{H}_0$ and $\mathbf{H}_1$ in terms of $\delta^\top\Sigma^{-1}\delta$, we see that the statistical distance is driven by the Euclidean length of δ. Consequently, we seek to transform the data in such a way that the dimension is reduced, while the length of the shift δ is mostly preserved upon passing to the transformed distributions. From this geometric point of view, it is natural to exploit the fact that random projections can simultaneously reduce dimension and approximately preserve lengths with high probability [14]. The use of projection-based test statistics has been considered previously in Jacob et al., [15], Clémençon et al. [10], and Cuesta-Albertos et al. [16].

At a high level, our method can be viewed as a two step procedure. First, a single random projection is drawn, and is used to map the samples from the high-dimensional space $\mathbb{R}^p$ to a low-dimensional space[1] $\mathbb{R}^k$, with $k := \lfloor n/2 \rfloor$. Second, the Hotelling T^2 test is applied to a new hypothesis testing problem, $\mathbf{H}_{0,\text{proj}}$ versus $\mathbf{H}_{1,\text{proj}}$, in the projected space. A decision is then pulled back to the original problem by simply rejecting $\mathbf{H}_0$ whenever the Hotelling test rejects $\mathbf{H}_{0,\text{proj}}$.

Formal testing procedure. Let $P_k^\top \in \mathbb{R}^{k\times p}$ denote a random projection with i.i.d. $N(0,1)$ entries, drawn independently of the data, where $k = \lfloor n/2 \rfloor$. Conditioning on the drawn matrix $P_k^\top$, the projected samples $\{P_k^\top X_1, \ldots, P_k^\top X_{n_1}\}$ and $\{P_k^\top Y_1, \ldots, P_k^\top Y_{n_2}\}$ are distributed i.i.d. according to $N(P_k^\top \mu_i, P_k^\top \Sigma P_k)$ respectively, with $i = 1, 2$. Since $n \geq k$, the projected data satisfy the usual conditions [17, p. 211] for applying the Hotelling T^2 procedure to the following new two-sample problem in the projected space $\mathbb{R}^k$:

$$\mathbf{H}_{0,\text{proj}} : P_k^\top \mu_1 = P_k^\top \mu_2 \;\text{ versus }\; \mathbf{H}_{1,\text{proj}} : P_k^\top \mu_1 \neq P_k^\top \mu_2. \tag{3}$$

For this projected problem, the Hotelling test statistic takes the form[2]

$$T_k^2 := \tfrac{n_1 n_2}{n_1+n_2}(\bar{X} - \bar{Y})^\top P_k (P_k^\top \widehat{\Sigma} P_k)^{-1} P_k^\top (\bar{X} - \bar{Y}),$$

where $\bar{X}$, $\bar{Y}$, and $\widehat{\Sigma}$ are as defined in Section 1. Lastly, define the critical value $t_\alpha := \frac{k\,n}{n-k+1} F^*_{k,n-k+1}(\alpha)$, where $F^*_{k,n-k+1}(\alpha)$ is the upper α quantile of the $F_{k,n-k+1}$ distribution [17].

It is a basic fact about the classical Hotelling test that rejecting $\mathbf{H}_{0,\text{proj}}$ when $T_k^2 \geq t_\alpha$ is a level-α test for the projected problem (3) (e.g., see Muirhead [17, p.217]). Inspection of the formula for T_k^2 shows that its distribution is the same under both $\mathbf{H}_0$ and $\mathbf{H}_{0,\text{proj}}$. Therefore, rejecting the original $\mathbf{H}_0$ when $T_k^2 \geq t_\alpha$ is also a level α test for the original problem (1). Likewise, we *define* this as the condition for rejecting $\mathbf{H}_0$ at level α in our procedure for (1). We summarize our procedure below.

[1] The choice of projected dimension $k = \lfloor n/2 \rfloor$ is explained in the preprint [13].

[2] Note that $P_k^\top \widehat{\Sigma} P_k$ is invertible with probability 1 when $P_k^\top$ has i.i.d. $N(0,1)$ entries.

1. Generate a single random matrix $P_k^\top$ with i.i.d. $N(0,1)$ entries.
2. Compute T_k^2, using $P_k^\top$ and the two sets of samples.
3. If $T_k^2 \geq t_\alpha$, reject $\mathbf{H}_0$; otherwise accept $\mathbf{H}_0$.

$(\star)$

Projected Hotelling test at level α for problem (1).

3 Main results and their consequences

This section is devoted to the statement and discussion of our main theoretical results, including a characterization of the asymptotic power function of our test (Theorem 1), and comparisons of asymptotic relative efficiency with state-of-the-art tests proposed in past work (Theorems 2 and 3).

3.1 Asymptotic power function

As is standard in high-dimensional asymptotics, we will consider a sequence of hypothesis testing problems indexed by n, allowing the dimension p, mean vectors μ_1 and μ_2 and covariance matrix Σ to implicitly vary as functions of n, with $n \to \infty$. We also make another type of asymptotic assumption, known as a *local alternative* [18, p.193], which is commonplace in hypothesis testing. The idea lying behind a local alternative assumption is that if the difficulty of discriminating between $\mathbf{H}_0$ and $\mathbf{H}_1$ is "held fixed" with respect to n, then it is often the case that most testing procedures have power tending to 1 under $\mathbf{H}_1$ as $n \to \infty$. In such a situation, it is not possible to tell if one test has greater asymptotic power than another. Consequently, it is standard to derive asymptotic power results under the extra condition that $\mathbf{H}_0$ and $\mathbf{H}_1$ become harder to distinguish as n grows. This theoretical device aids in identifying the conditions under which one test is more powerful than another. The following local alternative (**A1**), and balancing assumption (**A2**), are similar to those used in previous works [6–9] on problem (1). In particular, condition (**A1**) means that the KL-divergence between $N(\mu_1, \Sigma)$ and $N(\mu_2, \Sigma)$ tends to 0 as $n \to \infty$.

(**A1**) Suppose that $\delta^\top \Sigma^{-1} \delta = o(1)$.
(**A2**) Let there be a constant $b \in (0,1)$ such that $n_1/n \to b$.

To set the notation for Theorem 1, it is important to notice that each time the procedure $(\star)$ is implemented, a draw of $P_k^\top$ induces a new test statistic T_k^2. To make this dependence clear, recall $\theta := (\delta, \Sigma)$, and let $\beta(\theta; P_k^\top)$ denote the exact (non-asymptotic) power function of our level-α test for problem (1), induced by a draw of $P_k^\top$, as in $(\star)$. Another key quantity that depends on $P_k^\top$ is the KL-divergence between the projected sampling distributions $N(P_k^\top \mu_1, P_k^\top \Sigma P_k)$ and $N(P_k^\top \mu_2, P_k^\top \Sigma P_k)$. We denote this divergence by $\frac{1}{2}\Delta_k^2$, and a simple calculation shows that $\frac{1}{2}\Delta_k^2 = \frac{1}{2}\delta^\top P_k (P_k^\top \Sigma P_k)^{-1} P_k^\top \delta$.

Theorem 1. *Under conditions (**A1**) and (**A2**), for almost all sequences of projections $P_k^\top$,*

$$\beta(\theta; P_k^\top) - \Phi\left(-z_{1-\alpha} + \tfrac{b(1-b)}{\sqrt{2}}\, \sqrt{n}\, \Delta_k^2\right) \to 0 \;\; \textit{as} \;\; n \to \infty. \tag{4}$$

Remarks. Note that if $\Delta_k^2 = 0$, e.g. under $\mathbf{H}_0$, then $\Phi(-z_{1-\alpha} + 0) = \alpha$, which corresponds to blind guessing at level α. Consequently, the second term $(b(1-b)/\sqrt{2})\, \sqrt{n}\Delta_k^2$ determines the advantage of our procedure over blind guessing. Since Δ_k^2 is proportional to the KL-divergence between the projected sampling distributions, these observations conform to the intuition from Section 2 that the KL-divergence measures the discrepancy between $\mathbf{H}_0$ and $\mathbf{H}_1$.

3.2 Asymptotic relative efficiency (ARE)

Having derived an asymptotic power function for our test in Theorem 1, we are now in position to provide sufficient conditions for achieving greater power than two other recent procedures for problem (1): Srivastava and Du [7, 8] (SD), and Chen and Qin [9] (CQ). To the best of our knowledge, these works represent the state of the art[3] among tests for problem (1) with a known asymptotic power function under $(p, n) \to \infty$.

From Theorem 1, the asymptotic power function of our random projection-based test at level α is

$$\beta_{\text{RP}}(\theta; P_k^\top) := \Phi\left(-z_{1-\alpha} + \tfrac{b(1-b)}{\sqrt{2}}\sqrt{n}\,\Delta_k^2\right). \tag{5}$$

The asymptotic power functions for the CQ and SD testing procedures at level α are

$$\beta_{\text{CQ}}(\theta) := \Phi\left(-z_{1-\alpha} + \tfrac{b(1-b)}{\sqrt{2}}\tfrac{n\,\|\delta\|_2^2}{|\!|\!|\Sigma|\!|\!|_F}\right), \quad \text{and} \quad \beta_{\text{SD}}(\theta) := \Phi\left(-z_{1-\alpha} + \tfrac{b(1-b)}{\sqrt{2}}\tfrac{n\,\delta^\top D_\sigma^{-1}\delta}{|\!|\!|R|\!|\!|_F}\right).$$

Recall that $D_\sigma := \text{diag}(\Sigma)$, and R denotes the correlation matrix associated with Σ. The functions β_{CQ} and β_{SD} are derived under local alternatives and asymptotic assumptions that are similar to the ones used here to obtain β_{RP}. In particular, all three functions can be obtained allowing p/n to tend to an arbitrary positive constant or infinity.

A standard method of comparing asymptotic power functions under local alternatives is through the concept of *asymptotic relative efficiency* (ARE) *e.g.*, see van der Vaart [18, p.192]). Since Φ is monotone increasing, the term added to $-z_{1-\alpha}$ inside the Φ functions above controls the power. To compare power between tests, the ARE is simply defined via the ratio of such terms. More explicitly, we define $\text{ARE}\,(\beta_{\text{CQ}}; \beta_{\text{RP}}) := \left(\frac{n\,\|\delta\|_2^2}{|\!|\!|\Sigma|\!|\!|_F}\big/\sqrt{n}\Delta_k^2\right)^2$, and $\text{ARE}\,(\beta_{\text{SD}}; \beta_{\text{RP}}) := \left(\frac{n\,\delta^\top D_\sigma^{-1}\delta}{|\!|\!|R|\!|\!|_F}\big/\sqrt{n}\,\Delta_k^2\right)^2$.

Whenever the ARE is less than 1, our procedure is considered to have greater asymptotic power than the competing test—with our advantage being greater for smaller values of the ARE. Consequently, we seek sufficient conditions in Theorems 2 and 3 for ensuring that the ARE is small.

In the present context, the analysis of ARE is complicated by the fact that the ARE varies with n and depends on a random draw of $P_k^\top$ through Δ_k^2. Moreover, the quantity Δ_k^2, and hence the ARE, are affected by the orientation of δ with respect to the eigenvectors of Σ. In order to consider an average-case scenario, where no single orientation of δ is of particular importance, we place a prior on the unit vector $\delta/\|\delta\|_2$, and assume that it is uniformly distributed on the unit sphere of $\mathbb{R}^p$. We emphasize that our procedure ($\star$) does not rely on this assumption, and that it is only a device for making an average-case comparison. Therefore, to be clear about the meaning of Theorems 2 and 3, we regard the ARE as a function two random objects, $P_k^\top$ and $\delta/\|\delta\|_2$, and our probability statements are made with this understanding. We complete the preparation for our comparison theorems by isolating four assumptions with $n \to \infty$.

(**A3**) The vector $\frac{\delta}{\|\delta\|_2}$ is uniformly distributed on the p-dimensional unit sphere, independent of $P_k^\top$.
(**A4**) There is a constant $a \in [0, 1)$ such that $k/p \to a$.
(**A5**) The ratio $\frac{1}{\sqrt{k}}\ \text{tr}(\Sigma)/(p\,\lambda_{\text{min}}(\Sigma)) = o(1)$.
(**A6**) The matrix $D_\sigma = \text{diag}(\Sigma)$ satisfies $\frac{|\!|\!|D_\sigma^{-1}|\!|\!|_2}{\text{tr}(D_\sigma^{-1})} = o(1)$.

3.3 Comparison with Chen and Qin [9]

The next result compares the asymptotic power of our projection-based test with that of Chen and Qin [9]. The choice of $\epsilon_1 = 1$ below (and in Theorem 3) is the reference for equal asymptotic performance, with smaller values of ϵ_1 corresponding to better performance of random projection.

Theorem 2. *Assume conditions (***A3***), (***A4***), and (***A5***). Fix a number $\epsilon_1 > 0$, and let $c(\epsilon_1)$ be any constant strictly greater than $\frac{4}{\epsilon_1\,(1-\sqrt{a})^4}$. If the inequality*

$$n \geq c(\epsilon_1)\,\frac{\text{tr}(\Sigma)^2}{|\!|\!|\Sigma|\!|\!|_F^2} \tag{6}$$

holds for all large n, then $\mathbb{P}\left[\text{ARE}\,(\beta_{\text{CQ}}; \beta_{\text{RP}}) \leq \epsilon_1\right] \to 1$ as $n \to \infty$.

Interpretation. To interpret the result, note that Jensen's inequality implies that for any choice of Σ, we have $1 \leq \text{tr}(\Sigma)^2/\,|\!|\!|\Sigma|\!|\!|_F^2 \leq p$. As such, it is reasonable to interpret this ratio as a measure of

[3] Two other high-dimensional tests have been proposed in older works [6, 19, 20] that lead to the asymptotic power function β_{CQ}, but under more restrictive assumptions.

the *effective dimension* of the covariance structure. The message of Theorem 2 is that as long as the sample size n exceeds the effective dimension, then our projection-based test is asymptotically superior to CQ. The ratio $\mathrm{tr}(\Sigma)^2 / |\!|\!|\Sigma|\!|\!|_F^2$ can also be viewed as measuring the *decay rate* of the spectrum of Σ, with $\mathrm{tr}(\Sigma)^2 / |\!|\!|\Sigma|\!|\!|_F^2 \ll p$ indicating rapid decay. This condition means that the data has low variance in "most" directions in $\mathbb{R}^p$, and so projecting onto a random set of k directions will likely map the data into a low-variance subspace in which it is harder for chance variation to explain away the correct hypothesis, thereby resulting in greater power.

3.4 Comparison with Srivastava and Du [7, 8]

We now turn to comparison of asymptotic power with the test of Srivastava and Du (SD).

Theorem 3. *In addition to the conditions of Theorem 2, assume that condition (**A6**) holds. Fix a number $\epsilon_1 > 0$, and let $c(\epsilon_1)$ be any constant strictly greater than $\frac{4}{\epsilon_1 \, (1-\sqrt{a})^4}$. If the inequality*

$$n \geq c(\epsilon_1) \left(\frac{\mathrm{tr}(\Sigma)}{p}\right)^2 \left(\frac{\mathrm{tr}(D_\sigma^{-1})}{|\!|\!|R|\!|\!|_F}\right)^2 \tag{7}$$

holds for all large large n, then $\mathbb{P}\left[\mathrm{ARE}\left(\beta_{\mathrm{SD}}; \beta_{\mathrm{RP}}\right) \leq \epsilon_1\right] \to 1$ as $n \to \infty$.

Interpretation. Unlike the comparison with the CQ test, the correlation matrix R plays a large role in determining the relative efficiency between our procedure and the SD test. The correlation matrix enters in two different ways. First, the Frobenius norm $|\!|\!|R|\!|\!|_F$ is larger when the data variables are more correlated. Second, correlation mitigates the growth of $\mathrm{tr}(D_\sigma^{-1})$, since this trace is largest when Σ is nearly diagonal and has a large number of small eigenvalues. Inspection of the SD test statistic in [7] shows that it does not make any essential use of correlation. By contrast, our T_k^2 statistic *does* take correlation into account, and so it is understandable that correlated data enhance the performance of our test relative to SD.

As a simple example, let $\rho \in (0,1)$ and consider a highly correlated situation where all variables have ρ correlation will all other variables. Then, $R = (1-\rho)I_{p\times p} + \rho\mathbf{1}\mathbf{1}^\top$ where $\mathbf{1} \in \mathbb{R}^p$ is the all ones vector. We may also let $\Sigma = R$ for simplicity. In this case, we see that $|\!|\!|R|\!|\!|_F^2 = p + 2\binom{p}{2}\rho^2 \gtrsim p^2$, and $\mathrm{tr}(D_\sigma^{-1})^2 = \mathrm{tr}(I_{p\times p})^2 = p^2$. This implies $\mathrm{tr}(D_\sigma^{-1})^2 / |\!|\!|R|\!|\!|_F^2 \lesssim 1$ and $\mathrm{tr}(\Sigma)/p = 1$, and then the sufficient condition (7) for outperforming SD is easily satisfied in terms of rates. We could even let the correlation ρ decay at a rate of n^{-q} with $q \in (0, 1/2)$, and (7) would still be satisfied for large enough n. More generally, it is not necessary to use specially constructed covariance matrices Σ to demonstrate the superior performance of our method. Section 4 illustrates simulations involving *randomly selected* covariance matrices where T_k^2 is more powerful than SD.

Conversely, it is possible to show that condition (7) *requires* non-trivial correlation. To see this, first note that in the complete absence of correlation, we have $|\!|\!|R|\!|\!|_F^2 = |\!|\!|I_{p\times p}|\!|\!|_F^2 = p$. Jensen's inequality implies that $\mathrm{tr}(D_\sigma^{-1}) \geq \frac{p^2}{\mathrm{tr}(D_\sigma)} = \frac{p^2}{\mathrm{tr}(\Sigma)}$, and so $\left(\frac{\mathrm{tr}(\Sigma)}{p}\right)^2 \left(\frac{\mathrm{tr}(D_\sigma^{-1})}{|\!|\!|R|\!|\!|_F}\right)^2 \geq p$. Altogether, this shows if the data exhibits very low correlation, then (7) cannot hold when p grows faster than n. This will be illustrated in the simulations of Section 4.

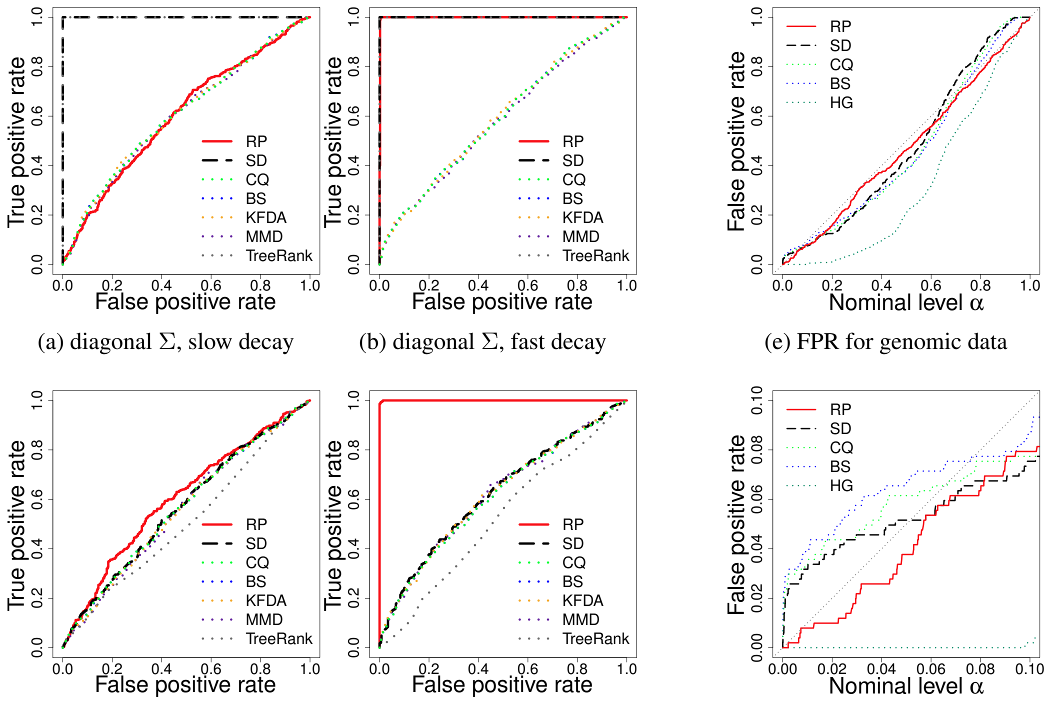
4 Performance comparisons on real and synthetic data

In this section, we compare our procedure to state-of-the-art methods on real and synthetic data, illustrating the effects of the different factors involved in Theorems 2 and 3.

Comparison on synthetic data. In order to validate the consequences of our theory and compare against other methods in a controlled fashion, we performed simulations in four settings: slow/fast spectrum decay, and diagonal/random covariance structure. To consider two rates of spectrum decay, we selected p equally spaced values between 0.01 and 1, and raised them to the power 20 for fast decay and the power 5 for slow decay. Random covariance structure was generated by specifying the eigenvectors of Σ as the column vectors of the orthogonal component of a QR decomposition of a $p \times p$ matrix with i.i.d. $N(0,1)$ entries. In all cases, we sampled $n_1 = n_2 = 50$ data points from two multivariate normal distributions in $p = 200$ dimensions, and repeated the process 500 times with $\delta = 0$ for $\mathbf{H}_0$, and 500 times with $\|\delta\|_2 = 1$ for $\mathbf{H}_1$. In the case of $\mathbf{H}_1$, δ was drawn uniformly from the unit sphere, as in Theorems 2 and 3. We fixed the total amount of variance by setting $|\!|\!|\Sigma|\!|\!|_F = 50$ in all cases. In addition to our random projection (RP)-based test, we implemented the methods of BS [6], SD [7], and CQ [9], all of which are designed specifically for problem (1) in the high-dimensional setting. For the sake of completeness, we also compare against recent non-parametric procedures for the general two-sample problem that are based on kernel methods (MMD) [11] and (KFDA) [12], as well as area-under-curve maximization (TreeRank) [10].

The ROC curves from our simulations are displayed in the left block of four panels in Figure 1. These curves bear out the results of Theorems 2 and 3 in several ways. First notice that fast spectral decay improves the performance of our test relative to CQ, as expected from Theorem 2. If we set $a = 0$ and $\epsilon_1 = 1$ in Theorem 2, then condition (6) for outperforming CQ is approximately $n \geq 75$ in the case of fast decay. Given that $n = 50 + 50 - 2 = 98$, the advantage of our method over CQ in panels (b) and (d) is consistent with condition (6) being satisfied. In the case of slow decay, the same settings of a and ϵ_1 indicate that $n \geq 246$ is sufficient for outperforming CQ. Since the ROC curve of our method is roughly the same as that of CQ in panels (a) and (c) (where again $n = 98$), our condition (6) is somewhat conservative for slow decay at the finite sample level.

To study the consequences of Theorem 3, observe that when the covariance matrix Σ is generated randomly, the amount of correlation is much larger than in the idealized case that Σ is diagonal. Specifically, for a fixed value of $\operatorname{tr}(\Sigma)$, the quantity $\operatorname{tr}(D_\sigma^{-1})/\,|\!|\!|R|\!|\!|_F$, is much smaller in in the presence of correlation. Consequently, when comparing (a) with (c), and (b) with (d), we see that correlation improves the performance of our test relative to SD, as expected from the bound in Theorem 3. More generally, the ROC curves illustrate that our method has an overall advantage over BS, CQ, KFDA, and MMD. Note that KFDA and MMD are not designed specifically for the $n \ll p$ regime. In the case of zero correlation, it is notable that the TreeRank procedure displays a superior ROC curve to our method, given that it also employs a dimension reduction strategy.



(a) diagonal Σ, slow decay

(b) diagonal Σ, fast decay

(e) FPR for genomic data

(c) random Σ, slow decay

(d) random Σ, fast decay

(f) FPR for genomic data (zoom)

Figure 1: Left and middle panels: ROC curves of several test statistics for two different choices of correlation structure and decay rate. (a) Diagonal covariance slow decay, (b) Diagonal covariance fast decay, (c) Random covariance slow decay, (d) Random covariance fast decay. Right panels: (e) False positive rate against p-value threshold on the gene expression experiment of Section 4 for RP ($\star$), BS, CQ, SD and enrichment test, (f) zoom on the p-value < 0.1 region.

Comparison on high-dimensional gene expression data. The ability to identify gene sets having different expression between two types of conditions, e.g., benign and malignant forms of a disease, is of great value in many areas of biomedical research. Likewise, there is considerable motivation to study our procedure in the context of detecting differential expression of p genes between two small groups of patients of sizes n_1 and n_2.

To compare the performance our T_k^2 statistic against competitors CQ and SD in this type of application, we constructed a collection of 1680 distinct two-sample problems in the following manner, using data from three genomic studies of ovarian [21], myeloma [22] and colorectal [23] cancers.

First, we randomly split the 3 datasets respectively into 6, 4, and 6 groups of approximately 50 patients. Next, we considered pairwise comparisons between all sets of patients on each of 14 biologically meaningful gene sets from the canonical pathways of MSigDB [24], with each gene set containing between 75 and 128 genes. Since $n_1 \simeq n_2 \simeq 50$ for all patient sets, our collection of two-sample problems is genuinely high-dimensional. Specifically, we have $14 \times (\binom{6}{2} + \binom{4}{2} + \binom{6}{2}) = 504$ problems under $\mathbf{H}_0$ and $14 \times (6 \cdot 4 + 6 \cdot 4 + 6 \cdot 6) = 1176$ problems under $\mathbf{H}_1$—assuming that every gene set was differentially expressed between two sets of patients with different cancers, and that no gene set was differentially expressed between two sets of patients with the same cancer type.[4]

A natural performance measure for comparing test statistics is the actual *false positive rate* (FPR) as a function of the nominal level α. When testing at level α, the actual FPR should be as close to α as possible, but differences may occur if the distribution of the test statistic under $\mathbf{H}_0$ is not known exactly (as is the case in practice). Figure 1 (e) shows that the curve for our procedure is closer to the optimal diagonal line for most values of α than the competing curves. Furthermore, the lower-left corner of Figure 1 (e) is of particular importance, as practitioners are usually only interested in p-values lower than 10^{-1}. Figure 1 (f) is a zoomed plot of this region and shows that the SD and CQ tests commit too many false positives at low thresholds. Again, in this regime, our procedure is closer to the diagonal and safely commits fewer than the allowed number of false positives. For example, when thresholding p-values at 0.01, SD has an actual FPR of 0.03, and an even more excessive FPR of 0.02 when thresholding at 0.001. The tests of CQ and BS are no better. The same thresholds on the p-values of our test lead to false positive rates of 0.008 and 0 respectively.

With consideration to ROC curves, the samples arising from different cancer types are dissimilar enough that BS, CQ, SD, and our method all obtain perfect ROC curves (no $\mathbf{H}_1$ case has a larger p-value than any $\mathbf{H}_0$ case). We also note that the hypergeometric test-based (HG) enrichment analysis often used by experimentalists on this problem [25] gives a suboptimal area-under-curve of 0.989.

5 Conclusion

We have proposed a novel testing procedure for the two-sample test of means in high dimensions. This procedure can be implemented in a simple manner by first projecting a dataset with a single randomly drawn matrix, and then applying the standard Hotelling T^2 test in the projected space. In addition to obtaining the asymptotic power of this test, we have provided interpretable conditions on the covariance matrix Σ for achieving greater power than competing tests in the sense of asymptotic relative efficiency. Specifically, our theoretical comparisons show that our test is well suited to interesting regimes where most of the variance in the data can be captured in a relatively small number of variables, or where the variables are highly correlated. Furthermore, in the realistic case of $(n, p) = (98, 200)$, these regimes were shown to correspond to favorable performance of our test against several competitors in ROC curve comparisons on simulated data. Finally, we showed on real gene expression data that our procedure was more reliable than competitors in terms of its false positive rate. Extensions of this work may include more refined applications of random projection to high-dimensional testing problems.

Acknowledgements. The authors thank Sandrine Dudoit, Anne Biton, and Peter Bickel for helpful discussions. MEL gratefully acknowledges the support of the DOE CSGF Fellowship under grant number DE-FG02-97ER25308, and LJ the support of Stand Up to Cancer. MJW was partially supported by NSF grant DMS-0907632.

[4] Although this assumption could be violated by the existence of various cancer subtypes, or technical differences between original tissue samples, our initial step of randomly splitting the three cancer datasets into subsets guards against these effects.

References

[1] E. L. Lehmann and J. P. Romano. *Testing statistical hypotheses*. Springer Texts in Statistics. Springer, New York, third edition, 2005.

[2] Y. Lu, P. Liu, P. Xiao, and H. Deng. Hotelling's T2 multivariate profiling for detecting differential expression in microarrays. *Bioinformatics*, 21(14):3105–3113, Jul 2005.

[3] J. J. Goeman and P. Bühlmann. Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics*, 23(8):980–987, Apr 2007.

[4] D. V. D. Ville, T. Blue, and M. Unser. Integrated wavelet processing and spatial statistical testing of fMRI data. *Neuroimage*, 23(4):1472–1485, 2004.

[5] U. Ruttimann et al. Statistical analysis of functional MRI data in the wavelet domain. *IEEE Transactions on Medical Imaging*, 17(2):142–154, 1998.

[6] Z. Bai and H. Saranadasa. Effect of high dimension: by an example of a two sample problem. *Statistica Sinica*, 6:311,329, 1996.

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

[8] M. S. Srivastava. A test for the mean with fewer observations than the dimension under non-normality. *Journal of Multivariate Analysis*, 100:518–532, 2009.

[9] S. X. Chen and Y. L. Qin. A two-sample test for high-dimensional data with applications to gene-set testing. *Annals of Statistics*, 38(2):808–835, Feb 2010.

[10] S. Clémençon, M. Depecker, and Vayatis N. AUC optimization and the two-sample problem. In *Advances in Neural Information Processing Systems (NIPS 2009)*, 2009.

[11] A. Gretton, K. M. Borgwardt, M. Rasch, B. Schölkop, and A.J. Smola. A kernel method for the two-sample-problem. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 513–520. MIT Press, Cambridge, MA, 2007.

[12] Z. Harchaoui, F. Bach, and E. Moulines. Testing for homogeneity with kernel Fisher discriminant analysis. In John C. Platt, Daphne Koller, Yoram Singer, and Sam T. Roweis, editors, *NIPS*. MIT Press, 2007.

[13] M. E. Lopes, L. J. Jacob, and M. J. Wainwright. A more powerful two-sample test in high dimensions using random projection. Technical Report arXiv: 1108.2401, 2011.

[14] S. S. Vempala. *The Random Projection Method*. DIMACS Series in Discrete Mathematics and Theoretical Computer Science. American Mathematical Society, 2004.

[15] L. Jacob, P. Neuvial, and S. Dudoit. Gains in power from structured two-sample tests of means on graphs. Technical Report arXiv: q-bio/1009.5173v1, 2010.

[16] J. A. Cuesta-Albertos, E. Del Barrio, R. Fraiman, and C. Matrán. The random projection method in goodness of fit for functional data. *Computational Statistics & Data Analysis*, 51(10):4814–4831, 2007.

[17] R. J. Muirhead. *Aspects of Multivariate Statistical Theory*. John Wiley & Sons, inc., 1982.

[18] A. W. van der Vaart. *Asymptotic Statistics*. Cambridge, 2007.

[19] A. P. Dempster. A high dimensional two sample significance test. *Annals of Mathematical Statistics*, 29(4):995–1010, 1958.

[20] A. P. Dempster. A significance test for the separation of two highly multivariate small samples. *Biometrics*, 16(1):41–50, 1960.

[21] R. W. Tothill et al. Novel molecular subtypes of serous and endometrioid ovarian cancer linked to clinical outcome. *Clin Cancer Res*, 14(16):5198–5208, Aug 2008.

[22] J. Moreaux et al. A high-risk signature for patients with multiple myeloma established from the molecular classification of human myeloma cell lines. *Haematologica*, 96(4):574–582, Apr 2011.

[23] R. N. Jorissen et al. Metastasis-associated gene expression changes predict poor outcomes in patients with dukes stage b and c colorectal cancer. *Clin Cancer Res*, 15(24):7642–7651, Dec 2009.

[24] A. Subramanian et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA*, 102(43):15545–15550, Oct 2005.

[25] T. Beissbarth and T. P. Speed. Gostat: find statistically overrepresented gene ontologies within a group of genes. *Bioinformatics*, 20(9):1464–1465, Jun 2004.