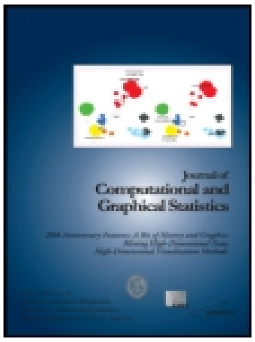


Journal of Computational and Graphical Statistics

ISSN: 1061-8600 (Print) 1537-2715 (Online) Journal homepage: http://www.tandfonline.com/loi/ucgs20

Direction-Projection-Permutation for High-Dimensional Hypothesis Tests

Susan Wei, Chihoon Lee, Lindsay Wichers & J. S. Marron

To cite this article: Susan Wei, Chihoon Lee, Lindsay Wichers & J. S. Marron (2016) Direction-Projection-Permutation for High-Dimensional Hypothesis Tests, Journal of Computational and Graphical Statistics, 25:2, 549-569, DOI: 10.1080/10618600.2015.1027773

To link to this article: http://dx.doi.org/10.1080/10618600.2015.1027773

Accepted author version posted online: 18 Apr 2015.
Published online: 10 May 2016.

Submit your article to this journal

Article views: 147

View related articles

View Crossmark data

Citing articles: 5 View citing articles

Full Terms & Conditions of access and use can be found at
http://www.tandfonline.com/action/journalInformation?journalCode=ucgs20

Download by: [University of Minnesota Libraries, Twin Cities] **Date:** 01 September 2016, At: 20:41

Direction-Projection-Permutation for High-Dimensional Hypothesis Tests

Susan WEI, Chihoon LEE, Lindsay WICHERS, and J. S. MARRON

High-dimensional low sample size (HDLSS) data are becoming increasingly common in statistical applications. When the data can be partitioned into two classes, a basic task is to construct a classifier that can assign objects to the correct class. Binary linear classifiers have been shown to be especially useful in HDLSS settings and preferable to more complicated classifiers because of their ease of interpretability. We propose a computational tool called direction-projection-permutation (DiProPerm), which rigorously assesses whether a binary linear classifier is detecting statistically significant differences between two high-dimensional distributions. The basic idea behind DiProPerm involves working directly with the one-dimensional projections of the data induced by binary linear classifier. Theoretical properties of DiProPerm are studied under the HDLSS asymptotic regime whereby dimension diverges to infinity while sample size remains fixed. We show that certain variations of DiProPerm are consistent and that consistency is a nontrivial property of tests in the HDLSS asymptotic regime. The practical utility of DiProPerm is demonstrated on HDLSS gene expression microarray datasets. Finally, an empirical power study is conducted comparing DiProPerm to several alternative two-sample HDLSS tests to understand the advantages and disadvantages of each method.

Key Words: Distance weighted discrimination; High-dimensional hypothesis test; High-dimensional low sample size; Linear binary classification; Permutation test; Two-sample problem.

1. INTRODUCTION

High-dimensional low sample size (HDLSS) datasets are becoming increasingly prominent in statistical applications. When the data can be partitioned into two classes, a basic

Susan Wei, Department of Statistics and Operations Research, University of North Carolina - Chapel Hill, NC 27599-3260 (E-mail: *susanwe@live.unc.edu*). Chihoon Lee, Assistant Professor, Department of Statistics and Operations Research, University of North Carolina - Chapel Hill, NC 27599-3260 and currently at Department of Statistics, Colorado State University, Fort Collins, CO 80523-1877 (E-mail: *chihoon@stat.colostate.edu*). Lindsay Wichers, Department of Environmental Sciences and Engineering, School of Public Health, University of North Carolina - Chapel Hill, NC 27599-3260 and currently at Environmental Media Assessment Group, MD B243-01, National Center for Environmental Assessment, Office of Research and Development, U.S. Environmental Protection Agency, Research Triangle Park, NC 27711 (E-mail: *wichers.lindsay@epa.gov*), J. S. Marron, Department of Statistics and Operations Research, University of North Carolina - Chapel Hill, NC 27599-3260 (E-mail: *marron@email.unc.edu*).

*© 2016 American Statistical Association, Institute of Mathematical Statistics,
and Interface Foundation of North America*

Journal of Computational and Graphical Statistics, Volume 25, Number 2, Pages 549–569
DOI: 10.1080/10618600.2015.1027773
Color versions of one or more of the figures in the article can be found online at *www.tandfonline.com/r/jcgs*.

task is to use the class labels to build a function that assigns data to the correct class, that is, classification. A popular classifier is the binary linear classifier that bases its decision on a linear combination of the features. Binary linear classifiers are preferable to more complicated classifiers such as random forests or neural networks in HDLSS settings for their ease of interpretability. If the variables are of the same scale, we can interpret the features with larger coefficients in the linear combination as being more "important," that is, they play a more prominent role in driving the separation between the two classes.

Linear classifiers are also known to find spurious linear combinations in HDLSS settings. For instance, a binary linear classifier could find, for two *identical* high-dimensional distributions, a linear combination of features such that the two classes appear to be very different. This is related to overfitting; when dimension exceeds sample size, a binary linear classifier can achieve 100% classification accuracy on the training data.

We propose a computational tool called direction-projection-permutation (DiProPerm) to rigorously assess whether a binary linear classifier is detecting a statistically significant difference between two high-dimensional distributions. This is accomplished by working directly with the one-dimensional projections of the data on the binary linear classifier, that is, the value of the linear combination. DiProPerm uses the projections to assess whether the original high-dimensional distributions are significantly different. The projection technique in DiProPerm has several advantages: (i) the outcome of the test can be related directly back to the binary linear classifier involved and (ii) DiProPerm borrows the strength of binary linear classifier in HDLSS settings to create a powerful hypothesis test.

We now provide a real world example, to be revisited in detail in Section 5, to further motivate the methodology. Two HDLSS breast cancer microarray datasets are considered. In each dataset, we consider the two-sample problem of testing for equal distribution. The panels in Figure 1 show, for each dataset, the one-dimensional projection of the data onto a binary linear classifier called distance weighted discrimination (further details are given in Section 2.1). Note that the one-dimensional projections in the UNCGEO dataset are better separated than in the UNCUP dataset. The DiProPerm test results, shown in Section 5, reveal the surprising conclusion that the UNCGEO classes are not actually significantly different while the UNCUP classes are very significantly different. The lesson here is that differences in lower dimensional visualizations do not always translate back to the original, high-dimensional, data distributions.

1.1 The Setup

Let $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ be independent random samples of $\mathbb{R}^d$-valued random vectors, $d \geq 1$ with multivariate distributions F_1 and F_2, respectively. We are interested in testing the null hypothesis of equal distributions

$$H_0 : F_1 = F_2 \quad \text{versus} \quad H_1 : F_1 \neq F_2. \tag{1}$$

1.2 Related Work

There are several two-sample multivariate tests for equal distributions in the literature based on inter-point statistics. One such family of tests is based on nearest neighbor (NN)

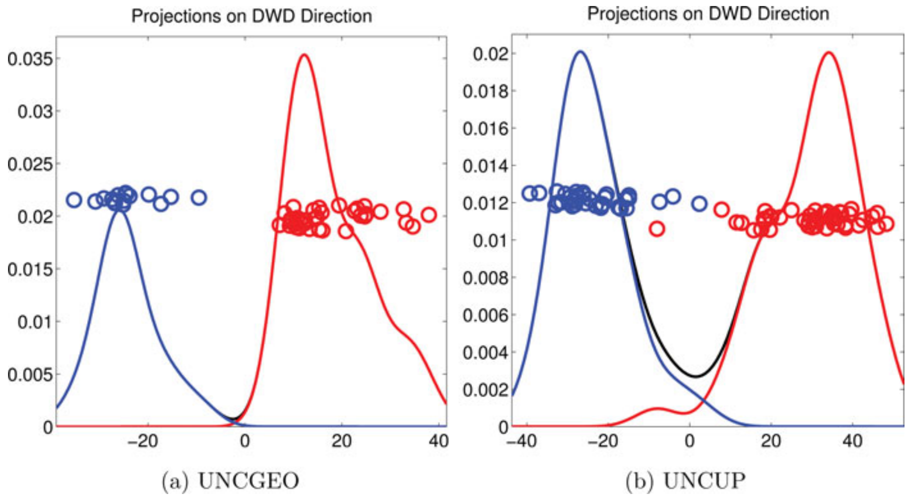


Figure 1. One-dimensional DWD projection plots for the UNCGEO dataset and the UNCUP dataset. The separation in the UNCGEO dataset is more visually pronounced than in the UNCUP dataset. Colors represent class labels.

coincidences (Bickel and Breiman 1983; Schilling 1986; Henze 1988). The test statistic is a function of the number of neighbors around a data point that belong to the same sample. The null distribution of the test statistic can either be determined using parametric theory or implemented nonparametrically via a permutation approach.

Another family of two-sample multivariate tests for equal distributions under HDLSS settings is based on pairwise distances (see Friedman and Rafsky 1979; Hall and Tajvidi 2002; Baringhaus and Franz 2004; Szekely and Rizzo 2004; Baringhaus and Franz 2010). The energy test statistic proposed in Szekely and Rizzo (2004), for example, is based on the Euclidean distance between pairs of sample elements. The significance of the energy test statistic is assessed using a permutation test. We will compare DiProPerm to the energy test later in Section 4. We choose to focus on the energy test rather than the NN tests because the latter critically depends on the choice of number of neighbors while the energy test has no tuning parameters.

A two-sample test for equal distributions that is quite similar to DiProPerm was recently proposed by Ghosh and Biswas (2013). We will describe in Section 4 details of the method and compare DiProPerm to it. We will demonstrate that DiProPerm has certain advantages over the Ghosh and Biswas method in part because the former makes use of all of the data while the latter holds out a subset during training.

It should also be noted that there are several two-sample tests for equal *means* in the literature. Many of these tests are based on extending the classic Hotelling T^2 test (Bai and Saranadasa 1996; Srivastava and Du 2008; Chen and Qin 2010). Essentially, they operate by replacing the covariance matrix in the Hotelling T^2 statistic by a diagonalized version. The critical values of the test statistic is often set using the limiting null distribution of the statistic letting both sample size and dimension go to infinity. The dimension d in Bai and Saranadasa (1996) is allowed to grow as long as $d/n \to y > 0$. In Srivastava and Du (2008), we have $n = O(p^{\psi})$ where $\frac{1}{2} < \psi \leq 1$. In Chen and Qin (2010), the dimension is also allowed to grow without any explicitly restriction between d and n directly. It is debatable whether this is a practical viewpoint since in HDLSS data, the sample size is very limited and additional data collection is often cost prohibitive.

Lopes, Jacob, and Wainwright (2011) claimed that the limited use of the covariance structure will cause the above procedures to suffer in power. They propose the random projection (RP) method, a two-sample test of equal means for Gaussian data. The procedure projects the high-dimensional data onto a *random* subspace of low enough dimension so that the traditional Hotelling T^2 statistic may be used. The resulting test statistic has a limiting F distribution in the classical asymptotic regime. The RP test critically assumes normality and equal covariances. We will compare DiProPerm to the RP test later in Section 4. Of particular interest is to see how the random projection scheme in the RP procedure compares with the carefully chosen projection direction in DiProPerm.

Finally, we will compare our method to the extreme value test in Cai, Liu, and Xia (2014), which addresses testing equality of means. Unlike the methods above, the extreme value test is not based on projections. The test statistic is based on a linear transformation of the data by the precision matrix, the inverse of the covariance matrix, which incorporates the correlations between the variables. The test is particularly powerful against sparse alternatives.

1.3 OVERVIEW

The outline of the article is as follows. In Section 2, the three-step DiProPerm procedure is presented in detail. The theoretical properties of DiProPerm under the HDLSS asymptotic regime whereby the dimension goes to infinity and sample size remains fixed are studied in Section 3. In particular, it is revealed that certain variations of DiProPerm are consistent and more importantly, consistency is a nontrivial property in HDLSS asymptotics, that is, there exists reasonable hypothesis tests whose power does not converge to 1 as dimension goes to infinity. In Section 4, we perform a Monte Carlo power study comparing DiProPerm to other two-sample HDLSS tests in a wide array of data settings. Finally, in Section 5, we follow up on the HDLSS breast cancer microarray datasets discussed at the beginning of the article.

2. METHODOLOGY

DiProPerm is a three-step procedure:

1. Direction—train a binary linear classifier on the class labels and find the normal vector to the separating hyperplane.
2. Projection—project data from both samples onto this normal vector and calculate a univariate two-sample statistic.
3. Permutation—assess the significance of this univariate statistic by a permutation test:

   (a) pool the two samples and permute the class labels

(b) recalculate direction based on permuted class labels

(c) project data onto this direction and recalculate the univariate two-sample statistic

For a level α test, we use a one-sided test and reject the null if the original test statistic is among the $100\alpha\%$ largest of the permuted statistics. Note that DiProPerm, being based on a permutation test, is an exact procedure, that is, the procedure is guaranteed to control the Type I error. Software for the DiProPerm procedure is available at

http://www.unc.edu/ marron/marron_software.html.

Further details on each of the DiProPerm steps is given below.

2.1 Direction

The first step of DiProPerm is based on a binary linear classifier. In particular, the normal vector to the separating hyperplane corresponding to the classifier is taken to be the projection direction in DiProPerm. Our classifier of choice is the distance weighted discrimination (DWD) classifier (Marron, Todd, and Ahn 2007), a powerful binary linear classifier in high dimensions. DWD seeks to find the separating hyperplane that minimizes the average inverse distance from data points to the separating hyperplane.

The support vector machine (SVM) is another popular binary linear classification method, see Hastie, Tibshirani, and Friedman (2001) for a good introduction. SVM seeks to find the separating hyperplane that maximizes the minimum distance from each data point to the hyperplane.

Both DWD and SVM are large margin-based classification methods. They can perform quite differently in HDLSS settings, however. Specifically, the SVM classifier suffers from severe data piling in this setting, that is, many projections pile onto the same value. Data piling is undesirable in the DiProPerm framework because it drastically reduces the information in the projected values. DWD was developed in part to overcome the data piling issue.

Figure 2 shows the projections of high-dimensional data ($d = 400$) onto DWD and SVM directions. The multivariate distributions considered are the multivariate t distribution with 2 degrees of freedom and the multivariate Gaussian distribution. We can see for both distributions, the SVM projections exhibit data piling while the DWD projections do not.

Neither the DWD and SVM projection directions have closed-form expressions. For this reason, it is challenging to study theoretical properties of DiProPerm tests based on these directions. In contrast, the centroid method is a simple but naive binary linear classifier where points are assigned to the class whose centroid is closest (Hastie, Tibshirani, and Friedman 2001). The normal vector to the separating hyperplane is the unit vector in the direction of the line segment connecting the centroids of each class, $(\bar{X} - \bar{Y})$.

We rarely use the centroid projection direction in practice because it is "all about the mean" and simple settings can be concocted where it performs much worse than either DWD or SVM. However, the centroid direction is most amenable for theoretical analysis, a fact we take advantage of in Section 3.

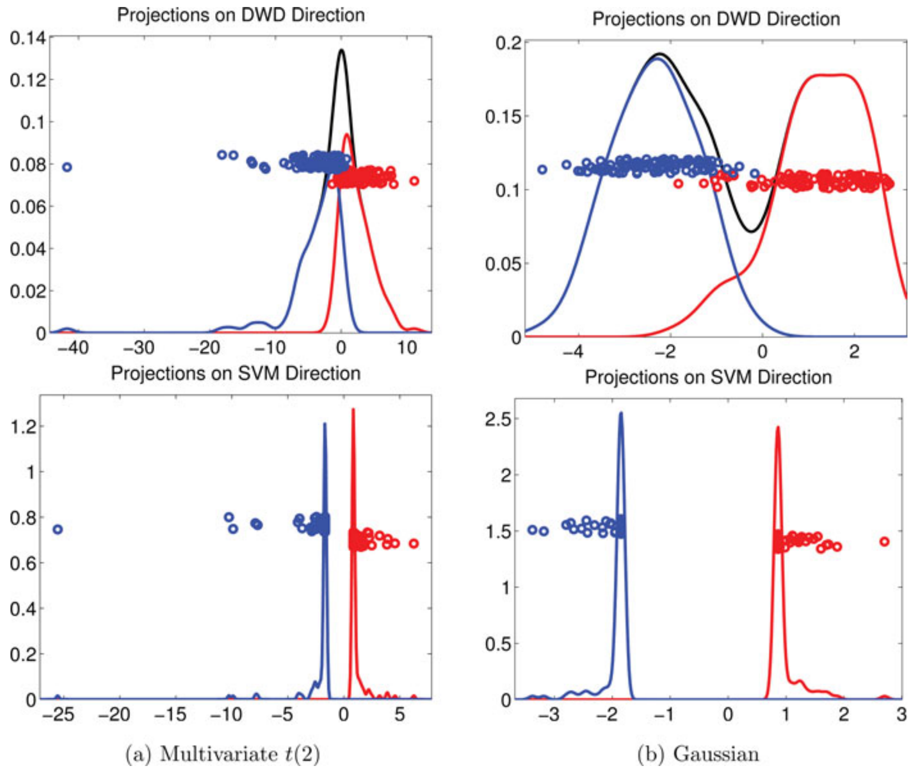


Figure 2. SVM projections versus DWD directions for multivariate $t(2)$ distribution and multivariate Gaussian distribution, $d = 400$. SVM projections (bottom row) exhibit data piling issue while DWD (top row) does not.

2.2 PROJECTION AND UNIVARIATE STATISTIC

In the second step of DiProPerm, the data are projected on the trained direction in the previous step and a univariate two-sample statistic is computed on the projected values. Let $\tilde{x}_1, \dots, \tilde{x}_m$ be the projected values in the first sample of X's and similarly, let $\tilde{y}_1, \dots, \tilde{y}_n$ be the projected values in the second sample of Y's. Here is a list of possible univariate two-sample statistics:

1. Difference of sample means (MD)

$$\frac{1}{m}\sum_{i=1}^{m}\tilde{x}_i - \frac{1}{n}\sum_{i=1}^{n}\tilde{y}_i$$

2. Two-sample t-statistic (t)

$$t = \frac{\text{MD}}{\sqrt{s_{\tilde{x}}^2/m + s_{\tilde{y}}^2/n}},$$

where $s_{\tilde{x}}^2$ is the unbiased estimator of the variance of the projected values $\tilde{x}$ and similarly for $s_{\tilde{y}}^2$.

3. Area under the curve (AUC), the curve being the receiver operating curve. The receiver operating curve is formed by plotting

$$\frac{1}{m}\sum_{i=1}^{m} 1\{\tilde{x}_i \leq c\}$$

versus

$$\frac{1}{n}\sum_{i=1}^{n} 1\{\tilde{y}_i \leq c\}$$

as c ranges over the values of the projected values in each sample. The AUC statistic is the area under this curve.

In Section 4, we will further explore the advantages and disadvantages of each of these statistics.

2.3 PERMUTATION

In the final step of DiProPerm, an approximate permutation test is conducted to assess the significance of the test statistic in the previous step. The empirical p-value, or the p-value simply, is calculated as the proportion of the statistics under permutation that exceeds the original test statistic.

For the permutation test to be exact, some randomization may have to be introduced particularly when both m and n are small. This is because when sample sizes are very small, the discreteness of the permutation distribution makes only certain p-values achievable. Randomization can be introduced if such is the case to make the test exact. For more details, see Ernst (2004).

3. THEORETICAL PROPERTIES OF DIPROPERM IN THE HDLSS ASYMPTOTIC REGIME

In this section, we study the limiting behavior of DiProPerm tests under the HDLSS asymptotic regime, that is, $d \to \infty$ for fixed n. Asymptotics in the classical setting (where $n \to \infty$ for fixed d) and the random matrix setting (where $d, n \to \infty$) are important as well. For this first article, however, we will restrict our attention to the HDLSS asymptotic regime.

First, a few words on notation. The DiProPerm test that uses direction A and univariate two-sample statistic B will henceforth be concatenated as A-B. For instance, the DiProPerm test that uses the centroid projection direction and the MD statistic will be concatenated to centroid-MD.

In the theoretical analysis, we will focus on the centroid projection direction for it has a simple closed-form expression. As for the univariate two-sample statistic, we will see in Section 4, the AUC statistic is favorable to the MD and t-statistics under certain settings. However, in this section we will focus only on the MD statistic and the two-sample t-statistic, which have simple closed-form expressions.

Consistency of hypothesis tests in the classical asymptotic regime is, in general, a rather trivial property, that is, most reasonable hypothesis tests will converge to 1 in power. We will show the same is not true in the HDLSS asymptotic regime. Examples of this can also be found in Biswas and Ghosh (2014).

Let $F_1 = N(\mu_x, \Sigma_x)$ and $F_2 = N(\mu_2, \Sigma_y)$ be two multivariate Gaussian distributions where the covariance matrices are spherical. Consider m draws from the former and n draws from the latter. For the alternative of unequal means $\mu_x \neq \mu_y$, it can be established that all DiProPerm tests are consistent under certain moment and mixing conditions using Theorem 3.1 of Ghosh and Biswas (2013). This is not surprising since binary linear classifiers are very adept at picking up signals in the mean. Furthermore, in higher dimensions, linear classifiers often outperform complex classifiers with complicated decision boundaries. Thus in the case of a mean effect, we expect DiProPerm to a consistent procedure.

A more interesting alternative is equal means $\mu_x = \mu_y$ but unequal covariances $\Sigma_x \neq \Sigma_y$. Theorem A.2 in Appendix A shows that the DiProPerm centroid-t test is consistent (as $d \to \infty$) under these settings and Theorem A.1 in Appendix A shows that the DiProPerm centroid-MD test is *not* consistent under these same settings. HDLSS geometric intuition of these results are given in Appendix B.

Remark 1. The normality and spherical covariance assumptions used in Theorems A.1 and A.2 in Appendix A may be relaxed at the expense of more complicated proofs. Specifically, we may relax them to the following assumptions used in Hall, Marron, and Neeman (2005), Jung and Marron (2009), and later in Ghosh and Biswas (2013), listed briefly here:

1. uniformly bounded fourth moments
2. ρ-mixing condition
3. $d^{-1}\sum_{k=1}^{d} \mathrm{var}(X^{(k)}) \to \sigma_1^2$, similarly for Y and $d^{-1}\sum_{k=1}^{d}\{E(X^{(k)}) - E(Y^{(k)})\}^2 \to \mu^2$.

The normality assumption can be replaced by the first two conditions above while the spherical covariance assumption can be replaced by the third condition above.

Theorem A.1 in Appendix A also has the surprising implication that the centroid-MD DiProPerm test, under equal sample sizes, is asymptotically valid for testing equality of means. This is surprising because permutation tests are, in general, not valid for testing equal means as the assumption of exchangeability of the samples may not hold under the null hypothesis. See Huang et al. (2006) and Chung and Romano (2013) for more details on this phenomenon.

4. COMPARISON WITH OTHER METHODS

We will compare the empirical power of DiProPerm with the RP test, the energy test, and the Ghosh and Biswas (GB) method. The empirical power is calculated using Monte

Carlo simulations and is calculated as the proportion of times a test rejects the null. The tests are conducted at significant level 0.05.

We will focus on the DWD direction in DiProPerm for reasons discussed in Section 2.1. We will consider all three univariate two-sample statistics presented in Section 2.2—the MD, two-sample *t*, and AUC statistic.

In the GB method, the data in each sample are first randomly split into two subsets, call them training and testing. A binary linear classifier is learned on the training subsets, that is, step 1 of DiProPerm. The data in the testing subsets are projected onto the linear classifier and a univariate two-sample statistic is calculated, that is, Step 2 of DiProPerm. Ghosh and Biswas (2013) provided two choices for the classifier—SVM and DWD—and two choices for the univariate two-sample statistic—the Kolmogorov–Smirnov (KS) statistic, and the Wilcoxon–Mann–Whitney (WMW) statistic. The test function for this random split is calculated according to the KS null distribution or the WMW null distribution. The final test function is averaged over many different random splits. As the final test function is usually a number strictly between 0 and 1, the test is implemented via randomization. We will compare DiProPerm to the version of the GB method, which uses DWD and the KS statistic, abbreviated as KS-DWD.

The extreme value test in Cai, Liu, and Xia (2014) is implemented using the test statistic proposed for the setting of an unknown precision matrix, which is not assumed to be sparse. The tuning parameter δ is set at the recommended value of $\delta = 2$.

The simulation settings considered in the empirical power study are summarized below, details of which are given later.

S1. Gaussian mixture.

S2. Normal versus *T*.

S3. Multivariate Gaussians with a large location shift, and small covariance difference.

S4. Multivariate Gaussians with a small location shift, and large covariance difference.

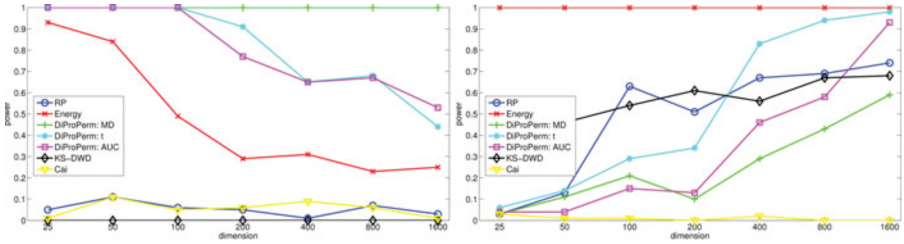
S5. Multivariate Gaussians, with a sparse location shift of $\log(d)/3$ in the first coordinate (as in Ghosh and Biswas 2013).

S6. Multivariate *T* distributions with 2 degrees of freedom, with a sparse location shift of $\log(d)/3$ in the first coordinate (as in Ghosh and Biswas 2013).

S7. Multivariate Cauchy distributions, with a sparse location shift of $\log(d)/3$ in the first coordinate (as in Ghosh and Biswas 2013).

In all simulations that follow, we generate 100 observations from each class and study the empirical power for a range of dimensions $d = 25, 50, 100, 200, 400, 800$, and 1600.

In Simulation, the data in each sample arise from a two-component Gaussian mixture, equally weighted. In the first sample, the first component has mean $(3, 30, 0, \dots, 0)$ and the second component has mean $(3, -30, 0, \dots, 0)$. The covariance matrix is the identity matrix in both components. In the second sample, we have the same setup with a different mean structure. The first component has mean $(-3, 30, 0, \dots, 0)$ and the second component has mean $(-3, -30, 0, \dots, 0)$. Figure 3(a) displays the empirical power of various methods considered here under this setting. The DiProPerm DWD-MD test performs



(a) Simulation S1, Gaussian mixtures (b) Simulation S2, normal versus T

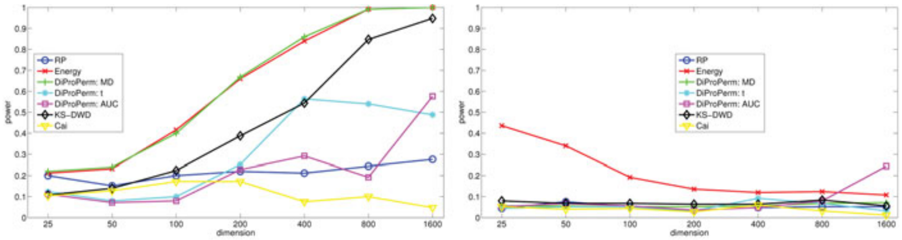
Figure 3. Non-Gaussian distributions.

excellently while the other two DiProPerm tests, DWD-t, and DWD-AUC, perform less well but still favorably to the other methods considered. The RP test and KS-DWD test of Ghosh and Biswas (2013) both perform quite poorly in this mixture setting. The large signal in the second coordinate of the mean vector in this simulation is a red herring, since the real action resides actually in the first coordinate. For this reason, tests based on mean and covariance quantities are likely to perform poorly (e.g., Bai and Saranadasa 1996; Srivastava and Du 2008; Chen and Qin 2010) and other similar tests for detecting high-dimensional mean effects.

In Simulation S2, the first sample arises from the standard multivariate Gaussian distribution. The second sample arises from a multivariate distribution with independent $t(5)$ marginals. This simulation setting was considered by Szekely and Rizzo (2004) that also proposed the energy test that can be seen to dominate the other methods for this setting.

In Simulation S3, we consider a setting where the mean effect grows stronger with dimension and covariance effect grows weaker with dimension. Let $F_1 = N(\mu_1, \Sigma_1)$ and $F_2 = N(\mu_2, \Sigma_2)$. We set μ_1 to be the zero vector and μ_2 to be zero in the first 25% of the coordinates and $1/\sqrt{n}$ in the rest. Let Σ be the covariance matrix with 1's along the main diagonal and 0.2 along the two diagonals off the main one. Let U be a d by d matrix with uniform$(0, 32/d^2)$ entries. Let $\delta = |\min(\text{ smallest eigenvalue of } \Sigma, \text{ smallest eigenvalue of } \Sigma + U)| + 0.05$. We set $\Sigma_1 = \Sigma + \delta I_d$ and $\Sigma_2 = \Sigma + U + \delta I_d$. This type of covariance alternative was considered by Cai, Liu, and Xia (2013). The empirical power of each method under this setting is shown in Figure 4(a). We see that the DiProPerm MD test and energy test perform very similarly and outperform all other methods. This is reasonable since the MD statistic is good at picking up a strong mean effect. The same is true for the energy test that is based on pairwise distances between points. The RP test also performs rather poorly in this setting even though the data are Gaussian and the mean effect is felt through many directions. This may be a sign that the RP test is very unrobust to the equal covariances assumption. The KS-DWD test also has decent performance but is not as powerful as DiProPerm. Both methods use the DWD projection direction, but the KS-DWD method uses only a subset of data to train DWD, which could lead to a decrease in power.

In Simulation S4, we consider a mean effect that grows weaker with dimension and a covariance effect that grows stronger with dimension. For this setting, we let μ_1 be the zero vector and μ_2 be $1/\sqrt{d}$ in all coordinates. The covariance matrices are formed as in



(a) Simulation S3, large mean small covariance alternative

(b) Simulation S4, small mean large covariance alternative

Figure 4. Gaussian distributions with mean and covariance alternatives.

Simulation 2 except U is a d by d matrix with uniform(0, 32) entries. The empirical power of each method under this setting is shown in Figure 4(b). The energy test outperforms all other methods by a large margin for low dimensions. However, all methods considered encounter difficulty as the dimension increases. For this setting, a test tailored for testing equal covariances may be necessary.

In Simulation S5, a setting studied in Ghosh and Biswas (2013), two standard multivariate Gaussian differ only in a location shift of $\log(d)/3$ in the first coordinate. Figure 5 shows the DiProPerm MD test, the energy test, and the KS-DWD test outperform all other methods and perform similarly to each other. The RP test performs poorly in this setting despite the fact that all assumptions for the procedure are satisfied, that is, normality and equal covariances. This is likely because the true signal is contained in the first coordinate direction but the RP method tries to sense many different random directions. In this setting, the covariance structure between the two samples is identical. The extreme value test proposed in Cai, Liu, and Xia (2014) has better performance in this setting than the simulation settings above. This may reflect the fact that the procedure in Cai, Liu, and Xia (2014) is for testing equality of means and is quite sensitive to violation of the assumption that the covariance structure is identical.

In Simulations S6 and S7, we look at multivariate distributions with heavier tails than the multivariate Gaussian. The distributions considered are the multivariate t distribution with 2

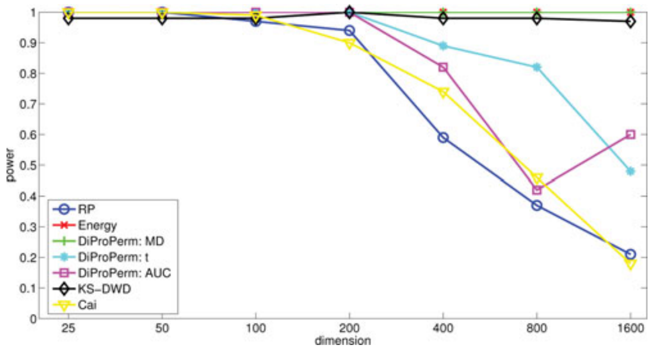
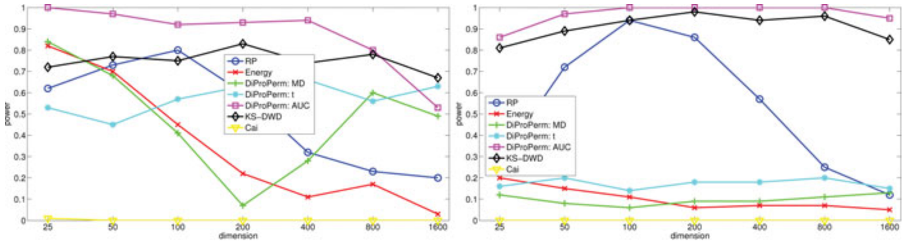


Figure 5. Simulation S5, multivariate Gaussian.



(a) Simulation S6, multivariate T (b) Simulation S7, multivariate Cauchy

Figure 6. Heavy-tail distributions with sparse location shift.

degrees of freedom and the multivariate Cauchy distribution. For both, we consider location shifts of $\log(d)/3$ in the first coordinate. The multivariate t and Cauchy distributions are generated with the correlation matrix set to the identity matrix; note that this does not imply independence.

Figure 6(a) and 6(b) shows the DiProPerm AUC outperforms all other methods for the two heavy-tailed distributions considered. The RP test performs decently for low dimensions. This is impressive since the RP procedure assumes normality and the data here are far from normal. However, the RP test seems to suffer a great loss in power as soon as sample size exceeds dimension. We also see that the energy test may have performed poorly because it is based on pairwise distances and thus is not robust to the heavy-tailedness of the multivariate distributions. The KS-DWD test performed well by maybe less powerful than DiProPerm because only part of the data is used to train DWD. The extreme value test by Cai, Liu, and Xia (2014) performs poorly in Simulations S6 and S7, which may reflect that the test is not robust to departures from Gaussianity or sub-Gaussianity.

5. DATA EXAMPLE

In this section, we revisit the analysis of the two HDLSS breast cancer microarray datasets introduced at the beginning of the article. The first dataset is denoted as UNCGEO and the second as UNCUP, following the naming convention of their source which can be found at *http://peroulab.med.unc.edu/*.

The UNCGEO dataset consists of the gene expression values of 9674 genes measured on 50 breast cancer patients at UNC. The UNCUP dataset looks at the same set of genes measured on 80 breast cancer patients in another study at UNC.

The UNCGEO patients are divided into two breast cancer subtypes: (i) Luminal A, and (ii) Luminal B. The UNCUP patients are divided into the groups: (i) Luminals (Luminal A and Luminal B), and (ii) HER and Basal. DiProPerm DWD-t is used to test for equal distributions of the gene expression between the groups in each dataset.

Recall Figure 1 that shows the data projected onto DWD directions for each dataset. The UNCGEO projections do not overlap at all whereas the UNCUP projections have a small amount of overlap. These projection plots suggest that the separation is better for Luminal A versus Luminal B in the UNCGEO dataset than for Luminals versus HER and Basal in the UNCUP dataset.

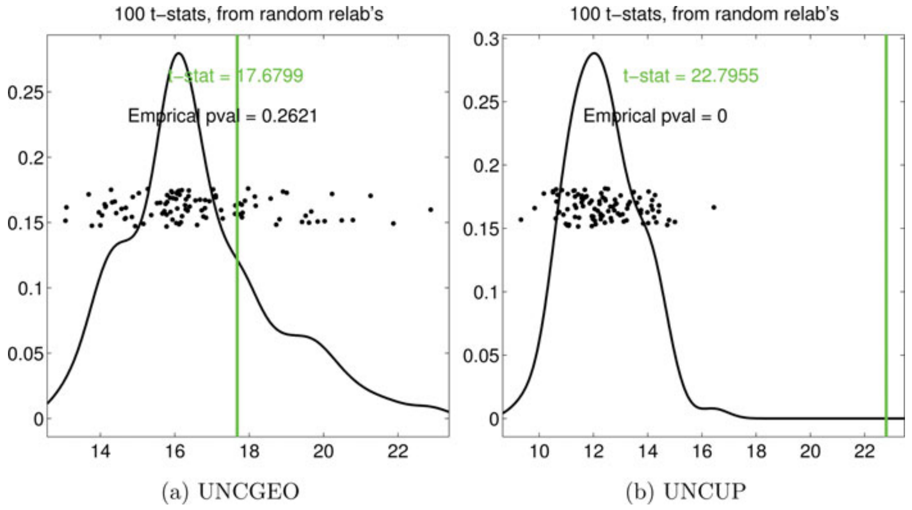


Figure 7. The DiProPerm t test result for each dataset is displayed. In the UNCGEO study, the difference between the Luminal A and Luminal B subgroups is not significant. In the UNCUP study, the difference between the Luminals and HER & Basal subgroups is very significant. This is surprising because the projection plots in Figure 1 suggest the contrary.

Figure 7 displays the DiProPerm test results. Each dot represents the test statistic resulting from a single permutation in the permutation test. The black curve is the kernel density estimate of the permutation statistics. The position of the original t-statistic is marked by a vertical dashed line. The empirical p-values show that the difference in the UNCGEO dataset is not significant while the difference in the UNCUP dataset is very significant. Note that we are not trying to directly compare the results of the two tests as they test different hypotheses.

This example illustrates that what may seem to be a visually striking separation in lower dimensional visualizations could well be an artifact of over-fitting or sampling variation.

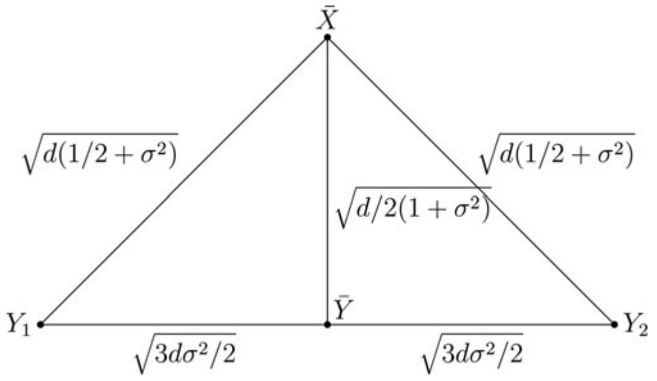


Figure 8. Plane generated by Y_1, Y_2, and $\bar{X}$ where $X_1, X_2 \sim F_1 = N(0, I_d)$ and $Y_1, Y_2 \sim F_2 = N(0, \sigma^2 I_d)$ for $\sigma^2 \neq 1$. Note that the projections of Y_1 and Y_2 onto $\bar{X} - \bar{Y}$ is close to the projection of $\bar{Y}$ onto $\bar{X} - \bar{Y}$. This has the implication that $s_{\bar{Y}}^2$ will be small.

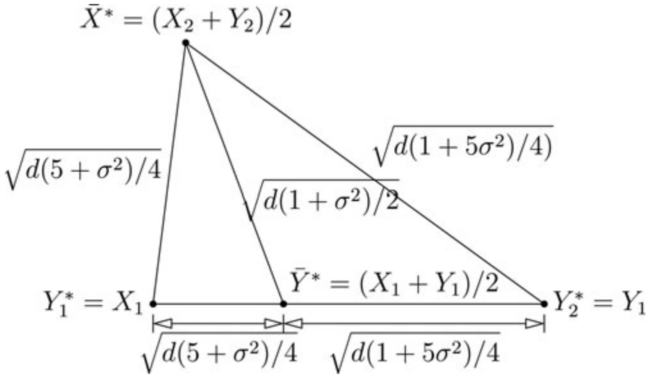

$\bar{X}^* = (X_2 + Y_2)/2$
$\sqrt{d(5+\sigma^2)/4}$
$\sqrt{d(1+5\sigma^2)/4)}$
$\sqrt{d(1+\sigma^2)/2}$
$\bar{Y}^* = (X_1 + Y_1)/2$
$Y_1^* = X_1$
$Y_2^* = Y_1$
$\sqrt{d(5+\sigma^2)/4}$
$\sqrt{d(1+5\sigma^2)/4}$

Figure 9. Plane generated by a particular permutation realization of X_1, X_2, Y_1, and Y_2. Note that the projections of Y_1^* and Y_2^* onto $\bar{X}^* - \bar{Y}^*$ is not close to the projection of $\bar{Y}^*$ onto $\bar{X}^* - \bar{Y}^*$. This has the implication that $s^2_{\bar{Y}^*}$ may be large.

Importantly, the hypothesis test results also agree with the biology. In particular, it is well known that Luminals have a very different gene expression signature from HER and Basal. On the other hand, the difference between Luminal A and Luminal B is less clear cut (Carey et al. 2006).

6. SUMMARY

In this article, we introduced DiProPerm, a hypothesis testing framework for testing equality of two multivariate distributions. DiProPerm hypothesis tests are nonparametric and exact. Furthermore, DiProPerm also ties in nicely with high-dimensional data visualization. In practice, we recommend using the DWD direction in DiProPerm. If one wishes to test for equality of means, the mean difference univariate statistic should be used. On the other hand, if testing the more general equality of distributions, the two-sample t-statistic and the AUC statistic both perform well.

APPENDIX A: HDLSS ASYMPTOTICS

In this section, we study the HDLSS asymptotic behavior of the centroid-MD and the centroid-t tests. We show the centroid-t is consistent but the centroid-MD is inconsistent under the alternative of equal means and unequal covariances.

Let Z be the pooled sample of X's and Y's. The centroid-MD test statistic, to be denoted by $T_{m,n}(Z)$, is the mean of the projections of the X's onto the unit vector in the direction of $\bar{X} - \bar{Y}$ minus the mean of the projections of the Y's onto the unit vector in the direction of $\bar{X} - \bar{Y}$:

$$T_{m,n}(Z) = T_{m,n}(X_1, \ldots, X_m, Y_1, \ldots, Y_n) \tag{A.1}$$

$$= \frac{1}{m}\sum_{i=1}^{m} X_i' \frac{(\bar{X} - \bar{Y})}{||\bar{X} - \bar{Y}||} - \frac{1}{n}\sum_{j=1}^{n} Y_j' \frac{(\bar{X} - \bar{Y})}{||\bar{X} - \bar{Y}||} \tag{A.2}$$

$$= ||\bar{X} - \bar{Y}||. \tag{A.3}$$

Theorem A.1. Let $X_1, \dots, X_m$ be an iid sample from the d-variate Gaussian distribution $N(\mu_X, \Sigma_x)$ and $Y_1, \dots, Y_n$ be an independent sample drawn iid from the d-variate Gaussian distribution $N(\mu_Y, \Sigma_y)$, where $\Sigma_X \neq \Sigma_Y$. If $m = n$, then the unconditional distribution and the permutation distribution of $T_{m,n}(Z)$ are equal under the null $\mu_X = \mu_Y$.

Proof. Under $\mu_X = \mu_Y$, $\bar{X} - \bar{Y}$ is distributed as

$$N(0, \Sigma_x/m + \Sigma_y/n) \tag{A.4}$$

and the permutation distribution of $\bar{X} - \bar{Y}$ is

$$\sum_{r=0}^{m} \frac{\binom{m}{r}\binom{n}{r}}{\binom{m+n}{m}} N\left(0, \frac{(m-r)\Sigma_x + r\Sigma_y}{m^2} + \frac{r\Sigma_x + (n-r)\Sigma_y}{n^2}\right). \tag{A.5}$$

If $m = n$, the expressions in (A.4) and (A.5) are the same, in which case the unconditional and permutation distribution of $T_{m,n}(Z)$ are also the same.

□

Notice the result in Theorem A.1 is based on finite samples. If the two samples are not independent, for instance, we would have to appeal to asymptotics to derive an approximation of (A.5).

The centroid-t-statistic, to be denoted by $U_{m,n}(Z)$, is the result of applying the unbalanced sample sizes, unequal variance two-sample t-test statistic (also known as Welch's t-test, Welch 1947) to the projections onto the centroid direction. Let $a \cdot b$ denote the standard dot product between two vectors in $\mathbb{R}^d$. The sample variances of the projected data can be expressed as

$$s_{\bar{X}}^2 = \frac{1}{m-1} \sum_{i=1}^{m} [(X_i - \bar{X}) \cdot (\bar{X} - \bar{Y})]^2$$

and

$$s_{\bar{Y}}^2 = \frac{1}{n-1} \sum_{i=1}^{n} [(Y_i - \bar{Y}) \cdot (\bar{X} - \bar{Y})]^2.$$

Define $S_{m,n}(Z) = S_{m,n}(X_1, \dots, X_m, Y_1, \dots, Y_n) = s_{\bar{X}}^2/m + s_{\bar{Y}}^2/n$. The centroid-$t$-statistic is

$$U_{m,n}(Z) = T_{m,n}(Z)/\{S_{m,n}(Z)\}^{1/2},$$

where $T_{m,n}(Z)$ is as defined above. We use the term "projected" rather loosely here since we have not normalized by $||\bar{X} - \bar{Y}||$. This is of no actual consequence since the two-sample t-statistic is scale invariant.

Theorem A.1 establishes the numerator of the centroid-t-statistic behaves the same in the permutation and unconditional world. The same is not true for the denominator of the centroid-t-statistic. The next result gives us a sense of just how far apart are the permutation and unconditional distributions of the denominator $S_{m,n}(Z)$.

Theorem A.2. Let $X_1, \dots, X_m$ be a sample from the d-variate Gaussian distribution $N(\mu_x, \sigma_x^2 I_d)$ and $Y_1, \dots, Y_n$ be an independent sample from the d-variate Gaussian distribution $N(\mu_y, \sigma_y^2 I_d)$, where $\sigma_x^2 \neq \sigma_y^2$ are scalars. Under $\mu_x = \mu_y$, we have

$$\frac{1}{d} S_{m,n}(Z) \xrightarrow{d} \left(\frac{\sigma_x^2}{m} + \frac{\sigma_y^2}{n}\right) \left\{\frac{1}{m-1}\frac{\sigma_x^2}{m}\chi^2(m-1) + \frac{1}{n-1}\frac{\sigma_y^2}{n}\chi^2(n-1)\right\}$$

as d goes to infinity. For the permuted version, we have for some nonzero constant c,

$$\frac{1}{d^2} S_{m,n}(Z_\pi) \to c \text{ in probability.}$$

The results of this theorem are surprising in that the denominator of the centroid-t-statistic is actually of different orders in the unconditional and permutation worlds. In particular, in the unconditional world $S_{m,n}(Z)$ grows like a random variable times d, while in the permutation world it grows like a constant times d^2.

To prove Theorem A.2, a few lemmas are needed.

Lemma A.1. Let $X_1, \dots, X_m$ be a sample from the d-variate Gaussian distribution $N(\mu_x, \sigma_x^2 I_d)$ and $Y_1, \dots, Y_n$ be an independent sample from the d-variate Gaussian distribution $N(\mu_y, \sigma_y^2 I_d)$, where $\sigma_x^2 \neq \sigma_y^2$. Let $\tilde{X}_k = X_k'(\bar{X} - \bar{Y})$. Let $\overline{\tilde{X}_{1:k-1}}$ be the sample mean of $\tilde{X}_1, \dots \tilde{X}_{k-1}$. Under $\mu_x = \mu_y$, we have, for $k = 2, \dots, m$

$$\frac{d^{-1/2}((\tilde{X}_k - \overline{\tilde{X}_{1:k-1}}))}{\left\{\frac{k}{k-1}\sigma_x^2(\sigma_x^2/m + \sigma_y^2/n)\right\}^{1/2}} \xrightarrow{d} N(0,1) \text{ as } d \to \infty.$$

Similarly we have

$$\frac{d^{-1/2}((\tilde{Y}_k - \overline{\tilde{Y}_{1:k-1}}))}{\left\{\frac{k}{k-1}\sigma_y^2(\sigma_x^2/m + \sigma_y^2/n)\right\}^{1/2}} \xrightarrow{d} N(0,1) \text{ as } d \to \infty$$

$k = 2, \dots, n.$

Proof. We can write $\tilde{X}_k - \overline{\tilde{X}_{1:k-1}}$ as a sum of products

$$\tilde{X}_k - \overline{\tilde{X}_{1:k-1}} = \sum_{p=1}^{d} (X_k - \bar{X}_{1:k-1})^{(p)} (\bar{X} - \bar{Y})^{(p)}, \tag{A.6}$$

where $X^{(p)}$ simply refers to the pth component in the d-dimensional vector X. The expectation of the summands in (A.6) is zero:

$$\begin{aligned} E(X_k - \bar{X}_{1:k-1})^{(p)}(\bar{X} - \bar{Y})^{(p)} &= E(X_k^{(p)}\bar{X}^{(p)}) - E(\bar{X}_{1:k-1}^{(p)}\bar{X}^{(p)}) \\ &\quad - E(X_k^{(p)}\bar{Y}^{(p)}) + E(\bar{X}_{1:k-1}^{(p)}\bar{Y}^{(p)}) \\ &= 0. \end{aligned}$$

Next we look at the variance of the summands. Recall for Gaussian data, zero covariance is equivalent to independence. We know the covariance between $(X_k - \bar{X}_{1:k-1})^{(p)}$ and $(\bar{X} - \bar{Y})^{(p)}$ is zero since the expectation of the latter is zero and the expectation of the product was shown above to be zero as well. Thus, each summand in (A.6) is the product of two independent variables. The variance of a product of independent variables, U and V, is

$$(EU)^2\text{var}(V) + (EV)^2\text{var}(U) + \text{var}(U)\text{var}(V). \tag{A.7}$$

Thus, we have

$$\text{var}(X_k - \bar{X}_{1:k-1})^{(p)}(\bar{X} - \bar{Y})^{(p)} = \text{var}(X_k - \bar{X}_{1:k-1})^{(p)}\text{var}(\bar{X} - \bar{Y})^{(p)}$$

$$= \frac{k}{k-1}\sigma_x^2(\sigma_x^2/m + \sigma_y^2/n).$$

By the central limit theorem, we have

$$\frac{d^{1/2}\left(\frac{1}{d}(\tilde{X}_k - \overline{\tilde{X}_{1:k-1}})\right)}{\left\{\frac{k}{k-1}\sigma_x^2(\sigma_x^2/m + \sigma_y^2/n)\right\}^{1/2}} \xrightarrow{d} N(0,1) \text{ as } d \to \infty$$

□

Lemma A.2. Let $X_1, \dots, X_m$ be a sample from the d-variate Gaussian distribution $N(\mu_x, \sigma_x^2 I_d)$ and $Y_1, \dots, Y_n$ be an independent sample from the d-variate Gaussian distribution $N(\mu_y, \sigma_y^2 I_d)$, where $\sigma_x^2 \neq \sigma_y^2$. Let π be a permutation of $\{1, \dots, N = m + n\}$. Let $\bar{Z}_\pi = (\bar{Z}_{\pi(1:m)} - \bar{Z}_{\pi(m+1:N)})$ be the centroid direction trained on the permuted labels determined by π. We have for $i = 1, \dots, m$,

$$E((Z_{\pi(i)} - \overline{Z_{\pi(1:m)}})^{(k)}\bar{Z}_\pi^{(k)})$$

is nonzero. Similarly, for $i = m + 1, \dots, N$, we have

$$E((Z_{\pi(i)} - \overline{Z_{\pi(m+1:N)}})^{(k)}\bar{Z}_\pi^{(k)})$$

is nonzero.

Proof. We prove the first statement. The second can be shown in a similar fashion. Let $P(n, k)$ denote the number of k permutations of n, that is,

$$P(n,k) = n \cdot (n-1) \cdot (n-2) \cdots (n-k+1).$$

We have for $i = 1, \dots, m$ and $k = 1, \dots, d$,

$$\begin{aligned}
E((Z_{\pi(i)} - \overline{Z_{\pi(1:m)}})^{(k)}\bar{Z}_\pi^{(k)}) &= E((Z_{\pi(i)} - \bar{Z}_{\pi(1:m)})^{(k)}(\bar{Z}_{\pi(1:m)} - \bar{Z}_{\pi(m+1:N)})^{(k)}) \\
&= EZ_{\pi(i)}^{(k)}\bar{Z}_{\pi(1:m)}^{(k)} - EZ_{\pi(i)}^{(k)}\bar{Z}_{\pi(m+1:N)}^{(k)} - E(\bar{Z}_{\pi(1:m)}^{(k)})^2 \\
&\quad + E\bar{Z}_{\pi(1:m)}^{(k)}\bar{Z}_{\pi(m+1:N)}^{(k)} \\
&= \frac{(EZ_{\pi(i)}^{(k)})^2}{m} + \frac{m}{m-1}\mu^2 - \mu^2 - E(\bar{Z}_{\pi(1:m)}^{(k)})^2 + \mu^2 \\
&= \frac{\operatorname{var}(Z_{\pi(i)}^{(k)}) + \mu^2}{m} + \frac{m}{m-1}\mu^2 - (\operatorname{var}(\bar{Z}_{\pi(1:m)}^{(k)}) + \mu^2) \\
&= \frac{\operatorname{var}(Z_{\pi(i)}^{(k)})}{m} - \operatorname{var}(\bar{Z}_{\pi(1:m)}^{(k)}) \\
&= \frac{m}{N}\left\{\frac{\sigma_x^2}{m} - \frac{1}{m^2}\frac{1}{w_1}\sum_{r=0}^{m-1} P(m-1,r)P(n,m-r)\right. \\
&\quad \left. \times [r\sigma_x^2 + (m-r)\sigma_y^2]\right\} \\
&\quad + \frac{n}{N}\left\{\frac{\sigma_y^2}{m} - \frac{1}{m^2}\frac{1}{w_2}\sum_{r=0}^{n-1} P(n-1,r)P(m,m-r)\right. \\
&\quad \left. \times [r\sigma_y^2 + (m-r)\sigma_x^2]\right\},
\end{aligned}$$

where w_1 and w_2 are the weights

$$w_1 := \sum_{r=0}^{m-1} P(m-1,r)P(n,m-r) \quad \text{and} \quad w_2 := \sum_{r=0}^{n-1} P(n-1,r)P(m,m-r).$$

Thus if $\sigma_x^2 \neq \sigma_y^2$, we have $E((Z_{\pi(i)} - \overline{Z_{\pi(1:m)}})^{(k)}\bar{Z}_\pi^{(k)})$ is nonzero.

□

Lemma A.3. Let Z_1, Z_2 be two random variables in $\mathbb{R}^d$ such that $Z_1^{(k)}Z_2^{(k)}$ are iid for $k = 1, \ldots, d$ and $E(Z_1^{(k)}Z_2^{(k)})$ exists and is finite. Then

$$\frac{1}{d^2}(Z_1 \cdot Z_2)^2 \to [E(Z_1^{(k)}Z_2^{(k)})]^2 \text{ in probability.}$$

Proof. By the law of large numbers, we have

$$\frac{1}{d}(Z_1 \cdot Z_2) \to E(Z_1^{(k)}Z_2^{(k)}) \text{ in probability.}$$

By continuous mapping theorem, we have

$$\frac{1}{d^2}(Z_1 \cdot Z_2)^2 \to [E(Z_1^{(k)}Z_2^{(k)})]^2 \text{ in probability.}$$

□

Proof of Theorem A.2. To prove the first part of Theorem A.2, we decompose $s^2_{\tilde{X}}$ and $s^2_{\tilde{Y}}$ each into a sum of independent variables. Let $\overline{\tilde{X}}_{k-1}$ be the sample mean of the first $k-1$ projections $\tilde{X}_1, \ldots \tilde{X}_{k-1}$. We will write $s^2_{\tilde{X}}$ in a recursive fashion. Define $s_1^2 = 0$. We will use the following recursive formula to define s_k^2 for $k = 2, \ldots, m$

$$(k-1)s_k^2 = (k-2)s_{k-1}^2 + \frac{k-1}{k}(\tilde{X}_k - \overline{\tilde{X}}_{k-1})^2. \tag{A.8}$$

Since s_{k-1}^2 is independent of $(\tilde{X}_k - \overline{\tilde{X}}_{k-1})^2$, this recursive viewpoint allows us to decompose $s^2_{\tilde{X}} = s_m^2$ into a sum of independent terms. Using the result in Lemma A.1 and the second-order Delta method, we have

$$\frac{\frac{1}{d}(\tilde{X}_k - \overline{\tilde{X}_{1:k-1}})^2}{\frac{k}{k-1}\sigma_x^2(\sigma_x^2/m + \sigma_y^2/n)} \xrightarrow{d} \chi^2(1) \text{ as } d \to \infty. \tag{A.9}$$

Inputting expression (A.9) into the recursion defined in (A.8) and exploiting the independence of the individual terms in $s^2_{\tilde{X}}$, we get

$$\frac{1}{d}s^2_{\tilde{X}} \xrightarrow{d} \frac{1}{m-1}\sigma_x^2\left(\frac{\sigma_x^2}{m} + \frac{\sigma_y^2}{n}\right)\chi^2(m-1) \text{ as } d \to \infty$$

Similarly, we can show for the sample of projections $\tilde{Y}_1, \ldots, \tilde{Y}_n$,

$$\frac{1}{d}s^2_{\tilde{Y}} \xrightarrow{d} \frac{1}{n-1}\sigma_y^2\left(\frac{\sigma_x^2}{m} + \frac{\sigma_y^2}{n}\right)\chi^2(n-1) \text{ as } d \to \infty$$

Thus, we have

$$\frac{1}{d}S_{m,n}(Z) = \frac{1}{d}\left(\frac{s^2_{\tilde{X}}}{m} + \frac{s^2_{\tilde{Y}}}{n}\right)$$

$$\xrightarrow{d} \frac{1}{m-1}\frac{\sigma_x^2}{m}\left(\frac{\sigma_x^2}{m}+\frac{\sigma_y^2}{n}\right)\chi^2(m-1)+\frac{1}{n-1}\frac{\sigma_y^2}{n}\left(\frac{\sigma_x^2}{m}+\frac{\sigma_y^2}{n}\right)\chi^2(n-1)$$

For the second part in Theorem A.2, we expand the sample variance of the projected values in the permuted group as follows:

$$\begin{aligned} s^2_{\tilde{Z}_{\pi(1:m)}} &= \frac{1}{m-1}\sum_{i=1}^{m}\left(\tilde{Z}_{\pi(i)} - \overline{\tilde{Z}_{\pi(1:m)}}\right)^2 \\ &= \frac{1}{m-1}\sum_{i=1}^{m}\left((Z_{\pi(i)} - \overline{Z_{\pi(1:m)}})\cdot(\bar{Z}_{\pi(1:m)} - \bar{Z}_{\pi(m+1:N)})\right)^2. \end{aligned}$$

Lemma A.2 shows $E(Z_{\pi(i)} - \bar{Z}_{\pi(1:m)})^{(k)}(\bar{Z}_{\pi(1:m)} - \bar{Z}_{\pi(m+1:N)})^{(k)}$ is nonzero. Now apply Lemma A.3 with $Z_1 = (Z_{\pi(i)} - \overline{Z_{\pi(1:m)}})$ and $Z_2 = (\bar{Z}_{\pi(1:m)} - \bar{Z}_{\pi(m+1:N)})$ to see that $\frac{1}{d^2}s^2_{\tilde{Z}_{\pi(1:m)}}$ converges in probability to a nonzero constant. A similar argument can be applied to $s^2_{\tilde{Z}_{\pi(m+1:N)}}$. Combining these results, it immediately follows that $\frac{1}{d^2}S_{m,n}(Z_\pi)$ converges in probability to a nonzero constant.

APPENDIX B: HDLSS GEOMETRY

In this section, we give some geometric intuition as to why the centroid-t-statistic behaves so differently in the permutation and the original world under equal means. Recall that under equal means the numerator in the centroid-t-statistic behaves similarly in the permutation world and the original world. Theorem A.2, however, shows that the denominator of the centroid-t is larger in the permuted world. This has the effect of making the unconditional distribution of the centroid-t-statistic larger than the permutation distribution.

To gain some intuition, consider the following toy HDLSS example. Suppose we observe $X_1, X_2 \sim F_1$ and $Y_1, Y_2 \sim F_2$ where $F_1 = N(0, I_d)$ and $F_2 = N(0, \sigma^2 I_d)$, $\sigma^2 \neq 1$. The points X_1, X_2, Y_1, Y_2 form the vertices of a tetrahedron in three-dimensional space. The two-dimensional plane generated by Y_1, Y_2, and $\bar{X}$ is shown in Figure 8.

Distances between elements of interest are calculated using HDLSS asymptotics in the manner of Hall, Marron, and Neeman (2005), which we briefly review here. Let $X \sim N(0, \sigma_x^2 I_d)$ and $Y \sim N(0, \sigma_y^2 I_d)$. We have simply by definition

$$||X - Y||^2/(\sigma_x^2 + \sigma_y^2) \sim \chi^2(d).$$

Then by the central limit theorem,

$$\sqrt{d}\left(\frac{||X-Y||^2/(\sigma_x^2+\sigma_y^2)}{\sqrt{2}d} - \frac{1}{\sqrt{2}}\right) \to N(0,1)$$

as $d \to \infty$. Applying the Delta Method, we get

$$\sqrt{d}\left(\frac{||X-Y||}{2^{1/4}\sqrt{(\sigma_x^2+\sigma_y^2)d}} - \frac{1}{2^{1/4}}\right) = O_P(1)$$

and thus

$$||X - Y|| = \sqrt{(\sigma_x^2+\sigma_y^2)d} + O_P(1).$$

In the diagrams, all distances have an additional $O_P(1)$ term that is not shown to avoid clutter. The geometric configuration in Figure 8 has the implication that $s^2_{\bar{Y}}$ is small. To see this, note the projections of Y_1 and Y_2 onto the centroid direction $\bar{X} - \bar{Y}$ is close to the projection of $\bar{Y}$ itself. A similar argument can be applied to show $s^2_{\bar{X}}$ is small.

Now let us look at what happens in the permutation world. Figure 9 shows the two-dimensional plane generated by the realization of a random permutation where $X_1^* = X_2$, $X_2^* = Y_2$ and $Y_1^* = X_1$ and $Y_2^* = Y_1$. Notice that the distance between Y_1^* and $\bar{X}^*$ is different than the distance between Y_2^* and $\bar{X}^*$. This has the effect of making $s^2_{\bar{Y}^*}$, the sample variance of the projections of Y_1^* and Y_2^*, large. To see this, note the projections of Y_1^* and Y_2^* onto the permuted centroid direction are not close to the projection of $\bar{Y}^*$. A similar argument can be applied to show $s^2_{\bar{X}^*}$, the sample variance of the projections of X_1^* and X_2^*, is large.

ACKNOWLEDGMENTS

The work presented in this article was supported in part by the NSF Graduate Fellowship and NIH grant T32 GM067553-05S1.

[Received August 2013. Revised February 2015.]

REFERENCES

Bai, Z., and Saranadasa, H. (1996), “Effect of High Dimension: By an Example of a Two Sample Problem,” *Statistica Sinica*, 6, 311–329. [551,558]

Baringhaus, L., and Franz, C. (2004), “On a New Multivariate Two-Sample Test,” *Journal of Multivariate Analysis*, 88, 190–206. [551]

——— (2010), “Rigid Motion Invariant Two-Sample Tests,” *Statistica Sinica*, 20, 1333–1361. [551]

Bickel, P. J., and Breiman, L. (1983), “Sums of Functions of Nearest Neighbor Distances, Moment Bounds, Limit Theorems and a Goodness of Fit Test,” *The Annals of Probability*, 11, 185–214. [551]

Biswas, M., and Ghosh, A. K. (2014), “A Nonparametric Two-Sample Test Applicable to High Dimensional Data,” *Journal of Multivariate Analysis*, 123, 160–171. [556]

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, 265–277. [558]

——— (2014), “Two-Sample Test of High Dimensional Means Under Dependence,” *Journal of the Royal Statistical Society*, Series B, 76, 349–372. [552,557,559,560]

Carey, L. A., Perou, C. M., Livasy, C. A., Dressler, L. G., Cowan, D., Conway, K., Karaca, G., Troester, M. A., Tse, C. K., Edmiston, S., Deming, S. L., Geradts, J., Cheang, M. C., Nielsen, T. O., Moorman, P. G., Earp, H. S., and Millikan, R. C. (2006), “Race, Breast Cancer Subtypes, and Survival in the Carolina Breast Cancer Study,” *The Journal of the American Medical Association*, 295, 2492–2502. [562]

Chen, S. X., and Qin, Y.-L. (2010), “A Two-Sample Test for High-Dimensional Data With Applications to Gene-Set Testing,” *The Annals of Statistics*, 38, 808–835. [551,558]

Chung, E. Y., and Romano, J. P. (2013), “Exact and Asymptotically Robust Permutation Tests,” *Annals of Statistics*, 41, 484–507. [556]

Ernst, M. D. (2004), “Permutation Methods: A Basis for Exact Inference,” *Statistical Science*, 19, 676–685. [555]

Friedman, J. H., and Rafsky, L. C. (1979), “Multivariate Generalizations of the Wald-Wolfowitz and Smirnov Two-Sample Tests,” *The Annals of Statistics*, 7, 697–717. [551]

Ghosh, A. K., and Biswas, M. (2013), “Exact Distribution-Free Two-Sample Tests Applicable to High Dimensional Data,” unpublished manuscript. [551,556,557,559]

Hall, P., Marron, J. S., and Neeman, A. (2005), “Geometric Representation of High Dimension, Low Sample Size Data,” *Journal of the Royal Statistical Society*, Series B, 67, 427–444. [556,567]

Hall, P., and Tajvidi, N. (2002), “Permutation Tests for Equality of Distributions in High-Dimensional Settings,” *Biometrika*, 89, 359–374. [551]

Hastie, T., Tibshirani, R., and Friedman, J. (2001), *The Elements of Statistical Learning*, New York: Springer-Verlag. [553]

Henze, N. (1988), “A Multivariate Two-Sample Test Based on the Number of Nearest Neighbor Type Coincidences,” *The Annals of Statistics*, 16, 772–783. [551]

Huang, Y., Xu, H., Calian, V., and Hsu, J. C. (2006), “To Permute or Not to Permute,” *Bioinformatics*, 22, 2244–2248. [556]

Jung, S., and Marron, J. (2009), “PCA Consistency in High Dimension, Low Sample Size Context,” *The Annals of Statistics*, 37, 4104–4130. [556]

Lopes, M., Jacob, L., and Wainwright, M. J. (2011), “A More Powerful Two-Sample Test in High Dimensions Using Random Projection,” in *Advances in Neural Information Processing Systems,* eds. M. Lopes, L. Jacob, and M. Wainwright, pp. 1206–1214. [552]

Marron, J., Todd, M. J., and Ahn, J. (2007), “Distance-Weighted Discrimination,” *Journal of the American Statistical Association*, 102, 1267–1271. [553]

Schilling, M. F. (1986), “Multivariate Two-Sample Tests Based on Nearest Neighbors,” *Journal of the American Statistical Association*, 81, 799–806. [551]

Srivastava, M. S., and Du, M. (2008), “A Test for the Mean Vector With Fewer Observations Than the Dimension,” *Journal of Multivariate Analysis*, 99, 386–402. [551,558]

Szekely, G. J., and Rizzo, M. L. (2004), “Testing for Equal Distributions in High Dimension,” *InterStat*, 5. [551,558]

Welch, B. L. (1947), “The Generalization of ‘Student’s’ Problem When Several Different Population Variances are Involved,” *Biometrika*, 34, 28–35. [563]