DiProPerm

Susan Wei

University of North Carolina - Chapel Hill

susanwe@live.unc.edu

October 9, 2012

Two Sample Problems - Equality of Means

(a) William Sealy Gosset: two sample t-test

(b) Harold Hotelling: Hotelling T^2 test

The Challenge

High Dimensional Low Sample Size (HDLSS) data. Hotelling T^2 completely breaks down when dimension exceeds sample size. Most existing methods for HDLSS data try to "fix" the classical Hotelling T^2 by some kind of diagonalization of the covariance matrix.

Statement of Problem

We observe $\{X_1, \ldots, X_m\}$ from distribution F_1 and $\{Y_1, \ldots, Y_n\}$ from distribution F_2. We are interested in testing

$$H_0 : F_1 = F_2 \text{ versus } H_1 : F_1 \neq F_2$$

and the weaker hypothesis

$$H_0 : \mu(F_1) = \mu(F_2) \text{ versus } H_1 : \mu(F_1) \neq \mu(F_2)$$

DiProPerm

- Three step framework: Direction-Projection-Permutation

1. project samples onto an appropriate direction
2. calculate univariate two sample statistic
3. assess significance using permutation test

DiProPerm - Direction

- Train binary linear classifier on the original class labels.
- Possibilities include Mean Difference direction, SVM, DWD, etc.

Di**Pro**Perm - Projection

- Project data onto the normal vector of the separating hyperplane
- Calculate univariate two-sample statistic on the projections
- Possibilities include two-sample t-statistic, difference of sample means, etc.

DiPro**Perm** - Permutation

- Assess significance of this univariate two-sample statistic by permutation test

1. Permute class membership
2. Re-train binary linear classifier
3. Re-calculate univariate two-sample statistic

- For level α test, reject H_0 if original statistic is among $100\alpha\%$ largest among all permuted statistics

Remarks

- PCA for Direction step?
- Why use permutation test in the last step?

Features of DiProPerm

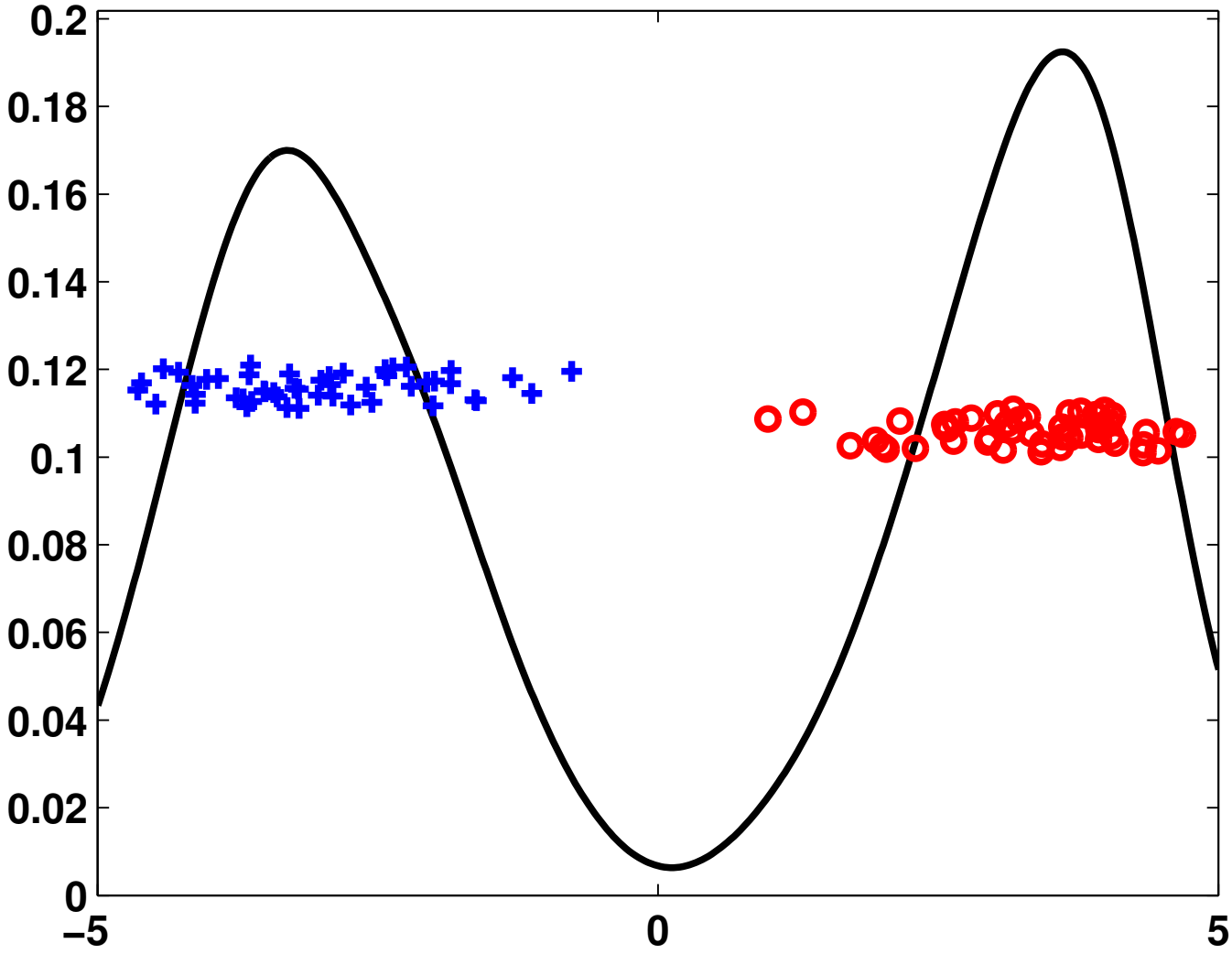
- Nonparametric test
- Intrinsically linked to data visualization
- Designed for high dimensional low sample size datasets

The Many Flavors of DiProPerm

Direction	**Univariate Statistic**
Mean Difference (MD)	Difference of Sample Means (MD)
SVM	Two-sample t-statistic (t)
DWD	Difference of Sample Medians
Fisher's LD	Area Under the Curve

Toy Example

$F_1 = F_2 =$ standard multivariate Normal with dimension 1000 and sample sizes $m = n = 50$



Toy Example Continued

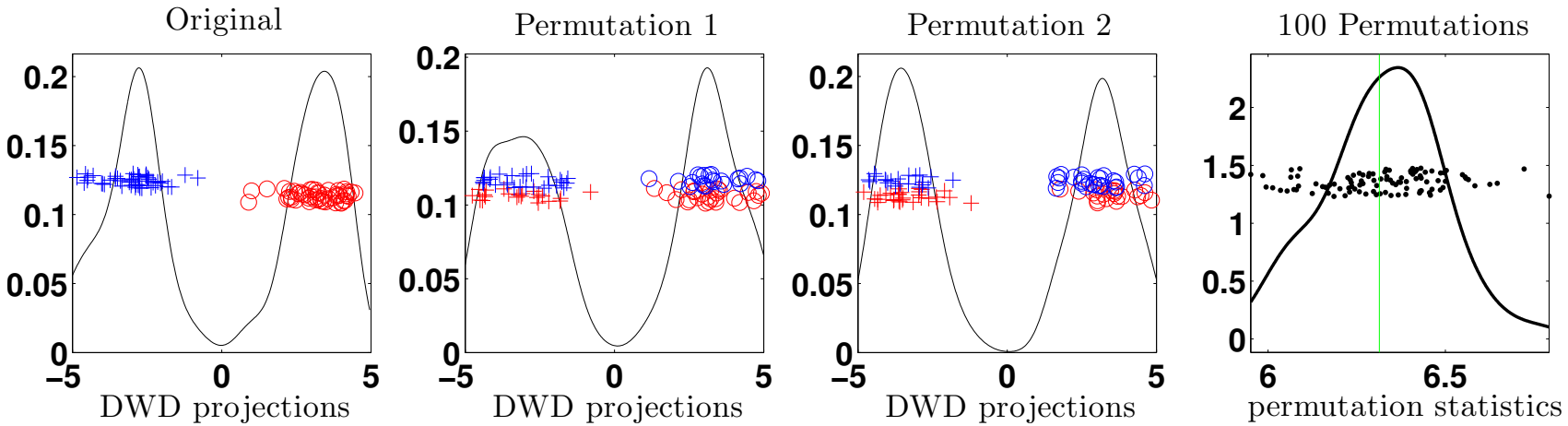


Figure: Colors represent original class labels. Symbols represent permuted class labels.

Validity

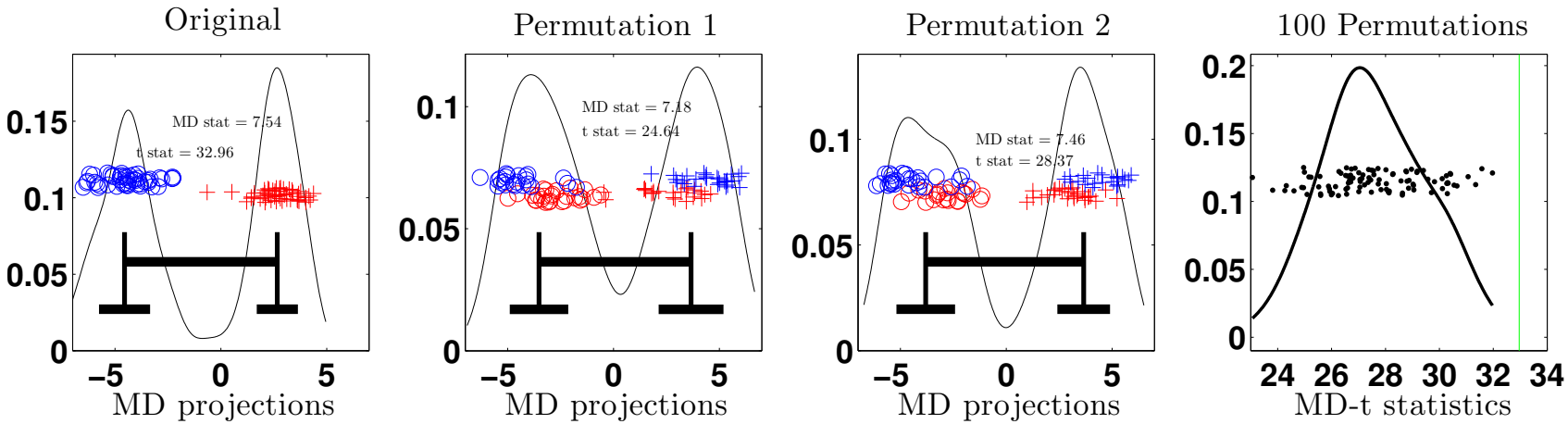
Idea originally developed by Chihoon Lee in 2007 consulting course. Dr. Marron's favorite flavor of DiProPerm is DWD-t, uses it to test equality of means. But DiProPerm is ultimately a permutation test. In general, can permutation tests be used to test equality of means?

Permutation Test

The null hypothesis is equality of distributions since only under this null assumption can labels be exchanged. If permutation test is used to test equality of means, it is very likely to obtain a test with seemingly high power which is actually invalid.
Then is it hopeless to use DiProPerm to test equality of means? It turns out this is possible with a careful choice of the univariate statistic in the Projection step of DiProPerm.

The choice of the univariate statistic

Simulation: $N(0, I_d)$ versus $t(5)^d$ for $d = 1000$.



MD statistic is similar across all three panels. Two-sample t-statistic is much higher in the first panel. This simulation suggests MD-MD test is valid for testing equality of means while MD-t is not.

Asymptotic Validity

Was able to provide a proof that MD-MD is asymptotically valid for testing equality of means while MD-t is not under the following conditions

- The distributions F_1 and F_2 are Gaussian with spherical covariance structure
- Asymptotic regime of dimension going to infinity for fixed sample size

Asymptotic Validity

The theorem predicts that the MD-t statistic is of the order $\sqrt{d}$ in the original world and of the order 1 in the permutation world.

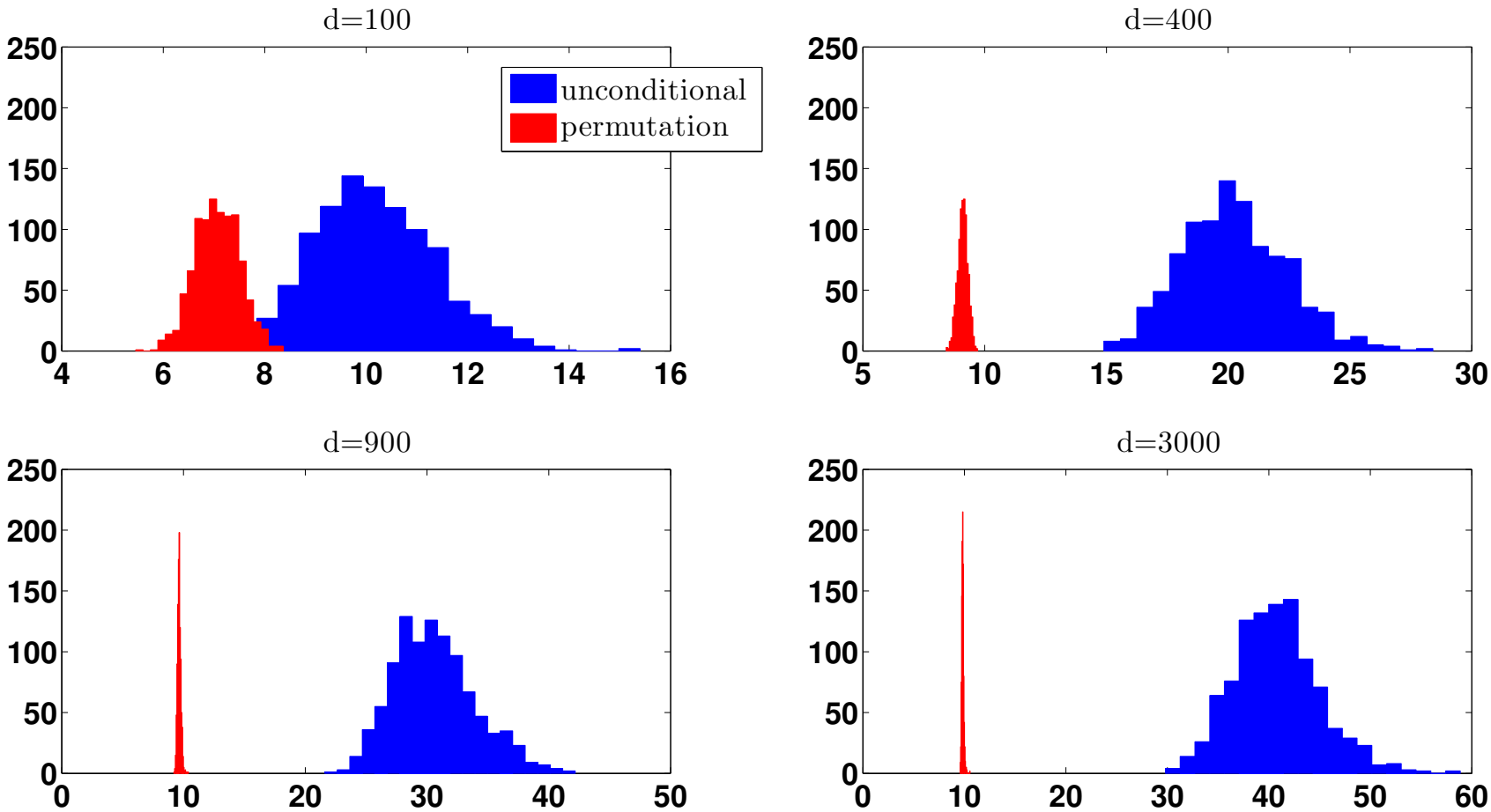


Figure: The unconditional and permutation distribution of the MD-t statistic for the distributions $F_1 = N(0, I_d)$ and $F_2 = N(0, 100 I_d)$, sample sizes $m = n = 50$.

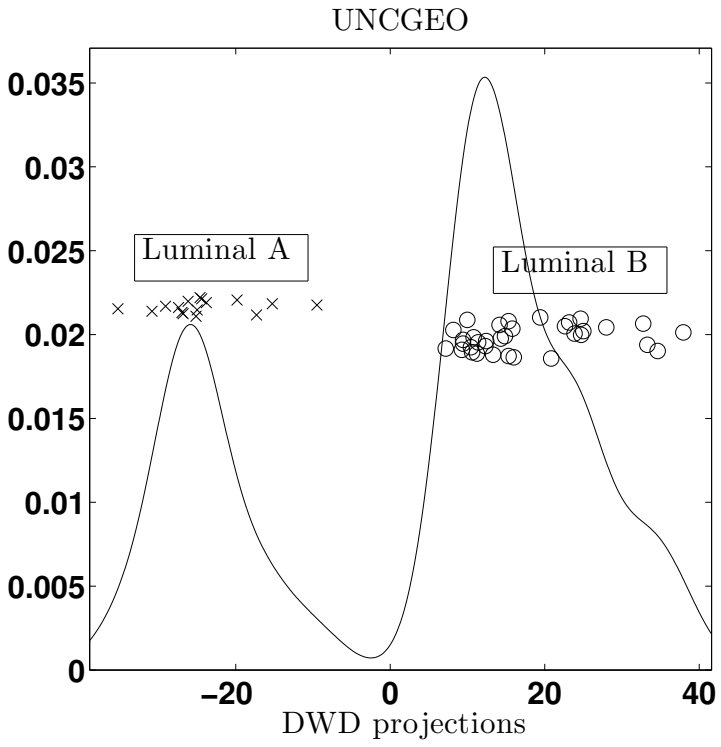
Asymptotic Validity

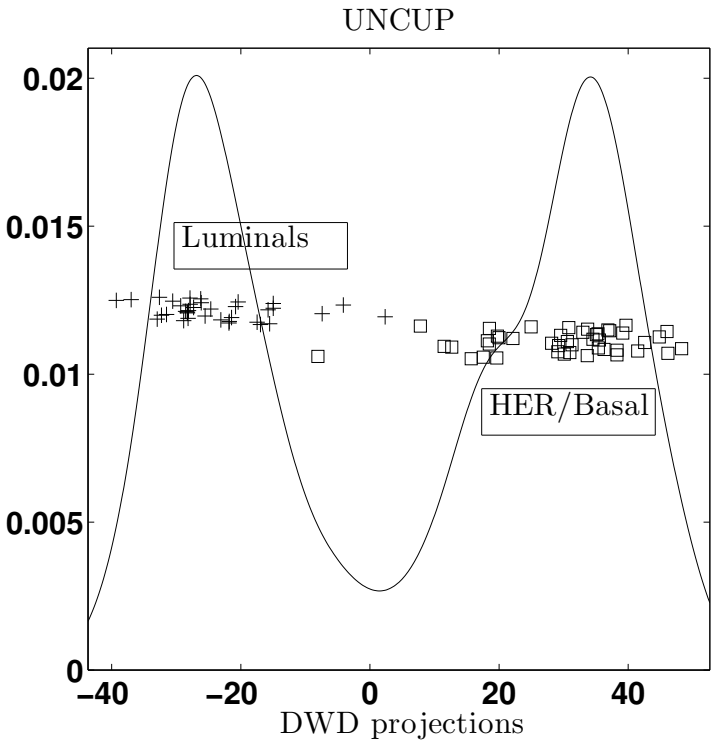
Optimistic that the theorem also extends to directions besides MD direction. Reason is all binary linear classifiers "look alike" in high dimensions.

Application

- Breast Cancer Dataset from Chuck Perou
- UNCGEO study: 50 patients
- UNCUP study: 80 patients
- 9674 genes measured on each patient

Application: Projection plots





Application: Test results