Journal of Computational and Graphical Statistics

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/ucgs20

# Measure of Strength of Evidence for Visually Observed Differences between Subpopulations

Xi Yang, Jan Hannig, Katherine A. Hoadley, Iain Carmichael & J.S. Marron

**To cite this article:** Xi Yang, Jan Hannig, Katherine A. Hoadley, Iain Carmichael & J.S. Marron (02 Nov 2023): Measure of Strength of Evidence for Visually Observed Differences between Subpopulations, Journal of Computational and Graphical Statistics, DOI: 10.1080/10618600.2023.2276113

**To link to this article:** https://doi.org/10.1080/10618600.2023.2276113

# Measure of Strength of Evidence for Visually Observed Differences between Subpopulations

Xi Yang[a], Jan Hannig[a], Katherine A. Hoadley[b], Iain Carmichael[c], J.S. Marron[a]

[a]Department of Statistics and Operations Research, University of North Carolina Chapel Hill, USA

[b]Department of Genetics, University of North Carolina Chapel Hill, USA

[c]Department of Statistics, University of California Berkeley, USA

*yangximath@gmail.com

***Abstract***

**For measuring the strength of visually-observed subpopulation differences, the Population Difference Criterion is proposed to assess the statistical significance of visually observed subpopulation differences. It addresses the following challenges: in high-dimensional contexts, distributional models can be dubious; in high-signal contexts, conventional permutation tests give poor pairwise comparisons. We also make two other contributions: Based on a careful analysis we find that a balanced permutation approach is more powerful in high-signal contexts than conventional permutations. Another contribution is the quantification of uncertainty due to permutation variation via a bootstrap confidence interval. The practical usefulness of these ideas is illustrated in the comparison of subpopulations of modern cancer data.**

*Keywords:* balanced permutations; confidence intervals; correlation adjustment; high dimension, population criterion difference.



## 1 Introduction

In the age of Big Data, many contexts involve analyzing and understanding relationships between multiple subpopulations. A fascinating and particularly deep example of this comes from cancer research as illustrated in Section 4, where we have gene expression data from five cancer types and a goal of understanding relationships between the cancer types (aka subpopulations). A common approach to understanding subpopulation differences is visualization

using Principal Component Analysis (PCA) Jolliffe (1986). While this method gives useful insights, visual approaches can be deceptive. This motivates our work to develop a method for quantification of the strength of the evidence for distinction between each pair of groups. Because of the very rich general structure of biological processes, classical statistical distributional models are woefully inadequate. Hence permutation testing methods, such as the DiProPerm test proposed by Wei et al. (2016), provide an appealing alternative.

The DiProPerm method measures strength of difference between any two subpopulations by projecting the data onto a direction aimed at separating them, and summarizing using the difference of the projected means as a statistic. Typical permutation p-values are calculated using the number of permuted summaries that lie outside the corresponding true data summary statistic. Bioinformatics data is frequently *high signal*, meaning there are usually very strong differences between subpopulations. The challenge of high signal situations is that there frequently are no permuted statistics outside that range, so the only conclusion is that the permutation p-value is $< \frac{1}{n_P}$, where $n_P$ is the number of permutations. From the classical viewpoint, that is strong evidence of all such differences being significant which is the end of the story. However, understanding the relationship between subpopulations is a different challenge. This motivates our invention of the *Population Difference Criterion* (PDC), which is a quantitative measure of separation between subpopulations that provides meaningful comparisons even in high dimensional and high signal contexts. A very important example of the value of PDC is in comparison of the many possible pre-processing operations that are routinely used for example in bioinformatics applications. In particular, better pre-processing methods are those which result in a larger PDC for a given set of subpopulations of interest.

A major contribution of this paper discussed in Section 2 is the discovery of a peculiar phenomenon that in high signal situations increasing signal strength can

actually entail the loss of statistical power as measured by the PDC. Detailed mathematical analysis reveals that this is caused by the traditional permutation scheme employed in DiProPerm Wei et al. (2016). This motivates our proposal of a non-standard *balanced permutation* scheme. Comparisons using simulated and real data show that balanced permutations provide more powerful results.

Yet another contribution of this paper appears in Section 3, where we propose confidence intervals that account for the Monte Carlo uncertainty in permutation testing. The value of quantifying that uncertainty for comparing multiple cancer subpopulations is demonstrated in Section 4. Discussion of controversies related to non-traditional permutation schemes can be found in Section 5.

## 2 Population Difference Criterion

Consider data from two potentially high-dimensional populations $X_1, \ldots, X_m$ and $Y_1, \ldots Y_n$. As explained in the introduction, a common way of visualizing the difference between the subpopulations uses projections on a given direction determined by a unit vector *w*, e.g., a PCA direction. A particularly useful visual direction for distinguishing subpopulations is the Distance Weighted Discrimination (DWD) direction vector *w* proposed by Marron et al. (2007). DWD solves an optimization problem that is formally stated in Appendix A. A potential drawback is that DWD can be relatively slow to compute. Therefore, we will also investigate a computationally faster version of DiProPerm based on projection onto the Mean Difference (MD) direction, i.e. $w = \frac{\bar{X}_m - \bar{Y}_n}{\| \bar{X}_m - \bar{Y}_n \|}$ – the direction pointing from the mean of one group to the mean of the other.

While such visualizations are suggestive, they can also be deceptive. As noted in Wei et al. (2016), rigorous quantification of visual differences can be surprisingly counter intuitive, because human intuition is not good at incorporating issues such as sample size and high dimensional variation into visual impression. An explicit example of this is shown in Figure 2.3 of the PhD dissertation of

Yang (2021). Hence it is very important to provide a quantitative measure of the strength of the evidence for visual subpopulation separation based on rigorous statistical inference.

An early version of a quantitative measure of this type was provided by the DiProPerm test (Wei et al. (2016)). The DiProPerm test statistic is the observed mean difference of the projected scores:

$$C = \left| \frac{1}{m} \sum_{j=1}^{m} \langle w, X_k \rangle - \frac{1}{n} \sum_{k=1}^{n} \langle w, Y_k \rangle \right|, \qquad (1)$$

where $\langle w, x \rangle$ is the inner product.

The *Population Difference Criterion* (PDC) is a quantitative measure of the differences between subpopulations that may be visually apparent in PCA scatter plots. Specifically,

$$PDC = \frac{C - EC}{\sqrt{\mathrm{Var}C}},$$

where the mean and variance of *C* are computed under the null model of no difference between subpopulations. Larger values of PDC represent stronger evidence of the difference between subpopulations.

Traditionally, the null distribution *EC* and $\mathrm{Var}C$ have been estimated using permutation based methods. In particular,

$$PDC = \frac{C - \bar{C}}{S}, \qquad (2)$$

where the mean $\bar{C}$ and standard deviation *S* are estimated using re-sampling of the class labels and re-projection of the data. Note that if the null distribution of *C* was Gaussian, the PDC would be the classical Z-score and was called that in

Wei et al. (2016). But that term is not used here as the null distributions of $C$ could be far from Gaussian.

### 2.1 Gaussian model

In this section, we study the behavior of the DiProPerm PDC using a basic two-class Gaussian model. Both classes are assumed to follow the multivariate normal distribution with means separated by $2g$, i.e.,

$$\begin{aligned} &\text{Class +1 } (X): \quad X_j \sim N_d(g \cdot u, \sigma^2 I), \quad j = 1, \ldots, m \\ &\text{Class -1 } (Y): \quad Y_k \sim N_d(-g \cdot u, \sigma^2 I), \quad k = 1, \ldots, n \end{aligned} \tag{3}$$

where $u \in \mathbb{R}^d$ is some unit vector, such as $[1/\sqrt{d} \cdots 1/\sqrt{d}]^T$ or $[1\ 0 \cdots 0]^T$, and $\sigma \in \mathbb{R}^+$. The actual direction of $u$ is irrelevant because the DiProPerm test is rotation invariant.

An important component of our mathematical analysis is the derivation of the distribution of the observed statistics (1) under the model (3). Using the MD direction

$$C^2 = \| \bar{X} - \bar{Y} \|^2 = \sum_{k=1}^{d} (\bar{X}^k - \bar{Y}^k)^2 \sim (\frac{1}{m} + \frac{1}{n}) \sigma^2 \chi_d^2 (\frac{4g^2}{(\frac{1}{m} + \frac{1}{n})})$$

and so

$$C = \| \bar{X} - \bar{Y} \| \sim \sigma \sqrt{\frac{1}{m} + \frac{1}{n}} \chi_d \left( \sqrt{\frac{4g^2}{(\frac{1}{m} + \frac{1}{n})}} \right). \tag{4}$$

Here $\chi_d(\lambda)$ is called the *non-central Chi distribution* (Johnson et al., 1972), the square root of the non-central Chi-square distribution with degrees of freedom $d$ and non-centrality parameter $\lambda^2$.

Under the null hypothesis $g = 0$ we have $EC = \sigma\sqrt{\frac{1}{m}+\frac{1}{n}}\sqrt{2}\frac{\Gamma((d+1)/2)}{\Gamma(d/2)}$ and $\mathrm{Var}C = \sigma^2(\frac{1}{m}+\frac{1}{n})(d - 2(\frac{\Gamma((d+1)/2)}{\Gamma(d/2)})^2)$. In general,

$$E(\chi_d(\lambda)) = \sqrt{\frac{\pi}{2}}\cdot L_{1/2}^{d/2-1}(\frac{-\lambda^2}{2}),$$

where $L_n^a(z)$ is the generalized Laguerre polynomial (Koekoek and Meijer, 1993). Then we have

$$E[||\bar{X} - \bar{Y}||] = \sigma\sqrt{\frac{1}{m}+\frac{1}{n}}\cdot\sqrt{\frac{\pi}{2}}\cdot L_{1/2}^{d/2-1}(-\frac{2g^2}{(\frac{1}{m}+\frac{1}{n})}) \qquad (5)$$

$$Var[||\bar{X} - \bar{Y}||] = \sigma^2(\frac{1}{m}+\frac{1}{n})(d + \frac{4g^2}{(\frac{1}{m}+\frac{1}{n})} - \frac{\pi}{2}\cdot(L_{1/2}^{d/2-1}(-\frac{2g^2}{(\frac{1}{m}+\frac{1}{n})}))^2) \qquad (6)$$

**2.2 Permutation distribution**

Since in practice, most people use the estimated PDC (2), it is important to study the behavior of the test statistic *C* under various permutation schemes. The classical approach is to randomly reshuffle the class labels $\{-1,1\}$ for all observations. This generates permuted data $X_{j,per},\ j = 1,\ldots,m$ and $Y_{k,per},\ k = 1,\ldots,n$. We call this approach *all permutations*. Let $C_i,\ i = 1,\cdots,N$ be independent realizations of the test statistic (1) computed using permuted data. Again, in the case of the MD direction $C_i = ||\bar{X}_{per} - \bar{Y}_{per}||,\quad i = 1,\cdots,N.$

When labels are randomly reshuffled, there is a random number $R_i$ of observations in each class that switch labels. Note that when all permutations are used, $R_i$ is a Hypergeometric random variable whose probability mass function is:

$$p_R(r) = P(R_i = r) = \frac{\binom{m}{r}\binom{n}{n-r}}{\binom{m+n}{m}}, \quad r = 0, 1, \ldots, \min(m, n).$$

The conditional distribution of $\bar{X}_{per} - \bar{Y}_{per}$ is (detailed derivation shown in Appendix B):

$$(\bar{X}_{per} - \bar{Y}_{per}) \mid R_i = r \sim N_d(2g \cdot (1 - \frac{r}{m} - \frac{r}{n}) \cdot u, \sigma^2 (\frac{1}{m} + \frac{1}{n}) I_d). \quad (7)$$

Thus, for a given permutation, $\frac{R_i}{m}$ is the proportion of the original Class -1 cases that are relabeled as the permuted Class +1, and $\frac{n - R_i}{n}$ is the proportion of the Class -1 cases that remain in the new Class -1. The difference between these 2 proportions, denoted as

$$\xi_i = \xi(R_i) = \frac{n - R_i}{n} - \frac{R_i}{m} = 1 - \frac{R_i}{m} - \frac{R_i}{n}, \quad (8)$$

quantifies the class balance of this permutation. Hence we call it the *coefficient of unbalance*. Those permutations with $\xi_i = 0$ are called *balanced permutations*.

To study the theoretical behavior of MD PDC we need to calculate the mean and variance of $\| \bar{X}_{per} - \bar{Y}_{per} \|$ under the model (3). First define the following notation:

$$g_r = g \cdot \xi(r) = 2g \cdot (1 - \frac{r}{m} - \frac{r}{n}), \qquad \lambda_r = \sqrt{\frac{4g_r^2}{\sigma^2 (\frac{1}{m} + \frac{1}{n})}}.$$

The unconditional distribution of $\bar{X}_{per} - \bar{Y}_{per}$ is a normal mixture which is related to Theorem 1 of Wei et al. (2016):

$$\bar{X}_{per} - \bar{Y}_{per} \sim \sum_{r=0}^{\min(m,n)} p_R(r) N_d(2 g_r \cdot u, \sigma^2 (\frac{1}{m} + \frac{1}{n}) I_d).$$

The permutation null distribution is

$$C_i = || \bar{X}_{per} - \bar{Y}_{per} || \sim \sigma \sqrt{\frac{1}{m} + \frac{1}{n}} \sum_{r=0}^{\min(m,n)} p_R(r) \chi_d(\lambda_r). \qquad (9)$$

Thus $C_i$ has a mixture distribution with the mixture component driven by the coefficient of unbalance $\xi_i$. Then we have

$$E(|| \bar{X}_{per} - \bar{Y}_{per} ||) = \sigma \sqrt{\frac{\pi}{2}} \cdot \sqrt{\frac{1}{m} + \frac{1}{n}} \sum_{r=0}^{m} p_R(r) L_{1/2}^{d/2-1}(\frac{-\lambda_r^2}{2}) \qquad (10)$$

$$Var(|| \bar{X}_{per} - \bar{Y}_{per} ||) = \sigma^2 (\frac{1}{m} + \frac{1}{n}) \sum_{r=0}^{m} p_R(r)(d + \lambda_r^2) - E(|| \bar{X}_{per} - \bar{Y}_{per} ||)^2 \qquad (11)$$

Note that, the sample sizes $m$, $n$ are inversely related to the sample variances.

Using (5), (10), (11) define the PDC function

$$f(m,n,d,g) := \frac{E(C) - E(C_i)}{\sqrt{Var(C_i)}} = \frac{E[|| \bar{X} - \bar{Y} ||] - E(|| \bar{X}_{per} - \bar{Y}_{per} ||)}{\sqrt{Var(|| \bar{X}_{per} - \bar{Y}_{per} ||)}}. \qquad (12)$$

This function provides important lessons about the behavior of the all permutation based PDC under the Gaussian model. In particular, careful examination of $f(m,n,d,g)$ viewed as a function of $g$ (with $m$, $n$, $d$ fixed) shows that the function first increases and then decreases, eventually converging to

$$\lim_{g \to \infty} f(m,n,d,g) = \frac{1 - \sum_{r=0}^{m} p_R(r) |1 - r/m - r/n|}{\sqrt{\frac{1}{m+n-1} - (\sum_{r=0}^{m} p_R(r) |1 - r/m - r/n|)^2}}. \qquad (13)$$

(see Appendix C). This behavior is a serious deficiency of the traditional all permutation approach since the power of the test fails to increase with stronger signal.

**2.3 Gaussian model simulation**

In this section, we demonstrate the behavior of the DiProPerm PDC using a simulation based on the Gaussian model (3). In each of our simulation scenarios we generate samples from the two populations of size $m = n = 100$. We consider three very different dimensions $d = 1, 10, 100,$ and 200 values of signal strength $g$ ranging equally between 0 and 20. The PDC values (2) are calculated using the mean and variance of $N = 100$ permutations. Each $d$, $g$ combination is replicated 100 times for the PDC calculated using MD and only 10 times using DWD due to the longer running time of DWD.

The results are summarized in Figure 1. The left and right panels show the PDC calculated using the MD and DWD directions respectively. In both panels the horizontal axis shows the signal strength $g$ and the vertical axis shows the DiProPerm PDC. The solid curves in both panels show local linear regression estimates (Fan and Gijbels, 1996) of the PDC samples vs signal strength $g$. The dashed curve on the left shows the theoretical PDC value $f(m, n, d, g)$ viewed as the function of $g$ computed in (12). Each dot is a single realization of the estimated PDC value, reflecting its variation due to randomness.

When using the MD direction and $d > 1$ (left panel), the PDC first goes up (as expected from increasing signal strength), then goes down (which is quite surprising), and finally converges to the limit given by (13). When using the DWD direction the PDC also levels off instead of increasing, as one would naïvely expect, with increasing signal strength.

The non-intuitive behavior of PDC computed using all permutations is further explored in Figure 2 by taking a deeper look at three particular data sets with $d = 100$, $g = 2, 4, 20$. These 3 values of $g$ represent increasing, peak, and decreasing

regions of the PDC computed using the MD direction. They are highlighted as black stars in Figure 1 and correspond to the 3 columns in Figure 2. The permuted distributions are shown in the top three panels of Figure 2 using a jitter plot (Tukey (1976)), where on the horizontal axis we plot the test statistic $C_i$, the independent realizations of the test statistic (1) computed using permuted data and the MD and DWD direction respectively, and on the vertical axis we plot a random height for visual separation. The colors of the dots in these 3 top panels represent the absolute value of coefficient of unbalance $|\xi_i|$ of the $i$th permutation, using the color bar shown in Figure 3. The black curves are kernel density estimates (KDE, i.e., a smooth histogram, Wand and Jones (1994)) of the distribution of the permuted test statistics (dots). The numbers on the vertical axes are the height of the kernel density estimate.

From left to right, the kernel density estimates become more skewed and multi-modal in agreement with the fact that the all permutation null distributions are mixtures of chi-distributions (9). This is particularly apparent in the top right panel. As the signal, $g$, gets stronger there is much more separation of colors based on $|\xi_i|$. In the top left panel, which has the weakest signal ($g = 2$), the colored dots are mostly mixed. In the top middle panel, as the signal strength ($g = 4$) increases, the colored dots separate more.

The multimodality is further explored in the bottom two rows which show the projections on the permuted directions for the smallest and largest $C_i$. In each case, the symbols represent the original class labels and the colors show the permuted labels, whose mean difference determines the direction. The $x$-axis is the projection scores on the permuted directions. The symbols in the middle are jitter plots and the heights of the symbols are random heights. The curves are kernel density estimates of the projection scores. The colors represent the permutated labels and symbols represent the original labels. Subdensities, corresponding to each permuted subpopulation, are shown using colors that correspond to the symbols. The $y$-axis shows the height of the KDE densities.

The middle row of Figure 2 shows the permutation with maximal $C_i$ for that $d$ and $g$ corresponding to the far-right circles in each top panel. Going from left to right, the permuted mean difference direction first separates the red/blue permuted class colors and then tends to separate the symbols (the original class labels). This direction essentially becomes the original mean difference direction of the non-permuted data for large $g$. This effect is usefully quantified by the angles between the observed mean difference and each permutation direction shown in each panel in the middle row. A large angle suggests a large discrepancy between the original mean difference and the corresponding permutation direction. The left panel $g = 2$ is separating the colors well and mixing up the symbols with a relatively large angle $56^\circ$, as intuitively expected from the permutation test. This results in a PDC reflecting no signal as expected from the permutation test. In the middle $g = 4$ panel, there is still some color separation but also a strong separation of the symbols, with a smaller angle, $33^\circ$. In the right $g = 20$ panel, the angle is very small, $8^\circ$, showing this direction is very close to the mean difference direction of the original data. Because of the true class difference and the large coefficient of unbalance, this results in PDC values that are much larger than would be expected under the null distribution of no signal ($g = 0$) which results in a strong loss of power.

The bottom three panels of Figure 2 show the permutation with minimal $C_i$ corresponding to the far-left stars in each top panel. However, because $|\xi_i| \approx 0$, the large $g$ doesn't affect these nearly balanced permutations as seen in the bottom panels, where the angles are relatively large and close to $90^\circ$ ($82^\circ, 79^\circ, 88^\circ$) as expected for random directions from the results of Hall et al. (2005). Hence the bottom panels show projections which are much more consistent with the null hypothesis of no signal.

**2.4 Proposed solution: Balanced Permutations**

As discussed above DiProPerm PDC has issues with power caused by the fact that the estimates of $EC$ and $\mathrm{Var}C$ under the null hypothesis is inflated when

using the traditional all permutations. In this section, we propose a solution to this issue by using only balanced permutations. Recall that a permutation is called balanced if $\xi_i = 1 - R^i \times (1/m + 1/n) \approx 0$, so the number of switches in labels is close to $\frac{mn}{m+n}$.

Figure 4 provides a simple example illustrating the difference between balanced and all permutations. There are 8 cases in each class shown as rows. The first row is colored using the real class labels followed by 7 permutations where symbols represent the true class labels and colors represent the permuted class labels as in the bottom 6 panels of Figure 2. The colors of the text on the right are in the spirit of the color bar in Figure 3 and the colored dots in the top panels in Figure 2. The top 3 permutations are all balanced permutations and in these cases solving the equation: $\xi_0 = 1 - R^0 \times (1/m + 1/n) = 0$ results in $R^0 = \frac{mn}{m+n} = \frac{8 \times 8}{8+8} = 4$. The bottom 4 permutations are all unbalanced. The original DiProPerm draws from all permutations, but the proposed improved DiProPerm only draws from balanced permutations as shown in Figure 4.

When there is a large separation between the centers of the two classes, using only the balanced permutations makes the mean of the permutation null distribution closer to zero. Therefore, it provides a more useful alternative distribution. In particular, under the Gaussian model (3), Equation (7) implies that for balanced permutations and the MD direction

$$(\overline{X}_{per} - \overline{Y}_{per}) \sim N_d(0, \sigma^2(\frac{1}{m} + \frac{1}{n})I_d).$$

Consequently,

$$E_b(\| \bar{X}_{per} - \bar{Y}_{per} \|) = \sqrt{2}\sigma \cdot \sqrt{\frac{1}{m} + \frac{1}{n}} \frac{\Gamma((d+1)/2)}{\Gamma(d/2)}$$

$$\mathrm{Var}_b(\| \bar{X}_{per} - \bar{Y}_{per} \|) = \sigma^2 (\frac{1}{m} + \frac{1}{n})[d - 2(\frac{\Gamma((d+1)/2)}{\Gamma(d/2)})^2],$$

which gives the balanced PDC curve as:

$$f_b(m,n,d,g) := \frac{E(C) - E_b(C_i)}{\sqrt{\mathrm{Var}_b(C_i)}} = \frac{E[\| \bar{X} - \bar{Y} \|] - E_b(\| \bar{X}_{per} - \bar{Y}_{per} \|)}{\sqrt{\mathrm{Var}_b(\| \bar{X}_{per} - \bar{Y}_{per} \|)}}. \quad (14)$$

Figure 5 compares PDCs computed using balanced vs. all permutations by adding the former to a part of Figure 1 for both MD and DWD directions. The lower curves labeled as *All permutations* are the same as the curves in Figure 1. The higher level curves, labeled as *Balanced permutations*, use colors and symbols analogous to Figure 1 with (14) replacing (12) for the dashed curve. Each dot is a single realization of the estimated balanced PDC value, reflecting its variation due to randomness. The two sets of curves give direct comparison between the original and proposed versions of the DiProPerm PDC. For small $g$ the PDCs overlap. When the signal reaches a certain level, the balanced PDCs continue increasing as expected (from the increased signal strength), and the all permutation PDCs reach a peak and then seem to decrease (MD) or stay constant (DWD). This indicates the balanced PDC is much more powerful than the all permutation PDC in the case of strong signals for both MD and DWD directions. A careful look at the axis labels shows that stronger signals are required to see this effect for DWD.

Next, we continue the investigation of the three cases studied in Figure 2. In Figure 6, the dashed curves are the derived theoretical null distribution using the MD direction, i.e. the scaled central $\chi_d$ distribution (4). The goodness of fit of that distribution is demonstrated by the red solid curves which are kernel density estimates of the red dots in the jitter plots, i.e. only the balanced permutations. The solid black curves are the kernel estimates of all permutations. The bottom

panels only show the red and black solid curves since the theoretical distribution using the DWD direction is much harder to derive. In both top and bottom panels, as the signal $g$ grows stronger, the distributions of all permutations (black curves) become more and more skewed but the distributions of balanced permutations (red dots/curves) stay the same and hence provide a much more useful null distribution. The skewness in the bottom panels is not as strong as in the top panels, indicating the DWD direction suffers less from the all permutations effect and has higher test power than using the MD direction.

## 3 Quantification of Permutation Sample Variation

While they provide useful comparisons between data sets, the estimated permutation p-values and PDCs inherit variation caused by random sampling from each set of permutations. In some cases, this variation can obscure important differences between classes which motivates careful quantification of this uncertainty using confidence intervals.

The DiProPerm PDC (2) is a random variable that depends on the permutation null distribution. Thus, a confidence interval for the PDC can be estimated by the upper and lower quantiles using bootstrap re-sampling methods. A general algorithm is based on $B$ repetitions. In our calculations $B = 100$:

1. Draw a $B \times N$ matrix where each row is a random sample (with replacement) from the $N$ permutations used in the original calculation of PDC. Calculate the sample means and variances of each row. This results in $B$ re-sampled means and variances, which are used to get $B$ re-sampled PDCs.
2. Find the upper and lower quantile of the PDCs based on the $B$ re-sampled PDCs in Step 2.

Note that this method is unrelated to the direction choices of DiProPerm, e.g., MD or DWD, and to the choice of balanced or all permutations.

Alternatively, as we discussed in Section 2 when the original data are close to normal, the permutation null distribution is a mixture the $\chi$ distributions (9). Thus, we can also estimate the distribution using the method of moments estimation based on the Welch–Satterthwaite approximation (Satterthwaite (1946)). This has been explored in the Section 3 of the PhD Dissertation Yang (2021). However, the normal assumption of the original data is often questionable, and hence the bootstrap re-sampling method is recommended.

## 4 TCGA Pan-Can Data

To demonstrate the proposed method, we consider gene expression for five different cancer types, one of which is very different from the rest and two of which are similar to each other. Here, we used a subset of the TCGA Pan-Cancer data representing 1523 cases from 5 cancer types including 12478 genes. The tissues came from different organs (hence different cancer types), and represent a useful cohort to illustrate our proposed method for quantitatively determining their level of similarity or dissimilarity (Hoadley et al., 2018; Hutter and Zenklusen, 2018). These cancer types will be contrasted in visualizations discussed below using the colors and symbols shown in Table 1.

Figure 7 shows the relationship between cancer types using a PCA scatter plot, i.e., the two-dimensional projection of the point cloud in $\mathbb{R}^{12478}$ onto the two directions of highest variation. The PC1 scores are plotted on the vertical axis and the PC2 scores on the horizontal axis. The liquid tumor LAML is very different from the rest, which are solid epithelial tumors. Within the epithelial tumors, READ and COAD appear visually quite overlapped, consistent with the fact that these cells come from organs in the same developmental process and often referred to as a single disease (colorectal cancer). The BLCA and BRCA are somewhat different but not as separated as BLCA and LAML.

Figure 8 shows differences between three representative pairs of the subpopulations using projections on the MD (top row) and DWD (bottom row)

directions. The subpopulations are indicated using symbols and colors described in Table 1. Each dot represents a projection of a case subject on the MD or DWD direction respectively. The value of the projection is displayed on the *x*-axis. The height of the dots are random for visual separation. The curves are kernel density estimates of the projection scores with subdensities corresponding to the subpopulations. The *y*-axis shows the height of KDE densities. The black text shows the corresponding all and balanced permutation PDC respectively. As discussed in Section 2.3 there is better visual separation of the subpopulations for DWD (bottom) than for MD (top) and the PDCs for balanced permutations are higher than all permutations. This effect is particularly pronounced for the BLCA versus LAML, where the signal is the strongest. Figure 8 shows broadly similar lessons to Figure 7: READ and COAD are rather overlapped (left panels); BRCA and BLCA are moderately different (middle panels) while LAML is very distinct from BLCA (right panels).

For all 10 pairs of TCGA cancer types, Figure 9 gives a comparison of the strength of separation using the PDC. The random permutation variability in estimating each PDC is reflected by a 95% confidence interval as developed in Section 3. Conventional single-sample confidence intervals are shown as thick lines, the thin lines are Bonferroni-adjusted for the fact that we have 10 intervals. The results based on MD are shown in black/gray and the results based on DWD are shown in blue/light blue. The PDCs are not the centers of the confidence intervals because the distributions of the permutation statistics are skewed. The PDCs computed using balanced permutations (circles) are much higher than the PDCs computed using all permutations (stars) showing the strong value of balanced permutations. Overall each DWD based PDC (blue) is higher than the corresponding MD based PDC (black) showing the utility of DWD over MD for distinguishing class differences in higher dimensions.

The PDC allows us to accurately quantify the strength of population difference in each pair. The confidence intervals allow us to statistically compare these

strengths of population difference across pairs. In Figure 9, all pairs involving LAML tend to have large PDCs, which is consistent with Figure 7 which shows that LAML (magenta) is the most distinct cancer type. Pairs including BRCA also have relatively large PDCs. This is consistent with the fact that BRCA has the largest sample size which leads to a smaller variance and thus stronger statistical significance. The PDCs for LAML vs. READ and LAML vs. BLCA and BRCA vs. COAD reflect similar amounts of population difference. Those PDCs are the smallest among all test pairs indicating the weakest difference among the considered comparisons as shown in Figure 9. This is consistent with the overlap of COAD and READ observed in Figures 7 and 8.

In cases with a strong signal, such as LAML vs. BRCA, LAML vs. COAD and BRCA vs. COAD, the balanced PDCs (gray/blue circles) are much larger than the corresponding PDCs computed using all permutations (gray/blue stars). This is consistent with the idea that when the signal is strong, all permutations will cause a loss of power (see Figure 5). When the signal is weak, such as COAD vs. READ, all and the balanced PDCs are small and similar to each other.

## 5 Discussion

Our recommendation of balanced permutations is somewhat opposite to the recommendation against balanced permutations in Southworth et al. (2009). They appeal to group theory and suggest that all permutations are generally superior to balanced permutations since balanced permutations tend to be anti-conservative, i.e. their reported p-values are too small. In particular, under their null hypothesis, the permutation distribution, e.g., distribution of the red dots in Figure 6 doesn't have enough extremely large values. Hemerik and Goeman (2018) provided adjustments that make the use of balanced permutations for p-value calculation valid. Appendix D derives an often negligibly small alternative adjustment to both types of permutation PDC that overcomes the anti-conservative problem.

Figure 1 reveals the strange behavior that under the alternative the power of the tests from all permutations can decrease as the signal strength increases. The much-improved power of balanced permutations is shown in Figure 5 where the balanced permutation power as measured by PDC is proportional to the signal strength. When the signal is weak, Figure 5 shows that the balanced and all permutations give very similar PDCs. Thus balanced permutations are superior to all permutations in large-signal cases which often arise in bioinformatics and have no or minor differences from all permutations in small-signal cases.

## 6 Acknowledgement

We thank anonymous reviewers whose constructive comments helped us to greatly improve the presentation of our results. Jan Hannig's research was supported in part by the National Science Foundation under Grant No. DMS-1916115, 2113404, and 2210337. J.S. Marron's research was partially supported by the National Science Foundation Grant No. DMS-2113404. Katherine A. Hoadley's research was partially supported by Grant No. U24 CA264021. The authors report there are no competing interests to declare.

## References

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3):273–297.

Fan, J. and Gijbels, I. (1996). *Local polynomial modelling and its applications*, volume 66 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, London.

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 (Statistical Methodology)*, 67(3):427–444.

Hemerik, J. and Goeman, J. (2018). Exact testing with random permutations. *Test*, 27(4):811–825.

Hoadley, K. A., Yau, C., Hinoue, T., Wolf, D. M., Lazar, A. J., Drill, E., Shen, R., Taylor, A. M., Cherniack, A. D., Thorsson, V., et al. (2018). Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer. *Cell*, 173(2):291–304.

Hutter, C. and Zenklusen, J. C. (2018). The cancer genome atlas: creating lasting value beyond its data. *Cell*, 173(2):283–285.

Johnson, N. L., Kotz, S., and Balakrishnan, N. (1972). *Continuous multivariate distributions*, volume 7. Wiley New York.

Jolliffe, I. T. (1986). Principal components in regression analysis. In *Principal component analysis*, pages 129–155. Springer.

Koekoek, R. and Meijer, H. (1993). A generalization of laguerre polynomials. *SIAM Journal on Mathematical Analysis*, 24(3):768–782.

Leone, F., Nelson, L., and Nottingham, R. (1961). The folded normal distribution. *Technometrics*, 3(4):543–550.

Marron, J. S., Todd, M. J., and Ahn, J. (2007). Distance-weighted discrimination. *Journal of the American Statistical Association*, 102(480):1267–1271.

Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. *Biometrics bulletin*, 2(6):110–114.

Southworth, L. K., Kim, S. K., and Owen, A. B. (2009). Properties of balanced permutations. *Journal of Computational Biology*, 16(4):625–638.

Tukey, J. W. (1976). Exploratory data analysis. 1977. *Massachusetts: Addison-Wesley*.

Wand, M. P. and Jones, M. C. (1994). *Kernel smoothing*. Chapman and Hall/CRC.

Wei, S., Lee, C., Wichers, L., and Marron, J. (2016). Direction-projection-permutation for high-dimensional hypothesis tests. *Journal of Computational and Graphical Statistics*, 25(2):549–569.

Yang, X. (2021). *Machine Learning Methods in Hdlss Settings*. PhD thesis, The University of North Carolina at Chapel Hill.



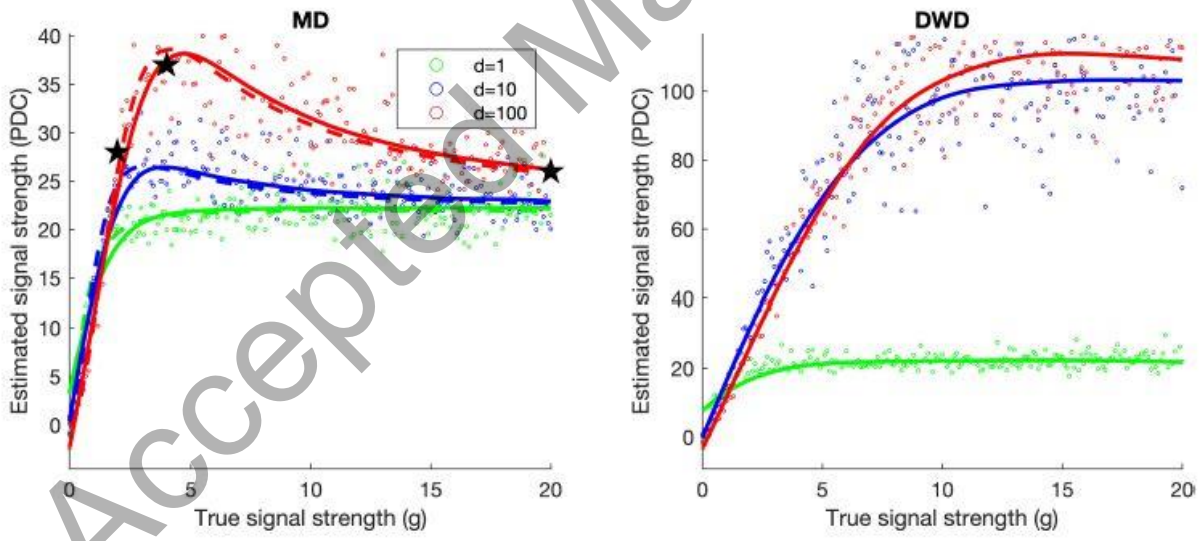


**Fig. 1** Test power as indicated by PDC, with MD on the left and DWD on the right, for different choices of *d* (shown with colors) and signal strength *g* (x-axes). The y-axes show the PDC from DiProPerm's results. The dashed curves in the left panel are the theoretical PDC (12). The solid curves in both panels are local linear regression fits of the Monte Carlo samples. Each dot is a single realization

of the estimated PDC value. Three representative cases studied in Figure 2 are highlighted in the left panel as black stars. We observe that PDC does not increase with signal strength $g$ as expected.

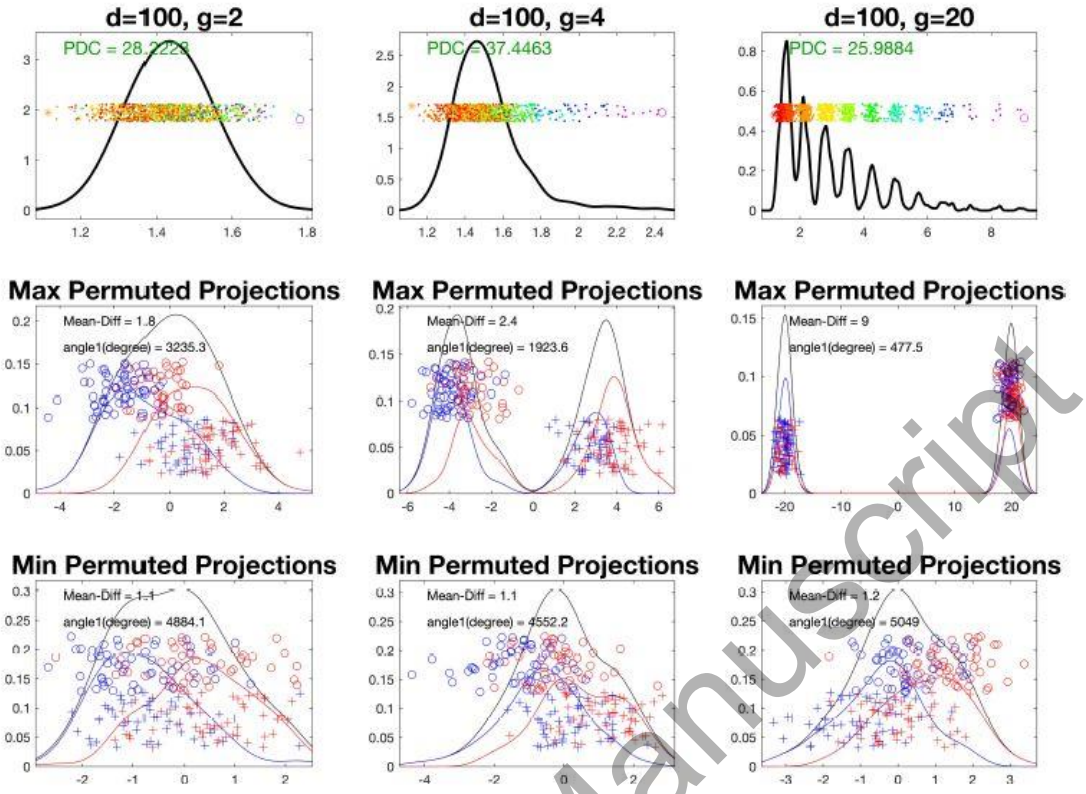


**Fig. 2** DiProPerm results for the MD direction and $d = 100$. Left: $g = 2$; middle: $g = 4$; right: $g = 20$ shown as black stars in Figure 1. Top panels are the permuted statistics $C_i$ and their kernel density estimates with PDCs values printed in green text. Middle and bottom panels are chosen permutations with colors representing the permuted labels and symbols representing the original labels; the middle row of panels have the largest permuted statistic (colored circles in the top panels); bottom panels are permutations with the smallest permuted statistic (colored stars in the top panels). Shows increasing skewness of the permutation distribution as $g$ grows due to permutation unbalance.

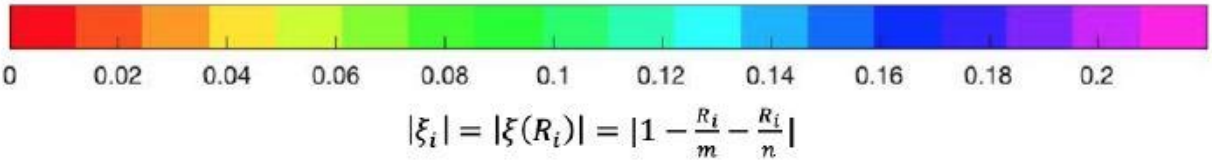


**Fig. 3** Colorbar used in the jitter plots in the top panels in Figure 2. Numbers represent the absolute value of the coefficient of unbalance $\xi_i$ defined in Equation (8) in Section 2.2.

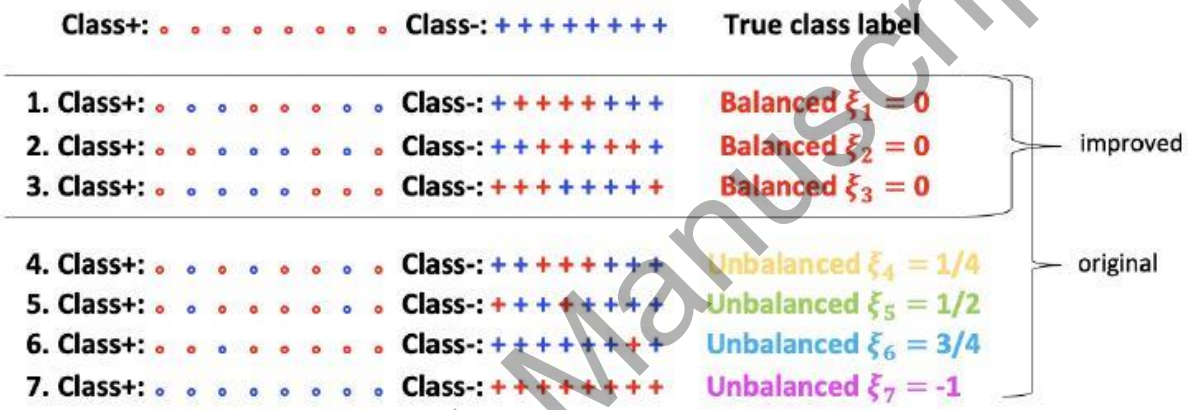


**Fig. 4** This figure shows 7 permutations from a simple example. The top row shows the true class labels and the rest of the rows show 7 different permutations represented as red and blue colored reassignments. The right column distinguishes between balanced and all permutations by coloring the text in the spirit of Figure 3.

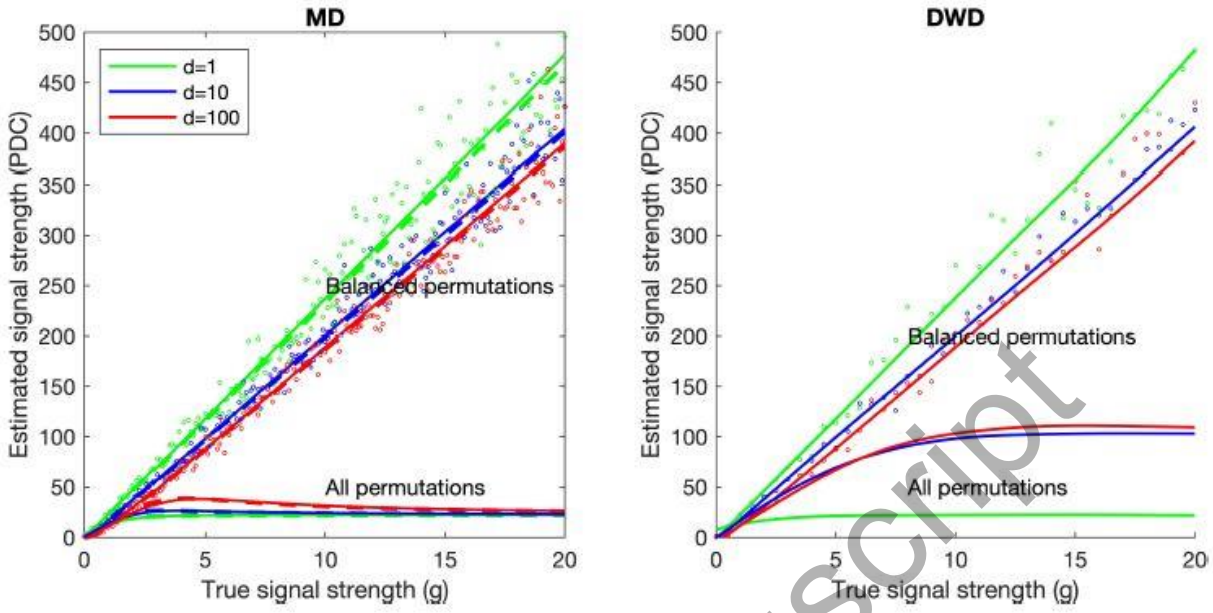


**Fig. 5** Realizations of the PDC for different choices of *d* for both all and balanced permutations based on DiProPerm. The lower curves in both panels, labeled as *All permutations*, are the same curves as Figure 1. The higher curves and dots in both panels, labeled as *Balanced permutations*, show the proposed DiProPerm using analogous colors and signs as in Figure 1. Unlike the all permutation PDCs, the balanced permutation PDCs keep growing with larger signal strength *g*.

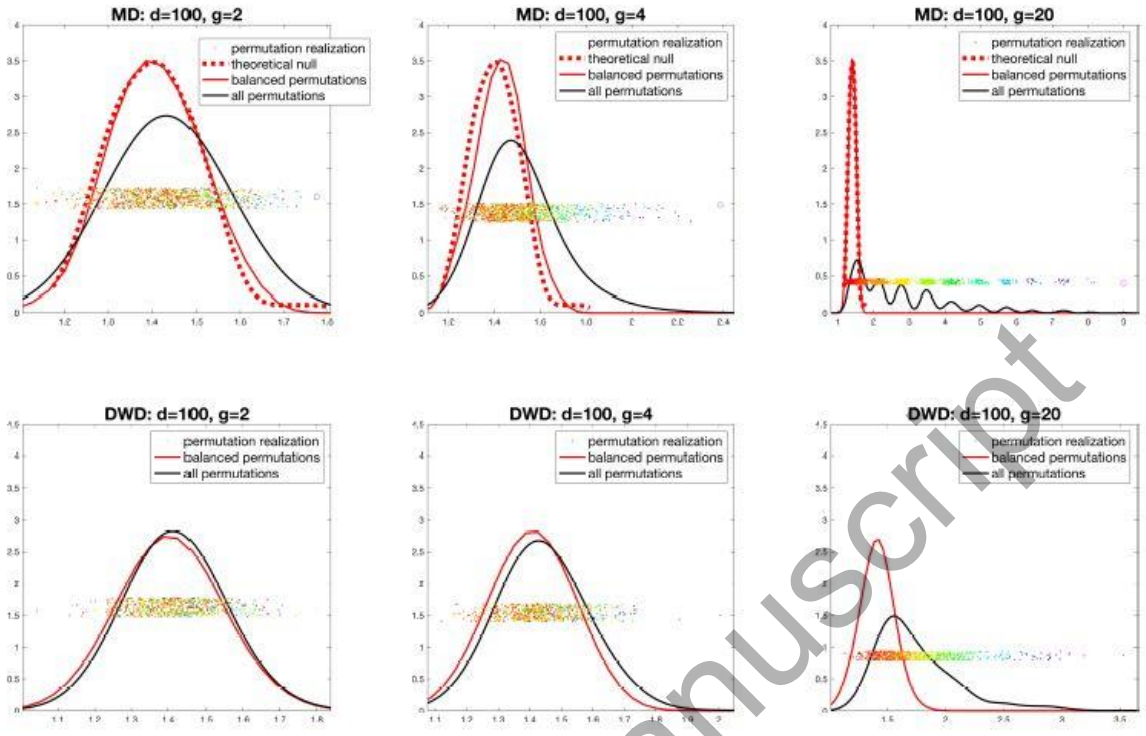


**Fig. 6** Null distributions of the three representative testing contexts studied in Figure 2. The MD and DWD directions are contrasted in the top and bottom panels. The dashed curves are the derived theoretical null distribution. The red solid curves are kernel density estimates of the balanced permutations (red dots in the jitter plots). The black curves are the kernel density estimates of all permutations (dots of all colors). The red curves do not change with *g* indicating a good estimate of the null distribution.

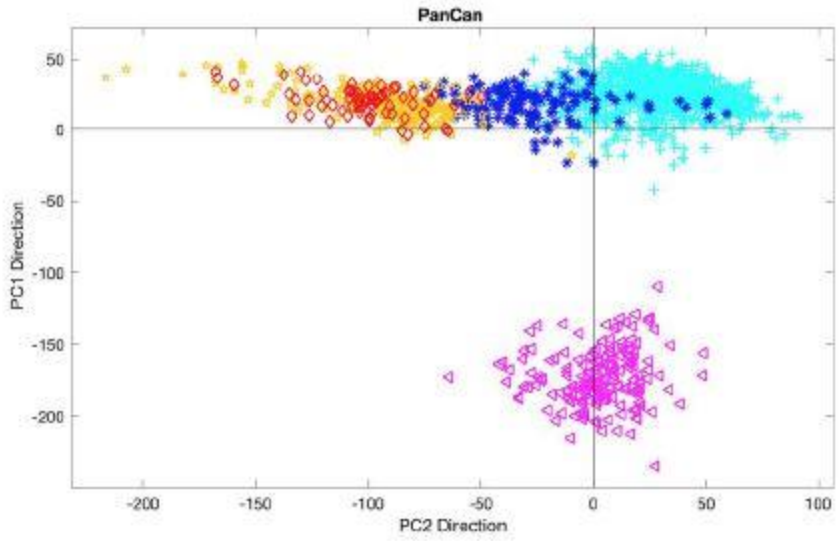


**Fig. 7** PCA scores scatter plot from TCGA Pan-Cancer gene expression data with symbols and colors in Table 1. As biologically expected LAML is much different from the rest and COAD and READ overlap in PCA space.

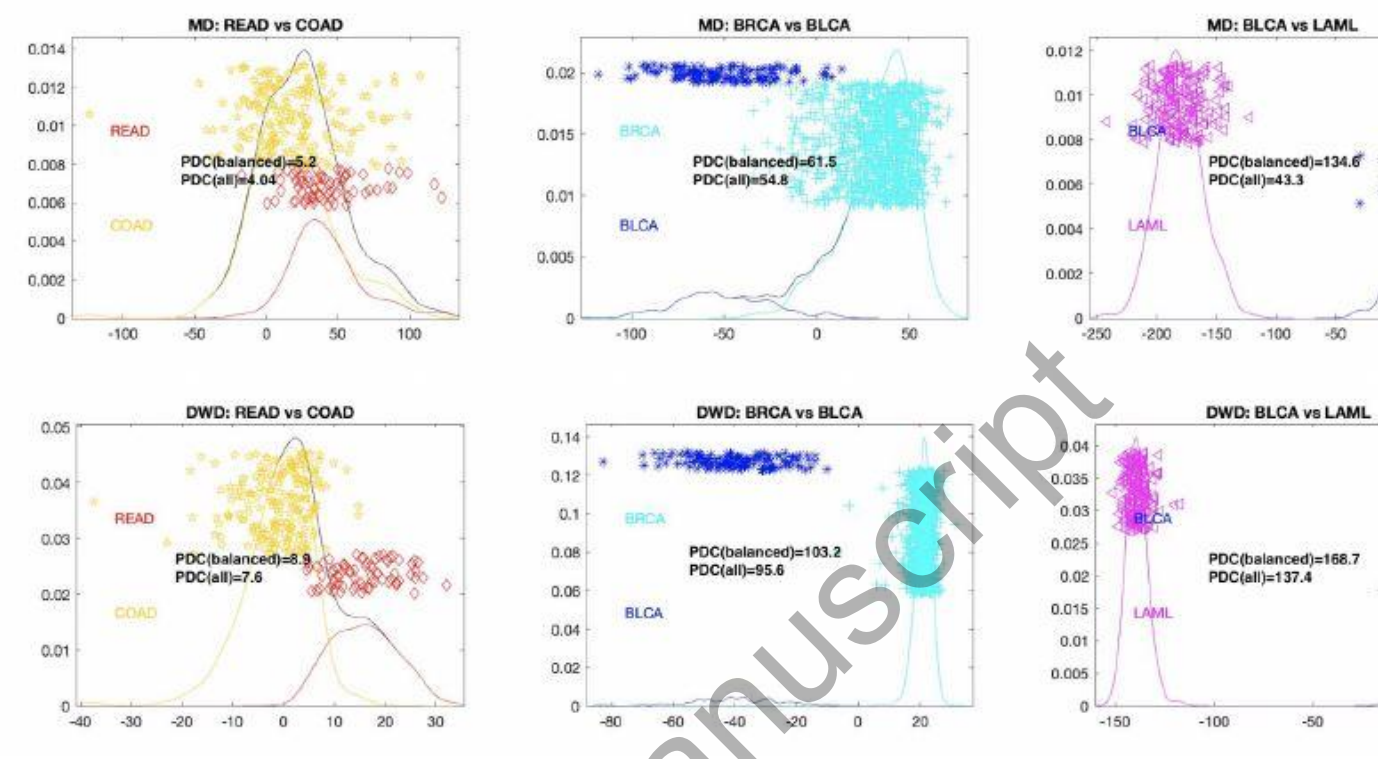


**Fig. 8** Distributions of projection scores on MD (top) and DWD (bottom) directions for the pairs: READ vs. COAD; BLCA vs. BRCA; LAML vs. BLCA. The *x*-axis is the value of projection scores, *y*-axis shows the height of the KDE estimates, and the black text shows the corresponding PDCs. The separation of the subpopulations on the top (MD) is similar to those for the bottom (DWD) but not as distinct. PDCs for balanced permutations are higher than for all permutations.

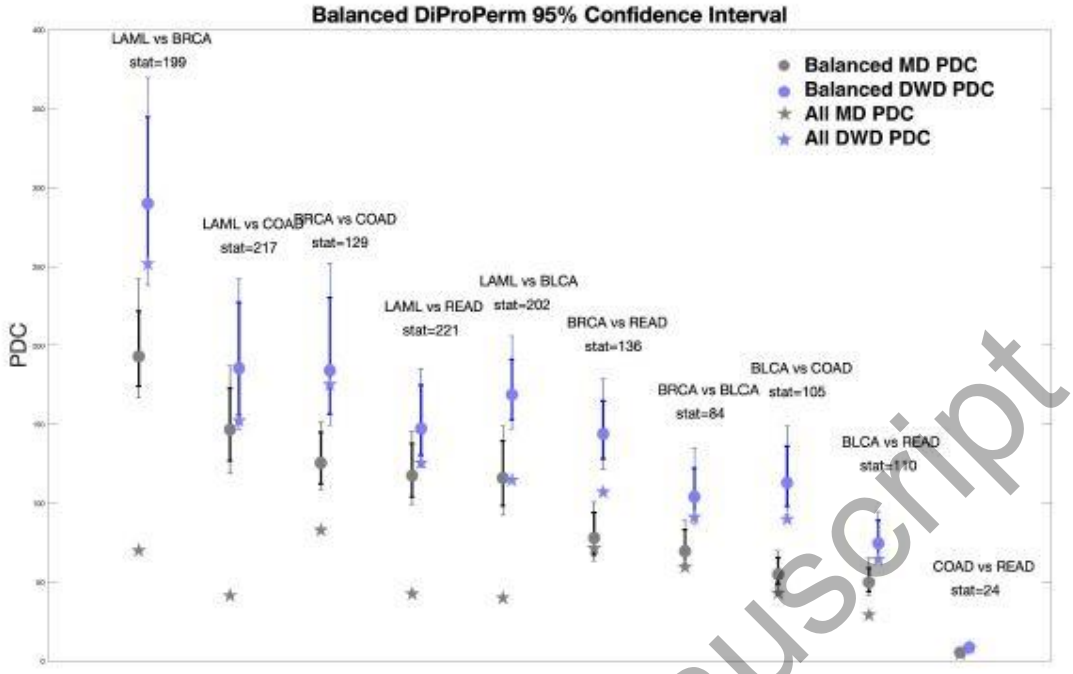


**Fig. 9** DiProPerm 95% confidence intervals for all 10 pairwise tests of 5 types of cancers from TCGA data. The thicker lines represent individual confidence intervals for each PDC and the thinner lines are the Bonferroni corrected confidence intervals. The circles and stars indicate the PDC estimates from balanced and all permutations respectively. The DWD-based PDCs are shown in blue/light blue and MD-based PDCs are shown in black/gray. The value of the test statistic is also printed at the top of each bar. This illustrates many effects discussed above.

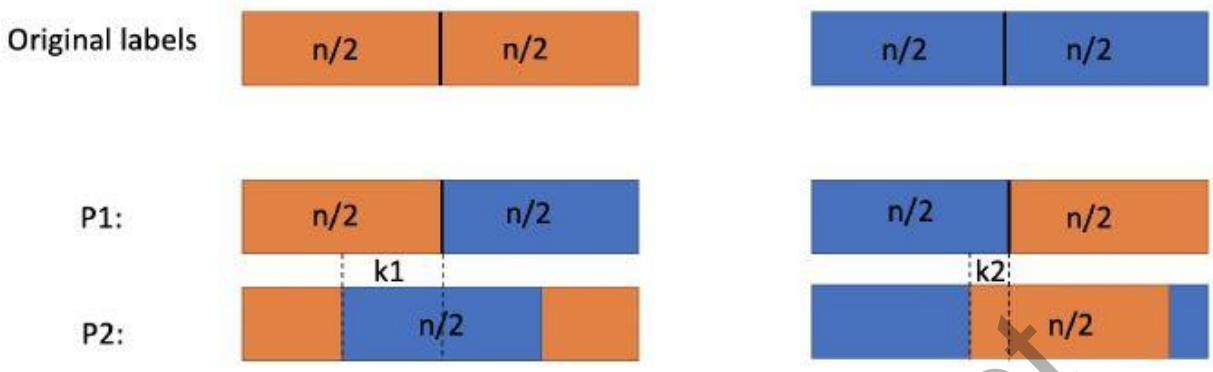


**Fig. 10** The first row represents the orginal class labels: orange vs. blue. The bottom two rows are two random balanced permutations.

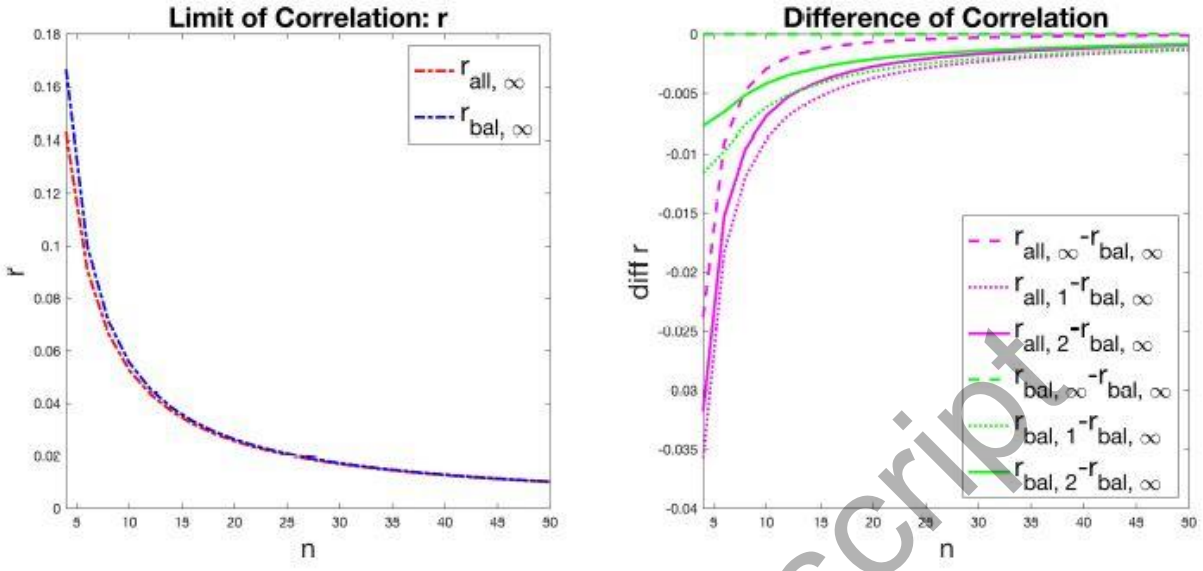


**Fig. 11** Correlation (*r*) as a function of *n* with different line types and colors indicating different *d* and balanced versus all permutations. In the left panel, when $d = \infty$, the red shows correlations of all permutations ($r_{all,\infty}$) with blue for balanced permutations ($r_{bal,\infty}$). These curves (decreasing rapidly) are very close to each other and $r_{bal,\infty} > r_{all,\infty}$. The right panel enables a more detailed study for *d* = 1, 2 in both cases, showing the difference between these correlations and the upper bound $r_{all,\infty}$. Correlations rapidly decrease as a function of *n* and increase slightly as a function of *d*. When *d* = 1, 2, correlations are already very close to the limit $d = \infty$, which is also the upper bound.

**Table 1** Abbreviated name, color, symbol (used in figures in this paper) and number of cases for each cancer type. Breast Cancer (BRCA) has the largest number of cases.

| Cancer | Abbreviation | Color | Symbol | Number |
|---|---|---|---|---|
| Acute Myeloid Leukemia | LAML | magenta | ◃ | 173 |
| Bladder Urothelial Carcinoma | BLCA | blue | * | 138 |
| Breast Cancer | BRCA | cyan | + | 950 |
| Colon Adenocarcinoma | COAD | yellow | ★ | 190 |
| Rectal Adenocarcinoma | READ | red | ◊ | 72 |