Biometrika (2019), **xx**, x, *pp*. 1–14
Printed in Great Britain

Supplementary material for 'A Nonparametric Approach to High-dimensional k-sample Comparison Problems'

BY SUBHADEEP MUKHOPADHYAY
Stanford University, Department of Statistics, Stanford, CA 94305, USA
deep@unitedstatalgo.com

AND KAIJUN WANG
Fred Hutchinson Cancer Research Center, Seattle, WA 98109, USA
kaijunwang.19@gmail.com

SUMMARY

This supplementary document contains thirteen sections. The first section provides an intuitive understanding of the special nonparametric polynomial basis called LP-basis. Next few sections discuss the performance of our proposed method under pairwise comparison, class imbalance problem, discrete data alternatives, correlation structure alternatives with fixed marginal distributions, and local alternatives. Then we examine some other practical details of our methodology, along with its application for geneset enrichment analysis. The last section provides a guide on how to use our R-package `LPKsample` based on Leukemia data.

S1. EMPIRICAL LP-BASIS: SHAPES AND NOMENCLATURE

Fig 1 depicts the shapes of first four empirical LP-basis functions for the variable `Start` of kyphosis dataset which are analyzed in Sec 3.2. This is a discrete variable, taking values from 1 to 18, denoting number of the topmost vertebra operated on a sample of $n = 81$ children who have had corrective spinal surgery. We display the basis function over unit interval by joining $\left\{\widetilde{F}(x_i), T_j(x_i; \widetilde{F})\right\}$ for $j = 1, \ldots, 4$ and $x_i \in \texttt{Unique}(x_1, \ldots, x_n)$. Note that they are discrete piecewise-constant orthonormal polynomials with respect to empirical measure $\widetilde{F}$.

The nomenclature issue: In nonparametric statistics, the letter L plays a special role to denote robust methods based on ranks and order statistics such as Quantile-domain methods. The connection of our polynomials with rank is evident from Theorem 1. With the same motivation, we use the letter L. On the other hand, P simply stands for Polynomials. Our custom-constructed basis functions are orthonormal polynomials of mid-rank transform instead of raw x-values; for more details see Parzen and Mukhopadhyay (2014). This serves two additional purposes: (i) injects robustness into our method (cf. Figures 4 (e,f) of the main paper), and (ii) we can define high-order polynomials without requiring the stringent condition of the existence of high-order moments of X, which are easily violated for heavy-tailed features.

S2. PAIRWISE COMPARISONS

Once the global k-sample null hypothesis of equality of high-dimensional distributions is rejected, we can conduct $k(k-1)/2$ pairwise comparisons to gain more insights into the possible alternative. This is straightforward under the Graph-based LP framework, as demonstrated below.

Consider the following example where we have $k = 4$ groups: Groups 1 and 2: $\mathcal{N}_d(0, I)$, Group 3: $\mathcal{N}_d(0.251_d, I)$, and Group 4: $\mathcal{N}_d(0, 1.5I)$. We simulated $n = 200$ samples from the 4 distributions, with $n_i = 50$ each, and dimension $d = 1000$. The primary objective is to test the equality of these 4 distributions.

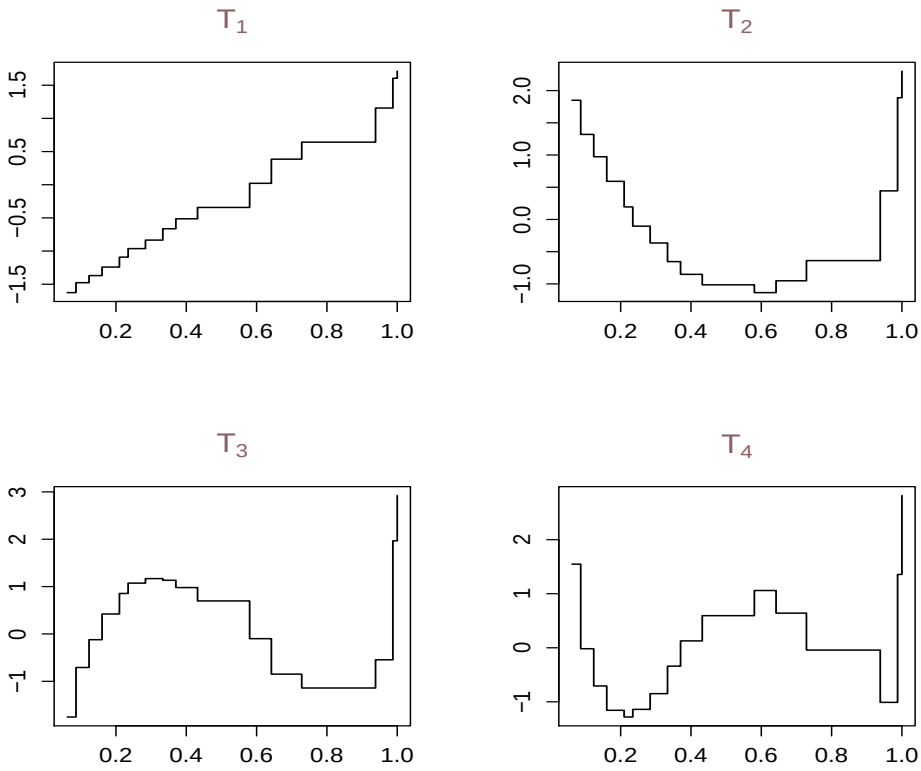

Fig. 1: The shapes of the first four empirical LP-basis functions $T_1(x;\widetilde{F}_X),\ldots,T_4(x;\widetilde{F}_X)$ of the variable `Start` in kyphosis dataset which are discussed in Sec 3.2 of the main paper. The `Start` is a discrete variable that denotes the number of the first vertebra operated on a sample of $n = 81$ children who have had corrective spinal surgery.

`Step 1`. Table 1 shows the results of global k-sample Graph-based LP test. This immediately indicates

Table 1: k-sample Graph-based LP test results. The global statistic along with its components are given.

Component	GLP	p-value
1*	0.967	8.01×10^{-37}
2*	0.946	5.89×10^{-36}
3	0.069	0.129
4	0.016	0.953
Overall	0.852	4.69×10^{-32}

that at least one of the population is different from others. Moreover, from that table we can also infer that the distributions are different with respect to location and scale, which exactly matches our data generating mechanism.

`Step 2`. Next step is to examine more carefully at the specific pattern of differences among the distributions. Table 2 shows all the 6 pairwise comparison results. We recorded the average number of rejections out of 100 trials for each pairwise comparison and displayed the estimated power.

Few conclusions:

- All methods correctly conclude that Group 1 and 2 are essentially from the same distribution.
- Friedman and Rafsky's method fails to distinguish between 1-4, 2-4 and 3-4, which further affirms its weakness for scale-alternatives.
- Our proposed method provides the correct conclusion in all six cases along with other method except that of Friedman and Rafsky's. But the most striking aspect of our proposed method lies in its ability to pin down the right cause(s) for rejecting each pair. For example consider the pairs 1-4 and 3-4, where our test indicates the differences exist in scale and location-scale. This spectacular insight into the possible alternative could be very useful for applied data scientists, as direct visualization of high-dimensional distributions is not possible.

Table 2: k-Sample pairwise comparison chart: average rejection for different methods. FR: Friedman and Rafsky's test; GEC: Generalized Edge Count test; RB: Rosenbaum's test; HP: Biswas' test.

Pair	FR	GEC	RB	HP	GLP				
					Overall	1	2	3	4
1-2	0.06	0.05	0.01	0.03	0.03	0.01	0.03	0.04	0.07
1-3	1.00	1.00	1.00	1.00	0.97	1.00	0.04	0.06	0.01
2-3	1.00	1.00	1.00	1.00	0.96	1.00	0.03	0.05	0.08
1-4	0.00	1.00	0.84	1.00	0.99	0.04	1.00	0.09	0.01
2-4	0.00	1.00	0.86	1.00	0.98	0.05	1.00	0.08	0.04
3-4	0.00	1.00	1.00	1.00	1.00	1.00	1.00	0.05	0.02

S3. CLASS IMBALANCE PROBLEM

Class imbalance is a common problem in real world datasets, where we observe disproportionate number of observations for different classes. It has been noted in Chen et al. (2017) that classical graph-based methods such as Friedman & Rafsky edge-count test (Friedman & Rafsky, 1979) is sensitive to unbalanced data. As a remedy to this problem, Chen et al. (2017) proposed the weighted edge-count test, which is one step further refinement of the generalized edge-count test (Chen & Friedman, 2017). So, naturally the question arises: whether our proposed method can automatically tackle this issue of class imbalance?

Table 3 investigates the performance for these 4 methods under the following cases: i) equal sample sizes, i.e., $n_1 : n_2 = 1 : 1$; ii) $n_1 : n_2 = 1 : 4$; iii) $n_1 : n_2 = 1 : 9$; and finally, iv) $n_1 : n_2 = 1 : 14$, an extreme imbalanced situation. We fix total sample size at $n = 100$, and the dimension at $d = 1000$. We focus on the location-alternative $\mathcal{N}_d(0, I)$ vs $\mathcal{N}_d(0.25\mathbf{1}, I)$, where Friedman & Rafsky's method is applicable. We use the R-package `gTests` with the option `test.type`="all". For our method we have used LP-graph kernel with $c = 0.1$. Table 3 shows that even though we have the same sample sizes in all of the four cases, just due to class imbalance, the performances of some of the methods get hampered dramatically. Friedman & Rafsky's method looses it efficiency by 80%! Our method performs surprisingly well even under extreme unbalance of $1 : 14$ scenario and outperforms the specialized weighted edge-count. Some

Table 3: Power comparison for unbalanced sample problem.

$n_1 : n_2$	FR	GEC	WEC	GLP
$1 : 1$	1.00	1.00	1.00	1.00
$1 : 4$	0.98	0.99	0.99	1.00
$1 : 9$	0.65	0.75	0.89	0.99
$1 : 14$	0.20	0.46	0.69	0.81

theoretical explanation of this automatic adaptability of our method can be found by noticing a striking

similarity in the weighting scheme of Chen et al. (2017, Eqs 3 and 4) and our Ncut-based partitioning scheme. The way normalized-cut modify the graph-cut metric (see Sec 2.4) based on inverse weighting by $\mathrm{Vol}(V_i) = \sum_{i \in V_i} \mathtt{deg}_i$, is very close to how weighted edge-count method corrects the Friedman and Rafsky's edge count statistic for unequal sample sizes.

S4. Discrete and Mixed Data Alternatives

Here we investigate the following four important cases: (i) the distributions are discrete, (ii) discrete features are contaminated with outliers; (iii) the features are mix of two-kinds of discrete distributions; and finally, (iv) the case where we have both discrete and continuous variables mixed, a very common situation in real world datasets. Note that, this framework is good enough to include the categorical variables via dummy coding, which refers to the process of coding a categorical variable into dichotomous variables. A variable with q categories is replaced by an indicator matrix $\mathbb{I}_{n \times q}$ with q columns where category memberships are indicated by the columns of zeros and ones. For example, we could code gender as a matrix with 2 columns where 1=female and 0=male. This is exactly how the categorical variables are included in a traditional regression analysis; see Agresti (2007, ch. 4 and 5) for more details. For the comparison in this section, we'll also consider the generalized edgecount test for discrete data (dGEC) proposed in Zhang and Chen (2017).

Simulation setting: We have generated two-sample datasets from the following four settings, with dimension $d = 200$, and group sizes $n_1 = n_2 = 50$. The results are summarized in Table 4.

- (a) Mean-shifted discrete distributions: $\mathrm{Poisson}(5)$ vs $\mathrm{Poisson}(5.5)$.
- (b) As in (a), except 2% discrete outliers are introduced by $\mathrm{Binomial}(40, 0.5)$
- (c) X_1: $\mathrm{Binomial}(10, .5)$; X_2: mixed distribution, $X_{2j} \sim \mathrm{Binomial}(10, .5)$ for $j = 1, \cdots, 100$, and $X_{2j} \sim \mathrm{Poisson}(5)$ for $j = 101, \cdots, 200$.
- (d) Both groups are mixed with continuous distributions.
  X_1: $X_{1j} \sim \mathcal{N}(0, I)$ for $j = 1, \cdots, 100$, and $X_{1j} \sim \mathrm{Binomial}(10, 0.1)$ for $j = 101, \cdots, 200$;
  X_2: $X_{2j} \sim \mathcal{T}_3(0, I)$ for $j = 1, \cdots, 100$, and $X_{2j} \sim \mathrm{Binomial}(10, 0.1)$ for $j = 101, \cdots, 200$

Table 4: Empirical power comparison for discrete and mixed data

Cases	GLP	FR	Rosenbaum	HP	GEC	dGEC
(a)	0.97	0.90	0.78	0.93	0.89	1.00
(b)	0.87	0.41	0.15	0.25	0.20	0.39
(c)	1.00	0.00	0.56	1.00	1.00	1.00
(d)	0.98	0.01	0.09	0.70	0.07	0.10

Compared to case (a), case (b) shows that all methods aside from our proposed test completely break down under outliers, even the discrete generalized edge count test, which shows our proposed method's robustness in the presence of outliers. The setting (c) presents a case with distribution-shift (not location shift), and it is clear Friedman & Rafsky's test fell short on these kind of testing problems; Rosenbaums' test also breaks down, with only about 50% rejection power. For case (d), where we have both discrete and continuous data, most of the existing methods fail miserably, apart from our proposed GLP method and the Biswas' test.

S5. Performance Under Various Correlation Structures

Here we investigate the cases where two groups have the same marginals but different correlation structures. Consider the two d-variate groups with distributions $X_1 \sim G_1(x)$, $X_2 \sim G_2(x)$. We set the

dimension d to be 500 while sample sizes are the same for both groups $n_1 = n_2 = 100$ and studied the following three types of covariance models:

- (a) Σ_1 is identity matrix; Σ_2 have $\sigma_{2,ij} = 1$ for $i = j$, and $\sigma_{2,ij} = 0.5$ for $i \neq j$.
- (b) Σ_1 is identity matrix; Σ_2 have a block diagonal shape where: $\sigma_{2,ij} = 1$ for $i = j$ and $\sigma_{2,ij} = 0.5$ for $5(k-1)+1 \leq i \neq j \leq 5k$, where $k = 1, \cdots, [d/40]$ is the number of blocks and $\sigma_{2,ij} = 0$ otherwise.
- (c) Σ_1 is identity matrix; Σ_2 have $\sigma_{2,ij} = \frac{1}{2}\left[(|i-j|+1)^{2\alpha} + (|i-j|-1)^{2\alpha} - 2(|i-j|)^{2\alpha}\right]$ with $\alpha = 0.9$.

Each structure is tested on both multivariate Gaussian setup: $G_1 = \mathcal{N}(0, \Sigma_1)$, $G_2 = \mathcal{N}(0, \Sigma_2)$ as well as multivariate T-distributions with degrees of freedom 3: $G_1 = \mathcal{T}_3(0, \Sigma_1)$, $G_2 = \mathcal{T}_3(0, \Sigma_2)$.

Table 5: Power comparison for two-sample data with same marginals but different correlation structures

Distribution	Cases	GLP	FR	Rosenbaum	HP	GEC
Gaussian	(a)	1.00	0.00	1.00	1.00	1.00
	(b)	0.94	0.00	1.00	1.00	0.97
	(c)	1.00	0.00	1.00	1.00	0.98
$\mathcal{T}_3$	(a)	1.00	0.38	1.00	1.00	0.76
	(b)	0.90	0.85	1.00	1.00	0.96
	(c)	0.99	0.66	1.00	1.00	0.80

For each setting, 100 simulations are performed and their average number of rejections are recorded in Table 5. As it is evident, our method achieves satisfactory performance in all of the above cases.

S6. Power Under Local Alternatives

We seek to investigate the power of different methods under local alternatives. In particular, we consider the following setup which is closely related to Bhattacharya (2017): $G_1 = \mathcal{N}_d(0, I)$ and $G_2 = \mathcal{N}_d(\frac{\delta}{\sqrt{n_1+n_2}}1_d, I)$. The setting was repeated 1000 times over a grid of 10 δ-values in $[0, 3]$ for (a) dimension $d = 20$ and (b) $d = 50$ with sample sizes $n_1 = 400$ and $n_2 = 1000$. The resulting power curves are displayed in Fig 2. Our proposed GLP statistic shows impressive local-power, in comparison to other graph-based methods. It is important to emphasize that all the competing methods have two common ingredients: they are based on Euclidean distances, and their counting cross-match statistic is a generalized version of the Wald-Wolfowitz run-type test. Interestingly, Bhattacharya (2017) showed that tests with these two characteristics suffer against $O(n^{-1/2})$ alternatives, as clear from Figure 2. On the other hand, the success of GLP comes from approaching the high-dimensional k-sample comparison problem an entirely different perspective using specially-designed data-driven kernel and spectral graph theory.

S7. Relationship with Permutation-based Rank Tests

Here we intend to investigate the connection between permutation-based rank tests and our approach. Although these methods are not directly related to graph-based nonparametric methodology, they do possess several attractive properties, which we briefly review below.

At a broad level, the Non-Parametric Combination based permutation testing methods operate as follows: (i) The global null-hypothesis is first broken down into a set of partial null hypotheses; (ii) For each partial null hypothesis an unbiased and consistent statistic is selected, depending on the alternative and data-type information, to compute the permutation p-values; (iii) At the final step, all the p-values associated with the partial tests are combined using an appropriate convex function (Birnbaum, 1954).

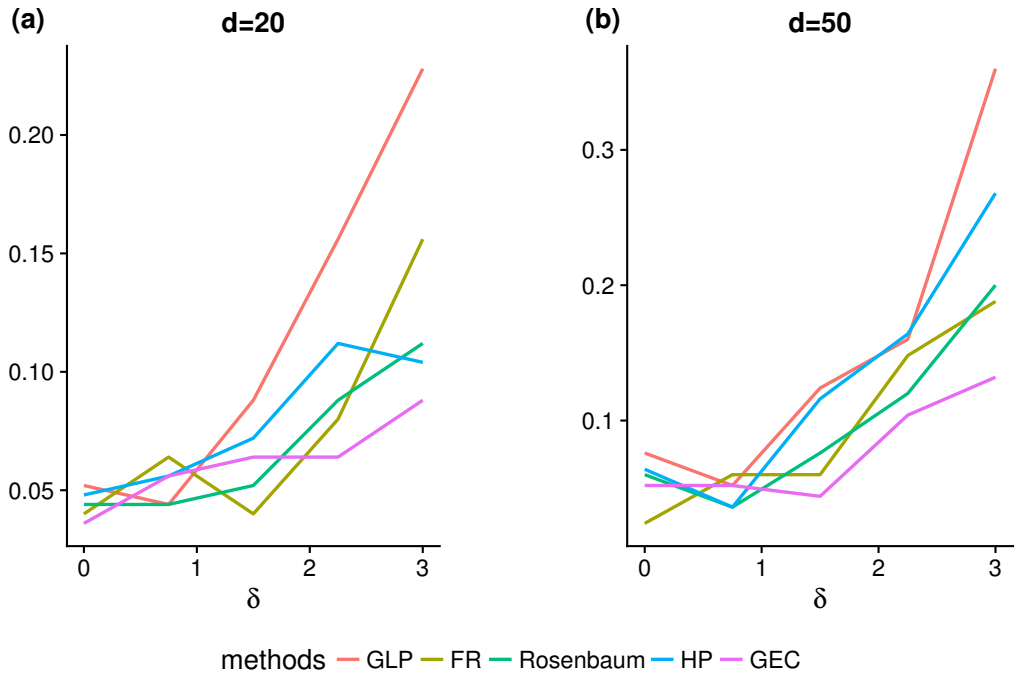

Fig. 2: Performance under local alternatives, discussed in Section S6.

Most popular ones are Fisher, Liptak, and Tippett combining functions. For an exhaustive treatment of Non-Parametric Combination based methods, see Pesarin and Salmaso (2010b) and Bonnini et al. (2014).

Few Remarks:

1. Because of its clever construction by breaking the original global hypothesis into several partial or sub-hypotheses Non-Parametric Combination can work on complex multivariate problems. For example, the multivariate d-dimensional k-sample testing problem can be broken down into a finite set of d sub-hypotheses $\cap_{h=1}^{d}\{F_{1h} \overset{\mathrm{d}}{=} \cdots \overset{\mathrm{d}}{=} F_{kh}\}$, where F_{jh} denotes the marginal distribution of j-th variable in the class h. Thus the multivariate problem boils down to d univariate k-sample problem. Note that H_0 is true if all the d partial hypotheses are jointly true. Each partial null hypothesis can be further broken down into $k(k-1)/2$ sub-hypotheses for pairwise comparisons; see Bonnini et al. (2014, ch. 3) for more details.

2. This flexibility comes at a price. The first practical challenge arises from its computational cost: $O(k^2 \times d \times B)$, where B is the number of permutation performed which is generally selected as 1000 (cf. R package `ICSNP`). Hence scalability for large-dimensional problems could be a major issue for this class of methods. In contrast, our Graph-based LP test performs just one omnibus test to check the global H_0.

3. Pesarin and Salmaso (2010b, ch. 6) describes how permutation-based Non-Parametric Combination technique can successfully tackle mixed data types including discrete, continuous, or even categorical. However, Non-Parametric Combination methodology requires appropriate partial test statistics for testing each sub-hypothesis H_{0i} against H_{1i}. For example, Wilcoxon rank-sum statistic for two-sample location test, Chi-square statistic if the covariate is discrete count, KruskalWallis if one needs to test $k > 2$ case for location difference etc. The word "separately" in the following excerpt highlight this point:

> "The extension to mixed variables, some nominal, some ordered categorical and some quantitative, is now straightforward. This can be done by separately combining the nominal, ordered categorical and quantitative variables, and

then by combining their respective p-values into one overall test. Details are left to the reader." Pesarin and Salmaso (2010b, Chapter 6.4)

Naturally, the whole testing process becomes data-type dependent. To apply this method for large-dimensional mixed data problems puts insurmountable challenge for a practitioner, as he/she has to pick the right test statistics after manually checking each variable type*. On the other hand, the main novelty of our graph-based LP procedure lies in developing a fully automatic multivariate test that does not require any data-type information from the user.

4. Non-Parametric Combination methodology arrives at the global test by combining the p-values. Whereas graph-based LP method, which is illustrated in P53 data of Sec 2.6, combines the principal kernels to produce the single master kernel function for the global testing. Also, it should be noted that the way we decompose the test statistics is very different from Non-Parametric Combination. Our component kernel functions capture ways in which the high-dimensional distributions can be different over two or more classes.

5. The unique advantage of our approach is its exploratory data analysis side which is generally outside the purview of permutation-based methods.

6. Next, we discuss the important case of stochastic dominance alternatives using two real data examples to highlight how both graph-based LP and Non-Parametric Combination yield almost identical conclusions.

Formulation. Let X_g, $g = 1, \ldots, k$ denote d-dimensional random vectors associated with group g. The multivariate monotonic stochastic order problem is concerned with the following testing problem:

$$H_0 : X_1 \overset{\mathrm{d}}{=} \cdots \overset{\mathrm{d}}{=} X_k \quad \text{versus} \quad H_1 : X_1 \overset{\mathrm{d}}{\succeq} \cdots \overset{\mathrm{d}}{\succeq} X_k, \tag{1}$$

where at least one inequality is strict. Interestingly, it can be shown that (Baccelli and Makowski, 1989; Davidov and Peddada, 2013) the d-dimensional stochastic dominance alternative H_1 holds if and only if $X_{j1} \overset{\mathrm{d}}{\succeq} \cdots \overset{\mathrm{d}}{\succeq} X_{jk}$, $j = 1, \ldots, d$. This decoupling result makes the problem in some sense "dimension-free." Consequently, one can rewrite (1) as

$$H_0 : \cap_{j=1}^{d} \big[X_{j1} \overset{\mathrm{d}}{=} \cdots \overset{\mathrm{d}}{=} X_{jk}\big] \quad \text{versus} \quad H_1 : \cup_{j=1}^{d} \big[X_{j1} \overset{\mathrm{d}}{\succeq} \cdots \overset{\mathrm{d}}{\succeq} X_{jk}\big] \tag{2}$$

by applying union-intersection principle. To put the discussion into context we now introduce two real data examples.

Example 1. Rat Tumor Data (Pesarin and Salmaso, 2010b Ch. 8.3). The `rats` data set contains $n = 36$ observations and $d = 17$ variables which denotes relative size of tumors over different time in total over $k = 4$ different classes. The control group was treated with a placebo; three other groups were treated with increasing doses of a taxol synthetic product. Researchers are interested in testing whether treatment significantly inhibits tumor growth with increasing dose levels: $X_1 \overset{\mathrm{d}}{\succeq} \cdots \overset{\mathrm{d}}{\succeq} X_4$.

Example 2. Economics Entrance Test Data (Bonnini et al. 2014, Table 1.9). The data consist of scores of a sample of $n = 20$ applicants on their mathematical skills and economic knowledge for enrolling in a university Economics course. There are 10 applicants coming from scientific studies backgrounds and 10 applicants from classical studies. We want to test whether the score distributions of the two populations of students are the same, against the alternative that the distribution of the population coming from a scientific high school is stochastically greater, i.e., $H_1 : X_1 \overset{\mathrm{d}}{\succeq} X_2$, where group 1: scientific studies, and group 2: students with classical studies background. Fig 3 indicates that score distribution of students from scientific studies is shifted toward greater values (mainly a location-shift) compared to group 2 classical studies students. Table 6 performs preliminary `GLP` analysis, to further validate that the only difference lies in location.

* One possibility to automate Non-Parametric Combination methodology for mixed data would be to use Eq (5), which adapt itself looking at the data–United Statistical Algorithm (Parzen & Mukhopadhyay, 2013). This fusion of LP and Non-Parametric Combination could be an interesting topic for future research direction.

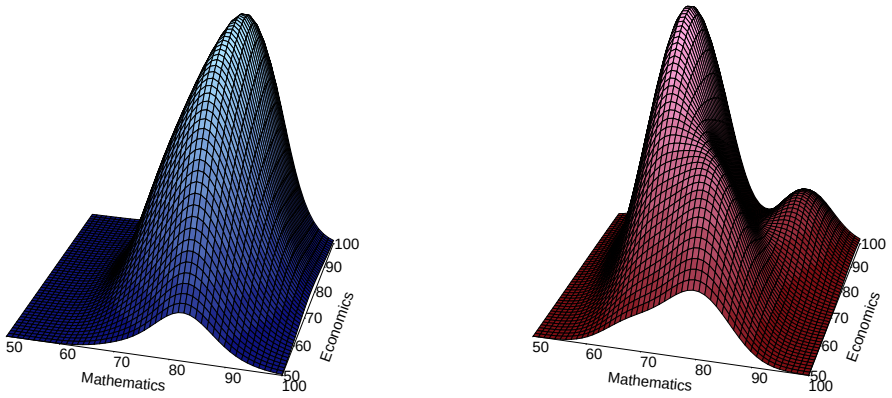

Fig. 3: Distribution of scores of Ex. 2: Scientific studies (blue); Classical studies (red).

Table 6: LP tables for Economics Examination data

Component	GLP	p-value
1*	0.275	0.019
2	0.042	0.361
Overall	0.275	0.019

Analysis: The k-sample testing for monotonic treatment effect can be reformulated as a correlation between $T_1(y; \widetilde{F}_Y)$ and $T_1(x; \widetilde{F}_X)$ which is shown in left panel of Fig 4. Our LP Hilbert correlation based trend test generalizes the celebrated Jonckheere-Terpstra test (Jonckheere, 1954) for mixed variables. This is required, as the rat data contains lots of ties, and our formulation automatically tackles that without any additional tuning. For testing the multivariate ordered alternative we thus propose our global statistic as $T_{\rm LP} = \min_{1\le j\le d} {\rm LP}[1, 1; Y, X_j]$. We use the permutation distribution of $T_{\rm LP}$ to compute the overall p-value. The sign of ${\rm LP}[1, 1; Y, X_j]$ indicates the direction of the treatment effect being positive or negative. For example, in the rat data it's sign is negative as taxol inhibit the tumor growth. See Fig 4.

Table 7: Rat tumor data, trend tests comparison. Both permutation and asymptotic p-values are reported for our $T_{\rm LP}$ trend detection statistic.

Variables	JT Test p-value	LP Trend Test		
		$T_{\rm LP}$	p-value (asymp)	p-value (perm.)
X_1	0.264	1.08	0.859	0.854
X_2	0.096	-2.67	0.038	0.032
X_3	0.011	-2.76	0.028	0.028
X_4	0.000	-4.05	0.000	0.000
X_5	0.000	-4.84	0.000	0.000
⋮	⋮	⋮	⋮	⋮
X_{17}	0.000	-4.45	0.000	0.000

Table 8: Admission data. Trend tests results.

Variables	JT Test p-value	LP Trend Test	
		T_{LP}	p-value
X_1 (Math)	0.018	-2.404	0.0081
X_2 (Econ)	0.912	-0.155	0.438

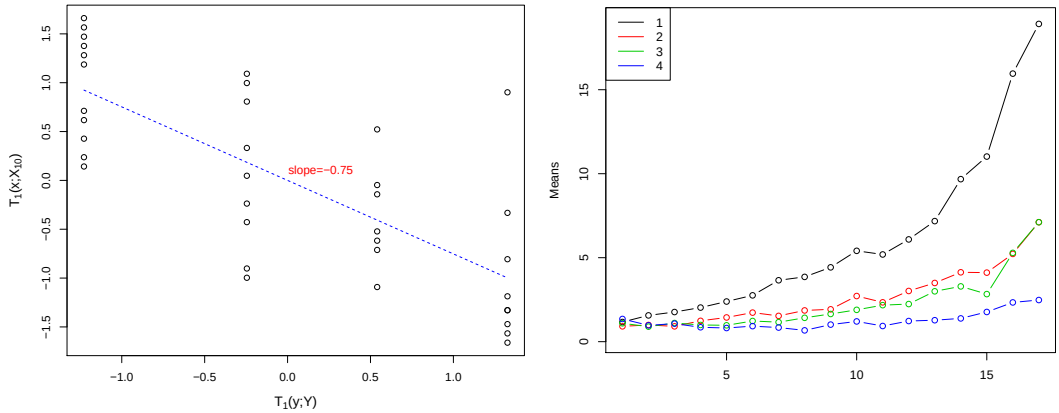

Fig. 4: Rat Tumor Data: The left panel shows the scatter plot of $\{T_1(y_i;Y), T_1(x_i;X_{10})\}$ for $i = 1,\ldots,36$. The slope is $\widehat{\mathrm{LP}}[1,1;Y,X_{10}]$ is -0.75. The right panel shows the group-specific means for each variables. Different colors represent different groups.

The results for the rat data is displayed in the Table 7. Key points are summarized below:

- The global k-sample multivariate LP-statistic for trend $T_{\mathrm{LP}} = \min_{1\le j\le d} \mathrm{LP}[1,1;Y,X_j] = -4.91$ with permutation p-value essentially 0. This strongly suggests the usefulness of the treatment as an anti-cancer agent.
- We have used $B = 5000$ permutations to compute the p-values, which are surprisingly close to their asymptotic counterparts, even for such a small sample settings (recall we had $n_1 = 11, n_2 = 9, n_3 = 7, n_4 = 9$, and dimension $d = 17$).
- Scientists are often interested to further understand which among the d variables, show a monotonic stochastic trend with higher dose level. In the context of `rat` data, all the variables, except X_1 show significant decrease in the tumor size as the dose of the taxol synthetic product increases. Note that X_1 indicates the relative size of the tumor at time point 1 – it could be too early for the effect of taxol to kick in.
- Economics entrance test data analysis virtually identical thus removed. In this case, only the mathematical scores show a significant increase for students with scientific studies backgrounds, which concurs with Fig 3; see Table 8 for more details.
- The authors Pesarin and Salmaso (2010b, Ch. 8.3, p. 276) arrived at identical conclusions for both rat and the Economics entrance test data using Non-Parametric Combination technique.

S8. Mood's Test for Scale Differences

It is important to keep in mind that the proper interpretation of Mood statistic as nonparametric tests for variances requires the unknown group medians to be equal. Moses (1963) and Marozzi (2013) noted that violation of this might hamper its interpretation and thus can affect final conclusion. Hence in practice, it is

recommended to standardize the variables by subtracting the group medians from each observation (when location difference exists) before performing the high-order component-tests. This location-alignment has been incorporated in all our reported results.

S9. MID-DISTRIBUTION TRANSFORMATION

It seems worthwhile to emphasize the fact that the mid-distribution function $F^{\rm mid}$ has been previously used as a device to define ranks for data with ties (Ruymgaart, 1980; Hudgens & Satten, 2002). Our theory, on the other hand, integrates $F^{\rm mid}$ in an important way to design the LP-transformation-based nonparametric modeling scheme, thereby broadening its utility for general purpose data analysis beyond treatment of ties in ranking problems. From a historical standpoint, the pioneer of this idea was Henry Oliver Lancaster (1961), who introduced mid-P-value, which was later formalized into the mid-distribution function by Parzen (1997). For that reason, many researchers, such as Alan Agresti, often refer it as Parzen's mid-distribution function.

S10. P53 GENESET ENRICHMENT ANALYSIS: FEW MORE DETAILS

This section has two main goals: (i) some graphical diagnostics to support the finding of "anthraxPathway" in Table 1 of the main paper, and (ii) how GLP technology can be used for multivariate geneset enrichment scoring– an important problem in genomics and proteomics applications.

Density Plot and Marginal Scoring

Since the "anthraxPathway" consists of two genes ($d = 2$), we can visually compare the shape of the density over the tumor and normal classes. Fig 5 clearly shows that the two bivariate distributions have significant tail-deviations, which again reinforce our findings of Section 2.7.

Next, we provide another justification. Given a genetic pathway, comprising of d genes, define the ℓ-th order LPcomens-based marginal enrichment score by

$$d^{-1}\sum_{j=1}^{d}\big|\,\mathrm{LP}[1,\ell;Y,\mathrm{Gene}_j]\big|^2.$$

Table 9 displays the values for two genesets "SA_G1_AND_S_PHASES" and "anthraxPathway" for $\ell = 1,\ldots,4$. From the table we can also read which components are enriched (denoted by '*') for that specific pathway.

Geneset	Components			
	1	2	3	4
(a)	4.300^* (0.371)	1.301 (0.164)	0.396 (0.022)	0.832 (0.061)
(b)	2.374 (1.563)	1.852 (1.309)	1.542 (1.049)	3.235^* (1.358)

Table 9: Displays the values of first four LPcomens-based marginal enrichment scores for (a) SA_G1_AND_S_PHASES and (b) anthraxPathway; standard errors are in parenthesis; '*' indicates the numbers that exceed the value $\chi^2_{1;0.9} = 2.7055$.

From Univariate to Multivariate Enrichment Scoring

The GLP statistic can also be used for multivariate nonparametric variable selection. In the present context of p53 geneset data, we can use it for Gene Set Enrichment Analysis. Our approach has a unique advantage for detecting differentially expressed groups of genes based on their multivariate distributions. This is a significant improvement over conventional methods (Subramanian et al., 2005; Efron & Tibshirani, 2007) which are based on aggregated univariate measures, thus can miss the 'joint' behavior of the

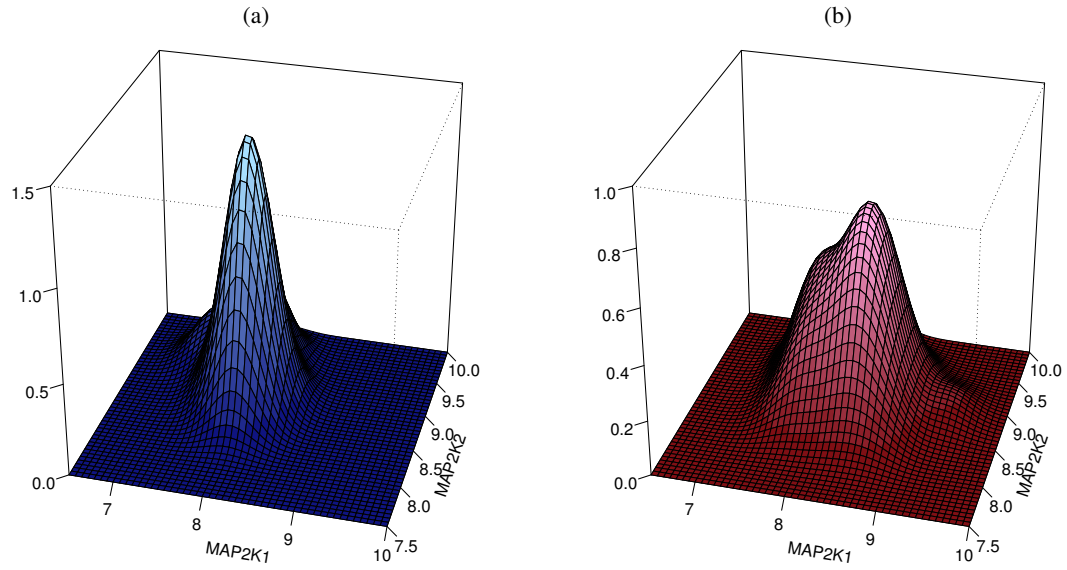

Fig. 5: The density of "anthraxPathway" geneset over two-groups: (a) normal cases and (b) tumor cases. Clearly $G_1 \neq G_2$, in particular the tumor cases have heavier tail than the normal ones.

genes. Ranking of top gene pathways based on their differential profiles is portrayed in Fig 6. The purpose of this exploratory graphical plot is to help biologists to understand which component-specific differential information (the question of 'how') is important in a specific geneset.

S11. Computational Time

Methods	n=100	n=250	n=500	n=1000
GLP	1.471 (0.162)	1.916 (0.207)	3.207 (0.432)	8.850 (0.398)
GEC	0.017 (0.004)	0.085 (0.042)	0.291 (0.123)	1.171 (0.217)
Rosenbaum	0.015 (0.006)	0.067 (0.030)	0.282 (0.122)	1.360 (0.243)
HP	0.039 (0.048)	0.186 (0.032)	0.844 (0.058)	4.281 (0.308)

Table 10: Computation time for location shift setting with $d = 100$ and varying sample sizes. Average runtime over 10 runs are reported along with the standard errors.

S12. Controlling Alpha

One may want to examine the performance under null to see whether the type-I error is controlled. For this purpose, we simulated sample points from $\mathcal{N}(0, I_{d_2})$ with dimension d ranging from 2 to $2^{10} = 1024$. The total sample size is $n = 200$, and these sample points are evenly assigned into two groups. Each case is simulated 100 times to approximate rejection rates of the proposed method as well as other comparing methods. These false rejection rates are displayed in Fig. 7. It is clear that for all these methods, type-I errors are controlled at the $\alpha = 0.05$ level.

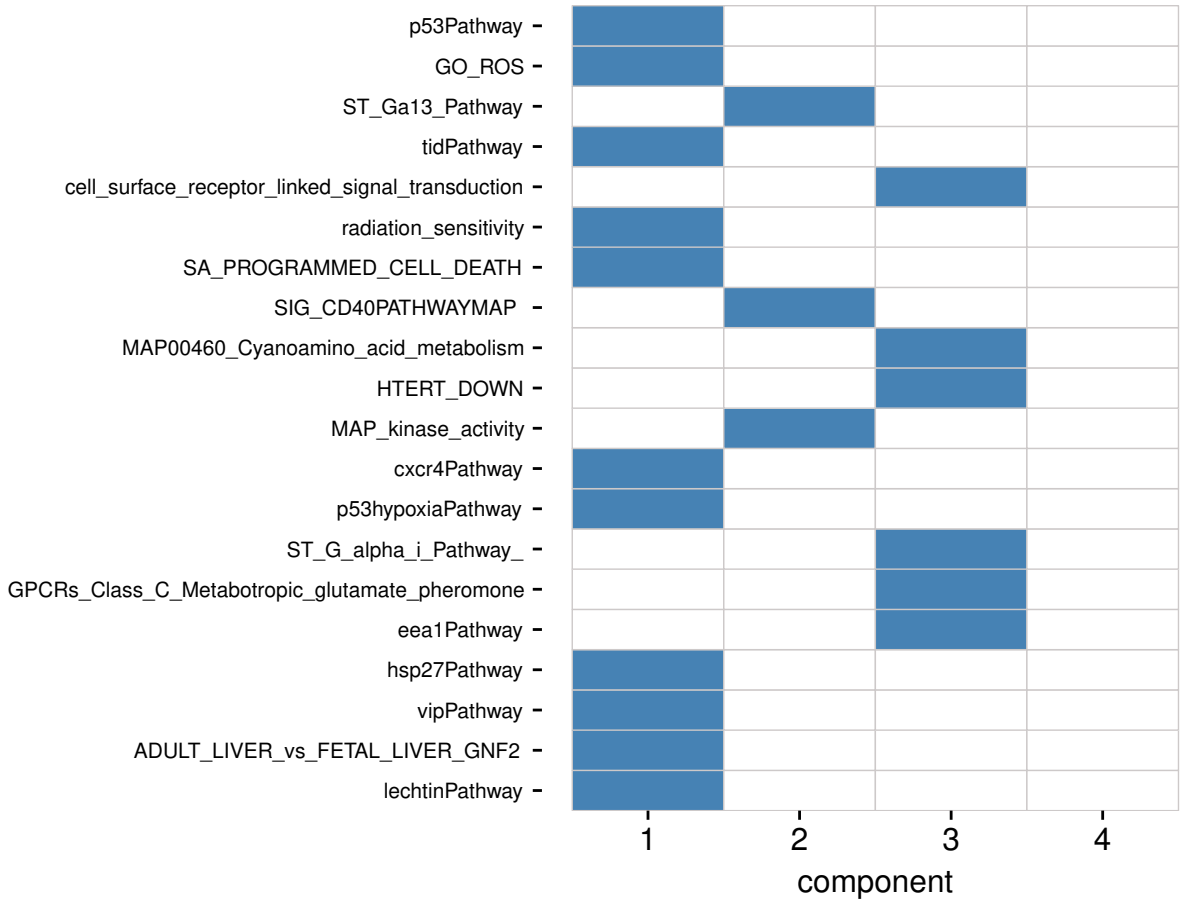

Fig. 6: Exploratory graphical plot of top 20 gene sets in p53 data; sorted from top to bottom. The cells with blue color indicates the significant components identified using GLP statistics for a specific geneset.

S13. SOFTWARE AND R-COMPUTATION

We provide an `R` package, `LPKsample`[†] that will perform all the tasks outlined in the paper. We now summarize the main functions and their usage.

The Leukemia data in our main paper will be used for demonstration. The following code describes the confirmatory phase using our algorithm.

```
#Phase I: Confirmatory Testing
#For Leukemia data in Table 3
 data(leukemia)
 attach(leukemia)
 leukemia.test<-GLP(X,class,components=1:4)
 leukemia.test$GLP
 #[1] 0.2092378
 leukemia.test$pval # overall p-value (Table 3)
 #[1] 0.0001038647
```

[†] Available online at https://CRAN.R-project.org/package=LPKsample

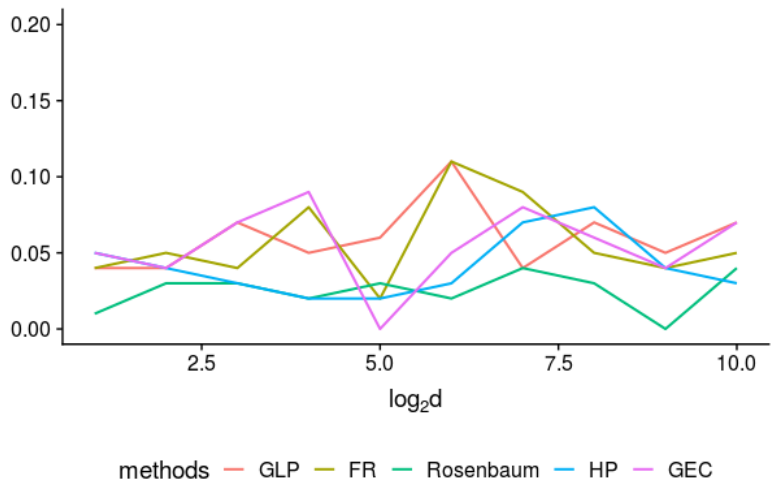

Fig. 7: Simulation result of Type-I error rate under null setting for all comparing methods.

`GLP` also provides an explanation for rejecting the null hypothesis. The following commands can be used to get the exploratory insights:

```
#Phase II: Exploratory Results
 leukemia.test$table  # rows as shown in Table 3
 #     component    comp.GLP       pvalue
 #[1,]         1 0.209237826 0.0001038647
 #[2,]         2 0.022145514 0.2066876581
 #[3,]         3 0.002025545 0.7025436476
 #[4,]         4 0.033361702 0.1211769396
```

At the final stage, a researcher might want to use those data-driven LP-transformed features to improve subsequent predictive models. The following code show how to extract that:

```
#Phase III: Predictive Modeling
leukemia.test<-GLP(X,class,components=1:4,return.LPT=TRUE)
X.new<-leukemia.test$LPT # LP-Transformed Features
#use "X.new" as input feature matrix to routines such as
#GLMNET, SVM, Random Forest, etc.
```

We hope this software will encourage applied data scientists to apply our method for their real problems.

REFERENCES

AGRESTI, A. (2007). An introduction to categorical data analysis, 2nd edn. Hoboken. NJ: John Wiley & Sons, Inc.

BACCELLI, F. & MAKOWSKI, A. M. (1989). Multidimensional stochastic ordering and associated random variables. *Operations Research* **37**, 478–487.

BHATTACHARYA, B. (2017). A General Asymptotic Framework for Distribution-Free Graph-Based Two-Sample Tests. *preprint arXiv:1508.07530*.

BIRNBAUM, A. (1954). Combining independent tests of significance. *Journal of the American Statistical Association*, **49**, 559–574.

BONNINI, S., CORAIN, L., MAROZZI, M., & SALMASO, L. (2014). Nonparametric hypothesis testing: rank and permutation methods with applications in R. John Wiley & Sons.

CHEN, H., CHEN, X. & SU, Y. (2017). A weighted edge-count two-sample test for multivariate and object data. *Journal of the American Statistical Association (forthcoming)* DOI: 10.1080/01621459.2017.1307757.

CHEN, H. & FRIEDMAN, J. H. (2017). A new graph-based two-sample test for multivariate and object data. *Journal of the American Statistical Association* **112**, 397–409.

DAVIDOV, O. & PEDDADA, S. (2013). The linear stochastic order and directed inference for multivariate ordered distributions. *Annals of statistics* **41**, 1.

EFRON, B. & TIBSHIRANI, R. (2007). On testing the significance of sets of genes. *The Annals of Applied Statistics* , 107–129.

FRIEDMAN, J. H. & RAFSKY, L. C. (1979). Multivariate generalizations of the wald-wolfowitz and smirnov two-sample tests. *The Annals of Statistics* , 697–717.

HOLLANDER, M., WOLFE, D. A. & CHICKEN, E. (2013). *Nonparametric statistical methods*. John Wiley & Sons.

HUDGENS, M. & SATTEN, G. (2002). Midrank unification of rank tests for exact, tied, and censored data. *Journal of Nonparametric Statistics* **14**, 569–581.

JONCKHEERE, A. R. (1954). A distribution-free k-sample test against ordered alternatives. *Biometrika* **41**, 133–145.

LANCASTER, H. (1961). Significance tests in discrete distributions. *Journal of the American Statistical Association* **56**, 223–234.

MAROZZI, M. (2013). Nonparametric simultaneous tests for location and scale testing: a comparison of several methods. *Communications in Statistics-Simulation and Computation* **42**, 1298–1317.

MOSES, L. E. (1963). Rank tests of dispersion. *The annals of mathematical statistics*, 973–983.

PARZEN, E. & MUKHOPADHYAY, S. (2014). LP Approach to Statistical Modeling. *Preprint arXiv:1405.2601*.

PARZEN, E. & MUKHOPADHYAY, S. (2013). LP Mixed Data Science: Outline of Theory. *Preprint arXiv:1311.0562* .

PARZEN, E. (1997). Concrete statistics. *Statistics in Quality, New York: Marcel Dekker* pp. 309–332.

PESARIN, F. & SALMASO, L. (2010a). Finite-sample consistency of combination-based permutation tests with application to repeated measures designs. *Journal of Nonparametric Statistics* **22**, 669–684.

PESARIN, F. & SALMASO, L. (2010b). *Permutation tests for complex data: theory, applications and software*. John Wiley & Sons.

RUYMGAART, F. H. (1980). A unified approach to the asymptotic distribution theory of certain midrank statistics. *Statistique non Parametrique Asymptotique*, pp. 1–18, J.P. Raoult (Ed.), Lecture Notes on Mathematics, N. 821, Springer, Berlin.

SUBRAMANIAN, A., TAMAYO, P., MOOTHA, V. K., MUKHERJEE, S., EBERT, B. L., GILLETTE, M. A., PAULOVICH, A., POMEROY, S. L., GOLUB, T. R., LANDER, E. S. et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences* **102**, 15545–15550.

ZHANG, J. & CHEN, H. (2017). Graph-Based Two-Sample Tests for Discrete Data. *arXiv:1711.04349*.