A Nonparametric Approach to High-dimensional K-sample Comparison Problem

BY SUBHADEEP MUKHOPADHYAY AND KAIJUN WANG

Department of Statistical Science, Temple University
Philadelphia, Pennsylvania, 19122, U.S.A.
deep@temple.edu  kaijun.wang@temple.edu

Summary

High-dimensional k-sample comparison is a common applied problem. We construct a class of easy-to-implement nonparametric distribution-free tests based on new tools and unexplored connections with spectral graph theory. The test is shown to possess various desirable properties along with a characteristic exploratory flavor that has practical consequences. The numerical examples show that our method works surprisingly well under a broad range of realistic situations.

Some key words: Graph-based nonparametrics; High-dimensional k-sample problem; Spectral graph partitioning; Distribution-free methods; High-dimensional exploratory analysis.

1. Introduction

1.1. The Problem Statement

The goal of this paper is to introduce some new theory for nonparametric multivariate k-sample comparison problems. Suppose k mutually independent random samples with sizes n₁, ..., nₖ are drawn, respectively, from the unknown distributions Gᵢ(x), i = 1, ..., k in \( \mathbb{R}^d \). We wish to test the following null hypothesis:

\[ H_0 : G_1(x) = \cdots = G_k(x) \text{ for all } x, \]

with the alternative hypothesis being \( H_A : G_r(x) \neq G_s(x) \) for some \( r \neq s \in \{1, \ldots, k\} \). This fundamental comparison problem frequently appears in a wide range of data-rich scientific fields: astrophysics, genomics, neuroscience, and chemometrics, just to name a few.

1.2. Existing Graph-based Nonparametric Approaches (k = 2)

Graphs are becoming an increasingly useful tool for nonparametric multivariate data analysis. Their significance has grown considerably in recent years due to its exceptional ability to tackle high-dimensional problems, especially when the dimension of the data is much larger than the sample size—a regime where the classical multivariate rank-based k-sample tests (Oja & Randles, 2004; Puri & Sen, 1971) are not even applicable. Lionel Weiss (1960) made an early attempt to generalize classical nonparametric two-sample procedures in the multivariate settings by utilizing nearest-neighbor-type graphs. Although the idea was intriguing, the main impediment to its use came from the fact that the null distribution of the proposed test statistic was distribution-dependent. The drawback was later addressed by Friedman & Rafsky (1979), providing the first practical graph-based algorithm, known as the edge-count test based on the idea of minimal spanning tree. This line of work led to a flurry of research, resulting in some promising new generalizations. For example, Chen & Friedman (2017) noted that the edge-count test has “low or even no power for scale alternatives when the dimension is moderate to high unless the sample size is astronomical due to the curse-of-dimensionality.” To counter this undesirable aspect the authors proposed a generalized edge-count test based on a clever Mahalanobis distance-based solution for two-sample location-scale alternatives. This was further extended in Chen et al. (2017) to tackle another shortcoming of the Friedman and Rafsky’s approach, namely the sample imbalance problem. The authors apply appropriate weights
to different components of the classical edge-count statistics to construct a weighted edge-count test. Rosenbaum (2005) proposed a distribution-free test based on minimum distance non-bipartite pairing, and Biswas et al. (2014) used the concept of shortest Hamiltonian path of a graph for two-sample testing.

Broadly speaking, all of the foregoing graph-based methods share the following characteristics:

- Define the pooled sample size \( n = n_1 + \cdots + n_k \). For \( i = 1, \ldots, n \), let \( x_i \) denotes the feature vector corresponding to the \( i \)-th row of the \( n \times d \) data matrix \( X \). Construct a weighted undirected graph \( G \) based on pairwise Euclidean distances between the data points \( x_i \in \mathbb{R}^d \) for \( i = 1, \ldots, n \).

- Compute a subgraph \( G^* \), containing certain subset of the edges from the original weighted graph \( G \) using some optimality criterion such as shortest Hamiltonian path or minimum non-bipartite matching or minimum weight spanning tree, which often requires sophisticated optimization routines like Kruskal’s algorithm or Derigs shortest augmentation path algorithm for efficient implementation.

- Compute cross-match statistics by counting the number of edges between samples from two different populations.

- The common motivation behind all of these procedures is to generalize univariate Wald-Wolfowitz runs test (Wald & Wolfowitz, 1940) for large dimensional problems.

The general idea here is to reformulate the multivariate comparison problem as a study of the structure of a specially designed graph that are constructed from the given data; see Section 2.3 for more details. We end this section with some key references to other related work, like Schilling (1986); Henze (1988); Hall & Tajvidi (2002); Rousson (2002); Gretton et al. (2012) and Bhattacharya (2017).

1-3. Some Desiderata and Goals

The main challenge is to design a nonparametric comparison test that continues to work for high-dimensional (and low-sample size) multi-sample case, and is distribution-free. In particular, an ideal nonparametric multivariate comparison test should

(D1.) be robust, not unduly influenced by outliers. This is an important issue, since detecting outliers is quite difficult in large-dimensional settings. Figs 4 (e) and (f) show that current methods are awfully vulnerable for datasets contaminated with even small percentage of outliers. Thus there is a need for robust nonparametric multivariate tests.

(D2.) allow practitioners to systematically construct tests for high-order alternatives, beyond conventional location-scale. Figs 4 (c) and (d) demonstrate the need for such a test. Complex ‘shape’ differences frequently arise in real-world data; see, for example: p53 geneset enrichment data of Section 2.7.

(D3.) be valid for any combination of discrete and continuous covariates—a common trait of real-world high dimensional data such as the Kyphosis data in Fig 5 (b) where existing methods perform poorly. The same holds true for Figs 4 (g) and (h), which further reinforce the need to develop mixed data algorithms.

(D4.) provide the user with some insight into why the test was rejected. In fact, data scientists need to be able to see not just the final result as a form of a single p-value, but also to understand and interpret why they are arriving at that particular conclusion, as in tables 1-4. These additional insights might lead to enhanced predictive analytics at the next modeling phase, as discussed in Sec. 3-3 for brain tumor data with \( k = 5, d = 5597 \), and \( n = 45 \).

(D5.) work for general \( k \)-sample problem. Surprisingly, all currently available graph-based comparison tests work only for two-sample problems. The main reason for this is that all the existing methods aim to generalize univariate run test to high dimension. Now, interestingly, \( k \)-sample generalization of run test is known to be hopelessly complicated even in univariate case—so much so that Mood (1940), in discussing the asymptotic distribution of run test for more than two samples, has not hesitated to say “that such a theorem would hardly be useful to the statistician, and the author does not feel that it would be worthwhile to go through the long and tedious details merely for the sake of completeness.” A convenient method that works for general \( k \)-sample would thus be highly valuable for practical reasons.

In this paper, we begin the process of considering an alternative class of nonparametric test based on new tools and ideas, which we hope will be useful in light of D1-D5.
Nonparametric High-dimensional K-sample Comparison

2. Theory and Methods

2.1. Nonlinear Data-Transformation

Suppose we are given \( \{(Y_i, X_i) : i = 1, \ldots, n\} \) where \( Y_i \in \{1, \ldots, k\} \) denotes the class membership index, and \( X_i \in \mathbb{R}^d \) is the associated multidimensional feature. \( n_g \) is the number of samples from class \( g \) and \( n = \sum_{g=1}^{k} n_g \). Our first task will be to provide a universal mechanism for constructing a new class of nonparametric data-graph kernel, useful for \( k \)-sample problems.

By \( \{F_j\}_{j=1}^{\infty} \) we denote the marginal distributions of a \( d \)-variate random vector. Given \( X_1, \ldots, X_n \) random sample from \( F_j \) construct the polynomials \( \{T_i(X; F_j)\}_{i \geq 1} \), for the Hilbert space \( L^2(F_j) \) by applying Gram-Schmidt orthonormalization on the set of functions \( \{\zeta, \tilde{F}_j, \ldots\} \):

\[
\zeta(x; F_j) = \frac{\sqrt{12}\left(\tilde{F}_j(x) - 1/2\right)}{\sqrt{1 - \sum_{x \in \mathcal{W}} \tilde{P}_j^3(x)}},
\]

where \( \tilde{\mu}_j \) and \( \tilde{F}_j \) respectively denote the empirical probability mass function, and distribution function; \( \tilde{F}_j(x) = \tilde{F}_j(x) - 0.5\tilde{\mu}_j(x) \) is known as the mid-distribution transform; \( \mathcal{W} \) denotes the set of all distinct observations of \( X \). We call this system of specially-designed orthonormal polynomials of mid-distribution transforms as empirical LP-basis. Two points should be emphasized. First, the number of empirical LP-basis \( m < \|\mathcal{W}\| \). As an example, for \( X \) binary, we can construct at the most one LP-basis. Second, the shape of these empirical-polynomials are not fixed, they are data-adaptive which make them inherently nonparametric by design. The nomenclature issue of LP-basis is discussed in Supplementary Material S1.

2.2. Theoretical Motivation

To appreciate the motivation behind this nonparametrically designed system of orthogonal functions, consider the univariate two-sample \((Y, X)\) problem. We start by deriving the explicit formulae of the LP-basis functions for \( X \) and \( Y \).

**Theorem 1.** Given independently drawn \( \{(Y_i, X_i), i = 1, \ldots, n = n_1 + n_2\} \) where \( Y \in \{0,1\} \) and \( X \in \mathbb{R} \) with the associated pooled rank denoted by \( R_i = \text{rank}(X_i) \), the first few empirical LP-polynomials for \( X \) and \( Y \) are given by:

\[
T_1(y_i; \tilde{F}_Y) = \begin{cases} \left\lfloor \frac{n_2}{n_1} \right\rfloor \text{ for } i = 1, \ldots, n_1 \\ \left\lfloor \frac{n_1}{n_2} \right\rfloor \text{ for } i = n_1 + 1, \ldots, n. \end{cases}
\]

\[
T_\ell(x_i; \tilde{F}_X) = \begin{cases} \frac{\sqrt{12}}{n} \left[R_i - \frac{n + 1}{2}\right] & \text{for } \ell = 1 \\ \frac{6\sqrt{3}}{n^2} \left[R_i - \frac{n + 1}{2}\right]^2 - \frac{\sqrt{7}}{2} & \text{for } \ell = 2. \end{cases}
\]

where \( \ell \) denotes the order of the polynomial.

**Definition 1.** Define LP-comeans as the following cross-covariance inner product

\[
LP[j, \ell; Y, X] = \mathbb{E}[T_j(Y; F_Y)T_\ell(X; F_X)], \quad j, \ell > 0.
\]

One can easily estimate the LP-comeans by substituting its empirical counterpart: \( \hat{LP}[j, \ell; Y, X] = n^{-1} \sum_{i=1}^{n} T_j(y_i; \tilde{F}_Y)T_\ell(x_i; \tilde{F}_X) \), which can be computed in \( \mathbb{R} \) by \( \text{cov}(\{T_j(y_i; \tilde{F}_Y), T_\ell(x_i; \tilde{F}_X)\}) \). This immediately leads to the following surprising identity.

**Theorem 2.** A compact LP-representation of the Wilcoxon and Mood Statistic is given by

\[
\hat{LP}[1, 1; Y, X] \equiv \text{Wilcoxon statistic for testing equality of location or means},
\]

\[
\hat{LP}[1, 2; Y, X] \equiv \text{Mood statistic for testing equality of variance or scale}.
\]
The proofs of Theorems 1 and 2 are deferred to the Appendix. Our LP-Hilbert space inner-product representation automatically produces ties-corrected linear rank statistics for $X$ discrete by appropriately standardizing by the factor $1 - \sum_{x \in X} \tilde{p}_X(x)$; cf. Chanda (1963), and Hollander et al. (2013, ch 4 p. 118). This implies that sum of squares of LP-comeans

$$\sum_{\ell > 0} |LP[1, \ell; Y, X]|^2$$

(5)
can provide credible and unified two-sample test to detect univariate distributional differences. Our goal in this paper will be to introduce a multivariate generalization of this univariate idea that is applicable even when the dimension exceeds the sample size – a highly non-trivial modeling task. The crux of our approach lies in interpreting the data-adaptive LP-basis transformations as specially designed nonlinear discriminator functions, which paves the way for extending it to encompass multivariate problems.

2.3. **LP Graph Kernel**

Here we are concerned with LP-graph kernel, the first crucial ingredient for constructing a graph associated with the given high-dimensional data.

**Definition 2.** Define the $l$-th order LP Gram Matrix $W_{l}^{LP} \in \mathbb{R}^{n \times n}$ as

$$W_{l}^{LP}(i, j) = \left( c + \langle \Phi_{l}^{LP}(x_i), \Phi_{l}^{LP}(x_j) \rangle \right)^2,$$

(6)
a polynomial kernel (degree=2) where $\Phi_{l}^{LP}: \mathbb{R}^{d} \rightarrow \mathbb{R}^{d}$ denotes the feature map in the LP space for $l = 1, 2, \ldots$

$$\Phi_{l}^{LP}: (x_1, x_2, \ldots, x_d)^T \mapsto \left\{ T_1(x_1, \tilde{F}_1), T_1(x_2, \tilde{F}_2), \ldots, T_l(x_d, \tilde{F}_d) \right\}^T,$$

(7)
as before the function $T_l(\cdot; \tilde{F}_j)$ denotes $l$-th order empirical LP-polynomial associated with $\tilde{F}_j$.

The positive symmetric kernel $W_{l}^{LP}: \mathbb{R}^{d} \times \mathbb{R}^{d} \rightarrow \mathbb{R}^{+}$ encodes the similarity between two $d$-dimensional data points in the LP-transformed domain. From $X$ and $W_{l}^{LP}$, one can thus construct a weighted graph $G = (V, W_{l}^{LP})$ of size $n$, where the vertices $V$ are data points $\{x_1, \ldots, x_n\}$ with edge weights given by LP-polynomial kernel $W_{l}^{LP}(x_i, x_j)$.

**Analysis pipeline:** $X \in \mathbb{R}^{n \times d} \rightarrow \ell$-th order LP-Transform $\rightarrow W_{l}^{LP} \in \mathbb{R}^{n \times n} \rightarrow G(V, W_{l}^{LP})$.

The resulting graph captures the topology of the high-dimensional data cloud in the LP-transformed domain. In the next section, we introduce a novel reformulation of the $k$-sample problem as a supervised structure learning problem of the learned LP-graph.

2.4. **Equivalence to Graph Partitioning**

Having reformulated the high-dimensional comparison as a graph-problem, we focus on understanding its structure. In the following, we describe an example to highlight how LP graph-based representation can provide rich insights into our $k$-sample learning objectives.

Consider a three-sample testing problem based on the following location-alternative model: $G_i = N_d(\delta_i 1_d, I_d)$ with $\delta_1 = 0, \delta_2 = 1.5, \delta_3 = 3$, dimension $d = 500$, and the sample sizes $n_i$’s are equal to 25 for $i = 1, 2$ and $3$. Fig 1 shows the inhomogeneous connections of the learned graph $G = (V, W_{l}^{LP})$. It is evident from the connectivity matrix that LP-edge density $W_{l}^{LP}(i, j)$ is significantly “higher” when both $x_i$ and $x_j$ coming from the same distribution, compare to different distributions. This implicitly creates a natural clustering or division of the vertices into $k$-groups (here $k = 3$) or communities. Naturally, under the null hypothesis (when all $G_i$’s are equal), we would expect to see one homogeneous (edge densities) graph of size $n$ with no community structure. In order to formalize this intuition of finding densely connected subgraphs, we need some concepts and terminology first.

The objective is to partition $V$ into $k$ non-overlapping groups of vertices $V_g$, $g = 1, \ldots, k$, where in our case $k$ is known. To better understand the principle behind graph partitioning, let us first consider the task of grouping the nodes into two clusters (i.e., the two-sample case). A natural criterion would be to
partition graph into two sets $V_1$ and $V_2$ such that the weight of edges connecting vertices in $V_1$ to vertices in $V_2$ is minimum. This can be formalized using the notion of graph cut:

$$\text{Cut}(V_1, V_2) = \sum_{i \in V_1, j \in V_2} W^{LP}(i, j).$$

By minimizing this cut value, one can optimally bi-partition the graph. However, in practice, minimum cut criteria does not yield satisfactory partitions and often produces isolated vertices due to the small values achieved by partitioning such nodes.

One way of getting around this problem is to design a cost function that prevents this pathological case. This can be achieved by normalizing the cuts by the volume of $V_i$, where $\text{Vol}(V_i) = \sum_{j \in V_i} \text{deg}_j$ and $\text{deg}_j$ is the degree of $i$-th node defined to be $\sum_{j=1}^{n} W^{LP}(i, j)$. Partitioning based on Normalized cut (Ncut) implements this idea by minimizing the following cost function:

$$\text{NCut}(V_1, \ldots, V_k) = \sum_{g=1}^{k} \frac{\text{Cut}(V_g, V - V_g)}{\text{Vol}(V_g)},$$

which was proposed and investigated by Shi & Malik (2000).

**THEOREM 3 (CHUNG, 1997).** Define the indicator matrix $\Psi = (\psi_1, \ldots, \psi_k)$ where

$$\psi_{j,g} = \begin{cases} \sqrt{\frac{\text{deg}_j}{\text{Vol}(V_g)}}, & \text{if } j \in V_g \\ 0, & \text{otherwise.} \end{cases}$$

for $j = 1, \ldots, n$ and $g = 1, \ldots, k$. Then the $k$-way Ncut minimization problem (8) can be equivalently rewritten as

$$\min_{V_1, \ldots, V_k} \text{Tr}(\Psi^T \mathcal{L} \Psi) \quad \text{subject to} \quad \Psi^T \Psi = I, \text{ and } \Psi \text{ as in (9)},$$

where $\mathcal{L}$ is known as the (normalized) Laplacian matrix given by $D^{-1/2} W^{LP} D^{-1/2}$ and $D$ is the diagonal matrix of vertex degrees.
This discrete optimization problem (10) is unfortunately NP-hard to solve. Hence, in practice relaxations are used for finding the optimal Ncut by allowing the entries of the $\Psi$ matrix to take arbitrary real values:

$$\min_{\Psi \in \mathbb{R}^{n \times k}} \text{Tr} (\Psi^T \mathcal{L} \Psi) \quad \text{subject to} \quad \Psi^T \Psi = I.$$  \hspace{1cm} (11)

The resulting trace minimization problem can be easily solved by choosing $\Psi$ to be the first $k$ generalized eigenvectors of $\mathbf{L} \mathbf{u} = \lambda \mathbf{D} \mathbf{u}$ as columns—Rayleigh-Ritz theorem. Spectral clustering methods (Von Luxburg, 2007, Sec. 5) convert the real-valued solution into a discrete partition (indicator vector) by applying $k$-means algorithms on the rows of eigenvector matrix $\mathbf{U}$; see Remark A1 of Appendix.

![Fig. 2: Laplacian spectral analysis of 3-sample location problem using LP-graph kernel (cf. Sec 2-4). The non-trivial singular values are shown in (a). The top ones ($\lambda_2, \lambda_3$) are marked with red diamonds. (b) and (c) display the dominant singular vectors ($u_2, u_3$); colors indicate the true group labels ($Y$).](image)

The fundamental function for dependence learning is the ‘normed joint density,’ pioneered by Hoeffding (1940), defined as the joint density divided by the product of the marginal densities:

$$\text{dep}(y_i, z_i; Y, Z) = \frac{p(y_i, z_i; Y, Z)}{p(y_i; Y)p(z_i; Z)},$$

which is a ‘flat’ function over the grid $\{1, \ldots, k\}^2$ under independence.
The test statistics (13) bear a surprising resemblance to (5), thus fulfilling a goal that we had set out to play the differences between the asymptotic p-values and permutation p-values for this chi-square null distribution in approximating p-values for finite samples. The boxplots in Figure 3 display the differences between the asymptotic p-values and permutation p-values for $G_1 = G_2 = N_d(0, I_d)$. The permutation p-values were computed by randomly shuffling the class-labels $Y$ 1000 times. Each simulation setting was repeated 250 times for all the combinations of $n_1$, $n_2$ and $d$. From the boxplots, we can infer that the theoretically predicted asymptotic p-values provide an excellent approximation to the permutation p-values for sample size $n \geq 100$. Furthermore, the limiting chi-square based p-values approximation continues to hold even for increasing dimension $d$.  

Theorem 4 implies that a test of independence is equivalent to determining whether all the $(k-1)^2$ parameters $\text{LP}[j, \ell; Y, Z]$ are zero. Motivated by this we propose our graph-based LP-nonparametric (abbreviated to GLP) test statistic:

$$\text{GLP statistic} = \sum_{j=1}^{k-1} \sum_{\ell=1}^{k-1} \left| \widehat{\text{LP}}[j, \ell; Y, Z] \right|^2. \tag{13}$$

The test statistics (13) bear a surprising resemblance to (5), thus fulfilling a goal that we had set out to achieve.

Theorem 5. Under the independence, the empirical LP-comeans $\widehat{\text{LP}}[j, \ell; Y, Z]$ have the following limiting null distribution as $n \to \infty$

$$\sqrt{n} \widehat{\text{LP}}[j, \ell; Y, Z] \overset{i.i.d.}{\sim} \mathcal{N}(0, 1).$$

Appendix A.3 contains the proof. Theorem 5 readily implies that the null distribution of our test statistic (13) is chi-square distribution with $(k-1)^2$ degrees of freedom. We seek to investigate the accuracy of this chi-square null distribution in approximating p-values for finite samples. The boxplots in Figure 3 display the differences between the asymptotic p-values and permutation p-values for $G_1 = G_2 = N_d(0, I_d)$. The permutation p-values were computed by randomly shuffling the class-labels $Y$ 1000 times. Each simulation setting was repeated 250 times for all the combinations of $n_1$, $n_2$ and $d$. From the boxplots, we can infer that the theoretically predicted asymptotic p-values provide an excellent approximation to the permutation p-values for sample size $n \geq 100$. Furthermore, the limiting chi-square based p-values approximation continues to hold even for increasing dimension $d$.  

\textbf{Definition 3.}  For given discrete $(Y, Z)$, the bivariate copula density $C : [0, 1]^2 \to \mathbb{R}_+$ is defined almost everywhere through

$$C(u, v; Y, Z) = \text{deep}[Q(u; Y), Q(v; Z); Y, Z] = \frac{p\{Q(u; Y), Q(v; Z); Y, Z\}}{p\{Q(u; Y)\}p\{Q(v; Z)\}}, \quad 0 \leq u, v \leq 1 \tag{12}$$

where $Q(\cdot)$ denotes the quantile function. It is not difficult to show that this quantile-domain copula density is a positive piecewise-constant kernel satisfying

$$\int_{[0,1]^2} C(u, v; Y, Z) \, du \, dv = \sum_{(i,j) \in \{1,\ldots,k\}^2} \int_{I_{ij}} C(u, v; Y, Z) \, du \, dv = 1,$$

where

$$I_{ij}(u, v) = \begin{cases} 1, & \text{if } (u, v) \in (F_Y(i), F_Y(i+1)] \times (F_Z(j), F_Z(j+1)] \\ 0, & \text{elsewhere.} \end{cases}$$

\textbf{Theorem 4.} The discrete checkerboard copula $C(u, v; Y, Z)$ satisfies the following important identity in terms of LP comeans between $Y$ and $Z$:

$$\int_{[0,1]^2} (C - 1)^2 \, du \, dv = \sum_{j=1}^{k-1} \sum_{\ell=1}^{k-1} \left| \text{LP}[j, \ell; Y, Z] \right|^2.$$

\textbf{Proof.} The key is to recognize that

$$\int_{[0,1]^2} C(u, v; Y, Z) T_j(Q(u; Y); F_Y) T_\ell(Q(v; Z); F_Z) \, du \, dv = \sum_{y,z} T_j(y; F_Y) T_\ell(z; F_Z) p(y, z; Y, Z),$$

which, using definition 1, can be further simplified as

$$E[T_j(Y; F_Y) T_\ell(Z; F_Z)] = \text{LP}[j, \ell; Y, Z].$$

Apply Parseval’s identity on the LP-Fourier copula density expansion to finish the proof.  

The test statistics (13) bear a surprising resemblance to (5), thus fulfilling a goal that we had set out to achieve.
Fig. 3: Boxplots of the difference between asymptotic p-values and p-values from 1000 permutations. Compared across different sample sizes $n_1$, $n_2$ and dimension $d$. Each setting was repeated 250 times.

2.6. Algorithm

This section describes the steps of our $k$-sample testing procedure that combines the strength of modern nonparametric statistics with spectral graph theory to produce an interpretable and adaptable algorithm.

**GLP: Graph-based Nonparametric K-sample Learning**

**Step 1.** Input: We observe the group identifier vector $Y$ and the data matrix $X \in \mathbb{R}^{n \times d}$. The features could be discrete, or continuous or even mixed.

**Step 2.** Construct LP-graph kernel $W_{\ell}^{LP} \in \mathbb{R}^{n \times n}$ using (6). The choice of $\ell$ depends on the type of testing problem: $\ell = 1$ gives $k$-sample test for mean, $\ell = 2$ for scale alternative and so on.

**Step 3.** Compute normalized Laplacian matrix $\mathcal{L}$ for the LP-learned graph $G = (V, W_{\ell}^{LP})$ by $D^{-1/2}W_{\ell}^{LP}D^{-1/2}$, where $D$ is the the diagonal degree matrix with elements $W_{\ell}^{LP}1_n$.

**Step 4.** Perform spectral decomposition of $\mathcal{L}$ and store leading nontrivial $k-1$ eigenvectors in the matrix $U \in \mathbb{R}^{n \times k-1}$.

**Step 5.** Apply k-means clustering by treating each row of $U$ as a point in $\mathbb{R}^{k-1}$ for Ncut community detection (11). Let $Z$, a vector of length $n$, denotes the cluster assignments obtained from the k-means.

**Step 6.** For each node of the constructed graph we now have bivariate random sample $(Y_i, Z_i)$ for $i \in V$. Perform correlation learning over graph by computing the GLP statistic (13). Compute the p-value using $\chi^2_{(k-1)/2}$ null distribution, as described in Sec 2.5.

**Step 7.** For complex multi-directional testing problems fuse the LP-graph kernels to create a super-kernel by $W^{LP} = \sum_{\ell} W_{\ell}^{LP}$.

(7a) Merging: As an example, consider the general location-scale alternatives—a targeted testing. Here we compute $W^{LP}$ by taking sum of $W_1^{LP} + W_2^{LP}$ and repeating the steps 3-6 for the testing.

(7b) Filtering: Investigators with no precise knowledge of the possible alternatives can combine informative $W_{\ell}^{LP}$ based on the p-value calculation of step 6 after adjustment for multiple comparisons to construct a tailored graph kernel, as in Tables 1, 3 and 4.

2.7. Illustration Using p53 Gene-set Data

Here we demonstrate the functionality of the proposed GLP algorithm using p53 geneset data (www.broad.mit.edu/gsea): it contains transcriptional profiles of 10, 100 genes in 50 cancer cell lines over two classes: $n_1 = 17$ classified as normal and $n_2 = 33$ as carrying mutations. The genes were cataloged by Subramanian et al. (2005) into 522 genesets based on known biological pathway informa
Nonparametric High-dimensional K-sample Comparison

For our illustration purpose we focus on two genesets: (a) “SA_G1_AND_S_PHASES” with number of genes $d = 14$, and (b) “anthraxPathway” with number of genes $d = 2$. Without any prior knowledge about the type of alternatives, we apply steps 2-6 of GLP algorithm to create the component-wise decomposition as depicted in Table 1. Our method decomposes the overall statistics-departure from the null hypothesis of equality of high-dimensional distributions-into different orthogonal components, providing insights into the possible alternative direction(s). For example, the pathway “SA_G1_AND_S_PHASES” shows location shift, whereas the output for “anthraxPathway” indicates the difference in the tails. This refined insights could be very useful for a biologist to better understand why any particular geneset is important.

Once we identified the informative directions, we compute the super-kernel (step 7 of our algorithm) by fusing $W^{LP} = \sum_{\ell \in \text{sig.}} W^{LP}_\ell$. In the case of “anthraxPathway”, this simply implies $W^{LP} \leftarrow W^{LP}_4$ and so on for other genesets. This combined $W^{LP}$ effectively packs all the principal directions (filtering out the uninteresting ones) into a single LP-graph kernel. At the final stage, we execute steps 3-5 using this combined kernel to generate the overall $k$-sample GLP statistic along with its p-value, shown in the bottom row of Table 1.

Table 1: GLP multivariate $k$-sample test for p53 data. The table shows the output of our GLP algorithm, given in Sec 2.6. The overall statistic provides the global $k$-sample confirmatory test and the individual ‘components’ give exploratory insights into how the multivariate distributions are different. The significant components based on p-values adjusted for multiple comparisons are marked with an asterisk ‘*’

| Component | (a) GLP p-value | (a) p-value | (b) GLP p-value | (b) p-value |
|-----------|----------------|-------------|----------------|-------------|
| 1         | 0.145          | 0.007*      | 3.84 $\times 10^{-4}$ | 0.890      |
| 2         | 0.045          | 0.136       | 0.003          | 0.720       |
| 3         | 0.002          | 0.754       | 0.045          | 0.136       |
| 4         | 0.034          | 0.196       | 0.131          | 0.011*      |
| overall   | 0.145          | 0.007       | 0.125          | 0.012       |

Interestingly, Subramanian et al. (2005) found the geneset “SA_G1_AND_S_PHASES” to be related to P53 function but missed “anthraxPathway,” the reason being they have used Kolmogorov-Smirnov test which is known to exhibit poor sensitivity for tail deviations; also see Supplementary Material S10.

Table 2: Comparing the five tests for two genesets: (a) SA_G1_AND_S_PHASES; (b) anthraxPathway. The dimensions are 14 and 2, respectively. The reported numbers are all p-values

| Geneset       | GLP  | FR  | Rosenbaum | HP   | GEC  |
|---------------|------|-----|-----------|------|------|
| (a)           | 0.007| 0.025| 0.061     | 0.041| 0.141|
| (b)           | 0.010| 0.327| 0.591     | 0.078| 0.904|

Methods abbreviation: FR is Friedman & Rafsky’s test; GEC is generalized edge count method by Chen & Friedman (2017); HP is Biswas’ shortest Hamiltonian path method. And GLP is the proposed nonparametric graph-based test.

Finally, it is also instructive to examine our findings in the light of other existing methods. This is done in Table 2. In the first case (a), where we have location difference, all methods have no problem finding the significance. However, interestingly enough, most of the competing methods start to fail for (b) where we have higher-order tail difference. This will be further clarified in the next section, where we present numerical studies to better understand the strengths and weaknesses of different methods.
3. NUMERICAL RESULTS

3.1. Power Comparison

We conduct extensive simulation studies to compare the performance of our method with four other methods: Friedman-Rafsky’s edge-count test, generalized edge-count test, Rosenbaum’s cross matching test and Shortest Hamiltonian Path method. All the simulations in this section are performed as follows: (i) We focus on the two-sample case, as all the others methods are only applicable in this setup ($k = 5$ case is reported in section 3.3) with sample sizes $n_1 = n_2 = 100$; (ii) each case is simulated 100 times to approximate the power of the respective test. The performance of the algorithms was examined under numerous realistic conditions, as we will shortly note.

In the first example, we investigate the location alternative case with two groups generated by $G_1 = \mathcal{N}(0, I_d)$ and $G_2 = \mathcal{N}(0.5, I_d)$ with dimension $d$ ranging from 2 to $2^{10} = 1024$. The result is shown in Fig. 4 (a). For small and medium dimensions, our proposed test performs the best; for moderately large dimensions all the methods are equally powerful. In example two, we examine the scale case by choosing $G_1 = \mathcal{N}(0, I_d)$ and $G_2 = \mathcal{N}(0, 1.5I_d)$. Here generalized edge-count test reaches the best performance, followed by our test; Friedman and Rafskys test completely breaks down. The third example is the general location-scale case $G_1 = \mathcal{N}(0, I_d)$ and $G_2 = \mathcal{N}(0.31 I_d, 1.3I_d)$. The estimated power function is shown in Fig 4 (c). Our method still displays high performance, followed by the Hamiltonian Path. Example four explores the tail alternative case: $G_1 = \mathcal{N}(0, I_d)$ and $G_2$ is $T_3(0, I_d)$ Student’s t-distribution with degrees of freedom 3. Not surprisingly, edge-count and generalized edge-count tests suffer the most, as they are not designed for these sorts of higher-order complex alternatives. Both, our approach and Hamiltonian test,
Fig. 5: Real data empirical power comparisons: (a) Ionosphere data; (b) Kyphosis; (c) Phoneme Data; and (d) Leukemia Data.

exhibit excellent performance, which also explains our Table 2 finding for “anthraxPathway.” Discrete data is also an important aspect of two-sample problems. We checked the case of location difference by generating the samples with $G_1 = \text{Poisson}(5)$ vs $G_2 = \text{Poisson}(5.5)$ in fifth example, which is depicted in Fig 4 (g). Here all methods perform equally well, especially for large dimensional cases. Example six delves into the important robustness issue. For that purpose, we introduce perturbation of the form $\epsilon_i \sim (1 - \eta)N(0, 1) + \eta N(\delta_i, 3)$ where $\delta_i = \pm 20$ with probability $1/2$ and $\eta$ varies from 0 to 0.3. Empirical power curves are shown in Figs 4 (e) and (f). Our proposed test shows extraordinary stability in both cases. The rest of the algorithms reveal their extreme sensitivity towards outliers, so much so that the presence of even .05% in location-alternative case can reduce the efficiency of the methods by nearly 80%, which is a striking number. In the final example we explore the interesting case of mixed alternatives. Here the idea is to understand the performance of the methods when different kinds of alternative hypotheses are mixed up. To investigate that we generate the first group from $G_1 = N(0, I_d)$, and for the alternative group we generate a portion (also 50 samples) from $N(0.31, I_{d_1})$ and another portion from $N(0, 1.3I_{d_2})$, where $d = d_1 + d_2$, and $r = d_2/d$. Fig 4 (h) shows that generalized edge-count test, Hamiltonian method and our test perform best. Additional simulation results are given in the Supplementary Materials S2-S6.

3.2. Analysis of Benchmark Datasets

Comparisons using the aforementioned methods are also done on benchmark datasets. For each dataset, we performed the testing procedures on subsets of the whole data, so that we can approximate the rejection power. For our resampling, (i) several sub-sample sizes are specified, and we pick randomly and evenly from the two groups in full data to form the subsets (ii) Each resampling is repeated 100 times. The results are shown on Figure 5.
The first example studied ionosphere data found on the UCI machine learning repository. The dataset is comprised of \(d = 34\) features and \(n = 351\) instances grouped into two classes: good or bad radar returns with \(n_1 = 225, n_2 = 126\). Re-samplings are performed with subsample size ranging from 10 to 60. Figure 5 (a) shows the proposed method have performance on par with Hamiltonian path method, while others fall behind in power. The next example is kyphosis laminectomy data, available in \texttt{gam R} package. It has discrete covariates whose range of values differ greatly. We observe 17 instances from the kyphosis group and 64 instances from the normal group. Figure 5 (b) shows that all the existing tests yield noticeably poor performance. Even more surprisingly, the power of these tests does not increase with sample size. Our proposed test significantly outperforms all other competitors here. Next, we consider the phoneme dataset with two groups ‘aa’ and ‘ao.’ The data have a dimension of \(d = 256\), and re-sampling subsample sizes range from 10 to 150. Here edge-count and Hamiltonian path methods show better performance, and our method also performs well, as shown in Figure 5 (c). Our final example is the Leukemia cancer spanning \(k\) sizes range from \(10\) to \(n\) fall behind in power. The next example is kyphosis laminectomy data, available in \texttt{gam R} package. It has discrete covariates whose range of values differ greatly. We observe 17 instances from the kyphosis group and 64 instances from the normal group. Figure 5 (b) shows that all the existing tests yield noticeably poor performance. Even more surprisingly, the power of these tests does not increase with sample size. Our proposed test significantly outperforms all other competitors here. Next, we consider the phoneme dataset with two groups ‘aa’ and ‘ao.’ The data have a dimension of \(d = 256\), and re-sampling subsample sizes range from 10 to 150. Here edge-count and Hamiltonian path methods show better performance, and our method also performs well, as shown in Figure 5 (c). Our final example is the Leukemia cancer gene expression data with \(d = 7128\) genes in \(n_1 = 47\) ALL (Acute lymphoblastic leukemia) samples and \(n_2 = 25\) AML (Acute myeloid leukemia) samples. Data are re-sampled using total sample sizes from 10 to 50. In the case of small subset size the competing methods methods show higher power with the exception of Rosenbaum’s test. Nevertheless, for moderately large sample sizes all methods show equal power. The excellent performance of the Friedman and Rafskys method for this dataset can be easily understood from GLP Table 3, which finds only the first component to be significant, i.e., location-only alternative.

### 3.3. K-sample: Brain Data and Enhanced Predictive Model

The brain data (Pomeroy et al., 2002) contains \(n = 42\) samples from \(d = 5597\) gene expression profiles spanning \(k = 5\) different tumor classes of the central nervous system with group sizes: \(n_1 = n_2 = n_3 = 10, n_4 = 8\) and \(n_5 = 4\). This dataset is available at http://www.broadinstitute.org/mpr/CNS. We use this dataset to illustrate our method’s performance in testing \(k\)-sample problems where other methods are not applicable.

To start with, we have no prior knowledge about the possible alternatives and thus we first compute the component-wise \(p\)-values using our \(k\)-sample learning algorithm described in Sec 2.6. The result is shown in Table 4, which finds GLP orders 1 and 2 as the informative directions. Following step 7b of our algorithm, we then fuse the significant LP-graph kernels to construct the super-kernel \(W_{LP}^\ell = \sum_{\ell=1}^2 W_{LP}^\ell\). Applying our spectral graph correlation algorithm on this specially tailored kernel \(W_{LP}^\ell\) yields the final row of Table 4. Combining all of these, we infer that the high-dimensional gene-expression distributions differ in locations and scales in five different tumor classes.

| Component | GLP      | \(p\)-value |
|-----------|----------|-------------|
| 1*        | 0.209    | \(1.04 \times 10^{-4}\) |
| 2         | 0.022    | 0.207       |
| 3         | 0.002    | 0.703       |
| 4         | 0.033    | 0.121       |
| Overall   | 0.209    | \(1.04 \times 10^{-4}\) |

Table 3: Leukemia GLP Chart

| Component | GLP      | \(p\)-value |
|-----------|----------|-------------|
| 1*        | 0.235    | 0.021       |
| 2*        | 1.933    | 5.83 \times 10^{-30} |
| 3         | 0.070    | 0.760       |
| 4         | 0.070    | 0.760       |
| Overall   | 0.233    | 0.022       |

Table 4: GLP chart for Brain Data

Consequently, our technique provides the global confirmatory testing result along with the insights into the possible reasons for rejecting the null hypothesis. This additional information, which is absent from extant technologies, allows us to go further than testing by designing enhanced predictive algorithms. To investigate this assertion, we designed two learning schemes based on two different feature matrices: (i) the original \(X \in \mathbb{R}^{n \times d}\), and the data-adaptive (ii) \(T \in \mathbb{R}^{n \times 2d}\) which is \([T_1 \mid T_2]\), where the \(j\)th column of the matrix \(T_j\) is simply \(T_\ell(x; F_j)\)--the \(\ell\)-th LP-transform of covariate \(X_j\). We use 75% of the original data as a training set and the rest as testing set; we repeat this process 250 times. The models are fitted by multinomial lasso-logistic regression; leave-one-out cross validation is used for selecting the tuning parameter. For comparison we use multi-class log-loss error given by \(- \frac{1}{n_{test}} \sum_{i=1}^{n_{test}} \sum_{j=1}^{k} y_{ij} \log(p_{ij})\).
Fig. 6: Logarithmic loss of the multinomial logistic regression with $k = 5$ classes, based on (left) original data matrix $X$ vs. (right) data-adaptive LP-transform matrix $\mathcal{T}$.

where $n_{\text{test}}$ is the size of the test set, $k$ is the number of groups, $\log$ is the natural logarithm, $y_{ij}$ is 1 if observation $i$ belongs to class $j$ and 0 otherwise; $p_{ij}$ is the predicted probability that observation $i$ belongs to class $j$. Figure 6 indicates that we can gain efficiency by incorporating LP-learned features in the classification model-building process. This aspect of our modeling makes it more powerful than merely testing the hypothesis. The user can use the output of our algorithm to upgrade the baseline predictive model. For discussion on R implementation, see Supplementary Materials S11.

4. DISCUSSION

This article provides some new graph-based modeling strategies for designing a nonparametric $k$-sample learning algorithm that is robust, increasingly automatic, and continues to work even when the dimension of the data is larger than the sample size. More importantly, it comes with an exploratory interface, which not only provides more insight into the problem but also can be utilized for developing a better predictive model at the next phase of data-modeling upon rejecting $H_0$. And for that reason, we call it a $k$-sample statistical learning problem—beyond simple testing, incorporating three modeling cultures: confirmatory, exploratory and predictive. In summing up, we must say that in designing the proposed algorithm, our priority has been flexibility as well as interpretation (D1-D5), in hopes of making it readily usable for applied data scientists. Our success comes from merging the strength of modern nonparametric statistics with the spectral graph analysis.

ACKNOWLEDGEMENT

The authors would like to thank the editor, associate editor, and reviewers for their insightful comments.

SUPPLEMENTARY MATERIAL

Supplementary material available at Biometrika online includes additional numerical and theoretical discussions. All datasets and the computing codes are available via open source R-software package LPKsample, available online at https://cran.r-project.org/package=LPKsample.
APPENDIX

A.1. Proof of Theorem 1

Recall $Y$ is binary with $\tilde{p}_Y(0) = n_1/n$ and $\tilde{p}_Y(1) = n_2/n$. Our goal is to find an explicit expression for the mid-transform function $T_1(y; \tilde{F}_Y) = \frac{\sqrt{12} \{ \tilde{F}_Y^{\text{mid}}(y) - 1/2 \}}{\sqrt{1 - \sum_{y \in \mathbb{Y}} p_Y^2(y)}}$.

We start by deriving the expression for the mid-distribution transform $\tilde{F}_Y^{\text{mid}}(y) = \tilde{F}_Y(y) - \frac{1}{2} \tilde{p}_Y(y)$:

$$\tilde{F}_Y^{\text{mid}}(y_i) = \begin{cases} -\frac{n_1}{2n} & \text{for } y_i = 0 \\ 1 - \frac{n_2}{2n} & \text{for } y_i = 1. \end{cases} \tag{A1}$$

Next we determine the re-normalizing factor $1 - \sum_{y \in \mathbb{Y}} \tilde{p}_Y^2(y) = 3n_1 n_2 / n^2$. Combining previous two results we obtain the empirical LP-basis for $Y$ as

$$T_1(y_i; \tilde{F}_Y) = \begin{cases} \sqrt{-\frac{n_1}{n^2} - \frac{n_1}{2n} - \frac{n_1}{2n}} & \text{for } i = 1, \ldots, n_1 \\ \sqrt{\frac{n_2}{n^2}} & \text{for } i = n_1 + 1, \ldots, n. \end{cases} \tag{A2}$$

For $X$ continuous, we now aim to derive its first two empirical LP-basis. As we will see these basis functions have a direct connection with ranks. Now note that $\tilde{F}_X^{\text{mid}}(x_i) = R_i - \frac{1}{2n}$, where $R_i = \text{rank}(x_i)$. Hence we immediately have

$$T_1(x_i; \tilde{F}_X) = \sqrt{\frac{12}{n^2} - \frac{1}{12}} \left( R_i - \frac{n + 1}{2} \right). \tag{A3}$$

This matches with the expression of $T_1(x; \tilde{F}_X)$ as given in (3), up to a negligible factor. Perform Gram-Schmidt orthonormalization of $\{T_1(x; \tilde{F}_X), T_2^2(x; \tilde{F}_X)\}$ to obtain the second empirical LP-basis of $X$. By routine calculation, we have

$$T_2(x_i; \tilde{F}_X) = 6\sqrt{5} \left\{ \left( \tilde{F}_X^{\text{mid}}(x_i) - 1/2 \right)^2 - \frac{1}{12} \right\}^{1/2}, \quad \text{for } i = 1, \ldots, n. \tag{A4}$$

Substituting the mid-transform function yields the desired result (3). This completes the proof.

A.2. Proof of Theorem 2

Applying Definition 1, we have

$$\tilde{\mathbb{L}}P[1, 1; Y, X] = n^{-1} \sum_{i=1}^{n} T_1(y_i; \tilde{F}_Y) T_1(x_i; \tilde{F}_X).$$

Substitute the expressions for the empirical LP-basis functions from Theorem 1 to verify that

$$\tilde{\mathbb{L}}P[1, 1; Y, X] = \frac{\sqrt{12}}{n^2} \left\{ - \frac{n_1}{n} \sum_{i=1}^{n_1} \left( R_i - \frac{n_1}{2} \right) + \sqrt{\frac{n_1}{n_2}} \sum_{i=n_1+1}^{n_1} \left( R_i - \frac{n_1 + 1}{2} \right) \right\}, \tag{A5}$$

which after some algebraic manipulation can be re-written as

$$\tilde{\mathbb{L}}P[1, 1; Y, X] = \frac{\sqrt{12}}{n^2 \sqrt{n_1 n_2}} \left\{ n_1 \sum_{i=n_1+1}^{n} R_i - n_2 \left( \frac{n(n+1)}{2} - R_i \right) \right\} \tag{A6}.$$
Complete the proof by noting that (A6) is in fact
\[ \sqrt{n} \hat{L}P[1; Y, X] = \sqrt{\frac{12}{n n_1 n_2}} \left\{ \sum_{i=n_1+1}^{n} R_i - \frac{n_2 (n+1)}{2} \right\}, \]
which is equivalent to the standardized Wilcoxon statistic up to a negligible factor \( \sqrt{\frac{n+1}{n}} \).
To derive the LP-representation of Mood statistic, we start with
\[ \hat{L}P[1, 2; Y, X] = n^{-1} \sum_{i=1}^{n} T_i(y_i; \bar{F}_Y) T_2(x_i; \bar{F}_X). \]
Proceeding as before we have
\[ \sqrt{n} \hat{L}P[1, 2; Y, X] = \frac{6 \sqrt{5}}{n^3} \left[ -\sqrt{\frac{n_2}{n_1}} \sum_{i=1}^{n} \left\{ (R_i - \frac{n+1}{2})^2 - \frac{5}{2} \right\} + \frac{n_1}{n_2} \sum_{i=n_1+1}^{n} \left\{ (R_i - \frac{n+1}{2})^2 - \frac{5}{2} \right\} \right]. \] (A7)
Routine calculations show that (A7) can be reduced to
\[ \sqrt{n} \hat{L}P[1, 2; Y, X] = \frac{\sqrt{180}}{n^3 n_1 n_2} \sum_{i=n_1+1}^{n} \left\{ (R_i - \frac{n+1}{2})^2 - \frac{n^2 + 2}{12} \right\}, \] (A8)
which is equivalent to the Mood statistic up to an asymptotically negligible factor. This proves the claim.

A.3. **Proof of Theorem 5**

Since under independence sample LP-comeans has the following weighted-average representation (where weights are marginal probabilities)
\[ \hat{L}P[j, \ell; Y, Z] = \sum_{1 \leq i_1 \leq \ldots \leq i_k \leq k} \hat{p}(i_1; Y) \hat{p}(i_2; Z) T_j(i_1; \bar{F}_Y) T_\ell(i_2; \bar{F}_Z), \quad j, \ell \in \{1, \ldots, k-1\}, \] (A9)
it is straightforward to show that in large samples the \( \hat{L}P[j, \ell; Y, Z] \) are independent and normally distributed by confirming \( |\sqrt{n} \hat{L}P[j, \ell; Y, Z]|^2 \) is the score statistic for testing \( H_0 : \hat{L}P[j, \ell; Y, Z] = 0 \) against \( \hat{L}P[j, \ell; Y, Z] \neq 0 \).

A.4. **Additional Remarks**

Remark A1. **Spectral relaxation**: Spectral clustering converts the intractable discrete optimization problem of graph partitioning (10) into a computationally manageable eigenvector problem (11). However, the eigenvectors of the Laplacian matrix \( U_{n \times k} \) will not in general be the desired piecewise constant form (9). Thus, naturally, we seek a piecewise constant matrix \( \hat{\Psi} \) closest to the “relaxed” solution \( U \), up to a rotation by minimizing the following squared Frobenius norm: \( \|U U^T - \hat{\Psi} \hat{\Psi}^T\|^2_F \). In an important result Zhang et al. (2008, Theorem 2) showed that minimizing this cost function is equivalent to performing k-means clustering on the rows of \( U \). This justifies why spectral clustering scheme is the closed tractable solution of the original NP-hard normalized cut problem, as described in Section 2.4.

Remark A2. **Joint covariate balance**: As a reviewer pointed out, researchers can use the proposed method for evaluating covariate balance in causal inference. The proposed GLP technology could be perfectly suitable for this task because: (i) it can tackle multivariate mixed data that is prevalent in observational studies, (ii) it goes beyond the simple mean difference and captures the distributional balance across exposed and non-exposed groups, and finally, (iii) it informs causal modeler the nature of imbalance—how the distributions are different within a principal stratum, which can be used to upgrade the logistic regression-based propensity-score matching algorithm (cf. Section 3.3) to reduce the bias.
REFERENCES

BHATTACHARYA, B. B. (2015). A general asymptotic framework for distribution-free graph-based two-sample tests. preprint arXiv:1508.07330.

BISWAS, M., MUKHOPADHYAY, M. & GHOSH, A. K. (2014). A distribution-free two-sample run test applicable to high-dimensional data. Biometrika 101, 913–926.

CHANDA, K. (1963). On the efficiency of two-sample mann-whitney test for discrete populations. The Annals of Mathematical Statistics, 612–617.

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.

CHUNG, F. R. (1997). Spectral graph theory, vol. 92. American Mathematical Soc.

COUILLET, R., BENAYCH-GEOGEES, F. et al. (2016). Kernel spectral clustering of large dimensional data. Electronic Journal of Statistics 10, 1393–1454.

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

GRETTON, A., BORGWARDT, K. M., RASCH, M. J., SCHÖLKOPF, B. & SMOLA, A. (2012). A kernel two-sample test. Journal of Machine Learning Research 13, 723–773.

HALL, P. & TAJVIDI, N. (2002). Permutation tests for equality of distributions in high-dimensional settings. Biometrika 89, 359–374.

HENZE, N. (1988). A multivariate two-sample test based on the number of nearest neighbor type coincidences. The Annals of Statistics, 772–783.

HOEFFDING, W. (1940). Massstabinvariante korrelationstheorie. Schriften des Mathematischen Seminars und des Instituts für Angewandte Mathematik der Universität Berlin 5, 179–233.

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

EL, KAROUI, N. (2010). The spectrum of kernel random matrices. The Annals of Statistics 38, 1–50.

MOOD, A. M. (1940). The distribution theory of runs. The Annals of Mathematical Statistics 11, 367–392.

OJA, H. & RANDLES, R. H. (2004). Multivariate nonparametric tests. Statistical Science, 598–605.

POMEROY, S. L., TAMAYO, P., GAASENBEEK, M. & STURLA, L. (2002). Prediction of central nervous system embryonal tumour outcome based on gene expression. Nature 415, 436–442.

PURI, M. L. & SEN, P. K. (1971). Nonparametric Methods in Multivariate Analysis. John Wiley & Sons, New York (US).

ROSENBAUM, P. R. (2005). An exact distribution-free test comparing two multivariate distributions based on adjacency. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67, 515–530.

ROUSSON, V. (2002). On distribution-free tests for the multivariate two-sample location-scale model. Journal of multivariate analysis 80, 43–57.

SCHILLING, M. F. (1986). Multivariate two-sample tests based on nearest neighbors. Journal of the American Statistical Association 81, 799–806.

SHI, J. & MALIK, J. (2000). Normalized cuts and image segmentation. Pattern Analysis and Machine Intelligence, IEEE Transactions on 22, 888–905.

SUBRAMANIAN, A., TAMAYO, P., MOOTH, V. K., MUKHERJEE, S., EBER, B. L., GILLETTE, M. A., PAULOLOVICH, 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.

VON LUXBURG, U. (2007). A tutorial on spectral clustering. Statistics and computing 17, 395–416.

VON LUXBURG, U., BELKIN, M. & BOUSQUET, O. (2008). Consistency of spectral clustering. The Annals of Statistics 36, 555–586.

WALD, A. & WOLFOWITZ, J. (1940). On a test whether two samples are from the same population. The Annals of Mathematical Statistics 11, 147–162.
WEISS, L. (1960). Two-sample tests for multivariate distributions. *The Annals of Mathematical Statistics* **31**, 159–164.

ZHANG, Z., JORDAN, M. I. (2008). Multiway spectral clustering: A margin-based perspective. *Statistical Science* **23**, 383–403.
Supplementary material for
‘A Nonparametric Approach to High-dimensional K-sample Comparison Problem’

BY SUBHADEEP MUKHOPADHYAY AND KAIJUN WANG

Department of Statistical Science, Temple University
Philadelphia, Pennsylvania, 19122, U.S.A.

deeptemple.edu kaijun.wang@temple.edu

SUMMARY

This supplementary document contains eleven 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 7 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 $\{\tilde{F}(x_i), T_j(x_i; \tilde{F})\}$ for $j = 1, \ldots, 4$ and $x_i \in \text{Unique}(x_1, \ldots, x_n)$. Note that they are discrete piecewise-constant orthonormal polynomials with respect to empirical measure $\tilde{F}$.

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: $N_d(0, I)$, Group 3: $N_d(0.251d, I)$, and Group 4: $N_d(0, 1.5I)$. 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.

Step 1. Table 5 shows the results of global $k$-sample Graph-based LP test. This immediately indicates that at least one of the population is different from others. Moreover, from that table we can also infer that...
Fig. 7: The shapes of the first four empirical LP-basis functions $T_1(x; \tilde{F}_X), \ldots, T_4(x; \tilde{F}_X)$ of the variable \textit{Start} in kyphosis dataset which are discussed in Sec 3-2 of the main paper. The \textit{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.

Table 5: $k$-sample Graph-based LP test results. The global statistic along with its components are given.

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

the distributions are different with respect to location and scale, which exactly matches our data generating mechanism.

\textbf{Step 2.} Next step is to examine more carefully at the specific pattern of differences among the distributions. Table 6 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 6: \( k \)-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 | GLP 1 | GLP 2 | GLP 3 | GLP 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 7 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 \( \mathcal{N}(0, I) \) vs \( \mathcal{N}(0.25, I) \), 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 7 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 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 \( \text{Vol}(V_i) = \sum_{v \in V_i} \text{deg}_v \), 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 8.

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

| 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) \( \Sigma_1 \) is identity matrix; \( \Sigma_2 \) have \( \sigma_{2,ij} = 1 \) for \( i = j \), and \( \sigma_{2,ij} = 0.5 \) for \( i \neq j \).
Nonparametric High-dimensional K-sample Comparison

- (b) \( \Sigma_1 \) is identity matrix; \( \Sigma_2 \) 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 \( k = 1, \ldots, [d/40] \) is the number of blocks and \( \sigma_{2,ij} = 0 \) otherwise.
- (c) \( \Sigma_1 \) is identity matrix; \( \Sigma_2 \) have \( \sigma_{2,ij} = \frac{1}{2} [(|i-j|+1)^{2\alpha} + (|i-j|-1)^{2\alpha} - 2(|i-j|)^{2\alpha}] \) 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 9: 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 9. 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 \( G_2 = \mathcal{N}_d(\delta \sqrt{n_1 + n_2} 1_d, I) \). The setting was repeated 1000 times over a grid of 10 \( \delta \)-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 8. 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 8. 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). Most popular ones are Fisher, Liptak, and Tippett combining functions. For an exhaustive treatment of Non-Parametric Combination based methods, see Pesaran 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 $\bigcap_{h=1}^{d} \{ F_{1h} \supseteq \cdots \supseteq F_{kh} \}$, 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, Kruskal-Wallis 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 2-7, 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:

\[
H_0 : X_1 \preceq \cdots \preceq X_k \quad \text{versus} \quad H_1 : X_1 \succeq \cdots \succeq X_k,
\]

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 \( X_{j1} \succeq \cdots \succeq X_{jk}, \ j = 1, \ldots, d \). This decoupling result makes the problem in some sense “dimension-free.” Consequently, one can rewrite (1) as

\[
H_0 : \cap_{j=1}^d [X_{j1} \preceq \cdots \preceq X_{jk}] \quad \text{versus} \quad H_1 : \cup_{j=1}^d [X_{j1} \succeq \cdots \succeq X_{jk}]
\]

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: \( X_1 \succeq \cdots \succeq X_4 \).

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., \( H_1 : X_1 \succeq X_2 \), where group 1: scientific studies, and group 2: students with classical studies background. Fig 9 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 10 performs preliminary GLP analysis, to further validate that the only difference lies in location.

Analysis: The \( k \)-sample testing for monotonic treatment effect can be reformulated as a correlation between \( T_1(y; F) \) and \( T_1(x; F) \) which is shown in left panel of Fig 10. 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

* 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.
additional tuning. For testing the multivariate ordered alternative we thus propose our global statistic as \( T_{LP} = \min_{1 \leq j \leq d} LP[1, 1; Y, X_j] \). We use the permutation distribution of \( T_{LP} \) to compute the overall p-value. The sign of \( LP[1, 1; Y, X_j] \) 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 10.

Table 11: Rat tumor data, trend tests comparison. Both permutation and asymptotic p-values are reported for our \( T_{LP} \) trend detection statistic.

| Variables | JT Test p-value | \( T_{LP} \) | LP Trend Test p-value (asymp) | LP Trend Test 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                         |

The results for the rat data is displayed in the Table 11. Key points are summarized below:

- The global k-sample multivariate LP-statistic for trend \( T_{LP} = \min_{1 \leq j \leq d} LP[1, 1; Y, X_j] = -4.91 \) with permutation p-value essentially 0. This strongly suggests the usefulness of the treatment as an anti-cancer agent.
Table 12: Admission data. Trend tests results.

| Variables  | JT Test p-value | LP Trend Test |
|------------|-----------------|---------------|
| $X_1$ (Math) | 0.018           | -2.404        |
| $X_2$ (Econ) | 0.912           | -0.155        |

Fig. 10: 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 $LP[1; 1; Y, X_{10}] = -0.75$. The right panel shows the group-specific means for each variables. Different colors represent different groups.

- 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 9; see Table 12 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.

88. 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.
It seems worthwhile to emphasize the fact that the mid-distribution function $F_{\text{mid}}$ 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 $F_{\text{mid}}$ 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 an 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.

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 11 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 $\ell$-th order LPcomens-based marginal enrichment score by

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

Table 13 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*     | 1.301      | 0.396      | 0.832      |
|         | (0.371)    | (0.164)    | (0.022)    | (0.061)    |
| (b)     | 2.374      | 1.852      | 1.542      | 3.235*     |
|         | (1.563)    | (1.309)    | (1.049)    | (1.358)    |

Table 13: 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 genes. Ranking of top gene pathways based on their differential profiles is portrayed in Fig 12. 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.
Nonparametric High-dimensional K-sample Comparison

Fig. 11: 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.

S11. 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.

```r
#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
```

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

```r
#Phase II: Exploratory Results
leukemia.test$table # rows as shown in Table 3
# component comp.GLPC 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
```

† Available online at https://CRAN.R-project.org/package=LPKsample
Fig. 12: 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.

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:

```r
# 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.

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

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.*