Check for updates

Received: 24 September 2019 | Revised: 23 December 2019 | Accepted: 24 December 2019

DOI: 10.1002/asmb.2508

SPECIAL ISSUE PAPER

WILEY

# Interpoint distances: Applications, properties, and visualization

**Reza Modarres** | **Yu Song**

Department of Statistics, George Washington University, Washington, District of Columbia,

**Correspondence**
Reza Modarres, Department of Statistics, George Washington University, Washington, DC.
Email: reza@gwu.edu

**Abstract**

This article surveys recent development on Euclidean interpoint distances (IPDs). IPDs find applications in many scientific fields and are the building blocks of several multivariate techniques such as comparison of distributions, clustering, classification, and multidimensional scaling. In this article, we explore IPDs, discuss their properties and applications, and present their distributions for several families, including the multivariate normal, multivariate Bernoulli, multivariate power series, and the unified hypergeometric distributions. We consider two groups of observations in $\mathbb{R}^d$ and present a simultaneous plot of the empirical cumulative distribution functions of the within and between IPDs to visualize and examine the equality of the underlying distribution functions of the observations.

**KEYWORDS**

classification, clustering, discrete, homogeneity

## 1 | INTRODUCTION

This article* is expository in nature and surveys some recent development on interpoint distances (IPDs), their applications in high dimensional multivariate analysis, their properties and visualization. Euclidean IPDs are the building blocks of several multivariate techniques such as comparison of distributions, clustering, classification, depth functions, correspondence analysis and multidimensional scaling. Analysis of point patterns, minimal spanning trees (MSTs), detection of spatial disease clusters, tests of the homogeneity of distributions, and many depth functions depend on the IPDs and their distributions.

Modarres[1,2] investigates the distribution of IPDs among the observations that are drawn from multivariate Bernoulli, and Poisson distributions, respectively. Marozzi[3] discusses multivariate tests based on IPDs and applies them to magnetic resonance images. For comparing means and variability of multivariate data, Marozzi et al[4] propose powerful tests based on IPDs. They present test statistics that are distribution free, unbiased, consistent, simple to compute, and applicable when the number of variables is much larger than the sample size. Statistics based on IPDs are very attractive for comparison of means of multivariate populations because one does not need to assume that observations are drawn from multivariate normal distributions. The distribution of the test statistic in most cases is obtained from random permutations under the null hypothesis. The permutation test is valid because under the null hypothesis that the distributions to be compared are the same, sample observations (the column vectors of the data matrix) are exchangeable.

*This article is dedicated to Professor Benjamin Kedem on the occasion of his 75th birthday.

 © 2020 John Wiley & Sons, Ltd. 

Song and Modarres[5] use IPDs to test for the homogeneity of multivariate mixture models. Guo and Modarres[6] present novel tests for the hypothesis of independence when the number of variables is larger than the number of vector observations. They show that two multivariate normal vectors are independent if and only if their IPDs are independent. Their proposed test statistics exploit different properties of the sample IPDs. IPDs have also been used extensively in shape analysis where shapes are based on the use of finite vectors of coordinates characterizing the shapes. Osada et al[7] explore the practical aspects of using the IPD in image analysis and Berrendero et al[8] identify a shape with the corresponding IPD distribution. Glick[9] proves that for classification or discrimination among arbitrary multivariate densities (assumed to be equally likely a priori), the Bayes discriminant's probability of correct classification is a linear transform of the separation measures whose lower and upper bounds are functions of all pairwise IPDs.

The rest of the article is organized as follows. Section 2 reviews applications of IPDs. We discuss some properties of IPDs and the Euclidean distance matrix in Section 3. Section 4 discusses the distribution of IPDs as quadratic forms and provides the distribution of IPDs from a multivariate Bernoulli distribution. Section 5 discusses the one- and two-sample IPDs from a multivariate normal distribution. We discuss the of IPDs from multivariate power series distribution (MPSD) and unified multivariate hypergeometric (UMHG) distribution families in Section 6. The last section discusses visualization of IPDs when comparing two high-dimensional distributions.

## 2 | APPLICATIONS

An important application of IPDs concerns tests of homogeneity of multivariate distributions. Under mild conditions, Maa et al[10] proved that two distributions are equal if and only if the IPDs within and between samples have the same univariate distribution for both continuous and discrete distributions. Several authors, including those of References 11-18, utilize IPDs in various ways to construct tests for the equality of distribution functions.

Given two sets of independently and identically distributed (i.i.d.) $d$-dimensional observations $\{\mathbf{X}_i\}_{i=1}^{N_x} \sim F_x$ and $\{\mathbf{Y}_j\}_{j=1}^{N_y} \sim F_y$, we are interested in testing the hypothesis $H_0 : F_x = F_y$ against general alternatives $H_a : F_x \neq F_y$. Let $\overline{d}_{(xx)} = \binom{N_x}{2}^{-1} \sum_{i=1}^{N_x} \sum_{j=i+1}^{N_x} ||\mathbf{X}_i - \mathbf{X}_j||$ and $\overline{d}_{(yy)} = \binom{N_y}{2}^{-1} \sum_{i=1}^{N_y} \sum_{j=i+1}^{N_y} ||\mathbf{Y}_i - \mathbf{Y}_j||$ represent the average IPD within the samples $\mathbf{X}$ and $\mathbf{Y}$, respectively. Let $\overline{d}_{(xy)} = \left(N_x N_y\right)^{-1} \sum_{i=1}^{N_x} \sum_{j=1}^{N_y} ||\mathbf{X}_i - \mathbf{Y}_j||$ represent the average IPD across the two samples. Baringhaus and Franz[14] prove that

$$\mathbb{E}||\mathbf{X}_1 - \mathbf{Y}_1|| - \frac{1}{2}\mathbb{E}||\mathbf{X}_1 - \mathbf{X}_2|| - \frac{1}{2}\mathbb{E}||\mathbf{Y}_1 - \mathbf{Y}_2|| \geq 0, \tag{1}$$

if and only if $F_x = F_y$ and construct the test statistic:

$$\mathrm{BF}(x, y) = \frac{N_x N_y}{N_x + N_y}\left[\overline{d}_{(xy)} - \frac{\binom{N_x}{2}}{2N_x^2}\overline{d}_{(xx)} - \frac{\binom{N_y}{2}}{2N_y^2}\overline{d}_{(yy)}\right]. \tag{2}$$

Rejection of $H_0$ is for large values of $\mathrm{BF}(x, y)$. The critical values are obtained by using the bootstrap method. Székely and Rizzo[15] propose an equivalent statistic to (2) for testing $H_0$ called the energy test.

Pointing out why BF statistic fails for a particular class of alternatives, Biswas and Ghosh[18] propose an alternative statistic, which is

$$\mathrm{BG}(x, y) = \left(\overline{d}_{(xx)} - \overline{d}_{(xy)}\right)^2 + \left(\overline{d}_{(yy)} - \overline{d}_{(xy)}\right)^2. \tag{3}$$

For small sample sizes, the critical values are found by using the permutation principle of testing. Biswas and Ghosh[18] provide the asymptotic distribution of BG for large sample sizes. The test statistic rejects $H_0$ for large values of $BG(x, y)$. The energy, BF, and BG statistics can be extended to $k$ groups by aggregating each statistic over $\binom{k}{2}$ comparisons. Both $\mathrm{BG}(x, y)$ and BF (energy) tests are rotation invariant and show good statistical power in testing for $H_0$. In general, BG statistic performs better for the scale alternatives and BF statistic performs better for the location alternatives.

Consider the basic method of comparing two groups for the equality of their multivariate means using the two-sample[19] $T^2$ test under several assumptions, including multivariate normality, equal covariance matrices, and $d < N_x + N_y - 2$. The $T^2$ statistic has many advantages such as invariance under linear transformations, known exact and asymptotic distributions, and being the uniformly most powerful invariant for testing the mean against two sided alternatives. It is powerful when $d$ is small compared with the sample size. However, it has two severe practical drawbacks, including lack of power for large $d$ and inapplicability due to the singularity of the covariance matrix when $d > N_x + N_y - 2$. Methods based on IPDs circumvent such difficulties because IPDs are always one dimensional irrespective of the number of variables under consideration.

Methods based on IPDs are competitive with traditional multivariate techniques in many applications. Moreover, IPDs provide a method of dealing with high-dimensional problems. These applications include classification,[20] tests of equality of distribution functions,[1] and tests of mixture distributions.[6] Guo and Modarres[21] offer tests of the equality of distribution functions for matrix distributions based on the Frobenius norm. Their method extends the work of References 14, 15, and 18 on the equality of vector distribution functions to the equality of matrix distributions. The energy statistic and BG test are applicable when dimension is greater than the sample size, competitive with Hotelling $T^2$ test under normality, and more powerful for heavy tailed distributions.

Suppose $\boldsymbol{\mu}_x$ and $\boldsymbol{\Sigma}_x$ are the mean vector and covariance matrix of distribution $F_x$, $\boldsymbol{\mu}_y$, and $\boldsymbol{\Sigma}_y$ are the mean vector and covariance matrix of distribution $F_y$. Biswas and Ghosh[18] assume that

(A1) There exist $\tau_x^2, \tau_y^2 > 0$ and $v$ such that $\mathrm{tr}\left(\boldsymbol{\Sigma}_x\right)/d \to \tau_x^2$, $\mathrm{tr}\left(\boldsymbol{\Sigma}_y\right)/d \to \tau_y^2$, and $||\boldsymbol{\mu}_x - \boldsymbol{\mu}_y||^2/d \to v^2$ as $d \to \infty$.
(A2) The fourth moments of the components of $\mathbf{X}_i$ and $\mathbf{Y}_j$ are uniformly bounded for $i = 1, \ldots, N_x, j = 1, \ldots, N_y$.
(A3) For $(U_t, V_t) = (X_{1t}, X_{2t}), (X_{1t}, Y_{1t}), (Y_{1t}, Y_{2t})$, $\sum_{t \neq s} \mathrm{Corr}\left[(U_t - V_t)^2, (U_s - V_s)^2\right]$ is of the order $o(d^2)$, where $1 \leq s, t \leq d$.

Biswas and Ghosh[18] prove that under(A1) to (A3), as $d \to \infty$,

(a) $d^{-\frac{1}{2}}||\mathbf{X}_i - \mathbf{X}_j|| \xrightarrow{p} \tau_x\sqrt{2}$ for $1 \leq i \leq j \leq N_x$;
(b) $d^{-\frac{1}{2}}||\mathbf{Y}_i - \mathbf{Y}_j|| \xrightarrow{p} \tau_y\sqrt{2}$ for $1 \leq i \leq j \leq N_y$;
(c) $d^{-\frac{1}{2}}||\mathbf{X}_i - \mathbf{Y}_j|| \xrightarrow{p} \sqrt{\tau_x^2 + \tau_y^2 + v^2}$ for $1 \leq i \leq N_x$ and $1 \leq j \leq N_y$.

Therefore, the combined sample points $N = N_x + N_y$ are asymptotically (as $d \to \infty$) located at the vertices of a N-polyhedron in $(N-1)$-dimensional space.

Classification is a second important application of IPD that we consider. The IPDs have been considered for discriminating stationary point processes Silverman and Brown.[22] The reason for considering IPDs for model discrimination is because their expected values and distributions can differ under different models. Suppose $\{\mathbf{X}_i\}_{i=1}^{N_x}$ and $\{\mathbf{Y}_j\}_{j=1}^{N_y}$ are two samples of independent $d$-dimensional random vectors, with known labels that are drawn from distribution $F_x$ and $F_y$. Let $\mathbf{Z}$ be a new observation to be classified. Dutta and Ghosh[23] illustrate a transformation based on the IPDs to classify $\mathbf{Z}$. For $N_x, N_y \geq 2$, these transformed data points are given as follows:

$$\mathbf{X}_i^* = \left(\frac{||\mathbf{X}_i - \mathbf{X}_1||}{\sqrt{d}}, \ldots, \frac{||\mathbf{X}_i - \mathbf{X}_{N_x}||}{\sqrt{d}}, \frac{||\mathbf{X}_i - \mathbf{Y}_1||}{\sqrt{d}}, \ldots, \frac{||\mathbf{X}_i - \mathbf{Y}_{N_y}||}{\sqrt{d}}\right)',$$
$$\mathbf{Y}_j^* = \left(\frac{||\mathbf{Y}_j - \mathbf{X}_1||}{\sqrt{d}}, \ldots, \frac{||\mathbf{Y}_j - \mathbf{X}_{N_x}||}{\sqrt{d}}, \frac{||\mathbf{Y}_j - \mathbf{Y}_1||}{\sqrt{d}}, \ldots, \frac{||\mathbf{Y}_j - \mathbf{Y}_{N_y}||}{\sqrt{d}}\right)'. \quad (4)$$

Using this transformation on an unlabeled observation $\mathbf{Z}$, we obtain

$$\mathbf{Z}^* = \left(\frac{||\mathbf{Z} - \mathbf{X}_1||}{\sqrt{d}}, \ldots, \frac{||\mathbf{Z} - \mathbf{X}_{N_x}||}{\sqrt{d}}, \frac{||\mathbf{Z} - \mathbf{Y}_1||}{\sqrt{d}}, \ldots, \frac{||\mathbf{Z} - \mathbf{Y}_{N_y}||}{\sqrt{d}}\right)', \quad (5)$$

for $1 \leq i \leq N_x$ and $1 \leq j \leq N_y$. Note that the $i$th component in $\mathbf{X}_i^*$ is 0, while the $(N_x + j)$th component in $\mathbf{Y}_j^*$ is 0. In this two-class problem, they consider an $N_x + N_y$-dimensional projection, whereas in the $k$-class problem, they consider an $N$-dimensional projection, where $N = N_1 + \cdots + N_k$. When $d$ is much larger than $N$, this transformation leads to a

substantial reduction in dimensionality. When the nearest neighbor classifier was used to classify $\mathbf{Z}^*$ into groups $\{\mathbf{X}^*\}_{i=1}^{N_x}$ and $\{\mathbf{Y}^*\}_{j=1}^{N_y}$, it correctly classified almost all test set observations. For a fixed value of $k$, the $k$ nearest neighbor classifier assigns the observation $\mathbf{Z}^*$ to the class having the maximum number of representatives in the set of $k$ labeled observations closest to $\mathbf{Z}^*$. Under certain assumptions on the distributions of $\{\mathbf{X}_i\}$ and $\{\mathbf{Y}_j\}$, the misclassification probability tends to 0 as $d$ tends to infinity.

Liao and Akritas[24] consider a test-based method to allocate an observation $\mathbf{z}$ to population $\pi_x$ or $\pi_y$. In the first step, one assumes that $\mathbf{Z}$ belongs to $\pi_x$ and tests the hypothesis that $H_0 : F_{x\cup z} = F_y$ using $N_x + 1$ observations from $\pi_x$ and $N_y$ observations from $\pi_y$. We denote the first test by $\mathbf{T}_1$. In the second step, one assumes that $\mathbf{z}$ belongs to $\pi_y$ and tests $H_0 : F_x = F_{y\cup z}$ using $N_x$ observations from $\pi_x$ and $N_y + 1$ observations from $\pi_y$. We denote the second test by $\mathbf{T}_2$. Intuitively, if $\mathbf{z}$ truly comes from $\pi_x$, placing it with $\pi_y$ samples will blur the difference between the two classes. This intuition provides the kernel of the test-based method. Suppose $\text{PV}_1$ is the $p$-value of $\mathbf{T}_1$ and $\text{PV}_2$ is the $p$-value of $\mathbf{T}_2$. The observation $\mathbf{Z}$ is classified as $\pi_x$ if $\text{PV}_1 < \text{PV}_2$. Guo and Modarres[20] utilize IPDs to measure the closeness of the samples and construct new rules for high-dimensional classification of discrete observations. Applicable to high-dimensional data, their method is nonparametric and uses test-based classification with permutation testing. They propose a modification of a test-based rule to use relative values with respect to the training samples baseline. The proposed method is compared with parametric methods, such as likelihood ratio rule and modified linear discriminate function, and nonparametric techniques such as support vector machine, nearest neighbor, and depth-based classification, under multivariate Bernoulli, multinomial (MN), and multivariate Poisson (MP) distributions.

The third application of IPD concerns detection of change points that may appear when analyzing time-ordered categorical tables. Suppose we observe a sequence of data vectors $\{\mathbf{X}_i\}_{i=1}^{N_x}$, indexed by some ordering, such as time or location. For detection of a single change point at $\tau$, one tests $H_0 : \mathbf{X}_i \sim F_0, i = 1, \ldots, N_x$, against the alternative, $H_a : \exists 1 \leq \tau < N_x$, such that $\mathbf{X}_i \sim F_1$, if $i > \tau$ and $F_0$, otherwise, where $F_0$ and $F_1$ are different probability distribution functions. Chen and Zhang[25] propose a test to detect change points based on the MST, which is constructed using Euclidean IPDs as weights. They use $G$ to refer to both the graph and its set of edges when the vertex set is implicitly obvious. The test statistics is $R_G(t) = \sum_{(i,j)\in G} \mathbb{1}_{\{\mathbb{1}_{\{i>t\}} \neq \mathbb{1}_{\{j>t\}}\}}$. Relative small values of $R_G(t)$ provide evidence against the null hypothesis. Other graphs-based methods, such as nearest neighborhood graph and minimum distance pairing graph, can also be used to define a test statistic in similar manner. IPDs are also important in detection of hotspots on disease and crime clusters where the underlying distributions are often discrete.[26] Atkinson et al[27] use IPDs to compare an observed sample with a high-dimensional simulated sample.

The last application we consider is cluster analysis. IPDs are the building blocks of many clustering strategies. The family of agglomerative clustering methods contains some of the most frequently used clustering methods. Starting with $n$ initial clusters of single observation vectors, an agglomerative clustering builds a tree structure by proceeding sequentially and joining pairs of clusters until all data points are grouped together in a single cluster (root of the tree). Lance and Williams[28] characterize a family of agglomerative clustering algorithms by defining the distance between a new cluster $Y_{(ij)}$ and any other cluster $Y_{(k)}$ with $d_{(ij)k} = \alpha_i d_{ik} + \alpha_j d_{jk} + \beta d_{ij} - \gamma \cdot |d_{ik} - d_{jk}|$, where $d_{ij}$s are the Euclidean distance between the $i$th and $j$th observations, and $\alpha_i, \alpha_j, \beta, \gamma$ are specified parameters that define a particular member of the family of agglomerative clustering. In this case, $Y_{(ij)}$ represents the cluster obtained by merging cluster $Y_{(i)}$ and $Y_{(j)}$, and $Y_{(k)}$ is the new cluster to be concerned. Define the parameter $\theta = (\alpha_i, \alpha_j, \beta, \gamma)$. Many well-known clustering methods are members of this family. For example, single linkage (or nearest-neighbor) uses $\theta = (1/2, 1/2, 0, -1/2)$, complete linkage (or furthest-neighbor) uses $\theta = (1/2, 1/2, 0, 1/2)$, and weighted average linkage uses $\theta = (1/2, 1/2, 0, 0)$.

## 3 | PROPERTIES

Suppose $\mathbf{X}$ and $\mathbf{Y}$ are $d$-dimensional random vectors that are drawn independently from some multivariate distribution $F_x$ and $F_y$, with mean $\boldsymbol{\mu}_x$, $\boldsymbol{\mu}_y$ and covariance matrix $\boldsymbol{\Sigma}_x$, $\boldsymbol{\Sigma}_y$, respectively. The Euclidean IPD between two i.i.d. observations $\mathbf{X}_1$ and $\mathbf{X}_2$ is defined as $d_{(xx)} = ||\mathbf{X}_1 - \mathbf{X}_2||$ and between $\mathbf{X}$ and $\mathbf{Y}$ is defined as $d_{(xy)} = d(\mathbf{X}, \mathbf{Y}) = ||\mathbf{X} - \mathbf{Y}||$. The IPD satisfies $d_{(xy)} = d_{(yx)}$, and $d_{(xy)} = 0$ if and only if $X_t = Y_t$ for all $t = 1, \ldots, d$.

The distance concentration phenomenon[29] describes the effect of high dimension $d$ on IPDs. Let $\eta = \frac{\sqrt{\text{Var}(\|\mathbf{X}_1 - \mathbf{X}_2\|)}}{\mathbb{E}(\|\mathbf{X}_1 - \mathbf{X}_2\|)}$ be the coefficient of variation of the IPDs. Under appropriate moment assumptions, one can show that $\eta$ tends to zero as $d$ tends to infinity. Using Chebyshev's inequality, Biau and Mason[30] show that for all $\epsilon > 0$, $\mathbb{P}\{|\frac{\|\mathbf{X}_1 - \mathbf{X}_2\|}{\mathbb{E}\|\mathbf{X}_1 - \mathbf{X}_2\|} - 1| \geq \epsilon\}$ approaches zero as $d$ tends to infinity. Hence, $\|\mathbf{X}_1 - \mathbf{X}_2\|$ approaches $\mathbb{E}\|\mathbf{X}_1 - \mathbf{X}_2\|$ as $d$ tends to infinity and the most

information about $\|\mathbf{X}_1 - \mathbf{X}_2\|$ is provided by $\mathbb{E}\|\mathbf{X}_1 - \mathbf{X}_2\|$. Furthermore, IPDs are more variable under dependence of the components than under their independence. Hall et al[31] study the geometry of a data cloud in high-dimension, low-sample size cases and show that the observations in each class have a tendency to lie deterministically at the vertices of a regular simplex, and the randomness in the data appears only as a random rotation of that simplex. The IPDs contain useful information about the separability between two distributions $F_x$ and $F_y$. Angiulli[32] discusses the concentration phenomenon and shows that IPDs are normally distributed when $d$ approaches infinity.

IPDs provide an alternative approach in high-dimension, low-sample-size problems. Li[33] shows that the asymptotic distribution of IPDs is multivariate normal under regularity conditions as $d$ becomes unbounded. Let $D_x = \|\mathbf{X}_1 - \mathbf{X}_2\|$, $D_y = \|\mathbf{Y}_1 - \mathbf{Y}_2\|$, and $D_{xy} = \|\mathbf{X}_1 - \mathbf{Y}_2\|$. Let $\mathbf{X}_1, \mathbf{X}_2 \sim \mathbb{N}(\boldsymbol{\mu}_x, \sigma_x^2 \mathbf{I}_d)$ and $\mathbf{Y}_1, \mathbf{Y}_2 \sim \mathbb{N}(\boldsymbol{\mu}_y, \sigma_y^2 \mathbf{I}_d)$, respectively. Let $\boldsymbol{\Delta} = \boldsymbol{\mu}_x - \boldsymbol{\mu}_y$ and $v^2 = \boldsymbol{\Delta}'\boldsymbol{\Delta}/d$. It is not difficult to show $D_x/\sqrt{d}$, $D_y/\sqrt{d}$, and $D_{xy}/\sqrt{d}$ converge to $\sqrt{2}\sigma_x$, $\sqrt{2}\sigma_y$, and $\sqrt{\sigma_x^2 + \sigma_y^2 + v^2}$, respectively, and that they have a joint trivariate normal distribution as $d$ tends to infinity.

A sample of $N_x$ independent vectors $\mathbf{X}_i$ define $N_x(N_x - 1)/2$ IPDs, of which any two IPDs are dependent if they have an index in common. Otherwise, they are independent. That is, a pair of IPDs $d(X_i, X_j)$ and $d(X_{i'}, X_{j'})$ are said to be independent if and only if $(i, j) \neq (i', j')$. IPDs are asymptotically pairwise independent as $N_x$ becomes large.[34] If the size $N_x$ of sample $\{\mathbf{X}\}_{i=1}^{N_x}$ gets large enough, the probability of sampling two dependent IPDs is approximately zero. To draw statistical inference about IPDs, one often needs to obtain a random sample from the distribution of IPDs. The IPDs between each pair of observations form a group of mutually independent IPDs. For example, if $N_x = 6$, one mutual independent group is $\{d_{(x)12}, d_{(x)34}, d_{(x)56}\}$.

**Lemma 1.** *Suppose* $\{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_{N_x}\}$ *is a set that contains* $N_x$ *independent observations and* $N_x$ *is even, one can obtain* $\frac{1}{2}N_x$ *independent pairs of IPDs. The total number of combination to obtain* $\frac{1}{2}N_x$ *independent pairs is* $\frac{N_x!}{2^{N_x/2}\left(\frac{N_x}{2}\right)!}$.

*Proof.* To obtain a mutually independent group of IPDs, one needs to make sure that none of the IPDs in this set share any index and there are $\frac{1}{2}N_x$ distinct pairs of index in total. Therefore, one can obtain only $\frac{1}{2}N_x$ independent pairs of IPDs from the dataset. To obtain $N_x/2$ independent pairs of IPD, the first two observations can be selected from the $N_x$ observations, then select another two from the remaining $N_x - 2$ observations, keep going until there are no remaining observations. The total number of combination to obtain this kind of group is $\frac{\binom{N_x}{2}\cdot\binom{N_x-2}{2}\cdots\binom{4}{2}\cdot\binom{2}{2}}{\left(\frac{N_x}{2}\right)!} = \frac{N_x!}{2^{N_x/2}\left(\frac{N_x}{2}\right)!}$. ∎

To compute the average squared IPD, $\bar{d}_{\mathbf{X}}^{\,2}$, one needs to find $\frac{N_x(N_x-1)}{2}$ squared IPDs so that the computational complexity is $O\left(N_x^2\right)$. There is another approach to compute $\bar{d}_{\mathbf{X}}^{\,2}$ with a lower computational complexity. Let $S_{\mathbf{X}}$ denote the sample covariance matrix of the squared IPDs. It follows that

$$\begin{aligned} \operatorname{tr}(S_{\mathbf{X}}) &= \operatorname{tr}\left[\frac{1}{N_x(N_x-1)}\sum_{1\le i<j\le d}\left(\mathbf{X}_i - \mathbf{X}_j\right)\left(\mathbf{X}_i - \mathbf{X}_j\right)'\right] \\ &= \frac{1}{N_x(N_x-1)}\sum_{1\le i<j\le d}\left(\mathbf{X}_i - \mathbf{X}_j\right)'\left(\mathbf{X}_i - \mathbf{X}_j\right) = \frac{1}{2}\bar{d}_{\mathbf{X}}^{\,2}. \end{aligned} \tag{6}$$

Hence, the average squared IPD is twice the trace of the sample covariance matrix. The computational complexity to find $\operatorname{tr}(S_{\mathbf{X}})$ is $O(N_x d)$. This computational complexity is smaller than $O\left(N_x^2\right)$ unless $d$ is larger than $N_x$.

Consider the $N_x \times N_x$ symmetric matrix of squared distances $D_{xx}(\mathbb{X})$ with elements $d_{(x)ij}^2$ when $i \neq j$ and zero on the main diagonal,

$$D_{xx}(\mathbb{X}) = \begin{array}{c} \\ X_1 \\ X_2 \\ X_3 \\ \\ X_{N_x} \end{array} \begin{array}{c} \begin{array}{ccccc} X_1 & X_2 & X_3 & \cdots & X_{N_x} \end{array} \\ \begin{pmatrix} 0 & d_{(x)12}^2 & d_{(x)13}^2 & \cdots & d_{(x)1N_x}^2 \\ d_{(x)21}^2 & 0 & d_{(x)23}^2 & \cdots & d_{(x)2N_x}^2 \\ d_{(x)31}^2 & d_{(x)32}^2 & 0 & \cdots & d_{(x)3N_x}^2 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ d_{(x)N_x1}^2 & d_{(x)N_x2}^2 & d_{(x)N_x3}^2 & \cdots & 0 \end{pmatrix} \end{array}. \tag{7}$$

One can write $D_{xx}(\mathbb{X}) \equiv U_x V_x$ where the $N_x \times 3$ matrix $U_x$ and the $3 \times N_x$ matrix $V_x$ are defined by

$$U_x = \begin{pmatrix} X_1'X_1 & 1 & -2X_1' \\ \cdots & \cdots & \cdots \\ X'_{N_x}X_{N_x} & 1 & -2X'_{N_x} \end{pmatrix}, \quad V_x = \begin{pmatrix} 1 & \cdots & 1 \\ X_1'X_1 & \cdots & X_{N_x}'X_{N_x} \\ X_1 & \cdots & X_{ZZN_x} \end{pmatrix}.$$

One can similarly show $D_{yy}(\mathbb{Y}) \equiv U_y V_y$, where $U_y$ is $N_y \times 3$ and $V_y$ is $3 \times N_y$. It is not difficult to show that

$$D(\mathbb{X}, \mathbb{Y}) = \begin{pmatrix} U_x \\ U_y \end{pmatrix} (V_x \; V_y) = \begin{pmatrix} D_{xx}(\mathbb{X}) = U_x \; V_x & D_{xy}(\mathbb{X}, \mathbb{Y}) = U_x \; V_y \\ D_{yx}(\mathbb{X}, \mathbb{Y}) = U_y V_x & D_{yy}(\mathbb{X}) = U_y \; V_y \end{pmatrix}. \tag{8}$$

Computational expense is a major obstacle when using IPDs with high-dimensional data. These decompositions are useful for parallel computation of the distance matrix.

**Lemma 2.** *Consider the off-diagonal elements of the $N_x \times N_x$ symmetric matrix of squared distances $D_{xx}(\mathbb{X})$ in Equation (7). One can show that*

$$\mathbb{E}\{\min(D_{xx}(\mathbb{X}))\} \leq 2\text{tr}(\boldsymbol{\Sigma}_x) \leq \mathbb{E}\{\max(D_{xx}(\mathbb{X}))\}, \tag{9}$$

*where $\boldsymbol{\Sigma}_x$ is the sample covariance matrix of squared IPDs.*

*Proof.* Note that the expected maximum of several random variables is never less than the maximum of their individual expected values. Similarly, the expected minimum of several random variables is never greater than the minimum of their individual expected values. Hence,

$$\mathbb{E}\{\max(D_{xx}(\mathbb{X}))\} \geq \max\left(\mathbb{E}d_{12}^2, \mathbb{E}d_{13}^2, \ldots, \mathbb{E}d_{(N_x-1)N_x}^2\right),$$

$$\mathbb{E}\{\min(D_{xx}(\mathbb{X}))\} \leq \min\left(\mathbb{E}d_{12}^2, \mathbb{E}d_{13}^2, \ldots, \mathbb{E}d_{(N_x-1)N_x}^2\right).$$

Hence, inequality (9) follows since $\mathbb{E}d_{ij}^2 = 2\text{tr}(\boldsymbol{\Sigma}_x)$ for $1 \leq i < j \leq N_x$. ∎

**Lemma 3.** *Let $N = N_x + N_y$ and consider the the off-diagonal elements of the $\binom{N}{2} \times \binom{N}{2}$ symmetric matrix of squared distances $D(\mathbb{X}, \mathbb{Y})$ in Equation (8) . One can similarly show that*

$$\mathbb{E}\{\max(D(\mathbb{X}, \mathbb{Y})\} \geq \max\{2\text{tr}(\boldsymbol{\Sigma}_x), 2\text{tr}(\boldsymbol{\Sigma}_y), \text{tr}(\boldsymbol{\Sigma}_x) + \text{tr}(\boldsymbol{\Sigma}_y) + \left(\boldsymbol{\mu}_x - \boldsymbol{\mu}_y\right)' \left(\boldsymbol{\mu}_x - \boldsymbol{\mu}_y\right)\},$$

$$\mathbb{E}\{\min(D(\mathbb{X}, \mathbb{Y})\} \leq \min\{2\text{tr}(\boldsymbol{\Sigma}_x), 2\text{tr}(\boldsymbol{\Sigma}_y), \text{tr}(\boldsymbol{\Sigma}_x) + \text{tr}(\boldsymbol{\Sigma}_y) + \left(\boldsymbol{\mu}_x - \boldsymbol{\mu}_y\right)' \left(\boldsymbol{\mu}_x - \boldsymbol{\mu}_y\right)\}.$$

Let $\mathbf{X}_1, \ldots, \mathbf{X}_{N_x+1}$ be points in $\mathbb{R}^d$ and let $\mathbf{1}_{N_x+1}$ be a column vector of 1s. Cayley and Menger[34,35] show that the squared volume of the $N_x$-simplex is defined in terms of the $(N_x + 1) \times (N_x + 1)$ Euclidean distance matrix $D_{xx}(\mathbb{X})$ as

$$\text{CM}(\mathbb{X}) = -\left(\frac{1}{N_x!}\right)^2 \begin{vmatrix} 0 & \mathbf{1}_{N_x+1} \\ \mathbf{1}'_{N_x+1} & D_{xx}(\mathbb{X}) \end{vmatrix}.$$

This expression is always nonnegative and zero only if the points are affinely dependent. Menger[36] shows that the nonnegativity of these expressions is, in fact, a sufficient condition for any symmetric matrix of nonnegative real numbers with zeros down the diagonal to be the matrix of squared distances among a set of points in Euclidean space. A Euclidean distance matrix is a matrix of squared Euclidean distances between the observations in the data matrix. The Euclidean distance geometry problem seeks to construct the configuration of the data matrix given information on pairwise IPDs. Since Euclidean distance matrices are symmetric, one can retrieve the original observation matrix by performing an eigenvalue decomposition.[37] Note that this method is invariant with respect to translation and rotation.

## 4 | DISTRIBUTION OF IPDS

We define the general form of the IPD in this section. Let $\mathbf{X}$ be a $d$-dimensional random vector with mean $\boldsymbol{\mu}$ and positive definite covariance matrix $\boldsymbol{\Sigma}$. Let $\mathbf{Z} = \boldsymbol{\Sigma}^{-\frac{1}{2}}(\mathbf{X} - \boldsymbol{\mu})$ and $\mathbf{A}$ be a $d \times d$ symmetric matrix, where $\boldsymbol{\Sigma}^{-\frac{1}{2}}$ is the symmetric positive-definite square root of matrix $\boldsymbol{\Sigma}^{-1}$. It follows that $\mathbb{E}(\mathbf{Z}) = \mathbf{0}$ and $\mathbb{C}\text{ov}(\mathbf{Z}) = \mathbf{I}_d$. Let $\mathbf{P}$ be a $d \times d$ orthogonal matrix which diagonalizes $\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{A}\boldsymbol{\Sigma}^{\frac{1}{2}}$, where $\mathbf{PP}' = \mathbf{I}_d$. That is, $\boldsymbol{P}'\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{A}\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{P} = \boldsymbol{\Lambda}$, where $\boldsymbol{\Lambda} = \text{diag}(\lambda_1, \ldots, \lambda_d)$, and $\lambda_1, \ldots, \lambda_d$ are the eigenvalues of $\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{A}\boldsymbol{\Sigma}^{\frac{1}{2}}$. Let $\mathbf{U} = \boldsymbol{P}'\boldsymbol{Z}$, and note that $\mathbf{Z} = \mathbf{PU}$, $\mathbb{E}(\mathbf{U}) = \mathbf{0}$ and $\mathbb{C}\text{ov}(\mathbf{U}) = \mathbf{I}_d$. Therefore, the nonsingular quadratic form in the random vector $\mathbf{X}$ associated with $\mathbf{A}$ is $Q(\mathbf{X}) = \boldsymbol{X}'\boldsymbol{A}\boldsymbol{X} = (\mathbf{U} + \mathbf{b})'\boldsymbol{\Lambda}(\mathbf{U} + \mathbf{b})$, where $\mathbf{U} = (U_1, \ldots, U_d)'$ and $\mathbf{b} = \boldsymbol{P}'\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\mu} = (b_1, \ldots, b_d)'$. Hence, we obtain the quadratic form and its expectation

$$Q(\mathbf{X}) = \boldsymbol{X}'\boldsymbol{A}\boldsymbol{X} = \begin{cases} \sum_{t=1}^{d} \lambda_t (U_t + b_t)^2, & \text{if } \boldsymbol{\mu} \neq \mathbf{0}, \\ \sum_{t=1}^{d} \lambda_t U_t^2, & \text{if } \boldsymbol{\mu} = \mathbf{0}, \end{cases} \tag{10}$$

$$\mathbb{E}[Q(\mathbf{X})] = \text{tr}(\boldsymbol{A}\boldsymbol{\Sigma}) + \boldsymbol{\mu}'\boldsymbol{A}\boldsymbol{\mu}. \tag{11}$$

Suppose $\mathbf{X}$ and $\mathbf{Y}$ are $d$-dimensional vectors independently drawn from distributions $F_x$ and $F_y$, with means $\boldsymbol{\mu}_x$, $\boldsymbol{\mu}_y$ and covariance matrices $\boldsymbol{\Sigma}_x$, $\boldsymbol{\Sigma}_y$, respectively. The squared Euclidean IPD between $\mathbf{X}$ and $\mathbf{Y}$ is defined as

$$d_{(xy)}^2 = d^2(\mathbf{X}, \mathbf{Y}) = ||\mathbf{X} - \mathbf{Y}||^2 = \sum_{t=1}^{d} (X_t - Y_t)^2. \tag{12}$$

For fixed vectors, the IPD satisfies $d_{(xy)} = d_{(yx)}$ and $d_{(xy)} = 0$ if and only if $\mathbf{x}_t = \mathbf{y}_t$ for all $t = 1, \ldots, d$. It is clear that $d^2(\mathbf{X}, \mathbf{Y}) = Q(\mathbf{X} - \mathbf{Y})$, where $\mathbf{A} = \mathbf{I}_d$. Since $\mathbb{E}(\mathbf{X} - \mathbf{Y}) = \boldsymbol{\mu}_x - \boldsymbol{\mu}_y$ and $\mathbb{C}\text{ov}(\mathbf{X} - \mathbf{Y}) = \boldsymbol{\Sigma}_x + \boldsymbol{\Sigma}_y$, by using Equation (10), the IPD can be expressed as

$$d^2(\mathbf{X}, \mathbf{Y}) = \sum_{t=1}^{d} \lambda_t (U_t + b_t)^2, \tag{13}$$

where $\mathbf{U} = \mathbf{P}'\left(\boldsymbol{\Sigma}_x + \boldsymbol{\Sigma}_y\right)^{-\frac{1}{2}}\left(\mathbf{X} - \mathbf{Y} - \boldsymbol{\mu}_x + \boldsymbol{\mu}_y\right)$, $\mathbf{b} = \mathbf{P}'\left(\boldsymbol{\Sigma}_x + \boldsymbol{\Sigma}_y\right)^{-\frac{1}{2}}\left(\boldsymbol{\mu}_x - \boldsymbol{\mu}_y\right)$, and $\mathbf{P}$ is a $d \times d$ orthogonal matrix, which diagonalizes $\boldsymbol{\Sigma}_x + \boldsymbol{\Sigma}_y$. The expected value is given by

$$\mathbb{E}\left[d^2(\mathbf{X}, \mathbf{Y})\right] = \text{tr}\left(\boldsymbol{\Sigma}_x + \boldsymbol{\Sigma}_y\right) + \left(\boldsymbol{\mu}_x - \boldsymbol{\mu}_y\right)'\left(\boldsymbol{\mu}_x - \boldsymbol{\mu}_y\right). \tag{14}$$

Mathai and Provost[38] derive the moment generating function of a quadratic form in normal random variables. One can use it to show that if $\mathbf{X} \sim N_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then

$$\mathbb{E}[Q(\mathbf{X})] = \text{tr}(\boldsymbol{A}\boldsymbol{\Sigma}) + \boldsymbol{\mu}'\boldsymbol{A}\boldsymbol{\mu}, \ \mathbb{V}\text{ar}[Q(\mathbf{X})] = 2\text{tr}(\boldsymbol{A}\boldsymbol{\Sigma})^2 + 4\boldsymbol{\mu}'\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}\boldsymbol{\mu}. \tag{15}$$

Suppose random vectors $\mathbf{X}$ and $\mathbf{Y}$ are drawn from discrete multivariate distributions. The squared IPD $d_{(xy)}^2$ can only be nonnegative integer values and the nonsquared IPD $d_{(xy)}$ can only be the square root of specific integer values in the range of squared IPD. Therefore, $\mathbb{P}\left(d_{(xy)} = \sqrt{v}\right) = \mathbb{P}\left(d_{(xy)}^2 = v\right)$, where $v$ is in the range of $d_{(xy)}^2$ depending on the underlying distribution.

While representations (10) and (13) are appropriate to determine the distribution of IPDs under continuous distributions, one needs other techniques for specific family of discrete distributions. We will now discuss the distribution of IPDs under four families of distributions: multivariate normal, Bernoulli, power series, and the unified hypergeometric distribution. The method of derivation is different due to the nature of data the distributions represent, binary, continuous, or categorical. Furthermore, both the one-sample and two-sample cases are considered. Within the MPSDs, only MN has

a restriction and other members are restriction-free. Within the unified hypergeometric distributions, only the negative MN (NGMN) has a restriction and other members are restriction free.

## 5 | MULTIVARIATE NORMAL IPDS

In this section, we obtain the distribution of IPDs from samples that are drawn from normal distributions. The distribution of IPDs in the one-sample case follows from representation (10) and in the two-sample case from Equation (13).

### 5.1 | One-sample IPDs

Suppose $\{\mathbf{X}_i\}_{i=1}^{N_x}$ is a sample of $d$-dimensional random vectors that are drawn from multivariate normal distribution with mean $\boldsymbol{\mu}_x$ and covariance matrix $\boldsymbol{\Sigma}_x = (\sigma)_{rs}$ for $r, s = 1, \ldots, d$, denoted by $\mathbf{X}_i \sim N_d\left(\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x\right)$. Then the distribution of IPD between any two random vectors within this sample is a linear combination of independent central chi-square random variables with 1 degree of freedom.

Since $\mathbb{E}\left(\mathbf{X}_i - \mathbf{X}_j\right) = \mathbf{0}$ and $\mathbb{C}\mathrm{ov}\left(\mathbf{X}_i - \mathbf{X}_j\right) = 2\boldsymbol{\Sigma}_x$, it follows from Equation (10) that $d^2\left(\mathbf{X}_i, \mathbf{X}_j\right) = \sum_{t=1}^{d} \lambda_t U_t^2$, where $\lambda_1, \ldots, \lambda_d$ are the eigenvalues of $2\boldsymbol{\Sigma}_x$ and $\mathbf{U} = \mathbf{P}'(2\boldsymbol{\Sigma}_x)^{-\frac{1}{2}}\left(\mathbf{X}_i - \mathbf{X}_j\right)$, $\mathbf{P}$ is an orthogonal matrix such that $\mathbf{P}'\left(2\boldsymbol{\Sigma}_x\right)\mathbf{P} = \mathrm{diag}(\lambda_1, \ldots, \lambda_d)$. Therefore, for $t = 1, \ldots, d$, $U_t \sim N(0, 1)$ and $U_t^2 \sim \chi^2(1)$. It is not difficult to show that the expected value and variance of the IPD $d^2\left(\mathbf{X}_i, \mathbf{X}_j\right)$ are $2\mathrm{tr}\left(\boldsymbol{\Sigma}_x\right)$ and $8\mathrm{tr}(\boldsymbol{\Sigma}_x)^2$, respectively. The covariance between any two dependent IPDs $d^2\left(\mathbf{X}_i, \mathbf{X}_j\right)$ and $d^2\left(\mathbf{X}_i, \mathbf{X}_l\right)$ is $2\mathrm{tr}(\boldsymbol{\Sigma}_x)^2$ for $i, j, l \in \{1, \ldots, N_x\}$, and their correlation is $\frac{1}{4}$.

### 5.2 | Two-sample IPDs

Suppose that $\{\mathbf{X}_i\}_{i=1}^{N_x}$ and $\left\{\mathbf{Y}_j\right\}_{j=1}^{N_y}$ are independent samples of $d$-dimensional random vectors that are drawn from $N_d\left(\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x\right)$ and $N_d\left(\boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y\right)$, respectively. The distribution of IPD between any $\mathbf{X}_i$ and $\mathbf{Y}_j$ for $i = 1, \ldots, N_x, j = 1, \ldots, N_y$, is a linear combination of independent noncentral chi-square random variables with 1 degree of freedom. It follows from (13) that $d^2\left(\mathbf{X}_i, \mathbf{Y}_j\right) = \sum_{t=1}^{d} \lambda_t (U_t + b_t)^2$, where $\lambda_t' s$ are eigenvalues of $\boldsymbol{\Sigma}_x + \boldsymbol{\Sigma}_y$, $\mathbf{U} = \mathbf{P}'\left(\boldsymbol{\Sigma}_x + \boldsymbol{\Sigma}_y\right)^{-\frac{1}{2}}\left(\mathbf{X}_i - \mathbf{Y}_j - \boldsymbol{\mu}_x - \boldsymbol{\mu}_y\right)$, and $\mathbf{b} = \mathbf{P}'\left(\boldsymbol{\Sigma}_x + \boldsymbol{\Sigma}_y\right)^{-\frac{1}{2}}\left(\boldsymbol{\mu}_x - \boldsymbol{\mu}_y\right)$. Note that $\mathbf{P}$ is an orthogonal matrix such that $\mathbf{P}'\left(\boldsymbol{\Sigma}_x + \boldsymbol{\Sigma}_y\right)\mathbf{P} = \mathrm{diag}(\lambda_1, \ldots, \lambda_d)$. Therefore, for $t = 1, \ldots, d$, $U_t \sim N(0, 1)$ and $(U_t + b_t)^2 \sim$ noncentral $\chi^2(1)$. It is not difficult to show that the expected value and variance of the IPD are

$$\mathbb{E}\left[d^2_{(xy)ij}\right] = \mathrm{tr}\left(\boldsymbol{\Sigma}_x + \boldsymbol{\Sigma}_y\right) + \left(\boldsymbol{\mu}_x - \boldsymbol{\mu}_y\right)'\left(\boldsymbol{\mu}_x - \boldsymbol{\mu}_y\right),$$
$$\mathbb{V}\mathrm{ar}\left[d^2_{(xy)ij}\right] = 2\mathrm{tr}\left(\boldsymbol{\Sigma}_x + \boldsymbol{\Sigma}_y\right)^2 + 4\left(\boldsymbol{\mu}_x - \boldsymbol{\mu}_y\right)'\left(\boldsymbol{\Sigma}_x + \boldsymbol{\Sigma}_y\right)\left(\boldsymbol{\mu}_x - \boldsymbol{\mu}_y\right).$$

## 6 | MULTIVARIATE BERNOULLI IPDS

This section reviews the results in Reference 1 and examines the distribution of IPDs within and between samples drawn from multivariate Bernoulli (MB) distributions. The random vector $\mathbf{X} = (X_1, \ldots, X_d)'$ has a $d$-variate Bernoulli distribution with probability mass function (p.m.f.) $\mathbf{P^{(d)}} = \mathbb{P}(X_1 = k_1, \ldots, X_d = k_d)$, where $k_i \in \{0, 1\}$ for $i = 1, \ldots d$. The mean vector is denoted by $\mathbf{P}$ with $p_i$ representing the $i$th element of $\mathbf{P}$ and the covariance matrix with $\boldsymbol{\Sigma}$. The vector of central moments is called the dependency vector as it captures all two-way or higher level dependencies that might exist between the elements of $\mathbf{X}$. The specification of the dependency vector is tantamount to determination of $\mathbf{P^{(d)}}$. The marginal distribution of any set of the components of $\mathbf{X}$ is MB with means, variances, and covariances obtained from the corresponding elements of $\mathbf{P}$ and $\boldsymbol{\Sigma}$. There are numerous applications of the MB distribution, including disease transmission and assessment of computer, social, communication, and financial networks where nodes in the network are represented by multivariate Bernoulli vectors with specified probability of functionality $\mathbf{P}$ and specified correlation matrix.

Also called the Hamming distance in information theory,[39] the squared IPDs of Bernoulli data are often used in clustering and classification algorithms. Suppose $\mathbf{X} = \{\mathbf{X}_i\}$ for $i = 1, \ldots, N_x$ is a sample of i.i.d. random vectors in $\{0, 1\}^d$

drawn from a MB distribution with mean vector $\mathbf{P}$, covariance matrix $\mathbf{\Sigma}$. We use the notation $\mathbf{X} \sim \text{MB}(\mathbf{P}, \mathbf{\Sigma})$. Similarly, suppose $\mathbf{Y} = \{\mathbf{Y}_j\}$ for $j = 1, \dots, N_y$ are $N_y$ independent draws from $\text{MB}(\mathbf{P}^*, \mathbf{\Sigma}^*)$. We also assume that the $\mathbf{X}$ and $\mathbf{Y}$ samples are independent.

## 6.1 | One-sample IPDs

The random vector $\mathbf{X}_i = (X_{i1}, \dots, X_{id})$ has a $d$-variate Bernoulli distribution with p.m.f. $\mathbf{P}^{(\mathbf{d})} = \mathbb{P}(X_{i1} = k_1, \dots, X_{id} = k_d)$, where $k_j \in \{0, 1\}$ for $j = 1, \dots, d$ and $i = 1, \dots, N_x$. The mean is denoted with $\mathbf{P}$, and the covariance with $\mathbf{\Sigma}$. We obtain the probability distribution of $d^2_{(x)ij}$, the squared distance between two randomly selected vectors $\mathbf{X}_i$ and $\mathbf{X}_j$, $i \neq j$, drawn from an $\text{MB}(\mathbf{P}, \mathbf{\Sigma})$. The proof of the following two theorems appears in Reference 1.

**Theorem 1.** *Let $\mathbf{\Omega} = \{1, \dots, d\}$ and $\mathbf{E} = \{t_1, \dots, t_k\}$ for $k = 0, \dots, d$. If the components of the MB vector are independent ($\mathbf{\Sigma}$ is diagonal), then the p.m.f. of $d^2_{(x)ij}$ is given by*

$$\mathbb{P}(d^2_{(x)ij} = k) = \sum_{\mathbf{E} \subseteq \mathbf{\Omega}} \prod_{t \in \mathbf{E}} 2p_t(1 - p_t) \prod_{t \notin \mathbf{E}} (p_t^2 + (1 - p_t)^2).$$

*When $\mathbf{\Sigma}$ is not diagonal, the p.m.f. of $d^2_{(x)ij}$ is given by*

$$\mathbb{P}(d^2_{(x)ij} = k) = \sum_{\mathbf{E} \subseteq \mathbf{\Omega}} \prod_{t \in \mathbf{E}} [2p_t(1 - p_t) + \sum_{l=t+1}^{k} b_l(p_l^2 + (1 - p_l)^2)] \times \prod_{t \notin \mathbf{E}} [(p_t^2 + (1 - p_t)^2) + \sum_{l=t+1}^{d} b_l p_l(1 - p_l)],$$

*where $b_l$ are constants that depend on the inverse of the covariance matrix.*

Note that $\mathbb{P}(d^2_{(x)ij} = k) = \mathbb{P}(\mathbf{X}_{it} \neq \mathbf{X}_{jt} \text{ in } k \text{ positions})$ where $k = 0, \dots, d$. Furthermore, $\mathbb{P}(\mathbf{X}_{it} = \mathbf{X}_{jt}) = p_t^2 + (1 - p_t)^2$ and $\mathbb{P}(\mathbf{X}_{it} \neq \mathbf{X}_{jt}) = 2p_t(1 - p_t)$ for any position $t$. Suppose $\mathbf{X}_i$, for $i = 1, \dots, N_x$, follows an exchangeable structure so that $\mathbf{P} = p\mathbf{1}$ and the correlations $\rho_{rs} = \rho$. Let $I_C = 1$ if $(i, j)$ and $(k, h)$ share an index, $1 \leq i < j \leq N_x$, $1 \leq k < h \leq N_x$, and zero, otherwise. One can verify the following:

- $\mathbb{E}\left(d^2_{(x)ij}\right) = 2dp(1 - p)$,
- $\mathbb{V}\text{ar}\left(d^2_{(x)ij}\right) = 2dp(1 - p)(1 - 2p(1 - p)) + 2\rho d(d - 1)p(1 - p)(2\rho p(1 - p) + (1 - 2p)^2)$,
- $\mathbb{C}\text{ov}\left(d^2_{(x)ij}, d^2_{(x)kh}\right) = \rho d^2 p(1 - p)(1 - 2p)^2 I_C$,
- $d^2_{(x)ij} \sim \text{binomial}(d, 2p(1 - p))$ when $\rho = 0$.

## 6.2 | Two-sample IPDs

Suppose $\{\mathbf{X}_i\}$ for $i = 1, \dots, N_x$ and $\{\mathbf{Y}_i\}$ for $i = 1, \dots, N_y$ are independent samples drawn from $\text{MB}(\mathbf{P}, \mathbf{\Sigma})$ and $\text{MB}(\mathbf{P}^*, \mathbf{\Sigma}^*)$, respectively. Let $\theta_{(xy)ijt} = 2p_t(1 - p_t)$.

**Theorem 2.** *Let $\mathbf{\Omega} = \{1, \dots, d\}$ and $\mathbf{E} = \{t_1, \dots, t_k\}$. For $k = 0, \dots, d$, the p.m.f. of $d^2_{(xy)ij}$ is given by*

$$\mathbb{P}(d^2_{(xy)ij} = k) = \sum_{\mathbf{E} \subseteq \mathbf{\Omega}} \prod_{t \in \mathbf{E}} [\theta_{(xy)ijt} + \sum_{l=t+1}^{k} b_l(1 - \theta_{(xy)ijl})] \prod_{t \notin \mathbf{E}} [1 - \theta_{(xy)ijt} + \sum_{l=t+1}^{d} b_l \theta_{(xy)ijl}].$$

*When $\mathbf{\Sigma}$ and $\mathbf{\Sigma}^*$ are diagonal, one obtains*

$$\mathbb{P}(d^2_{(xy)ij} = k)) = \sum_{\mathbf{E} \subseteq \mathbf{\Omega}} \prod_{t \in \mathbf{E}} (p_r + p_r^* - 2p_r p_r^*) \prod_{t \notin \mathbf{E}} (1 - (p_r + p_r^* - 2p_r p_r^*)).$$

For example, using the above two theorems,

$$\mathbb{P}(d^2_{(x)ij} \geq d^2_{(y)kh}) = \sum_{a=0}^{d}\sum_{b=0}^{a} \mathbb{P}(d^2_{(x)ij} = a)\mathbb{P}(d^2_{(y)kh} = b).$$

This is the probability that a randomly selected IPD from the $\mathbf{X}$ sample is equal or larger than a randomly selected IPD from the $\mathbf{Y}$ sample.

## 7 | MPSD AND UMHG DISTRIBUTIONS

In this section, we obtain the distribution of IPDs from samples that are drawn from MPSD and UMHG distributions as discussed in References 40 and 41. Categorical data are prevalent in practice. Discrete distributions arise in numerous circumstances including biology, ecology, physics, gene expression analysis, text mining, image analysis, and other scientific fields, either as sampling models or as models of contingency table for multiple counts. Certain families of the discrete multivariate distributions, namely, the MPSD and UMHG, have been applied in the investigations of these areas. The MPSD family is a discrete multivariate family of distributions, which includes the MN, NGMN, MP, and multivariate logarithmic (ML) distribution as prominent members. The prominent members of the UMHG family are the multivariate hypergeometric (MH), inverse hypergeometric (MIH), negative hypergeometric (MNH), negative MIH (MNIH), Pólya (MP), and the inverse Pólya (MIP) distributions.

### 7.1 | MPSD family of distributions

Noack[42] first introduces a class of discrete power series distribution. Sibuya et al[43] characterized the NGMN distribution and Patil[44] defines the multivariate analog of the power series distribution. Joshi and Patil[45] find certain structural properties of the power series distributions. Suppose $\mathbf{X} = (X_1, X_2, \dots, X_d)'$ is a $d$-dimensional random vector and $\boldsymbol{\theta} = (\theta_1, \dots, \theta_d)'$ is a parameter vector. Define $\boldsymbol{\Theta}$ to be the $d$-dimensional parameter space of $\boldsymbol{\theta}$. Let $\chi$ be a countable subset of a $d$-fold Cartesian product of the set of nonnegative integers. Set $\chi$ is $\chi = \{(x_1, x_2, \dots, x_d)'\} \subseteq \mathbb{N}_0^d$, where $\mathbb{N}_0$ is the set of nonnegative integers. Now, define the series function $f(\boldsymbol{\theta}) = \sum_{\mathbf{X}\in\chi} a(x_1, x_2, \dots, x_d)\theta_1^{x_1}\theta_2^{x_2}\dots\theta_d^{x_d}$, where the summation extends over $\chi$, and $\theta_i \geq 0$ with $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ so that $f(\boldsymbol{\theta})$ is finite and differentiable. Also, define the coefficient function $a(\mathbf{x})$ to be a positive-valued function of $\mathbf{x}$. The random vector $\mathbf{X}$ is said to have MPSD with range $\chi$, parameter $\boldsymbol{\theta}$, and series function $f(\boldsymbol{\theta})$, if it has p.m.f.

$$\mathbb{P}(\mathbf{X} = \mathbf{x}\,;\boldsymbol{\theta}) = \frac{a(x_1, \dots, x_d)\theta_1^{x_1}\dots\theta_d^{x_d}}{f(\theta_1, \dots, \theta_d)}, \text{given } (x_1, \dots, x_d)' \in \chi \text{ and } f(\theta_1, \dots, \theta_d) \neq 0, \tag{16}$$

where $a(\mathbf{x}) > 0$, if $\mathbf{x} \in \chi$, and $a(\mathbf{x}) = 0$, otherwise. We denote it as $\mathbf{X} \sim \text{MPSD}(\boldsymbol{\theta}_x, f_x(\boldsymbol{\theta}_x))$.

The MPSD is a discrete multivariate exponential type distribution with marginal p.m.f. for each component $X_j$, $j = 1, \dots, d$ as

$$\mathbb{P}(X_j = x_j) = \frac{\sum_{\{x_1,\dots,x_d\}\setminus x_j}\left(a(x_1, \dots, x_d)\prod_{i=1}^{d}\theta_i^{x_i}\right)}{f(\boldsymbol{\theta})} = \frac{a_j\left(x_j; \{\theta_1, \dots, \theta_d\} \setminus \theta_j\right)\theta_j^{x_j}}{f(\boldsymbol{\theta})}, \tag{17}$$

where $\{x_1, \dots, x_d\} \setminus x_j$ represents the relative complement of $\{x_j\}$ in $\{x_1, \dots, x_d\}$ and the coefficient function for marginal p.m.f. is $a_j\left(x_j; \{\theta_1, \dots, \theta_d\} \setminus \theta_j\right) = \sum_{\{x_1,\dots,x_d\}\setminus x_j}\left(a(x_1, \dots, x_d)\prod_{i\neq j}\theta_i^{x_i}\right)$. Hence, $X_j$ has a univariate power series distribution with the series function $f(\boldsymbol{\theta})$ as expanded in powers of $\theta_j$, other $\theta$s are treated as known. It is not difficult to obtain the mean vector and covariance matrix by using the $r$th factorial moment of $\mathbf{X}$.[44]

By taking different series function and reparametrization, we can obtain different multivariate distributions. For the purpose of this article, we illustrate the MN, NGMN, MP, and ML distributions. A sample of individuals from the population may be drawn with replacement or without replacement. Furthermore, sampling with replacement may be direct

**TABLE 1** MPSD with different parameters and series functions

| Distribution | Reparametrization | Series Function | Coefficient Function |
|---|---|---|---|
| MN | $\theta_i = \frac{p_i}{1-\sum_{i=1}^{d} p_i}$ | $\left(1+\sum_{i=1}^{d}\theta_i\right)^n$ | $\frac{n!}{\prod_{i=1}^{d} x_i!\left(n-\sum_{i=1}^{d} x_i\right)!}$ |
| NGMN | $\theta_i = p_i$ | $\left(1-\sum_{i=1}^{d}\theta_i\right)^{-k}$ | $\frac{\Gamma\left(\sum_{i=1}^{d} x_i+k\right)}{\prod_{i=1}^{d} x_i!\Gamma(k)}$ |
| MP | $\theta_i = \lambda_i$ | $e^{\sum_{i=1}^{d}\theta_i}$ | $1/\prod_{i=1}^{d} x_i!$ |
| ML | $\theta_i = p_i$ | $-\log\left(1-\sum_{i=1}^{d}\theta_i\right)$ | $\frac{\left(\sum_{i=1}^{d} x_i-1\right)!}{\prod_{i=1}^{d} x_i!}$ |

Abbreviations: ML, multivariate logarithmic; MP, multivariate Poisson; MPSD, multivariate power series distribution; MN, multinomial; NGMN, negative multinomial.

or inverse. Depending on the nature of the sampling procedure used, the random vector $\mathbf{X}$ is known to have an MN or NGMN distribution. MP and ML distributions can be obtained by taking limits and truncations of the first two distributions. Table 1 shows the corresponding series functions of these four distributions. By using Equation (17), one can show that the marginal distributions, respectively, are binomial, $B(n, p_t)$, with p.m.f. $\binom{n}{x_t} p_t^{x_t}(1-p_t)^{n-x_t}$, negative binomial, $\mathrm{NB}(k, p_t)$, with p.m.f. $\binom{k+x_t-1}{x_t} p_t^{x_t}(1-p_t)^k$, Poisson, $\mathrm{Poi}(\lambda_t)$, with p.m.f. $\frac{e^{-\lambda_t}\lambda_t^{x_t}}{x_t!}$, and logarithmic, $L(p_t)$, with p.m.f. $\frac{\log\left(1-\sum_{s=1}^{d} p_s+p_t\right)}{\log\left(1-\sum_{s=1}^{d} p_s\right)} \cdot \mathbb{1}_{\{x_t=0\}}$ and $\frac{1}{-x_t \log\left(1-\sum_{s=1}^{d} p_s\right)}\left(\frac{p_t}{1-\sum_{s=1}^{d} p_s+p_t}\right)^{x_t} \cdot \mathbb{1}_{\{x_t\geq 1\}}$.

## 7.2 | UMHG family of distributions

Janardan[46] investigates the problem of unifying several MH distributions, which arise in sampling with or without replacement from a finite population of $a$ individuals classified into $d$ categories. We first state the binomial coefficient $\binom{A}{B}$ for the nonintegral and negative values of $A$ and $B$ with $\binom{A}{B} = \frac{A!}{(A-B)!\ B!}$, where $A! = \int_0^\infty e^{-t}t^A dt$ if $A \in \mathbb{N}_0$ and $A! = \frac{(-1)^{A+1}}{(-A-1)!}$ if $A \in \mathbb{Z} \setminus \mathbb{N}_0$. We use $\mathbb{N}_0$ and $\mathbb{Z}$ to represent the sets of natural numbers and integers, respectively.

**Definition 1.** (UMHG distribution). Suppose $\mathbf{X} = (X_1, \ldots, X_d)'$ is a $d$-dimensional random vector with $X_t \in \chi = \mathbb{N}_0$ for $t = 1, \ldots, d$. Suppose further that $n \in \mathbb{N}_0$ and $\mathbf{a} = (a_1, \ldots, a_d)'$ is a $d$-dimensional vector with $a_t \in \mathbb{N}_0$ for $t = 1, \ldots, d$. The random vector $\mathbf{X}$ has a UMHG distribution if

$$\mathbb{P}(\mathbf{X} = \mathbf{x}\ ;\ n, \mathbf{a}, a) = \frac{\prod_{t=1}^{d}\binom{a_t}{x_t}}{\binom{a}{n}}, \tag{18}$$

where $\sum_{t=1}^{d} X_t = n$ and $\sum_{t=1}^{d} a_t = a$. We use the notation $\mathbf{X} \sim \mathrm{UMHG}_d(n, \mathbf{a}, a)$. The mean and covariance matrix of $\mathbf{X}$ are

$$\boldsymbol{\mu}_x = \frac{n\mathbf{a}}{a}, \quad (\boldsymbol{\Sigma}_x)_{ts} = \begin{cases} \frac{na_t}{a}\left(1-\frac{a_t}{a}\right)\left(\frac{a-n}{a-1}\right), \text{for } t = s; \\ \frac{-na_t a_s}{a^2}\left(\frac{a-n}{a-1}\right), \text{for } t \neq s. \end{cases} \tag{19}$$

The marginal p.m.f. for each component $X_t$, $t = 1, \ldots, d$ is given by

$$\mathbb{P}(X_t = x_t\ ;\ n, a, a_t) = \frac{\binom{a-a_t}{n-x_t}\binom{a_t}{x_t}}{\binom{a}{n}}, \tag{20}$$

and represented with $\mathrm{UHG}(n, a, a_t)$ the unified hypergeometric distribution.

**TABLE 2** UMHG with different parameters produces the MH, MIH, MP, and the MIP distributions

| Distribution | a | $a$ | n |
|---|---|---|---|
| MH | $a_t = N_t$, for $t = 1, \dots, d$, $N = \sum_{t=1}^{d} N_t$ | $N$ | $n$ |
| MIH | $a_1 = -N - 1$, $a_t = N_t$ for $t = 2, \dots, d$, $N = \sum_{t=1}^{d} N_t$ | $-N_1 - 1$ | $-k$ |
| MP | $a_t = -N_t$ for $t = 1, \dots, d$, $N = \sum_{t=1}^{d} N_t$ | $-N$ | $n$ |
| MIP | $a_1 = N - 1$, $a_t = -N_t$ for $t = 2, \dots, d$, $N = \sum_{t=1}^{d} N_t$ | $N_1 - 1$ | $-k$ |

Abbreviations: MIH, inverse hypergeometric; MH, multivariate hypergeometric; MIP, inverse Pólya; MP, Pólya; UMHG, unified multivariate hypergeometric.

Similar to the MN, the UMHG is a singular $d$-dimensional distribution because $x_d = n - \sum_{t=1}^{d-1} x_t$. Hence, the univariate analogue of the UMHG is the two-variate UMHG with p.m.f.

$$\mathbb{P}(X_1 = x_1, X_2 = x_2) = \frac{\binom{a_1}{x_1}\binom{a_2}{x_2}}{\binom{a}{n}} = \frac{\binom{a_1}{x_1}\binom{a-a_1}{n-x_1}}{\binom{a}{n}}, \tag{21}$$

which is the same as the marginal distribution in Equation (20). We obtain different multivariate distributions by using different values for **a**, $a$, and $n$. Table 2 shows the four prominent members of this family.

We will investigate the distribution of different types of IPDs within these two families in the following subsections.

## 7.3 | One-sample IPDs: Restricted and unrestricted vectors

We derive the distribution of squared IPD between i.i.d. random vectors in this subsection in two different circumstances. One with restricted support and one with unrestricted support. Note that we use i.i.d. to denote i.i.d. random variables or vectors throughout the article.

**Lemma 4.** *Suppose* $\mathbf{X}$ *and* $\mathbf{X}^*$ *are* $d$*-dimensional vectors that are drawn independently from* UMHG$(n, \mathbf{a}, a)$ *or* MN$(n, \mathbf{p})$ *distribution, then we obtain* $\sum_{t=1}^{d} X_t = \sum_{t=1}^{d} X_t^* = n$ *in Table 1 and Definition* 1.*. Hence, the maximum possible value of the IPD is* $2n^2$*. Moreover, one can express the IPD as*

$$d_{(x)}^2 = \sum_{t=1}^{d} (X_t - X_t^*)^2 = 2\sum_{t=1}^{d-1} (X_t - X_t^*)^2 + 2\sum_{s=1}^{d-2}\sum_{t=s+1}^{d-1} (X_s - X_s^*)(X_t - X_t^*). \tag{22}$$

*Hence, the squared IPD from either the UMHG or MN distribution can be only nonnegative even numbers. However, not all even numbers are included in its range. For example, when* $d = 3$ *and* $n = 2$*, the range of the squared IPD is* $\{0, 2, 6, 8\}$ *with 4 excluded. Therefore, we need to impose indicator functions on the p.m.f. of IPDs as shown in the following theorem.*

**Theorem 3** (One-sample IPD: Restricted vectors). *Suppose that* $\mathbf{X}$ *and* $\mathbf{X}^*$ *are* $d$*-dimensional random vectors that are drawn independently from* UMHG$(n, \mathbf{a}, a)$ *or* MN$(n, \mathbf{p})$ *distribution. Let* $\Phi = \{0, 1, \dots, n\}$*. The p.m.f. of the IPD between* $\mathbf{X}$ *and* $\mathbf{X}^*$ *is given by*

$$\mathbb{P}\left(d_{(x)}^2 = v\right) = \sum_{v \in \mathbf{H}_r} \sum_{\phi_1 \in \Phi} \cdots \sum_{\phi_d \in \Phi} \mathbb{P}(X_1 = \phi_1 + v_1) \cdot \mathbb{P}\left(X_1^* = \phi_1\right) \cdot \mathbb{1}_{\left\{\phi_t + v_t \geq 0, \sum_{t=1}^{d} \phi_t = n\right\}}$$
$$\times \prod_{t=2}^{d} \mathbb{P}(X_t = \phi_t + v_t | x_1, \dots, x_{t-1}) \cdot \mathbb{P}\left(X_t^* = \phi_t | x_1^*, \dots, x_{t-1}^*\right), \tag{23}$$

*where* $\mathbb{1}_{\left\{\phi_t + v_t \geq 0, \sum_{t=1}^{d} \phi_t = n\right\}}$ *equals 1 if* $\phi_t + v_t \geq 0$ *and* $\sum_{t=1}^{d} \phi_t = n$*, and equals 0, otherwise. For* $\mathbf{X}, \mathbf{X}^* \sim$ UMHG$(n, \mathbf{a}, a)$*, we have the distributions*

$$X_1 \sim \text{UHG}(n, a, a_1), X_t|x_1, \dots, x_{t-1} \sim \text{UHG}\left(n - \sum_{r=1}^{t-1} x_r, a - \sum_{r=1}^{t-1} a_r, a_t\right), \quad t = 2, \dots, d;$$

$$X_1^* \sim \text{UHG}(n, a, a_1), X_t^*|x_1^*, \dots, x_{t-1}^* \sim \text{UHG}\left(n - \sum_{r=1}^{t-1} x_r^*, a - \sum_{r=1}^{t-1} a_r, a_t\right), \quad t = 2, \dots, d.$$

*For $\mathbf{X}, \mathbf{X}^* \sim \text{MN}(n, \mathbf{p})$, the distributions of $X_t|x_1, \dots, x_{t-1}$ and $X_t^*|x_1, \dots, x_{t-1}$ are*

$$X_1 \sim B(n, p_1), X_t|x_1, \dots, x_{t-1} \sim B\left(n - \sum_{r=1}^{t-1} x_r, \frac{p_t}{1 - \sum_{r=1}^{t-1} p_r}\right), \quad t = 2, \dots, d;$$

$$X_1^* \sim B(n, p_1), X_t^*|x_1^*, \dots, x_{t-1}^* \sim B\left(n - \sum_{r=1}^{t-1} x_r^*, \frac{p_t}{1 - \sum_{r=1}^{t-1} p_r}\right), \quad t = 2, \dots, d.$$

*We define the set $\mathbf{H}_r = \left\{ v \mid v = \sum_{t=1}^d v_t^2, \sum_{t=1}^d v_t = 0, v_t \in \{-n, \dots, 0, \dots, n\} \right\}$ as the range of squared IPDs for both the UMHG and MN distributions. The expected value of the IPD is $\mathbb{E}\left(d_{(x)}^2\right) = 2tr\left(\mathbf{\Sigma}_x\right)$, where $\mathbf{\Sigma}_x$ is the covariance matrix of $\mathbf{X}$ that defined in Definition 1. for the UMHG and in Table 1 for the MN distribution.*

Note that $\sum_{v \in \mathbf{H}_r}$ and the set $\mathbf{H}_r$ appear in the rest of this article. They are all defined in the same way as in Theorem 3. To obtain set $\mathbf{H}_r$, we perform a full grid search for $v_t \in \{-n, \dots, 0, \dots, n\}$ that satisfy $\sum_{t=1}^d v_t = 0$ and $\sum_{t=1}^d v_t^2 = v \in \{0, \dots, 2n^2\}$.

Suppose $\mathbf{X}$ and $\mathbf{X}^*$ are the $d$-dimensional random vectors that are drawn independently from some discrete multivariate distribution $F_x$ with integer-valued unrestricted support. The mean and variance of $\mathbf{X}$ and $\mathbf{X}^*$ are $\boldsymbol{\mu}_x$ and $\mathbf{\Sigma}_x$, respectively. Let $T_{(x)}(t) = X_t - X_t^*$ for $t = 1, \dots, d$. According to Equation (12), the IPD can be expressed as

$$d_{(x)}^2 = \sum_{t=1}^d (X_t - X_t^*)^2 = \sum_{t=1}^d T_{(x)}^2(t). \tag{24}$$

Because $F_x$ has integer-valued support, then we have $T_{(x)}(t) \in \mathbb{Z}$, where $\mathbb{Z}$ is the set of all integers so that $T_{(x)}^2(t) \in \{0^2, 1^2, 2^2, 3^2, \dots\}$, the set of squared integers. Therefore, the range of $d_{(x)}^2$ is the set that contains all possible sums of squared integers up to dimension $d$.

**Lemma 5** (Difference between i.i.d. unrestricted vectors). *Let $\mathbf{T}_{(x)} = \mathbf{X} - \mathbf{X}^* = \left(X_1 - X_1^*, \dots, X_d - X_d^*\right)' = \left(T_{(x)}(1), \dots, T_{(x)}(d)\right)'$, the p.m.f. of $T_{(x)}^2(t)$ for $t = 1, \dots, d$ is*

$$\mathbb{P}\left(T_{(x)}^2(t) = v_t^2\right) = \left(1 + \mathbb{1}_{\{v_t \neq 0\}}\right) \sum_{\psi \in \Psi} \mathbb{P}\left(X_t = \psi + v_t\right) \mathbb{P}\left(X_t^* = \psi\right), \tag{25}$$

*where $\Psi$ is the range of $X_t$ for corresponding marginal distribution, $\mathbb{1}_{\{v_t \neq 0\}} = 1$ when $v_t > 0$ and $\mathbb{1}_{\{v_t \neq 0\}} = 0$, otherwise.*

**Theorem 4** (One-sample IPD: Unrestricted vectors). *The p.m.f. of IPD between observations $\mathbf{X}$ and $\mathbf{X}^*$ is given by*

$$\begin{aligned}\mathbb{P}\left(d_{(x)}^2 = v\right) &= \sum_{v \in \mathbf{H}_u} \prod_{t=1}^d \mathbb{P}\left(T_{(x)}^2(t) = v_t^2\right) \\ &= \sum_{v \in \mathbf{H}_u} \prod_{t=1}^d \left(1 + \mathbb{1}_{\{v_t \neq 0\}}\right) \sum_{\psi \in \Psi} \mathbb{P}(X_t = \psi + v_t) \cdot \mathbb{P}(X_t^* = \psi),\end{aligned} \tag{26}$$

*where $\sum_{v \in \mathbf{H}_u}$ is the sum taken over all $v \in \mathbf{H}_u$. We define set $\mathbf{H}_u = \{v \mid v = v_1^2 + \cdots + v_d^2, v_t \in \mathbb{N}_0, t = 1, \dots, d\}$ as the range of squared IPDs for the MPSD family except the MN distribution. The mean of the IPD depends on the underlying distribution since*

$$\mathbb{E}\left(d_{(x)}^2\right) = \text{tr}\left(2\mathbf{\Sigma}_x\right). \tag{27}$$

The set $\mathbf{H}_u$ contains values of $v = v_1^2 + \ldots + v_d^2$ based on all partitions of $v \geq 0$ into $d \geq 1$ parts $v_1^2, \ldots, v_d^2$. Note that $\sum_{v \in \mathbf{H}_u}$ and set $\mathbf{H}_u$ appear in the rest of this article and they are all defined in the same way as in Theorem 4. The value of $\mathbb{P}(X_t = \psi + v_t)$ depends on the marginal distribution of $\mathbf{X}$. A grid search can be used to compute $\mathbb{P}\left(d_{(x)ij}^2 = v\right)$. Since $v = v_1^2 + \cdots + v_d^2$ and $v_t \leq \sqrt{v}$ for $t = 1, \ldots, d$, then a full grid search would have time complexity $O\left(v^{d/2}\right)$. It is time-consuming to use a full grid search when the dimension $d$ is large. The number of compositions of $v$ into $d$ parts is well studied and the number of such partition is $\binom{v+d-1}{v}$, which is the number of ways to place $v \geq 0$ indistinguishable balls in $d \geq 0$ labeled cells. Nijenhuis and Wilf[47] illustrate an algorithm called NEXTCOM to list all such partitions. However, not all parts of each partition have integer square root. Therefore, one can use the algorithm NEXTCOM to find all partitions of $v$ into $d$ parts and then admit the partitions whose parts are perfect squares.

## 7.4 | Two-sample IPDs: Restricted and unrestricted vectors

Similarly to the IPD between i.i.d. vectors case, we also derive the distribution of squared IPD between i.n.i.d. random vectors in two different circumstances. One with restricted support and one with unrestricted support. Note that we use i.n.i.d. to denote independent and nonidentically distributed random variables or vectors throughout the article.

**Theorem 5** (Two-sample IPD: Restricted vectors). *Suppose* $\mathbf{X}$ *is a d-dimensional vector that is drawn from* $\text{UMHG}(n, \mathbf{a}_{(x)}, a_x)$ *or* $\text{MN}(n, \mathbf{p}_x)$ *distribution. Suppose further* $\mathbf{Y}$ *is a d-dimensional vector that is drawn from* $\text{UMHG}(n, \mathbf{a}_{(y)}, a_y)$ *or* $\text{MN}(n, \mathbf{p}_y)$ *distribution. Let* $\Phi = \{0, 1, \ldots, n\}$. *The p.m.f. of the IPD between* $\mathbf{X}$ *and* $\mathbf{Y}$ *is given by*

$$\mathbb{P}\left(d_{(xy)}^2 = v\right) = \sum_{v \in \mathbf{H}_r} \sum_{\phi_1 \in \Phi} \cdots \sum_{\phi_d \in \Phi} \mathbb{P}(X_1 = \phi_1 + v_1) \cdot \mathbb{P}(Y_1 = \phi_1) \cdot \mathbb{1}_{\left\{\phi_t + v_t \geq 0, \sum_{t=1}^{d} \phi_t = n\right\}} \times \prod_{t=2}^{d} \mathbb{P}(X_t = \phi_t + v_t | x_1, \ldots, x_{t-1}) \cdot \mathbb{P}(Y_t = \phi_t | y_1, \ldots, y_{t-1}), \tag{28}$$

*where set* $\mathbf{H}_r = \left\{ v \mid v = \sum_{t=1}^{d} v_t^2, \sum_{t=1}^{d} v_t = 0, v_t \in \{-n, \ldots, 0, \ldots, n\} \right\}$ *is the range of* $d_{(xy)}^2$. *For* $\mathbf{X} \sim \text{UMHG}(n, \mathbf{a}_{(x)}, a_x)$ *and* $\mathbf{Y} \sim \text{UMHG}(n, \mathbf{a}_{(y)}, a_y)$, *the distributions of random variables* $X_t | x_1, \ldots, x_{t-1}$ *and* $Y_t | y_1, \ldots, y_{t-1}$ *are*

$$X_1 \sim \text{UHG}(n, a_x, a_{(x)1}), X_t | x_1, \ldots, x_{t-1} \sim \text{UHG}\left(n - \sum_{r=1}^{t-1} x_r, a_x - \sum_{r=1}^{t-1} a_{(x)r}, a_{(x)t}\right);$$
$$Y_1 \sim \text{UHG}(n, a_y, a_{(y)1}), Y_t | y_1, \ldots, y_{t-1} \sim \text{UHG}\left(n - \sum_{r=1}^{t-1} y_r, a_y - \sum_{r=1}^{t-1} a_{(y)r}, a_{(y)t}\right),$$

*where* $a_x = \sum_{t=1}^{d} a_{(x)t}$ *and* $a_y = \sum_{t=1}^{d} a_{(y)t}$. *For* $\mathbf{X} \sim \text{MN}(n, \mathbf{p}_x)$ *and* $\mathbf{Y} \sim \text{MN}(n, \mathbf{p}_y)$, *the distribution of* $X_t | x_1, \ldots, x_{t-1}$ *and* $Y_t | y_1, \ldots, y_{t-1}$ *is*

$$X_1 \sim B(n, p_{(x)1}), X_t | x_1, \ldots, x_{t-1} \sim B\left(n - \sum_{r=1}^{t-1} x_r, \frac{p_{(x)t}}{1 - \sum_{r=1}^{t-1} p_{(x)r}}\right), \quad t = 2, \ldots, d;$$
$$Y_1 \sim B(n, p_{(y)1}), Y_t | y_1, \ldots, y_{t-1} \sim B\left(n - \sum_{r=1}^{t-1} y_r, \frac{p_{(y)t}}{1 - \sum_{r=1}^{t-1} p_{(y)r}}\right), \quad t = 2, \ldots, d.$$

*The expected value of* $d_{(xy)}^2$ *is*

$$\mathbb{E}\left(d_{(xy)}^2\right) = tr\left(\boldsymbol{\Sigma}_x + \boldsymbol{\Sigma}_y\right) + \left(\boldsymbol{\mu}_x - \boldsymbol{\mu}_y\right)' \left(\boldsymbol{\mu}_x - \boldsymbol{\mu}_y\right), \tag{29}$$

*where* $\boldsymbol{\mu}_x$ *and* $\boldsymbol{\Sigma}_x$ *are the mean and covariance matrix of* $\mathbf{X}$ *that were defined in Definition 1. for the UMHG and in Table* 1 *for the MN distribution and similarly for* $\mathbf{Y}$.

Suppose $\mathbf{X}$ and $\mathbf{Y}$ are $d$-dimensional random vectors that are drawn from some discrete multivariate functions $F_x$ and $F_y$, both with integer-valued unrestricted support. The mean of $\mathbf{X}$ and $\mathbf{Y}$ are $\boldsymbol{\mu}_x$ and $\boldsymbol{\mu}_y$, and the variance of $\mathbf{X}$ and $\mathbf{Y}$ are $\boldsymbol{\Sigma}_x$ and $\boldsymbol{\Sigma}_y$. According to Equation (12), the IPD between $\mathbf{X}$ and $\mathbf{Y}$ can be expressed as

$$d^2_{(xy)} = \sum_{t=1}^{d} (X_t - Y_t)^2 = \sum_{t=1}^{d} T^2_{(xy)}(t). \tag{30}$$

Because both $F_x$ and $F_y$ has integer-valued support, $T_{(x)}(t) \in \mathbb{Z}$. Therefore, the range of $d^2_{(xy)}$ is $\mathbf{H}_u = \left\{ v \mid v = \sum_{t=1}^{d} v_t^2, v_t \in \mathbb{N}_0, t = 1, \dots, d \right\}$.

**Lemma 6** (Difference between i.n.i.d. unrestricted vectors). *Let* $\mathbf{T}_{(xy)} = \mathbf{X} - \mathbf{Y} = (X_1 - Y_1, \dots, X_d - Y_d) = \left(T_{(xy)}(1), \dots, T_{(xy)}(d)\right)'$, *the p.m.f. of* $T^2_{(xy)}(t)$ *for* $t = 1, \dots, d$ *is*

$$\mathbb{P}\left(T^2_{(xy)}(t) = v_t^2\right) = \begin{cases} \sum_{\psi \in \Psi} \{\mathbb{P}(X_t = \psi + v_t) \cdot \mathbb{P}(Y_t = \psi)\} \\ + \sum_{\psi \in \Psi} \{\mathbb{P}(Y_t = \psi + v_t) \cdot \mathbb{P}(X_t = \psi)\}, & \text{if } v_t \neq 0, \\ \sum_{\psi \in \Psi} \mathbb{P}(X_t = \psi) \cdot \mathbb{P}(Y_t = \psi), & \text{if } v_t = 0, \end{cases} \tag{31}$$

*where* $\Psi$ *is the common support of* $F_x$ *and* $F_y$.

**Theorem 6** (Two-sample IPD: Unrestricted vectors). *The p.m.f. of the squared IPD between observations* $\mathbf{X}$ *and* $\mathbf{Y}$ *is given by*

$$\begin{aligned} \mathbb{P}\left(d^2_{(xy)} = v\right) &= \sum_{v \in \mathbf{H}_u} \prod_{t=1}^{d} \mathbb{P}\left(T^2_{(xy)}(t) = v_t^2\right) \\ &= \sum_{v \in \mathbf{H}_r} \left( \prod_{\{t \mid v_t = 0\}} \mathbb{P}(X_t - Y_t = 0) \prod_{\{t \mid v_t \neq 0\}} \left[\mathbb{P}(X_t - Y_t = v_t) + \mathbb{P}(X_t - Y_t = -v_t)\right] \right). \end{aligned} \tag{32}$$

*The probabilities* $\mathbb{P}(X_t - Y_t = 0)$, $\mathbb{P}(X_t - Y_t = v_t)$, *and* $\mathbb{P}(Y_t - X_t = v_t)$ *can be expressed as follows:*

$$\begin{aligned} \mathbb{P}(X_t - Y_t = v_t) &= \sum_{\psi \in \Psi} \mathbb{P}(X_t = \psi + v_t) \cdot \mathbb{P}(Y_t = \psi), \\ \mathbb{P}(Y_t - X_t = v_t) &= \sum_{\psi \in \Psi} \mathbb{P}(Y_t = \psi + v_t) \cdot \mathbb{P}(X_t = \psi), \\ \mathbb{P}(X_t - Y_t = 0) &= \sum_{\psi \in \Psi} \mathbb{P}(X_t = \psi) \cdot \mathbb{P}(Y_t = \psi). \end{aligned} \tag{33}$$

*The expected value of the IPD is*

$$\mathbb{E}\left(d^2_{(xy)}\right) = \operatorname{tr}\left(\boldsymbol{\Sigma}_x + \boldsymbol{\Sigma}_y\right) + \left(\boldsymbol{\mu}_x + \boldsymbol{\mu}_y\right)' \left(\boldsymbol{\mu}_x + \boldsymbol{\mu}_y\right). \tag{34}$$

In contrast with the p.m.f. of the IPD in Equation (26), the p.m.f. in Equation (32) does not have the term $\mathbb{1}_{\{v_t \neq 0\}}$. The reason is that $T_{(x)}(t) = X_t - X_t^*$ is symmetric since $X_t$ and $X_t^*$ are i.i.d., but $T_{(xy)}(t) = X_t - Y_t$ is not symmetric, since $X_t$ and $Y_t$ are only independent but not identically distributed.

## 7.5 | MN and NGMN IPDs

**Corollary 1** (One-sample IPD: MN vectors). *Suppose* $\mathbf{X}$ *and* $\mathbf{X}^*$ *are d-dimensional random vectors that are drawn independently from the MN*$(n, \mathbf{p})$ *distribution. Let* $\Phi = \{0, 1, \dots, n\}$ *and* $\mathbf{V}_r = \left\{ \mathbf{v} = (v_1, \dots, v_d)' \mid v_1^2 + \cdots + v_d^2 = v, v \in \mathbf{H}_r \right\}$, *the p.m.f. of the IPD between* $\mathbf{X}$ *and* $\mathbf{X}^*$ *is given by*

$$\mathbb{P}\left(d^2_{(x)}=v\right)=\sum_{\mathbf{v}\in\mathbf{V}_r}\sum_{\phi_1\in\Phi}\cdots\sum_{\phi_d\in\Phi}\mathbb{P}\left(X_1=\phi_1+v_1\right)\cdot\mathbb{P}\left(X_1^*=\phi_1\right)\cdot\mathbb{1}_{\left\{\phi_t+v_t\geq 0,\ \sum_{t=1}^{d}\phi_t=n\right\}}$$
$$\times\prod_{t=2}^{d}\mathbb{P}\left(X_t=\phi_t+v_t|x_1,\ldots,x_{t-1}\right)\cdot\mathbb{P}\left(X_t^*=\phi_t|x_1^*,\ldots,x_{t-1}^*\right),\tag{35}$$

*where* $\mathbb{1}_{\left\{\phi_t+v_t\geq 0,\sum_{t=1}^{d}\phi_t=n\right\}}$ *equals 1 if* $\phi_t+v_t\geq 0$ *and* $\sum_{t=1}^{d}\phi_t=n$*, and equals 0, otherwise. The distribution of* $X_t|x_1,\ldots,x_{t-1}$ *and* $X_t^*|x_1,\ldots,x_{t-1}$ *is*

$$X_1\sim B(n,p_1),\text{and } X_t|x_1,\ldots,x_{t-1}\sim B\left(n-\sum_{r=1}^{t-1}x_r,\frac{p_t}{1-\sum_{r=1}^{t-1}p_r}\right),\quad\text{for } t=2,\ldots,d;$$
$$X_1^*\sim B(n,p_1),\text{and } X_t^*|x_1^*,\ldots,x_{t-1}^*\sim B\left(n-\sum_{r=1}^{t-1}x_r^*,\frac{p_t}{1-\sum_{r=1}^{t-1}p_r}\right),\quad\text{for } t=2,\ldots,d,$$

*where* $B(n,p_1)$ *represents the binomial distribution with parameters n and* $p_1$*. The expected value of the IPD is* $\mathbb{E}\left(d^2_{(x)}\right)=2\sum_{t=1}^{d}n(p_t-p_t^2)$.

**Corollary 2** (Two-sample IPD: MN vectors). *Suppose* $\mathbf{X}$ *is a d-dimensional vector that is drawn from MN*$(n,\mathbf{p}_x)$ *distribution. Suppose* $\mathbf{Y}$ *is a d-dimensional vector that is drawn from MN*$(n,\mathbf{p}_y)$ *distribution independent with* $\mathbf{X}$*. Let* $\Phi=\{0,1,\ldots,n\}$*, the p.m.f. of the IPD between* $\mathbf{X}$ *and* $\mathbf{Y}$ *is given by*

$$\mathbb{P}\left(d^2_{(xy)}=v\right)=\sum_{\mathbf{v}\in\mathbf{V}_r}\sum_{\phi_1\in\Phi}\cdots\sum_{\phi_d\in\Phi}\mathbb{P}\left(X_1=\phi_1+v_1\right)\cdot\mathbb{P}\left(Y_1=\phi_1\right)\cdot\mathbb{1}_{\left\{\phi_t+v_t\geq 0,\ \sum_{t=1}^{d}\phi_t=n\right\}}$$
$$\times\prod_{t=2}^{d}\mathbb{P}\left(X_t=\phi_t+v_t|x_1,\ldots,x_{t-1}\right)\cdot\mathbb{P}\left(Y_t=\phi_t|y_1,\ldots,y_{t-1}\right),\tag{36}$$

*where the distribution of* $X_t|x_1,\ldots,x_{t-1}$ *and* $Y_t|y_1,\ldots,y_{t-1}$ *are*

$$X_1\sim B(n,p_{(x)1}),\text{and } X_t|x_1,\ldots,x_{t-1}\sim B\left(n-\sum_{r=1}^{t-1}x_r,\frac{p_{(x)t}}{1-\sum_{r=1}^{t-1}p_{(x)r}}\right),\quad\text{for } t=2,\ldots,d;$$
$$Y_1\sim B(n,p_{(y)1}),\text{and } Y_t|y_1,\ldots,y_{t-1}\sim B\left(n-\sum_{r=1}^{t-1}y_r,\frac{p_{(y)t}}{1-\sum_{r=1}^{t-1}p_{(y)r}}\right),\quad\text{for } t=2,\ldots,d.$$

*The expected value of IPD is* $\mathbb{E}\left(d^2_{(xy)}\right)=\sum_{t=1}^{d}n\left(p_{(x)t}+p_{(y)t}-2p_{(x)t}p_{(y)t}\right)$.

**Corollary 3** (One-sample IPD: NGMN vectors). *Suppose* $\mathbf{X}$ *and* $\mathbf{X}^*$ *are independent d-dimensional random vectors that are drawn from the NGMN distribution with parameters k and* $\mathbf{p}$ *defined in Table* 1*. The p.m.f. of the squared Euclidean IPD between* $\mathbf{X}$ *and* $\mathbf{X}^*$ *is given by*

$$\mathbb{P}\left(d^2_{(x)}=v\right)=\sum_{\mathbf{v}\in\mathbf{V}_u}\prod_{t=1}^{d}\left(1+\mathbb{1}_{\{v_t\neq 0\}}\right)\sum_{\psi=0}^{\infty}\text{NB}\left(v_t+\psi;k,\frac{p_t}{p_0}\right)\cdot\text{NB}\left(\psi;k,\frac{p_t}{p_0}\right),\tag{37}$$

*where* $\text{NB}\left(v_t+\psi;k,\frac{p_t}{p_0}\right)$ *is the negative binomial distribution with parameters k and* $\frac{p_t}{p_0}$ *that is evaluated at* $v_t+\psi$ *and* $p_0=1-\sum_{t=1}^{d}p_t$*. The expected value of the IPD is* $\mathbb{E}\left(d^2_{(x)}\right)=2k\sum_{t=1}^{d}\frac{p_t}{p_0}\left(1+\frac{p_t}{p_0}\right)$.

**Corollary 4** (Two-sample IPD: NGMN vectors). *Suppose* $\mathbf{X}$ *and* $\mathbf{Y}$ *are independent* $d$*-dimensional vectors that are drawn from* NGMM$(k, \mathbf{p}_x)$ *and* NGMN$(k, \mathbf{p}_y)$ *defined in Table* 1. *The p.m.f. of the squared Euclidean IPD between* $\mathbf{X}$ *and* $\mathbf{Y}$ *is given by*

$$\begin{aligned}\mathbb{P}\left(d^2_{(xy)}=v\right)=\sum_{\mathbf{v}\in\mathbf{V}_u}\Bigg[\prod_{t|v_t=0}\left\{\sum_{\psi=0}^{\infty}\mathrm{NB}\left(\psi;k,\frac{p_{(x)t}}{p_{(x)0}}\right)\cdot\mathrm{NB}\left(\psi;k,\frac{p_{(y)t}}{p_{(y)0}}\right)\right\}\\ \cdot\prod_{t|v_t\neq 0}\sum_{\psi=0}^{\infty}\left\{\mathrm{NB}\left(v_t+\psi;k,\frac{p_{(x)t}}{p_{(x)0}}\right)\cdot\mathrm{NB}\left(\psi;k,\frac{p_{(y)t}}{p_{(y)0}}\right)\right.\\ \left.+\mathrm{NB}\left(v_t+\psi;k,\frac{p_{(y)t}}{p_{(y)0}}\right)\cdot\mathrm{NB}\left(\psi;k,\frac{p_{(x)t}}{p_{(x)0}}\right)\right\}\Bigg],\end{aligned}\tag{38}$$

*where* $p_{(x)0}=1-\sum_{t=1}^{d}p_{(x)t}$ *and* $p_{(y)0}=1-\sum_{t=1}^{d}p_{(y)t}$. *The expected value of* $d^2_{(xy)}$ *is*

$$\mathbb{E}\left(d^2_{(xy)}\right)=\sum_{t=1}^{d}\left\{k(k+1)\left(\left(\frac{p_{(x)t}}{p_{(x)0}}\right)^2+\left(\frac{p_{(y)t}}{p_{(y)0}}\right)^2\right)-2k^2\frac{p_{(x)t}p_{(y)t}}{p^2_{(x)0}}+k\left(\frac{p_{(x)t}}{p_{(x)0}}+\frac{p_{(y)t}}{p_{(y)0}}\right)\right\}.\tag{39}$$

Figure 1 displays the p.m.f. of IPD between i.i.d. vectors from an MN distribution with $d=6$ and parameters $n=5$ and $\mathbf{p}=(0.10,0.15,0.30,0.10,0.15,0.20)'$. The p.m.f. is slightly skewed to the right with a short tail compared to NGMN IPD. The expected value of the IPD is 8.05. Figure 2 displays the p.m.f. of IPD between i.i.d. vectors from an NGMN distribution with $d=6$ and parameters $k=5$ and $\mathbf{p}=(0.05,0.15,0.25,0.10,0.15,0.05)'$. The p.m.f. is also skewed to the right, with a long tail. The expected value of the squared IPD is 49.60.

The plot of p.m.f. of IPDs between i.n.i.d. vectors from MN distributions appears in Figure 3. Both vectors have parameter $n=5$, one has $\mathbf{p}=(0.05,0.15,0.25,0.15,0.15,0.25)'$ and the other one has $\mathbf{p}=(0.14,0.27,0.19,0.12,0.21,0.07)'$. The p.m.f. is skewed to the right with a short tail. The expected value of the IPD is 8.38. Figure 4 displays the p.m.f. of IPDs between i.n.i.d. vectors from NGMN distributions. Both vectors have parameter $k=5$, one has $\mathbf{p}_x=(0.05,0.15,0.25,0.10,0.15,0.05)'$ and the other one has $\mathbf{p}_y=(0.04,0.17,0.19,0.09,0.21,0.07)'$. The p.m.f. is skewed to the right with a long tail. The expected value of the IPD is 57.21.

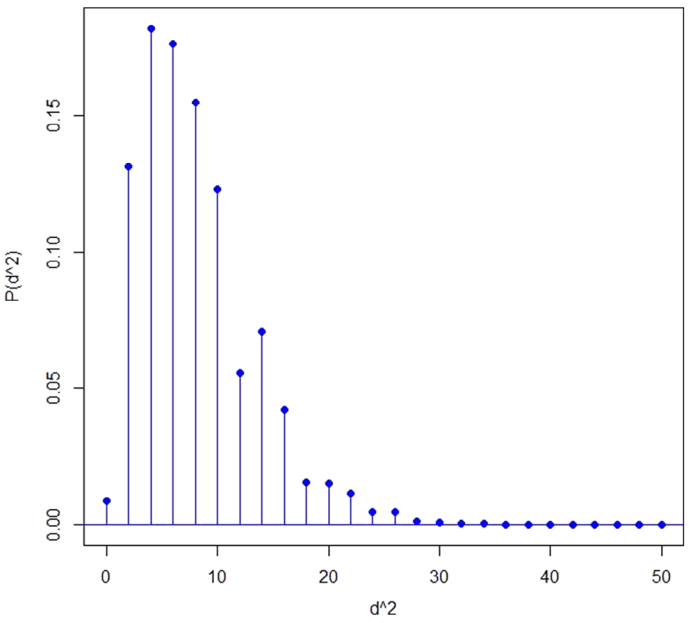



**FIGURE 1** Multinomial interpoint distance [Colour figure can be viewed at wileyonlinelibrary.com]

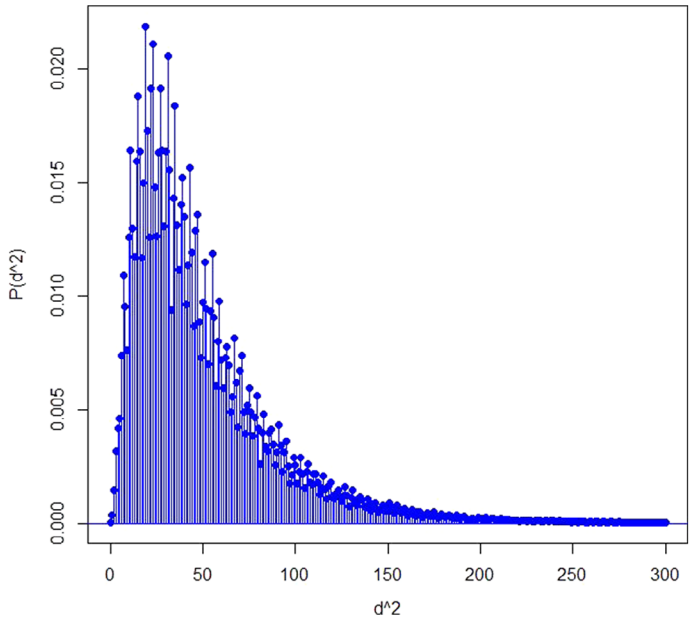



**FIGURE 2** Negative multinomial interpoint distance [Colour figure can be viewed at wileyonlinelibrary.com]

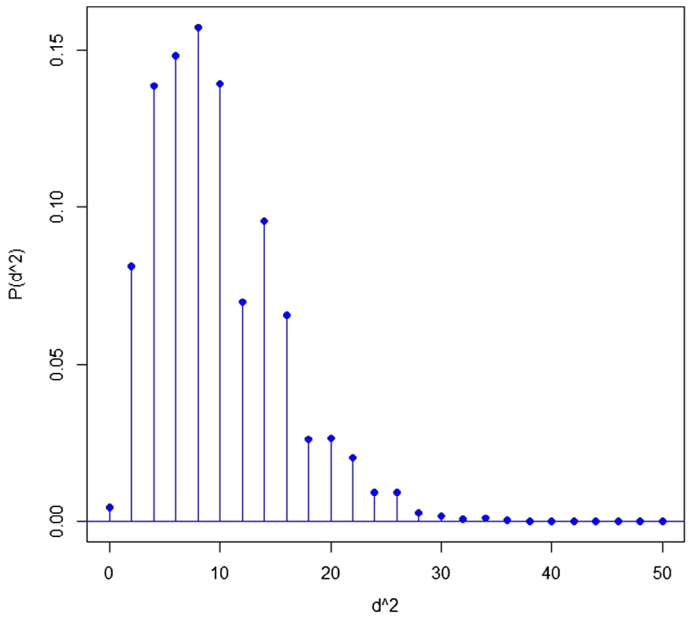



**FIGURE 3** Multinomial interpoint distance of i.n.i.d. vectors [Colour figure can be viewed at wileyonlinelibrary.com]

## 8 | VISUALIZATION

This section reviews the results in References 48 and uses the IPDs to assess and visualize the null hypothesis $H_0 : F_x = F_y$ against $H_a : F_x \neq F_y$. While data visualization contributes to our understanding of complex multivariate data, high-dimensional datasets present a major challenge for data visualization. High-dimensional data refer to situations where the dimension $d \to \infty$ as the sample size $N \to \infty$ and ultrahigh-dimensional data include situations when $d$ grows at a nonpolynomial rate in sample size $N$, namely, $\log(d) = O(N^{\alpha})$ for $\alpha > 0$. Many methods of displaying multivariate data have been suggested. Two general techniques for visualizing multivariate observations are iconic and geometric data representations. The iconic representation uses icons such as glyphs, stars, sunflowers, or Chernoff faces. Glyphs are small icons that represent data points in $\mathbb{R}^d$. Differences and similarities in the structure, form, or appearance of the icon can reveal important data characteristics.

We introduce a method for the visualization of high-dimensional datasets using IPDs and provide a basis for their comparison. We are interested in a visual aid to examine the null hypothesis $H_0 : F_x = F_y$ against $H_a : F_x \neq F_y$. Both

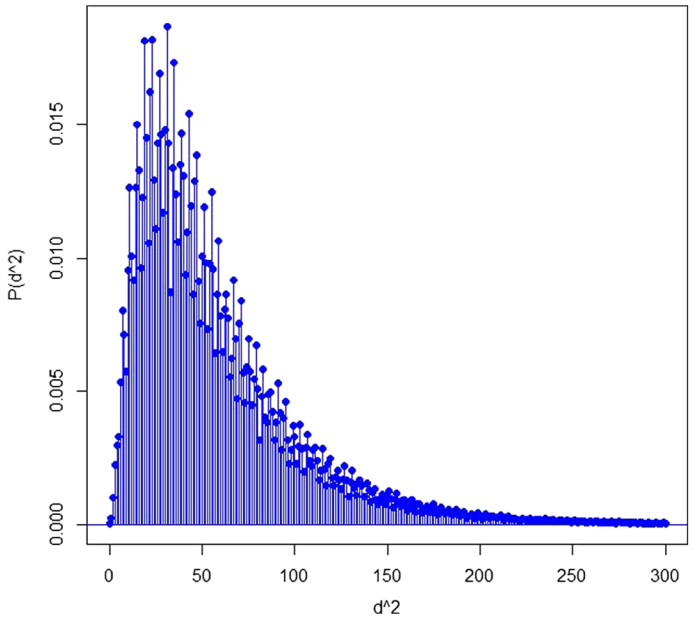



**FIGURE 4** Negative multinomial interpoint distance of i.n.i.d. vectors [Colour figure can be viewed at wileyonlinelibrary.com]

parametric and nonparametric tests have been proposed to investigate $H_0$. Some approaches reduce $H_0$ and assume $F_x$ and $F_y$ are equal except for their centers and/or scales.

Let $\mathcal{Q}_x$, $\mathcal{Q}_y$, and $\mathcal{Q}_{xy}$ denote the DFs of $D_x = \|\mathbf{X}_1 - \mathbf{X}_2\|$, $D_y = \|\mathbf{Y}_1 - \mathbf{Y}_2\|$, and $D_{xy} = \|\mathbf{X}_1 - \mathbf{Y}_2\|$, respectively. Maa et al[10] prove that $F_x = F_y$ if and only if the IPDs within and between samples have the same univariate distribution. Hence, instead of testing $H_0 : F_x = F_y$, we can consider an equivalent null hypothesis $H_0' : \mathcal{Q}_x = \mathcal{Q}_y = \mathcal{Q}_{xy}$, which holds if and only if $F_x = F_y$. Let $\mu_{F_xF_x}$ and $\mu_{F_yF_y}$ represent the expected value of within sample IPDS , respectively, and $\mu_{F_xF_y}$ denotes the expected value of the between samples IPD. We compute the average IPD of $\mathbb{X}$ and $\mathbb{Y}$ using $\overline{d}_{(x)} = \frac{2}{N_x(N_x-1)} \sum_{i=1}^{N_x-1} \sum_{j=i+1}^{N_x} d_{(x)ij}$ and $\overline{d}_{(y)} = \frac{2}{N_y(N_y-1)} \sum_{i=1}^{N_y-1} \sum_{j=i+1}^{N_y} d_{(y)ij}$, respectively. The average IPD between $\mathbb{X}$ and $\mathbb{Y}$ is $\overline{d}_{(xy)} = \frac{1}{N_xN_y} \sum_{i=1}^{N_x} \sum_{j=1}^{N_y} d_{(xy)ij}$. We estimate $\mu_{F_xF_x}$, $\mu_{F_yF_y}$, and $\mu_{F_xF_y}$ with $\overline{d}_{(x)}$, $\overline{d}_{(y)}$, and $\overline{d}_{(xy)}$, respectively.

Let $R = \max D(\mathbb{X}, \mathbb{Y}) - \min D(\mathbb{X}, \mathbb{Y})$ denote the range of all IPDs that appear below the main diagonal of $D(\mathbb{X}, \mathbb{Y})$. We will divide the range into $s$ evaluation points denoted by $\delta(t)$ for $t = 1, \dots s$. Denote the cumulative distributions of $d_{(x)ij}$, $d_{(y)ij}$, and $d_{(xy)ij}$ evaluated at $\delta(t)$ by $H_x(t) = \mathbb{P}(d_{(x)ij} \le \delta(t))$, $H_y(t) = \mathbb{P}(d_{(y)ij} \le \delta(t))$, and $H_{xy}(t) = \mathbb{P}(d_{(xy)ij} \le \delta(t))$, respectively. Let $I(.)$ denote the indicator function and estimate the DFs by

$$\hat{H}_x(t) = \frac{1}{m_x} \sum_{i=1}^{N_x-1} \sum_{j=i+1}^{N_x} I(d_{(x)ij} \le \delta(t)),$$

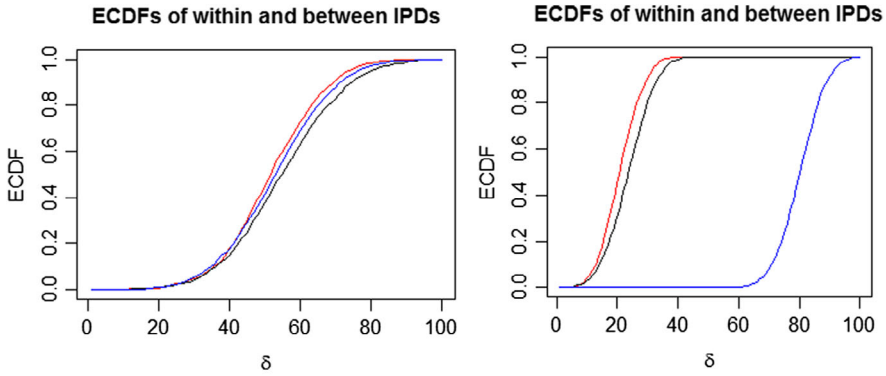
$$\hat{H}_y(t) = \frac{1}{m_y} \sum_{i=1}^{N_y-1} \sum_{j=i+1}^{N_y} I(d_{(y)ij} \le \delta(t)),$$

$$\hat{H}_{xy}(t) = \frac{1}{m_{xy}} \sum_{i=1}^{N_x} \sum_{i=1}^{N_y} I(d_{(y)ij} \le \delta(t)). \tag{40}$$

We obtain a simultaneous plot of

$$(\delta(t), \hat{H}_x(\delta(t))), (\delta(t), \hat{H}_y(\delta(t))), (\delta(t), \hat{H}_{xy}(\delta(t))), \quad \text{for } t = 1, \dots, s. \tag{41}$$

With applications in statistical decision theory, clinical trials, experimental design, and portfolio analysis, stochastic orderings of random vectors are of interest when comparing several multivariate distributions. Giovagnoli and Wynn[49] define the multivariate dispersive order (weak D-ordering) for two random vectors $\mathbf{X}$ and $\mathbf{Y}$ in $\mathbb{R}^d$ as follows. Define $\mathbf{X} \le_D \mathbf{Y}$ if and only if $D_x \le_{St} D_y$, where $\le_{St}$ is the stochastic ordering between two random variables. One can show that $D_x \le_{St} D_y$ holds if and only if $\mathcal{Q}_x(t) \ge \mathcal{Q}_y(t)$ for all $t \in \mathbb{R}$. The hypothesis $K_0 : \mathbf{X} \le_D \mathbf{Y}$ is equivalent to $K_0' : \mathcal{Q}_x(t) \ge \mathcal{Q}_y(t)$




(A) $F_x = F_y \sim \mathbb{N}(\mathbf{0}, \mathbf{I}_d)$. (B) $F_x \sim \mathbb{N}(\mathbf{0}, \mathbf{I}_d)$ and $F_y \sim \mathbb{N}(\mathbf{1}, \mathbf{I}_d)$.

**FIGURE 5** A, Empirical CDFs of $\mathbb{X}$ interpoint distances (IPDs) (red, top), $\mathbb{Y}$ IPDs (black, bottom), and between sample IPDs (blue, middle) when $F_x = F_y \sim \mathbb{N}(\mathbf{0}, \mathbf{I}_d)$. B, Empirical CDFs of $\mathbb{X}$ IPDs (red, next to top), $\mathbb{Y}$ IPDs (black, top), and between sample IPDs (blue, bottom) when $F_x \sim \mathbb{N}(\mathbf{0}, \mathbf{I}_d)$ and $F_y \sim \mathbb{N}(\mathbf{1}, \mathbf{I}_d)$ [Colour figure can be viewed at wileyonlinelibrary.com]

for all $t \in \mathbb{R}$. It follows from $\mathbf{X} \leq_D \mathbf{Y}$ that $\mathbb{E}(r(D_x)) \leq \mathbb{E}(r(D_y))$ for all nondecreasing functions $r$ on $[0, \infty)$ and $\text{tr}(\mathbb{C}\text{ov}(\mathbf{X})) \leq \text{tr}(\mathbb{C}\text{ov}(\mathbf{Y}))$. Furthermore, the dispersion order $\leq_D$ is location and rotation free. If $D_x \leq_{St} D_y$, then $\mathbf{\Gamma X} + \mathbf{a} \leq_D \mathbf{\Lambda Y} + \mathbf{b}$ for all orthogonal matrices $\mathbf{\Gamma}$ and $\mathbf{\Lambda}$ and for all vectors $\mathbf{a}$ and $\mathbf{b}$.

## 8.1 | Shift alternatives

When the $\mathbb{X}$ and $\mathbb{Y}$ groups differ in location $\boldsymbol{\mu}$ only, that is, $F_x(X) = F_y(Y - \boldsymbol{\mu})$, we have $\mathcal{Q}_x = \mathcal{Q}_y \geq \mathcal{Q}_{xy}$ with equality holding if and only if $\boldsymbol{\mu} = 0$. If $\boldsymbol{\mu} = 0$, then the result follows since $F_x = F_y$ is equivalent to $\mathcal{Q}_x = \mathcal{Q}_y = \mathcal{Q}_{xy}$. One can verify that the within-sample IPDs are invariant with respect to location shift so that $\mathcal{Q}_x = \mathcal{Q}_y$. Moreover, the between-sample IPDs will become larger than within sample IPDs when $\boldsymbol{\mu} \neq 0$ since $\mathbf{X} - \mathbf{Y}$ is no longer centered at zero. Hence, $\mathcal{Q}_x = \mathcal{Q}_y \geq \mathcal{Q}_{xy}$. As the following example shows, when ECDFs $\hat{H}_x(t)$ and $\hat{H}_y(t)$ are closer together than the empirical cumulative distribution function (ECDF) $\hat{H}_{xy}(t)$, that is, $\mathcal{Q}_x = \mathcal{Q}_y \geq \mathcal{Q}_{xy}$, one can say that the two distributions $F_x$ and $F_y$ have a location shift difference. To illustrate the utility of such simultaneous display of the ECDFs, we present the following examples where $d = 10\,000$, $s = 100$, and $N_x = N_y = 50$.

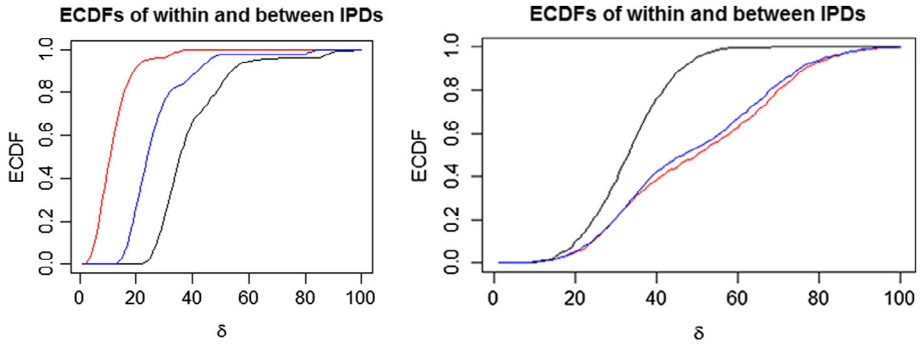
We generate observations from $\mathbb{N}(\mathbf{0}, \sigma_x^2 \mathbf{I}_d)$ and $\mathbb{N}(\boldsymbol{\mu}, \sigma_y^2 \mathbf{I}_d)$. Figure 5A displays the simultaneous plots when $H_0 : F_x = F_y$ is true, with $\sigma_x^2 = \sigma_y^2 = 1$ and $\boldsymbol{\mu} = 0$.

The ECDFs of the three IPDs concentrate in a narrow band since $H_0' : \mathcal{Q}_x = \mathcal{Q}_x = \mathcal{Q}_{xy}$ is true. Equivalently, $D_x \overset{(L)}{=} D_y \overset{(L)}{=} D_{xy}$, where $\overset{(L)}{=}$ stands for equality in law. Figure 5B displays the simultaneous plot under multivariate normal distributions with $\sigma_x^2 = \sigma_y^2 = 1$ and $\boldsymbol{\mu} = 1$. Since $H_0 : F_x = F_y$ is false, we expect the ECDFs to differ. We see $\mathcal{Q}_x = \mathcal{Q}_y \geq \mathcal{Q}_{xy}$. Equivalently, $D_x \overset{(L)}{=} D_y \leq_{St} D_{xy}$ with probability 1 under a location shift.

## 8.2 | Scale and shape alternatives

When the $\mathbb{X}$ and $\mathbb{Y}$ groups differ in scale only, that is, $F_x(X) = F_y(\frac{1}{\sigma} Y)$, we have $\min(\mathcal{Q}_x, \mathcal{Q}_y) \leq \mathcal{Q}_{xy} \leq \max(\mathcal{Q}_x, \mathcal{Q}_y)$ with equality holding if and only if $\sigma = 1$. If $\sigma = 1$, then the result follows since $F_x = F_y$ is equivalent to $\mathcal{Q}_x = \mathcal{Q}_y = \mathcal{Q}_{xy}$. Furthermore, $\mathcal{Q}_y \leq \mathcal{Q}_{xy}$ and $\mathcal{Q}_{xy} \leq \mathcal{Q}_x$ when $\sigma < 1$. Similarly, $\mathcal{Q}_x \leq \mathcal{Q}_{xy}$ and $\mathcal{Q}_{xy} \leq \mathcal{Q}_y$ when $\sigma > 1$ Equivalently, $D_x \leq_{St} D_{xy} \leq_{St} D_y$ if $\sigma > 1$ and $D_y \leq_{St} D_{xy} \leq_{St} D_x$ if $\sigma < 1$ with probability 1 under a scale change. To see the effects of a change in scale, consider Figure 6A that displays the simultaneous plots under multivariate normal groups when $F_x \neq F_y$, with $\sigma_x^2 = 1, \sigma_y^2 = 2$ and $\boldsymbol{\mu} = \mathbf{0}$.

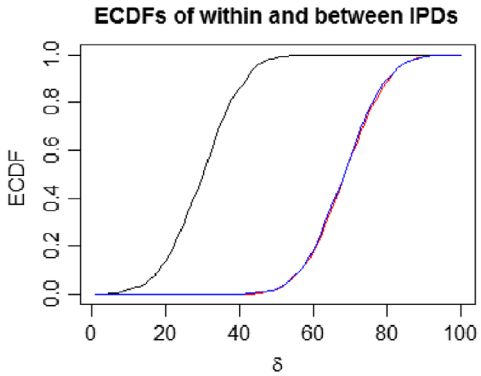
Here, the three ECDFs are clearly separated with ECDF of $D_{xy}$ in the middle. That is, $D_x \leq_{St} D_{xy} \leq_{St} D_y$. To see the effects of changing the shape of the distributions on the IPDs, consider Figure 6B that displays the simultaneous plots when $\mathbb{X}$ has the distribution of mixture $g_x(\mathbb{X}) = 0.5\mathbb{N}(\mathbf{0}, \mathbf{I}_d) + 0.5\mathbb{N}(\mathbf{1}, \mathbf{I}_d)$. The $\mathbb{Y}$ group is $\mathbb{N}(\mathbf{0}, \mathbf{I}_d)$. Here, the ECDF of the IPDs in the $\mathbb{Y}$ group is well-separated from those of $\mathbb{X}$ group, which is close to the ECDF of the between IPDs. We see the ordering, $D_x \leq_{St} D_{xy} \leq_{St} D_y$.




(A) Scale Change. (B) Shape Change.

**FIGURE 6** A, Empirical CDFs of $\mathbb{X}$ sample interpoint distances (IPDs) (red, top), $\mathbb{Y}$ sample IPDs (black, bottom), and between sample IPDs (blue, middle) when $F_x \sim \mathbb{N}(\mathbf{0}, \mathbf{I}_d)$ and $F_y \sim \mathbb{N}(\mathbf{0}, 2\mathbf{I}_d)$. B, Empirical CDFs of $\mathbb{X}$ sample IPDs (red, bottom), $\mathbb{Y}$ sample IPDs (black, top), and between sample IPDs (blue, next to bottom) when $F_x \sim 0.5\mathbb{N}(\mathbf{0}, \mathbf{I}_d) + 0.5\mathbb{N}(1, \mathbf{I}_d)$ and $F_y \sim \mathbb{N}(\mathbf{0}\ , \mathbf{I}_d)$ [Colour figure can be viewed at wileyonlinelibrary.com]

**FIGURE 7** The effects of a change in both location and scale, ECDFs of the interpoint distances (IPDs) of the multivariate Bernoulli distributions with independent components and probability vectors $\mathbf{p_x} = 0.5\mathbf{1}$ and $\mathbf{p_y} = 0.6\mathbf{1}$ are compared. The ECDF of the $\mathbb{Y}$ sample is on the top (black) and the ECDF of the $\mathbb{X}$ sample (red) almost coincides with that of $D_{xy}$ (blue). We see the relationship $D_y \leq_{St} D_{xy} = D_x$ [Colour figure can be viewed at wileyonlinelibrary.com]




To see the effects of a change in both location and scale, consider Figure 7 that displays the simultaneous plots for multivariate Bernoulli distributions with independent components and probability vectors $\mathbf{p}_x = p\mathbf{1}$ and $\mathbf{p}_y = q\mathbf{1}$, where $p = 0.5$ and $q = 0.6$. Here, the distributions of $X$ and $Y$ have different means and variances. Under independence of the components $D_X^2 \sim B(d, 2p(1-p), D_Y^2 \sim B(d, 2q(1-q)$, and $D_{XY}^2 \sim B(d, p+q-2pq)$. It follows that the stochastic order of the within- and between-sample IPDS depends on $p$ and $q$. The ECDF of the $\mathbb{X}$ sample is on the top and the ECDF of the $\mathbb{Y}$ sample almost coincides with that of the ECDF of $D_{xy}$. We see the relationship $D_y \leq_{St} D_{xy} = D_x$. Moreover, with $d = 10\,000$, we obtain the IPD means $\mathbb{E}(D_x) = \mathbb{E}(D_{xy}) = 5000$ and $\mathbb{E}(D_y) = 4800$.

## ORCID

*Reza Modarres* https://orcid.org/0000-0003-1240-6027

## REFERENCES

1. Modarres R. On the interpoint distances of Bernoulli vectors. *Stat Prob Lett*. 2013;84:215-222.
2. Modarres R. Multivariate Poisson interpoint distances. *Stat Prob Lett*. 2015;112:113-123.
3. Marozzi M. Multivariate tests based on interpoint distances with application to magnetic resonance imaging. *Stat Methods Med Res*. 2014;25(6). https://doi.org/10.1177/0962280214529104.
4. Marozzi M, Mukherjee M, Kalina J. Interpoint distance tests for high-dimensional comparison studies. 2019. https://doi.org/10.1080/02664763.2019.1649374.
5. Song Y, Modarres R. Interpoint distance test of homogeneity for multivariate mixture models. *Int Stat Rev*. 2019a;87(3):613-638. https://doi.org/10.1111/insr.12332.
6. Guo L, Modarres R. Nonparametric Tests of Independence Based on Interpoint Distances, Technical Report. Washington, DC, Department of Statistics, George Washington University, 2019a.
7. Osada R, Funkhouser T, Chazelle B, Dobkin D. Shape distributions. *ACM Trans Graph*. 2002;21(4):807-832.

8. Berrendero JR, Cuevas A, Pateiro-López B. Shape classification based on interpoint distance distributions. *J Multivar Anal*. 2016;146:237-247.
9. Glick N. Measurements of separation among probability densities or random variables. *Can J Stat*. 1975;3(2):267-276.
10. Maa JF, Pearl DK, Bartoszyński R. Reducing multidimensional two-sample data to one-dimensional interpoint comparisons. *Ann Stat*. 1996;24(3):1069-1074.
11. Friedman JH, Rafsky LC. Multivariate generalizations of the Wald-Wolfowitz and Smirnov two-sample tests. *Ann Stat*. 1979;7(4):697-717.
12. Henze N, Penrose MD. On the multivariate runs test. *Ann Stat*. 1999;27(1):290-298.
13. Hall P, Tajvidi N. Permutation tests for equality of distributions in high-dimensional settings. *Biometrika*. 2002;89(2):359-374.
14. Baringhaus L, Franz C. On a new multivariate two-sample test. *J Multivar Anal*. 2004;88(1):190-206.
15. Székely GJ, Rizzo ML. Energy statistics: statistics based on distances. *J Stat Plan Infer*. 2013;143:1249-1272.
16. Liu Z, Modarres R. A triangle test for equality of distribution functions in high dimensions. *J Nonparametric Stat*. 2011;22:1-11.
17. Jurečková J, Kalina J. Nonparametric multivariate rank tests and their unbiasedness. *Bernoulli*. 2012;18(1):229-251.
18. Biswas M, Ghosh AK. A nonparametric two-sample test applicable to high dimensional data. *J Multivar Anal*. 2014;123:160-171.
19. Hotelling H. Stability in competition. *Econ J*. 1929;39:41-57.
20. Guo L, Modarres R. Interpoint distance classification of high dimensional discrete observations. *Int Stat Rev*. 2018;87(2):191-206. https://doi.org/10.1111/insr.12281.
21. Guo L, Modarres R. Testing the equality of matrix distributions. *JISS*. 2019b. https://doi.org/10.1007/s10260-019-00477-7.
22. Silverman B, Brown T. Short distances, at triangles and Poissonlimits. *J Appl Probab*. 1978;15(4):816-826.
23. Dutta S, Ghosh AK. On some transformations of high dimension, low sample size data for nearest neighbor classification. *Mach Learn*. 2016;102(1):57-83.
24. Liao SM, Akritas M. Test-based classication: A linkage between classication and statistical testing. *Statistics and Probability Letters*. 2007;77(12):1269-1281.
25. Chen H, Zhang N. Graph-based change-point detection. *Ann Stat*. 2015;43(1):139-176.
26. Patil GP, Modarres R. Hotspot detection with bivariate data. *J Stat Plan Infer*. 2007;137(11):3643-3654.
27. Atkinson EN, Brown BW, Thompson, JR. Parallel algorithms for fixed seed simulation based parameter estimation. Paper presented at: Computer Science and Statistics. Proceedings of the 21st Symposium on the Interface; 1989, 256-261.
28. Lance GN, Williams WT. A general theory of classificatory sorting strategies: 1. Hierarchical systems. *Comput J*. 1967;9(4):373-380.
29. Francois D, Wertz V, Verleysen M. The concentration of fractional distances. *IEEE Trans Knowl Data Eng*. 2007;19:873-886.
30. Biau, G, Mason, DM. High-dimensional p-Norms. 1967. arXiv:1311.0587v1 [math.ST].
31. Hall P, Marron JS, Neeman A. Geometric representation of high dimension, low sample size data. *J Royal Stat Soc Ser B*. 2005;67(3):427-444.
32. Angiulli F. On the behavior of intrinsically high-dimensional spaces: distances, direct and reverse nearest neighbors and hubness. *J Mach Learn Res*. 2018;18:1-60.
33. Li J. Asymptotic normality of interpoint distances for high-dimensional data with applications to the two-sample problem. *Biometrika*. 2019;105(31):529-546.
34. Shurygin AM. Using interpoint distances for pattern recognition. *Pattern Recog Image Anal*. 2006;16(4):726-729.
35. Cayley A. A theorem in the geometry of position. *Cambridge Math J*. 1841;II:267-271.
36. Menger K. New foundation of Euclidean geometry. *Am J Math*. 1931;53:721-745.
37. Mardia KV, Kent JT, Bibby JM. *Multivariate Analysis*. London, UK: Academic Press; 1979.
38. Mathai AM, Provost SB. *Quadratic Forms in Random Variables*. New York, NY: Marcel Derkker, INC.; 1992.
39. Hamming RW. Error detecting and error correcting codes. *Bell Syst Tech J*. 1950;29(2):147-160.
40. Song Y, Modarres R. Multivariate power series interpoint distances. 2019b. Submitted.
41. Song Y, Modarres R. Unified multivariate hypergeometric interpoint distances. *Statistics*. 2019c;53(4):921-942. https://doi.org/10.1080/02331888.2019.1618857.
42. Noack A. A class of random variables with discrete distributions. *Ann Math Stat*. 1950;21(1):127-132.
43. Sibuya M, Yoshimura I, Shimizu R. Negative multinomial distribution. *Ann Inst Stat Math*. 1964;16(1):409-426.
44. Patil GP. On sampling with replacement from populations with multiple characters. *Indian J Stat Ser B*. 1968;30(3/4):355-366.
45. Joshi SW, Patil GP. Certain structural properties of the sum-symmetric power series distributions. *Indian J Stat Ser A*. 1971;33(2):175-184.
46. Janardan KG. Chance mechanisms for multivariate hypergeometric models. *Indian J Stat Ser A*. 1972;35(4):465-478.
47. Nijenhuis A, Wilf HS. *Combinatorial Algorithms*. 2nd ed. New York, NY: Academic Press; 1978.
48. Modarres R. Graphical Comparison of High Dimensional Distributions. Technical Report. Washington DC, Department of Statistics. George Washington University, 2019.
49. Giovagnoli A, Wynn HP. Multivariate dispersion orderings. *Statist Probab Lett*. 1995;22:325-332.


**How to cite this article:** Modarres R, Song Y. Interpoint distances: Applications, properties, and visualization. *Appl Stochastic Models Bus Ind*. 2020;36:1147–1168. https://doi.org/10.1002/asmb.2508