## Title

Nonparametric tests for homogeneity of species assemblages: a data depth approach.

**Permalink**
https://escholarship.org/uc/item/5v92g48s

**Journal**
Biometrics, 67(4)

**ISSN**
0006-341X

**Authors**
Li, Jun
Ban, Jifei
Santiago, Louis S

**Publication Date**
2011-12-01

**DOI**
10.1111/j.1541-0420.2011.01573.x

Peer reviewed

Biometrics 67, 1481–1488
December 2011

DOI: 10.1111/j.1541-0420.2011.01573.x

# Nonparametric Tests for Homogeneity of Species Assemblages: A Data Depth Approach

**Jun Li,[1,*] Jifei Ban,[1] and Louis S. Santiago[2]**

[1]Department of Statistics, University of California, Riverside, California 92521, U.S.A.
[2]Department of Botany & Plant Sciences, University of California, Riverside, California 92521, U.S.A.
*email: jun.li@ucr.edu

Summary. Testing homogeneity of species assemblages has important applications in ecology. Due to the unique structure of abundance data often collected in ecological studies, most classical statistical tests cannot be applied directly. In this article, we propose two novel nonparametric tests for comparing species assemblages based on the concept of data depth. They can be considered as a natural generalization of the Kolmogorov–Smirnov and the Cramér-von Mises tests (KS and CM) in this species assemblage comparison context. Our simulation studies show that the proposed test is more powerful than other existing methods under various settings. A real example is used to demonstrate how the proposed method is applied to compare species assemblages using plant community data from a highly diverse tropical forest at Barro Colorado Island, Panama.

Key words: Data depth; *DD*-plot; Nonparametric tests; Permutation tests; Species richness.

## 1. Introduction

Testing homogeneity across different species assemblages is important in ecology because it provides crucial information about the spatial and temporal stability of ecosystems. One typical type of data collected in ecological studies is abundance data, which consists of counts of abundances of individual species in each sampling unit. For example, as part of the Barro Colorado Island forest dynamics research project, a study was carried out to investigate spatial differences between two highly diverse tropical forest census plots from Barro Colorado Island, Panama. Each of the two plots, which were 1 hectare in size, was divided into twenty-five 20 m × 20 m quadrats. Counts of each individual species were then recorded in all of the 25 quadrats. Based on those species abundance data, one fundamental ecological question is whether the two species assemblages differ significantly. In this study, a total of 159 tree species was observed in the two plots. Therefore, if we treat the vector of the counts of all 159 tree species in each of the quadrats as an observation in the sample, the data we have consists of two 159-dimensional samples with both sample sizes being 25. Our task is essentially to compare the distributions of the species abundance data from the two plots based on these two samples.

Typically for abundance data, dimensionality, which is equal to the number of species, is often high (in our case it is 159), and zeros are common due to the rarity of some species, making it difficult to find a satisfactory parametric model for such data. Thus, a nonparametric testing procedure is more desirable when comparing species assemblages given abundance data. Furthermore, for abundance data, measures such as Bray–Curtis distance (Bray and Curtis, 1957) are usually preferred to Euclidean distance for describing the dissimilarity between observations (Faith, Minchin, and Belbin, 1987; Clarke, 1993). Therefore, a nonparametric testing procedure that can incorporate such measures would be the most appropriate to carry out the comparison between species assemblages.

In the literature there have been some approaches which can incorporate distance measures into the comparison procedure for multivariate outcomes (e.g., Gower and Krzanowski, 1999; McArdle and Anderson, 2001; Reiss et al., 2010). Most of them are based on so-called "analysis of distance," which partitions the variation inherent in distance matrices, analogous to the well-known multivariate analysis of variance. Similar to multivariate analysis of variance, those approaches were motivated by testing equal means among distributions, and therefore are only sensitive to the location differences among distributions. In practice, the distributions of abundance data from different species assemblages may differ in other characteristics. In this article, we propose two novel nonparametric tests, both of which have the flexibility to incorporate any desired distance measure and are also capable of detecting any distributional differences between species assemblages. More specifically, the two tests are derived based on the concept of data depth. Because the data depth we use is based on any distance measure between observations, it can be directly applied to abundance data and at the same time is capable of incorporating any desired distance measure for abundance data. Based on this distance-based depth, we also employ the so-called two-dimensional *DD*-plot (Liu, Parelius, and Singh, 1999) to visualize the difference between species assemblages. This graphical tool serves as further motivation for our two proposed tests for species assemblage comparisons. The two tests can be considered as the analogues of the classical KS and CM tests in a species assemblage comparison context. The analogue of the CM test is shown to have more power than other existing nonparametric tests for a variety of alternative hypotheses.

The rest of this article is organized as follows. In Section 2, we briefly review the general concept of data depth, and then introduce the special notion of data depth that we use in this article, distance-based depth. In Section 3, we demonstrate the use of *DD*-plot for graphical comparison of two species assemblages. In Section 4, we describe the two proposed nonparametric testing procedures. Simulation studies are carried out to evaluate the performance of the proposed tests in Section 5. In Section 6, we demonstrate the application of the proposed procedures by revisiting the species abundance data from the two tropical forest census plots in Barro Colorado Island, Panama. Finally, we provide concluding remarks in Section 7.

## 2. A Distance-Based Data Depth

A data depth is a measure of how central or how outlying a given point is with respect to a multivariate data cloud or its underlying distribution. The word *depth* was first used by Tukey (1975) for picturing data. Since then, many different notions of data depth have been proposed for capturing different probabilistic features of multivariate data. Among the most popular choices of data depths are Mahalanobis depth (Mahalanobis, 1936; Hu et al., 2009), half-space depth (Hodges, 1955; Tukey, 1975), simplicial depth (Liu, 1990), projection depth (Stahel, 1981; Donoho, 1982; Donoho and Gasko, 1992; Zuo, 2003), etc. More discussion on different notions of data depth can be found in Liu et al. (1999), Zuo and Serfling (2000), and Mizera (2002).

In the last two decades, data depth has provided many new and powerful nonparametric tools for multivariate data (see, e.g., Liu et al., 1999; Li and Liu, 2004, 2008). However, due to the discrete nature of the abundance data and the special distance measure required between the observations, most existing depths in the literature cannot be directly applied to abundance data. This motivates us to explore a distance-based depth, the idea of which was briefly mentioned in Bartoszynski, Pearl, and Lawrence (1997). The definition of the distance-based depth is given below.

Definition (Distance-based depth). *Let* $\mathbf{X} = \{X_1, \ldots, X_n\}$ *be a random sample from* $F$*, where* $F$ *is a distribution of any type. The distance-based depth at* $x$ *w.r.t.* $F$ *is defined as*

$$\begin{aligned} D_F(x) = {} & \Pr\{d(X_1, X_2) > \max[d(X_1, x), d(X_2, x)]\} \\ & + \frac{1}{2}\Pr\{d(X_1, X_2) = d(X_1, x) > d(X_2, x)\} \\ & + \frac{1}{2}\Pr\{d(X_1, X_2) = d(X_2, x) > d(X_1, x)\} \\ & + \frac{1}{3}\Pr\{d(X_1, X_2) = d(X_1, x) = d(X_2, x)\}, \end{aligned}$$

*and the sample version is*

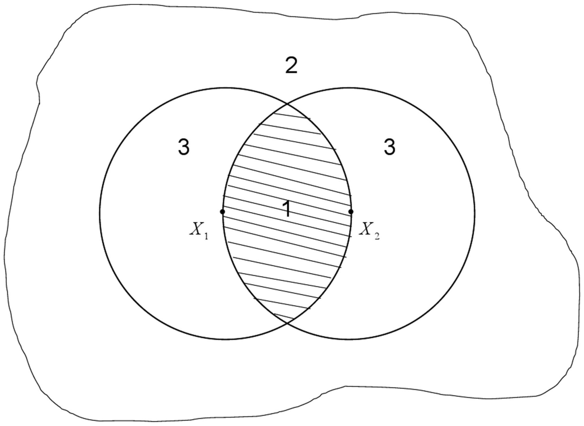

**Figure 1.** $B(X_i, X_j)$ in two-dimensional case.

$$\begin{aligned} D_{F_n}(x) = \frac{1}{\binom{n}{2}} \Bigg( & \sum_{i<j} I\{d(X_i, X_j) > \max[d(X_i, x), d(X_j, x)]\} \\ & + \frac{1}{2}\sum_{i<j} I\{d(X_i, X_j) = d(X_i, x) > d(X_j, x)\} \\ & + \frac{1}{2}\sum_{i<j} I\{d(X_i, X_j) = d(X_j, x) > d(X_i, x)\} \\ & + \frac{1}{3}\sum_{i<j} I\{d(X_i, X_j) = d(X_i, x) = d(X_j, x)\} \Bigg), \end{aligned}$$

*where* $d(x, y)$ *is any suitably chosen distance measure between* $x$ *and* $y$*, and* $I\{A\}$ *is the indicator function which takes 1 if* $A$ *is true and 0 otherwise.*

In the above definition, $\Pr\{d(X_1, X_2) > \max[d(X_1, x), d(X_2, x)]\} (\equiv p_1)$ represents the probability that the side joining $X_1$ and $X_2$ is the longest in a triangle with vertices $X_1$, $X_2$, and $x$. Similarly, we can define

$$p_2 = \Pr\{d(X_1, X_2) < \min[d(X_1, x), d(X_2, x)]\}$$

and

$$\begin{aligned} p_3 = \Pr\{ & \min[d(X_1, x), d(X_2, x)] < d(X_1, X_2) \\ & < \max[d(X_1, x), d(X_2, x)]\}, \end{aligned}$$

which represent the probabilities that the side joining $X_1$ and $X_2$ is the shortest or middle in the triangle with vertices $X_1$, $X_2$, and $x$. If we consider the case in $\Re^2$ and Euclidean distance as the distance measure, given $X_1$ and $X_2$, we can form two circles, each having one of the points as the center and the other on the circle, as shown in Figure 1. The radiuses of both circles are equal to the Euclidean distance between $X_1$ and $X_2$, $d(X_1, X_2)$. We denote region $k$ ($k = 1, 2, 3$) in Figure 1 by $B_k(X_1, X_2)$. Then the probability $p_k$ ($k = 1, 2, 3$) is equivalent to the probability of $x$ falling into $B_k(X_1, X_2)$. Similarly,

$$\begin{aligned} Pr\{d(X_1, x_2) = d(X_1, x) > d(X_2, x)\} + Pr\{d(X_1, X_2) \\ = d(X_2, x) > d(X_1, x)\} \end{aligned}$$

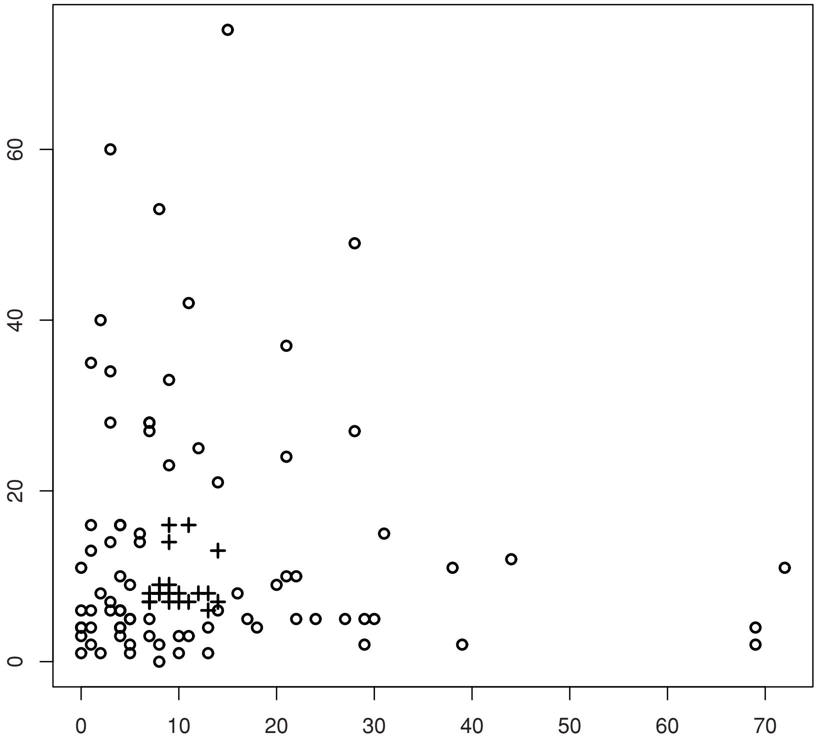

**Figure 2.** A bivariate Poisson-lognormal sample with the 20% deepest points.

calculates the probability of $x$ falling on the boundary between $B_1(X_1, X_2)$ and $B_3(X_1, X_2)$, and

$$Pr\,\{d(X_1, X_2) = d(X_1, x) = d(X_2, x)\}$$

calculates the probability of $x$ falling on the boundary between $B_1(X_1, X_2)$, $B_2(X_1, X_2)$, and $B_3(X_1, X_2)$. Splitting these probabilities evenly among their adjacent regions has led to the fractions 1/2 and 1/3 in the definition of the above distance-based depth. As a result, the distance-based depth $D_F(x)$ can be considered as the probability of $x$ falling into $B_1(X_1, X_2)$ and its boundary.

Given a sample $\mathbf{X} = \{X_1, \ldots, X_n\}$ in $\Re^2$, the sample distance-based depth, $D_{F_n}(x)$, has a similar interpretation and it calculates the proportion of $B_1(X_i, X_j)$ $(i, j = 1, \ldots, n,\ i \neq j)$ and its boundary containing $x$. For any point $x$ in $\Re^2$, if $x$ is near the center of the data cloud, $x$ should be contained in many of $B_1(X_i, X_j)$ and its boundary generated from the sample. On the other hand, if $x$ is relatively near the outskirts, we would expect that $x$ is contained by only a few of $B_1(X_i, X_j)$ and its boundary. In higher dimensions or with other distance measures being used, the value of the above depth has similar interpretations. Therefore, the above notion of depth provides a reasonable measure of "depth" of $x$ w.r.t. the data cloud $\{X_1, \ldots, X_n\}$.

Because any distance measure can be used in the above definition of distance-based depth, it can be directly applied to our species abundance data using any desired distance measures between observations. Based on this distance-based depth, for any given abundance data sample $\{X_1, \ldots, X_n\}$, we can calculate the depth values $D_{F_n}(X_i)$, and then order the $X_i$'s according to their descending depth values. This gives rise to a natural center-outward ordering of the sample points. As an example and for demonstration purposes, we assume that there are only two species in the species assemblage. The counts of the two species from 100 sampling units are generated from a bivariate Poisson-lognormal distribution (Aitchison and Ho, 1989), where the sample is drawn from a bivariate Poisson with mean $(\lambda_1, \lambda_2)$ being random draws from bivariate lognormal distribution. To facilitate the exposition, we denote the general multivariate Poisson-lognormal distribution as $PL(\mu, \Sigma)$, where $\mu$ and $\Sigma$ are the parameters of the multivariate lognormal distribution. In ecology, for this type of data, Euclidean distance is generally not considered appropriate. Instead, measures such as Bray–Curtis distance (Bray and Curtis, 1957) are preferred. The Bray–Curtis distance for sample points $X_l = (X_{l1}, X_{l2}, \ldots, X_{lp})'$ and $X_{l'} = (X_{l'1}, X_{l'2}, \ldots, X_{l'p})'$ is defined as,

$$d_{ll'} = \frac{\sum_{k=1}^{p} |X_{lk} - X_{l'k}|}{\sum_{k=1}^{p} (X_{lk} + X_{l'k})},$$

and $d_{ll'} = 0$ if both $X_l$ and $X_{l'}$ equal $\mathbf{0}_p$, where $\mathbf{0}_p$ is the vector of $p$ zeros. Figure 2 shows the simulated data ordering based on the distance-based depth when Bray–Curtis distance is used. In the plot, "+" marks the deepest 20% of the observations.

## 3. *DD*-plot: A Graphical Comparison of Species Assemblages

In this section, we demonstrate how the so-called *DD*-plot (depth versus depth plot) can be used to provide a graphical tool for comparisons of species assemblages. The *DD*-plot was

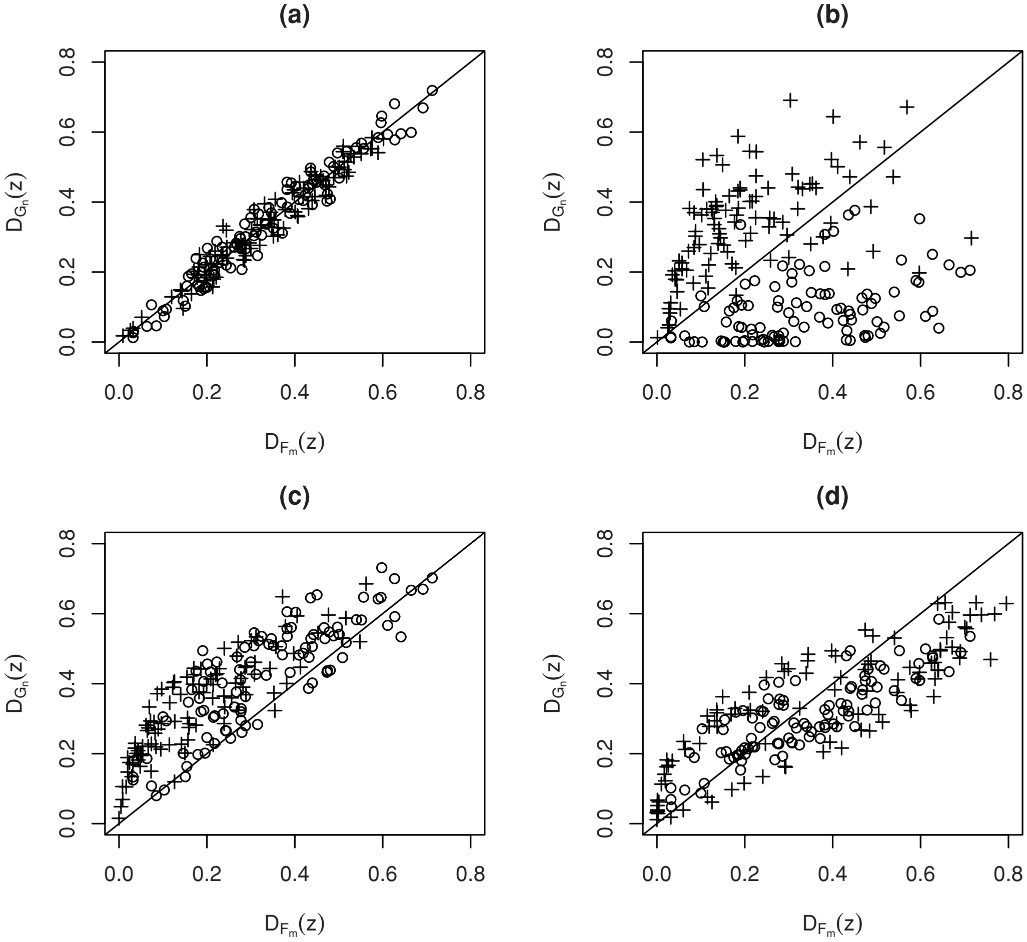

**Figure 3.** *DD*-plots: (a) $F = G = PL(\mathbf{1}_{10}, I_{10})$; (b) $F = PL(\mathbf{1}_{10}, I_{10})$ and $G = PL(2\mathbf{1}_{10}, I_{10})$; (c) $F = PL(\mathbf{1}_{10}, I_{10})$ and $G = PL(\mathbf{1}_{10}, 2I_{10})$; and (d) $F = PL(\mathbf{1}_{10}, I_{10})$ and $G = PL(\mathbf{1}_{10}, 0.8\mathbf{1}_{10}\mathbf{1}_{10}' + 0.2I_{10})$. In all the plots, the circles represent the observations from $F$ and the pluses represent the observations from $G$.

first introduced by Liu et al. (1999) for graphical comparisons of two continuous multivariate distributions. Based on our newly adopted distance-based depth in Section 2, the *DD*-plot can now be directly applied to our species abundance data. Let $\{X_1, \ldots, X_m\}(\equiv \mathbf{X})$ and $\{Y_1, \ldots, Y_n\}(\equiv \mathbf{Y})$ be the abundance data from two species assemblages, respectively. The *DD*-plot is constructed by

$$DD(F_m, G_n) = \{(D_{F_m}(z), D_{G_n}(z)), z \in \mathbf{X} \cup \mathbf{Y}\}, \qquad (1)$$

where $D_{F_m}(z)$ and $D_{G_n}(z)$ are the sample distance-based depths w.r.t. samples $\mathbf{X}$ and $\mathbf{Y}$, respectively.

From the construction of the above *DD*-plot, we can see that if the distributions of the abundance data from the two species assemblages are the same, all the data points in the *DD*-plot should be concentrated along the 1:1 correspondence line as shown in Figure 3a. Here the abundance data $\mathbf{X}$ and $\mathbf{Y}$ from the two species assemblages are generated from the same distribution $PL(\mathbf{1}_{10}, I_{10})$, where $\mathbf{1}_d$ is a vector of $d$ ones, and $I_d$ is the $d$-dimensional identity matrix. If the two species assemblages are different, the *DD*-plot would exhibit a noticeable departure from the 1:1 correspondence line as shown in Figure 3b–d. Here the abundance data $\mathbf{X}$ and $\mathbf{Y}$ from the two species assemblages are generated from two different distributions. More specifically, $\mathbf{X}$ is generated from $PL(\mathbf{1}_{10}, I_{10})$ in all the plots, whereas $\mathbf{Y}$ is generated from $PL(2\mathbf{1}_{10}, I_{10})$, $PL(\mathbf{1}_{10}, 2I_{10})$, and $PL(\mathbf{1}_{10}, 0.8\mathbf{1}_{10}\mathbf{1}_{10}' + 0.2I_{10})$, respectively. To make the difference between the two samples more visible, unlike the *DD*-plot originally used in Liu et al. (1999), where the observations from different samples were not distinguished, we use different symbols to indicate different memberships of the observations in the *DD*-plot. For example, in all the plots in Figure 3, the circles represent the observations from $\mathbf{X}$, and the pluses represent the observations from $\mathbf{Y}$. In all the plots, Bray–Curtis distance is used in calculating the distance-based depths, and $m$ and $n$ are set as 100.

In general, if the distributions of abundance data from the two species assemblages mainly differ in location, the *DD*-plot would have a leaf-shaped figure as the one in Figure 3b, because the deepest point with respect to one sample will not be the deepest point with respect to the other sample and therefore will have relatively smaller depth value with respect to that sample. If the two distributions mainly have different scales, for example, $G$ is more spread out than $F$, then the depth of any point with respect to $G$ would be no less than its depth with respect to $F$. In such a case, the *DD*-plot would have an early-half-moon-shaped figure arching above the diagonal line as the one in Figure 3c. How other distributional differences are associated with particular patterns of deviation from the 1:1 correspondence line in the *DD*-plot can be interpreted in a similar way.

As we can see from the above plots, the *DD*-plot based on the distance-based depth provides a simple diagnostic tool for visual comparison of two species assemblages.

## 4. Tests of Homogeneity of Species Assemblages

We again denote the abundance data from two species assemblages by $\{X_1, \ldots, X_m\}$ and $\{Y_1, \ldots, Y_n\}$. We assume that they are random samples from the underlying distributions $F$ and $G$, respectively. The comparison of the two species assemblages can be formulated as the following hypothesis testing problem,

$$H_0 : F = G \text{ v.s. } H_1 : F \neq G \tag{2}$$

As noted in the previous section, when the two species assemblages are identical, i.e., $F = G$, we would expect all the points in the *DD*-plot clustered along the 1:1 correspondence line. In other words, $D_{F_m}(z)$ and $D_{G_n}(z)$ should be approximately the same for all the observations from the pooled sample $\mathbf{X} \cup \mathbf{Y}$. If there is a difference between the two species assemblages, $D_{F_m}(z)$ and $D_{G_n}(z)$ would be different from each other. Therefore, the difference between $D_{F_m}(z)$ and $D_{G_n}(z)$ from all of the observations can be used as an indicator of heterogeneity of the two species assemblages. Motivated by this observation, we propose the following two test statistics for hypothesis testing problem (2), which can be considered as a natural generalization of KS and CM tests in this species assemblage comparison context:

- KS type test statistic:

$$T_{KS} = \sup_{z \in \mathbf{X} \cup \mathbf{Y}} |D_{F_m}(z) - D_{G_n}(z)| \tag{3}$$

- CM type test statistic:

$$T_{CM} = \sum_{z \in \mathbf{X} \cup \mathbf{Y}} [D_{F_m}(z) - D_{G_n}(z)]^2 \tag{4}$$

Define

$$p_{KS} = P_{H_0}(T_{KS} \geqslant T_{KS}^{\text{obs}}), \text{ and } p_{CM} = P_{H_0}(T_{CM} \geqslant T_{CM}^{\text{obs}}),$$

where $T_{KS}^{\text{obs}}$ and $T_{CM}^{\text{obs}}$ are the observed values of $T_{KS}$ and $T_{CM}$, respectively, based on the given sample $\mathbf{X} \cup \mathbf{Y}$. Then $p_{KS}$ and $p_{CM}$ are the $p$-values of the proposed two tests. To determine their values directly from the null distributions of $T_{KS}$ and $T_{CM}$ is not trivial. Instead, we proceed and use the permutation method to approximate $p_{KS}$ and $p_{CM}$. More specifically, we randomly permute the pooled sample $\mathbf{X} \cup \mathbf{Y}$ $B$ times. Here $B$ is sufficiently large. For each permutation, we treat the first $m$ elements as the $X$-sample and the remaining elements as the $Y$-sample. We denote the outcome of the $i$th permutation by $\mathbf{X}_i^* = \{X_{i1}^*, \ldots, X_{in}^*\}$, and $\mathbf{Y}_i^* = \{Y_{i1}^*, \ldots, Y_{in}^*\}$, for $i = 1, \ldots, B$. For each $\mathbf{X}_i^* \cup \mathbf{Y}_i^*$, we evaluate the corresponding $T_{KS}$ and $T_{CM}$ values (following (3) and (4)), denoted, respectively, by $T_{i,KS}^*$ and $T_{i,CM}^*$, $i = 1, \ldots, B$. Then $p_{KS}$ and $p_{CM}$ can be approximated, respectively, by

$$\hat{p}_{KS} = \frac{1 + \sum_{i=1}^{B} I\left\{T_{i,KS}^* \geqslant T_{KS}^{\text{obs}}\right\}}{B + 1},$$

and

$$\hat{p}_{CM} = \frac{1 + \sum_{i=1}^{B} I\left\{T_{i,CM}^* \geqslant T_{CM}^{\text{obs}}\right\}}{B + 1},$$

(see, e.g., Fay, Kim, and Hachey, 2007). In the following, we refer to our permutation tests based on $T_{KS}$ and $T_{CM}$ as a depth-based KS test and a depth-based CM test, respectively.

## 5. Simulation Study

In this section, we conduct several simulation studies to evaluate the performance of our proposed two tests. In particular, we compare our tests with two tests available in the literature, which can also be applied to the species assemblage comparison context.

The first one is the test proposed by Nettleton and Banerjee (2001) (NB hereafter), which applied the testing procedure of Friedman and Rafsky (1979) to compare distributions of random vectors with categorical components. Let $\mathbf{Z} = \{Z_1, \ldots, Z_{m+n}\}$ denote the pooled sample $\mathbf{X} \cup \mathbf{Y}$. The NB test statistic is defined as

$$T_{NB} = \sum_{i=1}^{m+n} I\{\text{the nearest neighbor of } Z_i \text{ belongs to different sample}\},$$

where the nearest neighbor of $Z_i$ is the one which minimizes $\delta(Z_i, Z_k)$, $k = 1, \ldots, i-1, i+1, \ldots, m+n$, and $\delta(\cdot, \cdot)$ is any distance measure which is appropriate for the application. The test rejects $H_0$: $F = G$ if $T_{NB}$ is too small.

The second test we will consider was proposed by Hall and Tajvidi (2002) (HT hereafter). Again we consider the pooled sample $\mathbf{Z}$. We define $M_i(j)$ as the number of observations being from sample $\mathbf{Y}$ in the neighborhood of $X_i$, where the neighborhood is bounded by a circle with center at $X_i$ and radius as the distance between $X_i$ and its $j$th nearest neighbor. Similarly, we define $N_i(j)$ as the number of observations being from sample $\mathbf{X}$ in the neighborhood of $Y_i$, where the neighborhood is bounded by a circle with center at $Y_i$ and radius as the distance between $Y_i$ and its $j$th nearest neighbor. Under $H_0$, it can be shown that

$$E_0(M_i(j)) = \frac{nj}{m+n-1} \text{ and } E_0(N_i(j)) = \frac{mj}{m+n-1}.$$

Define the deviations of $M$ and $N$ from their expected values under $H_0$ as

$$\begin{aligned} DM_i(j) &= \left|M_i(j) - \frac{nj}{m+n-1}\right| \\ \text{and } DN_i(j) &= \left|N_i(j) - \frac{mj}{m+n-1}\right|. \end{aligned}$$

The HT test statistic is then defined as

$$T_{HT} = \frac{1}{m} \sum_{i=1}^{m} \sum_{j=1}^{n} DM_i(j)^\gamma w_1(j) + \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{m} DN_i(j)^\gamma w_2(j),$$

where $w_1(j)$ and $w_2(j)$ denote nonnegative weights and $\gamma$ is some positive value. Like the NB test, the HT test can be based on any distance measure. The test rejects $H_0$: $F = G$ if $T_{HT}$ is too large. Based on the simulation studies reported

**Table 1**

*Simulated power for different tests using the samples from $F = PL(\boldsymbol{\mu}_F, \Sigma_F)$ and $G = PL(\boldsymbol{\mu}_G, \Sigma_G)$ where $\boldsymbol{\mu}_F = \mathbf{1}_{10}$, $\boldsymbol{\mu}_G = \mu\boldsymbol{\mu}_F$, and $\Sigma_F = \Sigma_G = 0.5\mathbf{1}_{10}\mathbf{1}'_{10} + 0.5I_{10}$*

| | $T_{NB}$ | $T_{HT}$ | $T_{KS}$ | $T_{CM}$ |
|---|---|---|---|---|
| $\mu = 1$ | 0.037 | 0.057 | 0.045 | 0.039 |
| $\mu = 1.1$ | 0.035 | 0.053 | 0.073 | 0.069 |
| $\mu = 1.2$ | 0.048 | 0.117 | 0.111 | 0.117 |
| $\mu = 1.3$ | 0.064 | 0.176 | 0.224 | 0.245 |
| $\mu = 1.4$ | 0.102 | 0.319 | 0.380 | 0.422 |
| $\mu = 1.5$ | 0.128 | 0.454 | 0.537 | 0.592 |
| $\mu = 1.6$ | 0.176 | 0.613 | 0.696 | 0.769 |
| $\mu = 1.7$ | 0.243 | 0.713 | 0.820 | 0.860 |
| $\mu = 1.8$ | 0.346 | 0.838 | 0.919 | 0.946 |
| $\mu = 1.9$ | 0.435 | 0.929 | 0.973 | 0.985 |
| $\mu = 2$ | 0.549 | 0.960 | 0.987 | 0.995 |

**Table 2**

*Simulated power for different tests using the samples from $F = PL(\boldsymbol{\mu}_F, \Sigma_F)$ and $G = PL(\boldsymbol{\mu}_G, \Sigma_G)$ where $\boldsymbol{\mu}_F = \boldsymbol{\mu}_G = \mathbf{1}_{10}$, $\Sigma_F = 0.5\mathbf{1}_{10}\mathbf{1}'_{10} + 0.5I_{10}$, and $\Sigma_G = \sigma\Sigma_F$*

| | $T_{NB}$ | $T_{HT}$ | $T_{KS}$ | $T_{CM}$ |
|---|---|---|---|---|
| $\sigma = 1.1$ | 0.038 | 0.060 | 0.051 | 0.050 |
| $\sigma = 1.2$ | 0.060 | 0.079 | 0.068 | 0.078 |
| $\sigma = 1.3$ | 0.046 | 0.137 | 0.129 | 0.148 |
| $\sigma = 1.4$ | 0.063 | 0.196 | 0.163 | 0.198 |
| $\sigma = 1.5$ | 0.072 | 0.271 | 0.220 | 0.303 |
| $\sigma = 1.6$ | 0.087 | 0.375 | 0.310 | 0.404 |
| $\sigma = 1.7$ | 0.079 | 0.444 | 0.371 | 0.487 |
| $\sigma = 1.8$ | 0.105 | 0.544 | 0.427 | 0.569 |
| $\sigma = 1.9$ | 0.114 | 0.633 | 0.527 | 0.687 |
| $\sigma = 2$ | 0.131 | 0.719 | 0.632 | 0.748 |

in HT, several different choices of weight functions and $\gamma$ values do not have significant effects on the power of the test. Therefore, in our simulation study, we set $\gamma = 1$ and $w_1(j) = w_2(j) = 1$.

To compare our proposed tests with the NB and HT tests in various settings, we first generated $m = n = 30$ random observations from $F = PL(\boldsymbol{\mu}_F, \Sigma_F)$ and $G = PL(\boldsymbol{\mu}_G, \Sigma_G)$, where $\boldsymbol{\mu}_F = \mathbf{1}_{10}$, $\Sigma_F = 0.5\mathbf{1}_{10}\mathbf{1}'_{10} + 0.5I_{10}$, $\boldsymbol{\mu}_G = \mu\boldsymbol{\mu}_F$ and $\Sigma_G = \sigma\Sigma_F$ with $\mu$ and $\sigma$ being manipulated according to different settings. All of the tests were then carried out through permutation. The number of permutations was set to be 999. The significance level was set at 0.05. Again, we chose Bray–Curtis distance as the distance measure in all of the tests. Table 1 shows the simulated power for the four tests under different choices of $\mu$ with $\sigma$ being fixed at 1, i.e., $\Sigma_F = \Sigma_G$. Table 2 shows the simulated power for different choices of $\sigma$ with $\mu$ being fixed at 1, i.e., $\boldsymbol{\mu}_F = \boldsymbol{\mu}_G$. The results were based on 1000 simulations. As we can see from the tables, our depth-based CM test outperforms the other three tests in both settings. When $\Sigma_F = \Sigma_G$, our depth-based KS test is ranked as the second, outperforming both NB and HT tests. When $\boldsymbol{\mu}_F = \boldsymbol{\mu}_G$, the KS test is slightly worse than the HT test. In both settings, the NB test has the lowest power.

Our second simulation study is to investigate the powers of the four tests for comparing samples from different distribution families. Recall the Poisson-Lognormal distribution is essentially a Poisson-Lognormal mixture. Similarly, we can also consider Poisson-gamma mixture and Poisson–Weibull mixture. We refer to those mixtures as Poisson-gamma distribution and Poisson–Weibull distribution. For simplicity, we choose the mixing distribution in the Poisson-gamma (Poisson–Weibull) as a multivariate distribution with independent gamma (Weibull) distributed marginals. Therefore, we can denote the Poisson-gamma and Poisson–Weibull distributions by $PG(\mathbf{a}, \boldsymbol{\theta})$ and $PW(\mathbf{b}, \boldsymbol{\lambda})$, respectively, where $(\mathbf{a}, \boldsymbol{\theta})$ and $(\mathbf{b}, \boldsymbol{\lambda})$ are the shape and scale parameter vectors for the gamma and Weibull marginals, respectively. In the simulation, we chose $F$ as $PL(\mathbf{1}_{10}, I_{10})$, and $G$ as $PG(0.582\mathbf{1}_{10}, 7.701\mathbf{1}_{10})$ or $PW(0.772\mathbf{1}_{10}, 3.851\mathbf{1}_{10})$. The shape and scale parameters in the Poisson-gamma and Poisson–Weibull distribution were chosen to make them have the same componentwise mean and variance as those in $PL(\mathbf{1}_{10}, I_{10})$. Table 3 shows the power of the four tests when comparing the samples from different distribution families. Again, our depth-based CM test is the best among the four.

**Table 3**

*Simulated powers for comparing samples from different distribution families*

| | $T_{NB}$ | $T_{HT}$ | $T_{KS}$ | $T_{CM}$ |
|---|---|---|---|---|
| $G = PG(0.582\mathbf{1}_{10}, 7.701\mathbf{1}_{10})$ | 0.097 | 0.611 | 0.475 | 0.748 |
| $G = PW(0.772\mathbf{1}_{10}, 3.851\mathbf{1}_{10})$ | 0.062 | 0.476 | 0.345 | 0.561 |

We also carry out some additional simulation studies to evaluate the performance of the tests when the underlying multivariate distributions are continuous and Euclidean distance is used in our distance-based depth. Please see Web Tables 1–6 for results of those studies. Similar to what we observe from Tables 1–3, the depth-based CM test performs best among all the tests and outperforms the depth-based KS test in most of the cases. This may be explained by the fact that the depth-based CM test considers all of the differences of depth from the data points, whereas the depth-based KS test only considers the maximal difference of depth among the data points.

## 6. Real Application

In this section, we revisit the species abundance data from the two tropical forest census plots from Barro Colorado Island, Panama, briefly described in the "Introduction." The two highly diverse plots were located within 1 km of each other and represent 100- to 400-year-old lowland tropical forest. In both plots, species identity was determined and location within the plot was recorded for all woody stems ⩾10 mm diameter at 1.5 m height (Condit, 1998; Hubbell et al., 1999; Hubbell, Condit, and Foster, 2005).

As mentioned in the "Introduction," our task is to compare the two species assemblages based on the two abundance data samples, which are 159-dimensional and both have 25 observations each. Before we apply our tests to the data, we first use the *DD*-plot described in Section 3 to visualize the difference of these two species assemblages. Figure 4 shows the corresponding *DD*-plot based on the distance-based depth by using Bray–Curtis distance. In the plot, the circles represent

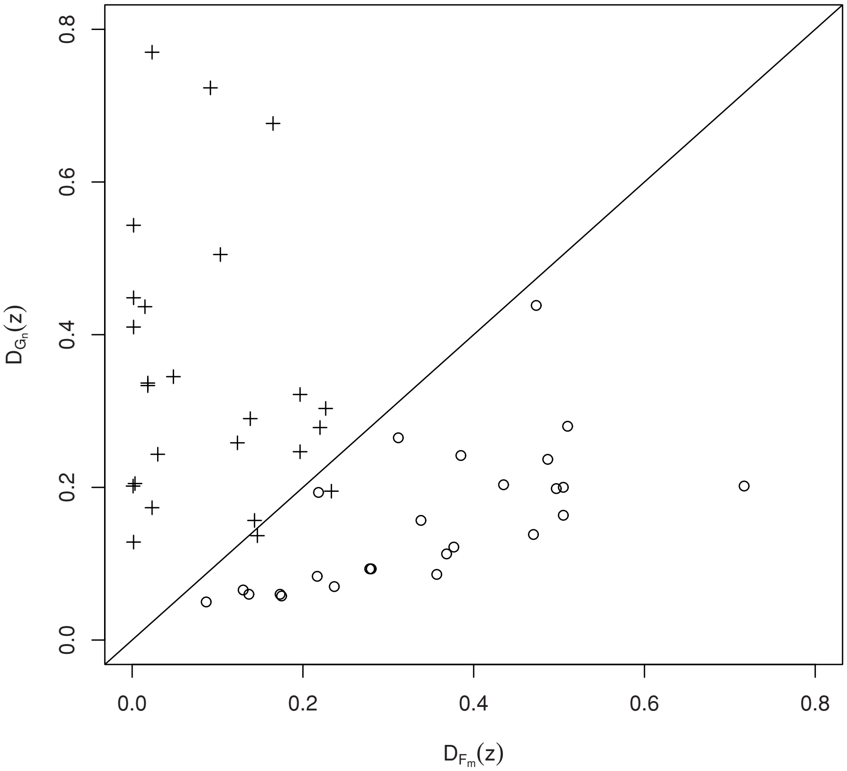

**Figure 4.** *DD*-plot for the samples from the two tropical forest census plots on Barro Colorado Island, Panama.

the observations from one census plot and the pluses represent those from the other. From the plot, it clearly suggests that there is a location difference between the distributions of the species abundance data in these two census plots. Both of our depth-based KS and CM tests yield $p$-values 0.001, which further confirms that the distributions of these two plots are indeed different.

## 7. Concluding Remarks

In this article, we present a data depth approach to the problem of comparing species assemblages given abundance data. It is completely nonparametric and does not require any knowledge of the underlying distribution. Results from simulation studies have shown that our depth-based CM test performs very well and has better power than other alternatives under different settings. Furthermore, the use of the *DD*-plot that motivated our tests also provides an easy graphical tool for visualizing the difference of species assemblages.

Although the proposed tests were motivated by the species assemblages comparison problem in ecology and were demonstrated mostly by examples of count data, they are very flexible and can be easily applied to other applications with different data types. For example, this approach could be applied to comparing samples of functional data, or samples of image data, because properly defined distance measures are usually available for these types of data and distance-based depth, which is capable of incorporating any desired distance measure, makes our approach applicable for a wide range of applications. It is worth pointing out that our proposed depth-based KS and CM tests can be also paired with any other data depths which are suitable for the particular application. For example, to compare samples of functional data, we may base our KS or CM tests on the depth proposed by Lopez-Pintado and Romo (2009) for functional data.

## 8. Supplementary Materials

Web Tables referenced in Section 5 are available under the Paper Information link at the *Biometrics* website `http://www.biometrics.tibs.org`.

## Acknowledgements

The Barro Colorado Island forest dynamics research project was made possible by National Science Foundation grants to Stephen P. Hubbell, support from the Center for Tropical Forest Science, the Smithsonian Tropical Research Institute, the John D. and Catherine T. MacArthur Foundation, the Mellon Foundation, the Celera Foundation, and numerous private individuals, and through the hard work of over 100 people from 10 countries over the past two decades. The plot project is part of the Center for Tropical Forest Science, a global network of large-scale demographic tree plots. We thank the editor, the associate editor, and two referees for their thoughtful comments, which have helped improve our article. JL thanks NSF for supporting her research through Grant DMS-0907655.

## References

Aitchison, J. and Ho, C. H. (1989). The multivariate Poisson-log normal distribution. *Biometrika* **76,** 643–653.

Bartoszynski, R., Pearl, D. K., and Lawrence, J. (1997). A multidimensional goodness-of-fit test based on interpoint distances. *Journal of the American Statistical Association* **92,** 577–586.

Bray, J. R. and Curtis, J. T. (1957). An ordination of the upland forest communities of southern Wisconsin. *Ecological Monographs* **27,** 325–349.

Clarke, K. R. (1993). Nonparametric multivariate analyses of changes in community structure. *Australian Journal of Ecology* **18,** 117–143.

Condit, R. (1998). *Tropical Forest Census Plots*. New York: Springer-Verlag.

Donoho, D. L. (1982). *Breakdown properties of multivariate location estimators*. Ph.D. Thesis, Harvard University.

Donoho, D. L. and Gasko, M. (1992). Breakdown properties of location estimates based on half-space depth and projected outlyingness. *Annals of Statistics* **20,** 1803–1827.

Faith, D. P., Minchin, P. R., and Belbin, L. (1987). Compositional dissimilarity as a robust measure of ecological distance. *Vegetatio* **69,** 57–68.

Fay, M. P., Kim, H. J., and Hachey, M. (2007). On using truncated sequential probability ratio test boundaries for Monte Carlo implementation of hypothesis tests. *Journal of Computational and Graphical Statistics* **16,** 946–967.

Friedman, J. H. and Rafsky, L. C. (1979). Multivariate generalizations of the Wald-Wolfowitz and Smirnov two-sample tests. *Annals of Statistics* **7,** 697–717.

Gower, J. C. and Krzanowski, W. J. (1999). Analysis of distance for structured multivariate data and extensions to multivariate analysis of variance. *Applied Statistics* **48,** 505–519.

Hall, P. and Tajvidi, N. (2002). Permutation tests for equality of distributions in high-dimensional settings. *Biometrika* **89,** 359–374.

Hodges, J. (1955). A bivariate sign test. *The Annals of Mathematical Statistics* **26,** 523–527.

Hu, Y., Wang, Y., Wu, Y., Li, Q., and Hou, C. (2009). Generalized Mahalanobis depth in the reproducing kernel Hilbert space. To appear in *Statistical Papers*. DOI: 10.1007/s00362-009-0265-1.

Hubbell, S. P., Foster, R. B., O'Brien, S. T., Harms, K. E., Condit, R., Wechsler, B., Wright, S. J., and Loo de Lao, S. (1999). Light gap disturbances, recruitment limitation, and tree diversity in a neotropical forest. *Science* **283,** 554–557.

Hubbell, S. P., Condit, R., and Foster, R. B. (2005). Barro Colorado Forest Census Plot Data. `http://ctfs.arnarb.harvard.edu/webatlas/datasets/bci`, last accessed February 2010.

Li, J. and Liu, R. Y. (2004). New nonparametric tests of multivariate locations and scales using data depth. *Statistical Science* **19,** 686–696.

Li, J. and Liu, R. Y. (2008). Multivariate spacings based on data depth: I. Construction of nonparametric multivariate tolerance regions. *Annals of Statistics* **36,** 1299–1323.

Liu, R. Y. (1990). On a notion of data depth based on random simplices. *Annals of Statistics* **18,** 405–414.

Liu, R. Y., Parelius, J. M., and Singh, K. (1999). Multivariate analysis by data depth: Descriptive statistics, graphics and inference. *Annals of Statistics* **27,** 783–840.

Lopez-Pintado, S. and Romo, J. (2009). On the concept of depth for functional data. *Journal of the American Statistical Association* **104,** 718–734.

Mahalanobis, P. (1936). On the generalized distance in statistics. *Proceedings of the National Academy India* **12,** 49–55.

McArdle, B. H. and Anderson, M. J. (2001). Fitting multivariate models to community data: A comment on distance-based redundancy analysis. *Ecology* **82,** 290–297.

Mizera, I. (2002). On depth and deep points: A calculus. *Annals of Statistics* **30,** 1681–1736.

Nettleton, D. and Banerjee, T. (2001). Testing the equality of distributions of random vectors with categorical components. *Computational Statistics and Data Analysis* **37,** 195–208.

Reiss, P. T., Stevens, M. H. H., Shehzad, Z., Petkova, E., and Milham, M. P. (2010). On distance-based permutation tests for between-group comparisons. *Biometrics* **66,** 636–643.

Stahel, W. (1981). Robust Schaetzungen: Infinitesmale Optimalitaet und Schaetzungen von Kovarianzmatrizen (Robust estimation: Infinitesimal optimality and covariance matrix estimators). Ph.D. Thesis, ETH Zurich.

Tukey, J. (1975). Mathematics and picturing data. *Proceedings of the 1975 International Congress of Mathematics* **2,** 523–531.

Zuo, Y. J. (2003). Projection-based depth functions and associated medians. *Annals of Statistics* **31,** 1460–1490.

Zuo, Y. J. and Serfling, R. (2000). General notions of statistical depth function. *Annals of Statistics* **28,** 461–482.

*Received June* 2010. *Revised November* 2010.
*Accepted December* 2010.