UNIVERSITY OF CALIFORNIA
RIVERSIDE

Comparison of Species Assemblages Using Date Depth and Mixture Model

A Dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Applied Statistics

by

Jifei Ban

September 2012

Dissertation Committee:

Dr. Jun Li, Chairperson
Dr. Thomas Girke
Dr. Daniel R. Jeske

The Dissertation of Jifei Ban is approved:

Committee Chairperson

University of California, Riverside

Acknowledgments

I would like to express my sincere gratitude to my thesis advisor, Professor Jun Li, who has spent much of her valuable time guiding me on my thesis research. She has shared with me numerous valuable ideas and insights, and provided me with many critical advices. From her, I have learnt what a real scholar should be like. She has also been extremely understanding and supportive. She has always been patient with me and showed great tolerance of my occasional naiveness. I am very fortunate to have her as my advisor.

I would also like to express my appreciation to Professor Thomas Girke and Professor Daniel R. Jeske for serving on my oral exam committee and dissertation committee. Professor Girke taught me Advanced Genomics&Bioinformatics and introduced me to Biocluster, which has proved to be very useful to my research. Professor Jeske has influenced me positively by his work ethic, passion, and dedication to statistics, his career, and his department.

My special thanks go to Professor Richard Arnott, who has generously shared his time, wisdom, and experience with me and given me great guidance and encouragement in life and study.

I am also grateful to every faculty and staff member in the Department of Statistics at University of California-Riverside. It has been an enjoyable experience studying and doing research here for five years.

At last, I would like to thank my family and my friends for their unconditional support whenever I need them. Without them, I would not have been here today.

The text appearing in Chapter 2 of this thesis, in part or in full, is a reprint of the material as is appears in *Biometrics* (2011), 67(4):1481–1488. The co-author Jun

Li listed in that publication directed and supervised the research which forms the basis for this thesis. The co-author Louis S. Santiago listed in that publication contributed empirical data. 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.

The text appearing in Chapter 3 and Appendix A, in part or in full, is a reprint of the material submitted as a working paper for publication to *Statistica Sinica*. The co-author Jun Li listed in that working paper directed and supervised the research which forms the basis for this thesis. We thank Professor Robert K. Colwell and Robin L. Chazdon for generously sharing the Bosques Project data.

To my parents

ABSTRACT OF THE DISSERTATION

Comparison of Species Assemblages Using Date Depth and Mixture Model

by

Jifei Ban

Doctor of Philosophy, Graduate Program in Applied Statistics
University of California, Riverside, September 2012
Dr. Jun Li, Chairperson

Comparing species assemblages at different times and locations provides useful information regarding ecosystems. Due to the unique structure of abundance data often collected in species assemblages, appropriate statistical tests are needed. In this thesis, we propose two types of tests, one is data depth based nonparametric test, the other is zero-inflated Poisson mixture model based test. These two types of tests are developed for different testing objectives and under different assumptions. We use simulation to demonstrate that their performance is better than that of some existing tests. We also discuss the differences between these two tests.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

A species assemblage refers to the species that exist in a particular habitat, e.g., tree species in a tropical forest in Panama. Recently there have been many studies on comparison of species assemblages. Comparing species assemblages will not only provide scientists information about how ecosystems vary in both temporal and spatial manners, but also give them insights into how ecosystems are responding to environmental factors (e.g., Chao *et al.*, 2006; Wells *et al.*, 2007). For example, Warwick *et al.* (1990) analyzed the Indonesian coral assemblages in 1981-1988 to study the effect of the El Niño event occurred in 1982-1983. Based on their findings, Warwick and Clarke (1993) concluded that increased variability in coral species assemblages may be a sign of increased environmental stress. By comparing dung beetle species assemblages between protected areas and adjacent pasturelands in a mediterranean savanna landscape, Numa *et al.* (2012) concluded that management activities such as plowing and the use of veterinary substances affect soil structure and dung quality and could be important factors that alter dung beetle assemblages on traditional farms, which play an important role in decomposition, seed dispersal, and control of vertebrate parasites.

The data we study in this thesis is called abundance data. It consists of counts of individual species in each sampling unit. The sampling unit we often encounter is called quadrat. A quadrat is a square used in ecology to isolate sample, which is suitable for sampling species. Here is an example of abundance data. As part of the Barro Colorado Island (BCI) 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 25 20m × 20m quadrats. Counts of each individual species were then recorded in all of the 25 quadrats. A total of 159 tree species were observed in the two plots. Furthermore, we can present the collected data as a matrix. With the columns representing different species observed and the rows representing different quadrats, the element at the j-th row and the k-th column of the matrix will just stand for the counts of the j-th species observed in the k-th quadrat. In the above example, the abundance data was collected from multiple quadrats, therefore we refer to this type of data as abundance data from multiple quadrats. In this thesis, we focus on the study of abundance data from multiple quadrats.

Based on those species abundance data, one fundamental ecological question is whether the two joint distributions of the counts of all the observed species from two species assemblages differ significantly. That is essentially a problem of testing homogeneity of two multivariate distributions. Typically for abundance data, dimensionality, which is equal to the number of species, is often high (e.g., for the BCI data mentioned previously 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. Gower and Krzanowski (1999), McArdle and Anderson (2001), and Reiss *et al.* (2010) proposed nonparametric distance-based tests for detection of location difference of the distributions of abundance data from species assemblages. In practice, the distributions of abundance data from different species assemblages may differ in other characteristics. Therefore we propose more general tests to test whether the underlying distributions of two species assemblages are anyhow different. To solve this problem, we take advantage of data depth, transform the high dimensional data into one dimensional depth values, and then develop two tests: one is an analogue of the Kolmogorov-Smirnov test, and the other is an analogue of the Cramér von-Mises test. We use permutation to decide the rejection regions. Power comparison is conducted between our proposed tests and some existing tests. We also discuss how to employ the DD-plot (Liu, Parelius, and Singh, 1999) to visualize the difference between species assemblages. Additionally, we mention that this data depth based method is not limited to the application to discrete count data.

To demonstrate the real application of our data depth based tests, we asked for empirical data from our collaborator, Professor Louis S. Santiago from the Department of Botany and Plant Sciences at University of California-Riverside. At the beginning, Professor Santiago provided us with abundance data from two geographically distant tree species assemblages, which have few species in common and are not good for the purpose of demonstration of the data depth based tests. However, from Professor Santiago, we know that, instead of tracking changes of abundance of individual species, sometimes ecologists are more interested in comparing species diversity of those species assemblages in order to understand how species organize themselves into local communities. For example, conventionally people compare species assemblages to study the variations in temporal or spatial patterns of species diversity by visually looking at species accumulation curves (e.g., Pipan and Culver, 2007), i.e., plots of expected number of observed species versus sampling effort, and dominance-diversity curves (e.g., McGill et al., 2007), i.e., plots of abundance of species versus species rank in abundance. But the results are inevitably subjective. Therefore a rigorous statistical test for comparing species diversity across species assemblages is needed. This leads to the second part of our research, the mixture model based test.

In the literature, mixture models are popular choices to model the above ecological data due to their capabilities to account for heterogeneity among species (e.g., Ord and Whitmore, 1986; Bunge and Fitzpatrick, 1993; Chao and Bunge, 2002; Böhning and Schön, 2005; Mao and Colwell, 2005; Mao, 2006). Additionally, using mixture model does not require the information of species identity, which makes it a great choice for the testing objective of comparison of species diversity across species assemblages with few species in common. In this thesis, we propose a zero-inflated Poisson mixture model to compare species diversity across species assemblages. We assume whether a species appears or not in a sampling quadrat follows a Bernoulli distribution, and if a species does appear in a sampling quadrat, the count of the individual species in that sampling quadrat follows a zero-truncated Poisson distribution, and the parameters of the two distributions are also random variables following unknown distributions. It can be seen later that this model is able to handle zero-inflated data, which is very common in the study of species assemblages since most species are rarely observed. We develop a test statistic from the model and prove it asymptotically follows a χ^2 distribution. We also propose an adjustment using eigenvalue decomposition of a covariance matrix to overcome the difficulty of numerically calculating the inverse matrix. Furthermore, a bootstrap testing procedure is proposed to best approximate the distribution of our test statistic.

The rest of this thesis is organized as follows. In Chapter 2, we first briefly introduce the concept of data depth. Then we introduce two data depth based nonparametric tests for comparison of species assemblages, and compare the power of the two tests to that of other existing tests. Finally we look at the real data application. In Chapter 3, we first introduce the zero-inflated Poisson mixture model, and then develop statistical tests from it for comparison of species assemblages. We perform simulation studies to evaluate our tests, and finally we will also see some real examples. In Chapter 4, we will discuss the differences between the two types of tests we proposed. We will discuss some possible future research work in Chapter 5.

Chapter 2

Data depth based tests

2.1 Introduction

In this chapter, we will develop two data depth based statistical tests for testing homogeneity of multivariate distributions and apply them on the problem of comparison of species assemblages.

Testing whether the two joint distributions of the counts of all the observed species from two species assemblages differ significantly is essentially the problem of testing homogeneity of two multivariate distributions. Let us take the BCI data mentioned in Chapter 1 as an example. 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. The 159 species are uniquely labeled so that there is one to one correspondence of species between the two species assemblages. The testing problem is actually to compare the multivariate distribution represented by the first group of 25 vectors with the multivariate distribution represented by the second group of 25 vectors.

As we mentioned in Chapter 1, since dimensionality is high and zeroes are common for abundance data, a nonparametric testing procedure is often adopted for such problem. 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 which 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 chapter, 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 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 Kolmogorov-Smirnov and Cramér-von Mises tests in a species assemblage comparison context. The analogue of the Cramér-von Mises test is shown to have more power than other existing nonparametric tests for a variety of alternative hypotheses.

The rest of this chapter is organized as follows. In Section 2.2, we briefly review the general concept of data depth, and then introduce the special notion of data depth we use in this thesis, distance-based depth, in Section 2.3. In Section 2.4, we demonstrate the use of *DD*-plot for graphical comparison of two species assemblages. In Section 2.5, we describe the two proposed nonparametric testing procedures. Simulation studies are carried out to evaluate the performance of the proposed tests in Section 2.6. In Section 2.7, 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 2.8 and briefly mention the multiple sample versions of the proposed tests. Part of this chapter is based on our published article, Li, Ban and Santiago (2011).

2.2 Background: data depth

The tests we will propose are based on data depth. Therefore, to begin with, we briefly introduce the concept of data depth in this section.

Roughly speaking, the data depth of a point measures how "deep" or how "central" that a point lies within a data cloud or w.r.t. its underlying distribution. The smaller the data depth value is, the more outlying the point is w.r.t. the data cloud or its underlying distribution. In the last two decades, different notions of data depth have been proposed. Next we briefly introduce several existing notions of data depth.

Mahalanobis (1936) introduced a distance between two points x and y in $\mathbf{R}^d$, w.r.t. a positive definite $d \times d$ matrix M, as

$$d^2{}_M(x, y) = (x - y)' M^{-1} (x - y).$$

Based on this *Mahalanobis distance*, one can define a *Mahalanobis depth* (MHD),

$$MHD_F(x) = (1 + d^2{}_{\Sigma(F)}(x, \mu(F)))^{-1},$$

where F is a given distribution and $\mu(F)$ and $\Sigma(F)$ are any corresponding location and covariance measures, respectively. The sample version of Mahalanobis depth is

$$MHD_{F_n}(x) = (1 + (x - \bar{X})' S^{-1} (x - \bar{X}))^{-1},$$

where $\bar{X}$ is the sample mean and S is the sample covariance matrix.

Turkey (1975) proposed a *halfspace depth.* The *halfspace depth* (HD) of a point x in $\mathbf{R}^d$ w.r.t. a probability distribution F in $\mathbf{R}^d$ is defined as the minimum probability mass carried by any closed halfspace containing x, that is,

$$HD_F(x) = inf\{P(H : H \text{ is a closed halfspace}, x \in H)\}, x \in \mathbf{R}^d.$$

The sample version of halfspace depth is

$$HD_{F_n}(x) = \min(\#\{i : X_i \in H \text{ and } x \in H\})/n.$$

Liu (1990) introduced a notion of *simplicial depth.* Namely, the *simplicial depth* (SD) of a point x in $\mathbf{R}^d$ w.r.t. a probability distribution F on $\mathbf{R}^d$ is defined to be the probability that x belongs to a random simplex in $\mathbf{R}^d$, that is,

$$SD_F(x) = P(x \in S[X_1, \ldots, X_{d+1}]), x \in \mathbf{R}^d,$$

where $X_1, \ldots, X_{d+1}$ is a random sample from F and $S[X_1, \ldots, X_{d+1}]$ denotes the d-dimensional simplex with vertices $X_1, \ldots, X_{d+1}$, that is, the set of all points in $\mathbf{R}^d$ that are convex combinations of $X_1, \ldots, X_{d+1}$. If F is unknown and we have observations $X_1, \ldots, X_n$, then the sample version of the simplicial depth is

$$SD_{F_n}(x) = \binom{n}{d+1}^{-1} \sum_{(*)} I(x \in S[X_{i_1}, \ldots, X_{i_{d+1}}]),$$

which measures how deep x is within the data cloud $\{X_1, \ldots, X_n\}$. The function $F_n(\cdot)$ denotes the empirical distribution of $\{X_1, \ldots, X_n\}$ and $(*)$ runs over all possible subsets of $\{X_1, \ldots, X_n\}$ of size $(d+1)$.

Zuo and Serfling (2000) also mentioned L^p depth $(p > 0)$. It is defined as

$$L^p D_F(x) = (1 + E \parallel x - X \parallel_p)^{-1},$$

where $\parallel \cdot \parallel$ denotes the L^p norm. The sample version of it is

$$L^p D_{F_n}(x) = (1 + \sum_{i=1}^{n} \parallel x - X_i \parallel_p /n)^{-1}$$

See Liu *et al.* (1999) and Zuo and Serfling (2000) for more notions of data depth.

All these data depths we mentioned here characterize different aspects of the geometric structure of the underlying distribution or sample of data. For Mahalanobis depth, the depth value of each sample point is determined only by its quadratic distance to the sample mean, so the depth contour expands only when the distance of the points on it to the sample mean increases. So this depth will not reflect the structure of the asymmetric distribution. However, for Simplicial depth, it measures the probability that each sample point is covered by the simplex in the data cloud. So it will reflect the relative position of each point.

Now we use the same notation $D_{F_n}(x)$ to denote whatever data depth that is a bounded and nonnegative mapping from $\mathbf{R}^d \times \mathcal{F}$ to $\mathbf{R}$ unless indicated otherwise. For a given sample $\{X_1, \ldots, X_n\}$, we calculate all the depth values $D_{F_n}(X_i)$ and then order the X_i's according to their descending depth values. The sample point with jth largest depth value is denoted by $X_{[j]}$. Then we obtain the sequence $\{X_{[1]}, X_{[2]}, \ldots, X_{[n]}\}$ which is the depth order statistics of X_i's, with $X_{[1]}$ being the deepest point, and $X_{[n]}$ the most outlying point. Here, a larger rank is associated with a more outlying position w.r.t. the underlying distribution F. At last, we will have a center-outward ordering of the sample points by using the data depth.

Based on data depth and its induced center-outward ordering, many useful statistical tools were developed. People developed data depth based descriptive statistics such as location "center", scale curve, skewness, depth contours, and quantiles to characterize multivariate distributions (e.g., Liu, Parelius and Singh, 1999). Data depth was also used to develop tools for statistical inference such as confidence region construction and hypothesis testing (e.g., Liu and Singh, 1997; Li and Liu, 2004). It has been shown that data depth can be applied in the following areas such as multivariate control chart construction, tolerance region construction, and classification (e.g., Liu, 1995; Gkosh and Chaudhuri, 2005; Li and Liu, 2008; Li *et al.*, 2012).

2.3 A distance-based data depth

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 *et al.* (1997). The definition of the distance-based depth is given below.

Definition (Distance-based depth) Let $\mathbf{X} = \{X_1, ..., X_n\}$ 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

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

and the sample version is

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

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.

To demonstrate that the above definition can be used to quantify the centrality of multivariate data points w.r.t. any multivariate data cloud, we first consider $\mathbf{X}$ as a random sample in $\mathbf{R}^2$, and Euclidean distance as the distance measure. Given any two data points X_i and X_j, we can form two circles, each having one of the points as the center and the other on the circle, as shown in Figure 2.1. The radiuses of both circles are equal to the Euclidean distance between X_i and X_j, $d(X_i, X_j)$. We denote the shaded area in Figure 2.1 by $B(X_i, X_j)$. Then the event $A = \{x \,|\, d(X_i, X_j) > \max\left[d(X_i, x), d(X_j, x)\right]\}$ is equivalent to $E = \{x \,|\, x \in B(X_i, X_j)\}$. Therefore, the above $D_{F_n}(x)$ calculates the proportion of the intersections $B(X_i, X_j)$ containing x. For any point x in $\mathbf{R}^2$, if x is deep inside or near the center of the data cloud, x should be

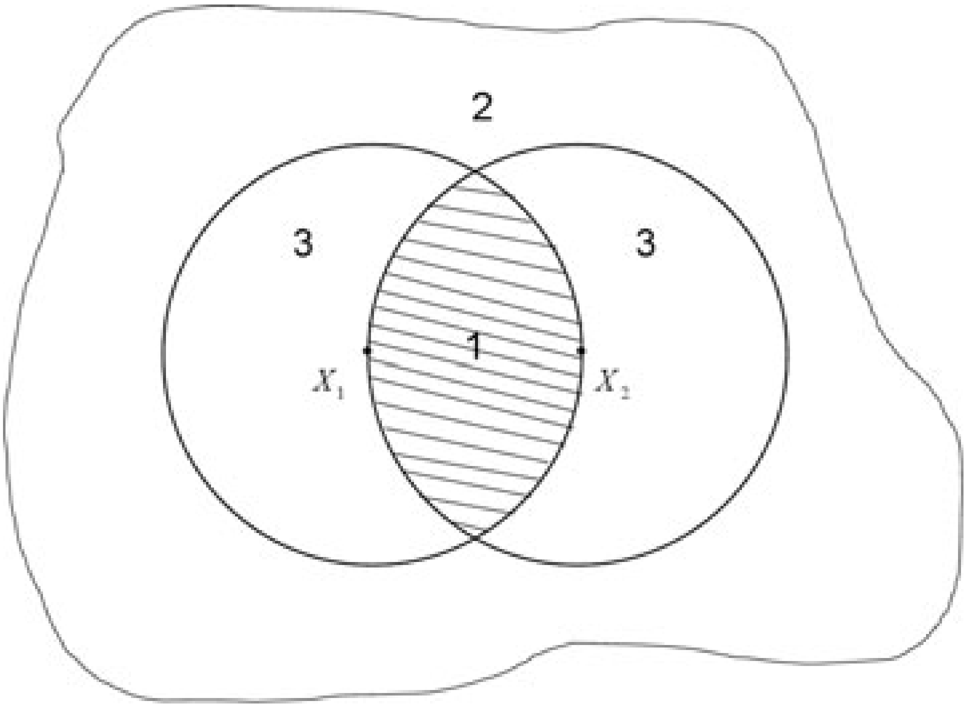

Figure 2.1: $B(X_i, X_j)$ in two dimensional case

contained in many of the intersections $B(X_i, X_j)$ 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 the intersections $B(X_i, X_j)$. 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 $\{X_1, \cdots, X_n\}$.

Since 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 $\{X_1, \cdots, X_n\}$, 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 purpose, 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 (λ_1, λ_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 μ and Σ are the parameters of the multivariate lognormal distribution. In ecology, for this type of data, Euclidean distance is not generally considered appropriate. Instead, measures such as Bray-Curtis distance (Bray and Curtis, 1957) are preferred. The Bray-Curtis distance for sample points $X_l = (X_{l1}, X_{l2}, ..., X_{lp})'$ and $X_{l'} = (X_{l'1}, X_{l'2}, ..., X_{l'p})'$ 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 2.2 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.

2.4 *DD*-plot: a graphical comparison of species assemblages

In this section, we demonstrate how the so-called *DD*-plot (depth vs depth plot) can be used to provide a graphical tool for comparisons of species assemblages. The *DD*-plot was 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 2.3, the *DD*-plot can now be directly applied to our species abundance data. Let $\{X_1, ..., X_m\}(\equiv \mathbf{X} \subset \mathbf{R}^c\ )$ and $\{Y_1, ..., Y_n\}(\equiv \mathbf{Y} \subset \mathbf{R}^c)$ be the abundance

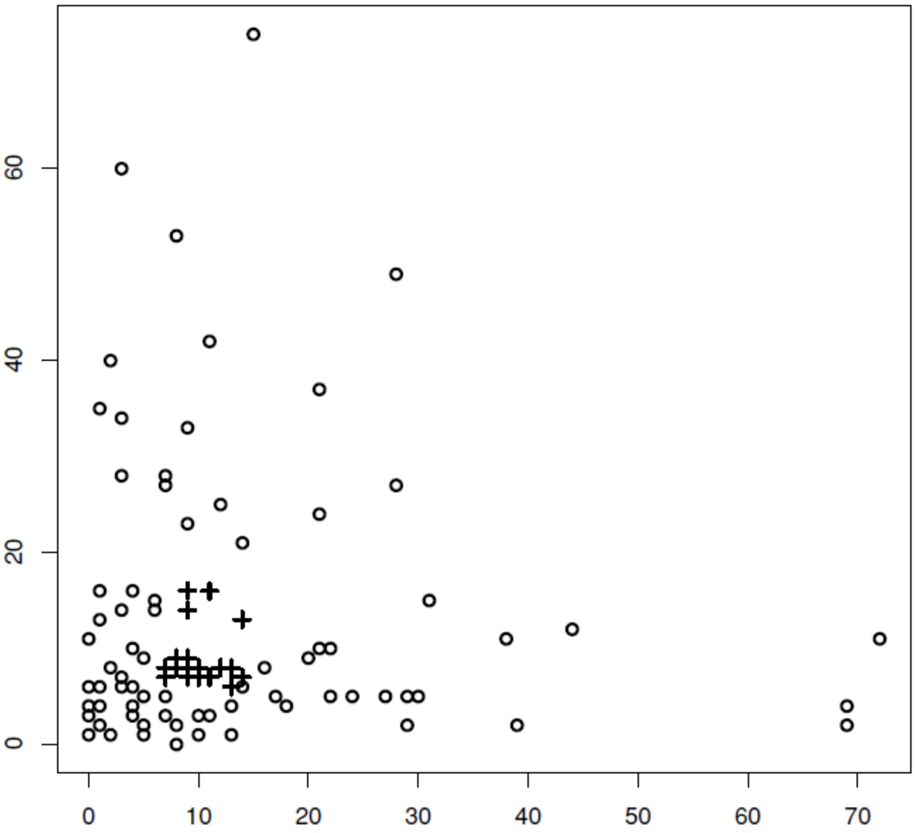

Figure 2.2: A bivariate Poisson-lognormal sample with the 20% deepest points

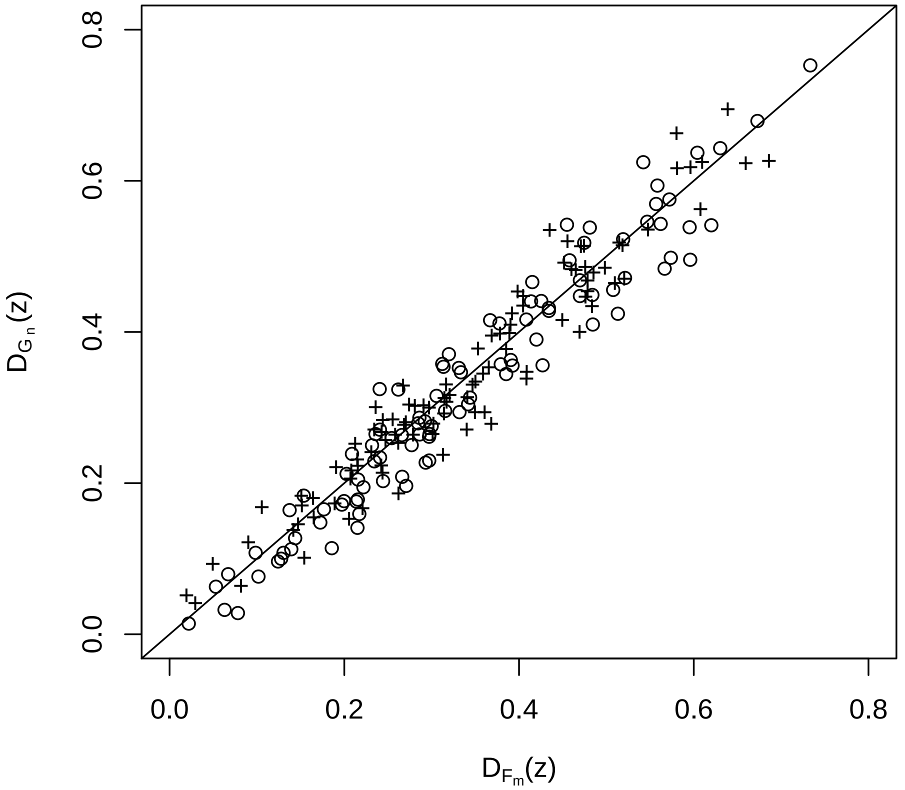

Figure 2.3: *DD*-plot: $F = G = PL(\mathbf{1}_{10}, I_{10})$.

data from two species assemblages respectively, where c is the total number of observed species in the two species assemblages. The *DD*-plot is constructed by

$$DD(F_m, G_n) = \{(D_{F_m}(z), D_{G_n}(z)), z \in \mathbf{X} \cup \mathbf{Y}\}, \tag{2.1}$$

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

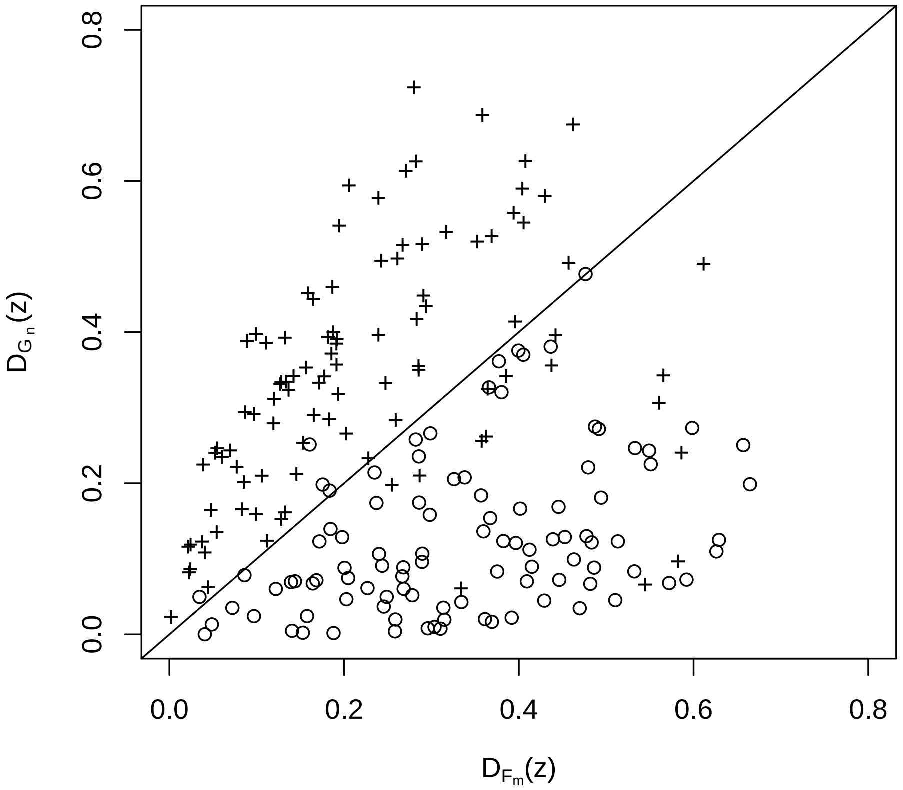

Figure 2.4: *DD*-plot: $F = PL(\mathbf{1}_{10}, I_{10})$ and $G = PL(2\mathbf{1}_{10}, I_{10})$.

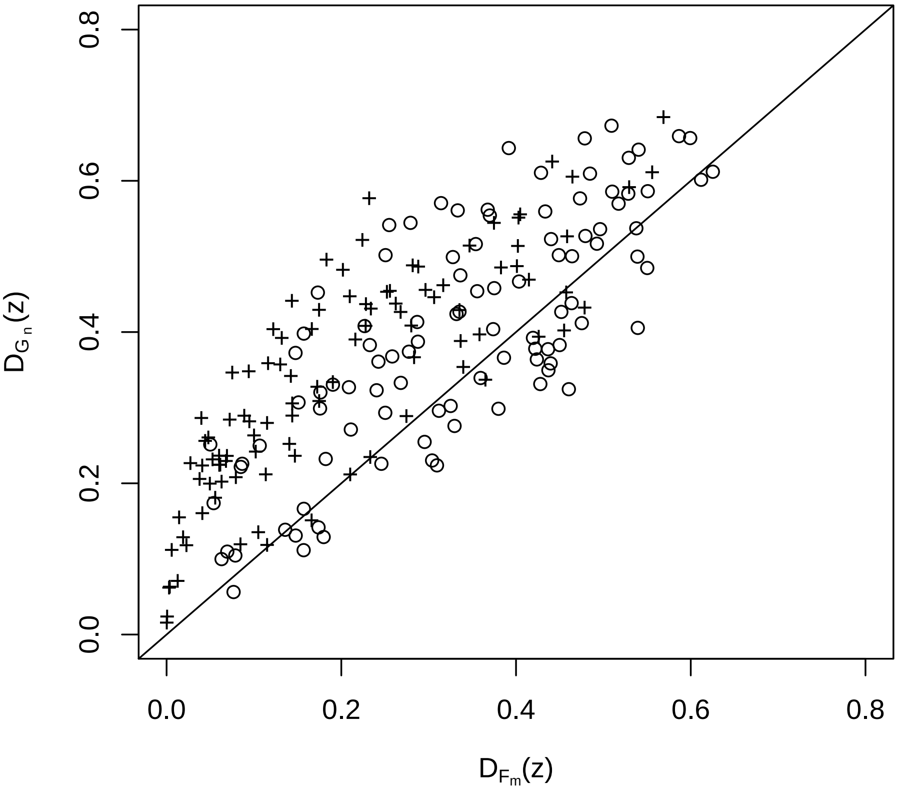

Figure 2.5: DD-plot: $F = PL(\mathbf{1}_{10}, I_{10})$ and $G = PL(\mathbf{1}_{10}, 2I_{10})$.

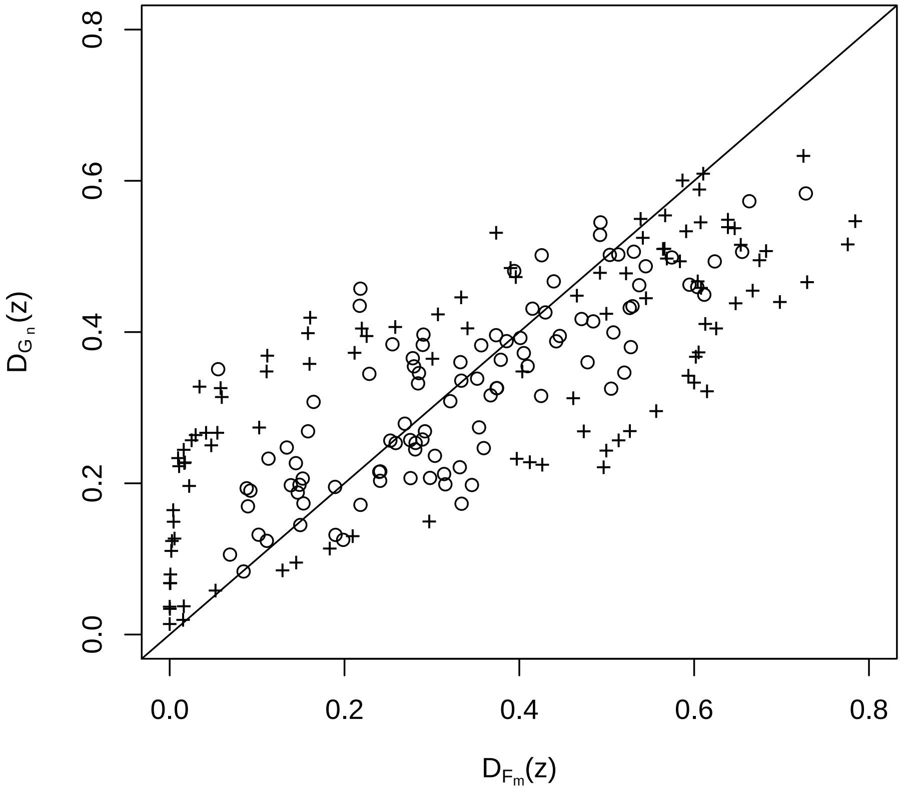

Figure 2.6: DD-plot: $F = PL(\mathbf{1}_{10}, I_{10})$ and $G = PL(\mathbf{1}_{10}, 0.8\mathbf{1}_{10}\mathbf{1}_{10}' + 0.2I_{10})$.

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 2.3. 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 Figures 2.4, 2.5, and 2.6. 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, while $\mathbf{Y}$ 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 Figures 2.3-2.6, the circles represent the observations from $\mathbf{X}$, and the pluses represent the observations from $\mathbf{Y}$. In Figures 2.3-2.6, 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 2.4, 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 2.5. 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.

2.5 Tests of homogeneity of species assemblages

In univariate case, suppose $X_1, X_2, \ldots, X_m$ and $Y_1, Y_2, \ldots, Y_n$ are independent random samples drawn from continuous distributions F and G, respectively. We would like to test

$$H_0 : F = G \text{ v.s. } H_a : F \neq G \tag{2.2}$$

Let $\mathbf{Z} = \{X_1, X_2, \ldots, X_m, Y_1, Y_2, \ldots, Y_n\}$. The Kolmogrov-Smirnov statistic is

$$D_{m,n} = \sup_{z \in \mathbf{Z}} |F_m(z) - G_n(z)|,$$

where F_m and G_n are the empirical distribution functions of the first and the second sample respectively. The null hypothesis is rejected at level α if $\sqrt{\frac{mn}{m+n}} D_{m,n} > \mathscr{K}_\alpha$, where $\mathscr{K}_\alpha$ is found from $Pr(\mathscr{K} \leq \mathscr{K}_\alpha) = 1 - \alpha$ and $\mathscr{K}$ follows the Kolmogorove distribution. The Cramér von-Mises statistic is

$$\omega^2_{m,n} = \sum_{z \in \mathbf{Z}} (F_m(z) - G_n(z))^2.$$

The null hypothesis is rejected when $\omega^2_{m,n}$ is too large. Next, we will discuss how similar tests are developed based on data depth given multivariate data.

In this chapter, from here on, we denote the abundance data from two species assemblages by $\{X_1, ..., X_m\}$ and $\{Y_1, ..., Y_n\}$. 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,

$$H_0 : F = G \text{ v.s. } H_a : F \neq G \tag{2.3}$$

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 (2.3), which can be considered as a natural generalization of Kolmogorov-Smirnov test and Cramér von-Mises test in this species assemblage comparison context:

- Kolmogorov-Smirnov (KS) type test statistic:

$$T_{KS} = \sup_{z \in \mathbf{X} \cup \mathbf{Y}} |D_{F_m}(z) - D_{G_n}(z)| \tag{2.4}$$

- Cramér-von Mises (CM) type test statistic:

$$T_{CM} = \sum_{z \in \mathbf{X} \cup \mathbf{Y}} [D_{F_m}(z) - D_{G_n}(z)]^2 \tag{2.5}$$

Define

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

where T_{KS}^{obs} and T_{CM}^{obs} are the observed value 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 $\mathbf{X}_i^* = \{X_{i1}^*, \cdots, X_{in}^*\}$, and $\mathbf{Y}_i^* = \{Y_{i1}^*, \cdots, Y_{in}^*\}$, 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 (2.4) and (2.5)), 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

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

and

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

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.

2.6 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 Dan Nettleton and T. Banerjee (2001) (NB thereafter), which can be used to test the equality of distributions of random vectors with categorical components. It is a specialization of the testing procedure proposed by Friedman and Rafsky (1979). We find their tests are not limited to categorical data, since the test allow one to define his own distance function. They define that each data point or random vector is linked to its nearest neighbor(s). The test statistics is the number of linkages that connects observations or random vectors from two different distributions. Let $\mathbf{Z} = \{Z_1, ..., Z_{m+n}\}$ denote the pooled sample $\mathbf{X} \cup \mathbf{Y}$. The NB's 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 $d(Z_i, Z_k)$, $k = 1, ..., i-1, i+1, ..., m+n$, and $d(\cdot, \cdot)$ is any distance measure which is appropriate for the application. The test rejects $H_0 : F = G$ if T_{NB} is too small, since if H_0 is false, the two groups of observations are more likely to form two clusters in space, increasing the probability that observations from the same group are linked and decreasing the probability that observations from different groups are linked. The rejection region can be determined by permutation.

The second test we will consider was proposed by Hall and Tajvidi (2002) (HT thereafter). Again consider the pooled sample $\mathbf{Z}$. They 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, they 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

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

The HT's 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 non-negative weights and γ is some positive value. Like the NB's test, the HT's test can be based on any distance measure. The test rejects $H_0 : F = G$ if T_{HT} is too large. The rejection region can be determined by permutation. Based on the simulation studies reported in HT (2002), several different choices of weight functions and γ 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's 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 μ and σ being manipulated according to different settings. All of the tests were then carried out through permutation. The number of permutations was set to be 1000. The significance level was set at 0.05. Again, we chose Bray-Curtis distance as the distance measure in all the tests. Table 2.1 shows the simulated power for the four tests under different choices of μ with σ being fixed at 1, i.e., $\Sigma_F = \Sigma_G$. Table 2.2 shows the simulated power for different choices of σ with μ 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's tests. When $\boldsymbol{\mu}_F = \boldsymbol{\mu}_G$, the KS test is slightly worse than the HT's test. In both settings, the NB's test has the lowest power.

Our second simulation study is to investigate the powers of the four tests for

Table 2.1: 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}
$\mu = 1$	.066	.046	.052	.053
$\mu = 1.1$	.072	.062	.082	.071
$\mu = 1.2$	.082	.110	.121	.122
$\mu = 1.3$	.091	.211	.242	.264
$\mu = 1.4$	.118	.307	.361	.418
$\mu = 1.5$	.178	.447	.573	.618
$\mu = 1.6$	.219	.592	.722	.766
$\mu = 1.7$	.308	.740	.839	.876
$\mu = 1.8$	.410	.853	.922	.938
$\mu = 1.9$	.510	.931	.974	.984
$\mu = 2$	.620	.965	.987	.995

Table 2.2: 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}
$\sigma = 1.1$	.066	.065	.051	.046
$\sigma = 1.2$	.071	.076	.092	.092
$\sigma = 1.3$	.076	.129	.117	.132
$\sigma = 1.4$	.093	.193	.169	.201
$\sigma = 1.5$	.103	.254	.224	.288
$\sigma = 1.6$	.107	.339	.289	.365
$\sigma = 1.7$	.120	.446	.398	.497
$\sigma = 1.8$	.128	.527	.480	.578
$\sigma = 1.9$	.168	.613	.547	.667
$\sigma = 2$	.188	.692	.618	.755

Table 2.3: Simulated powers for comparing samples from different distribution families.

	T_{NB}	T_{HT}	T_{KS}	T_{CM}
$G = PG(0.582\mathbf{1}_{10}, 7.701\mathbf{1}_{10})$	.13	.611	.475	.74
$G = PW(0.772\mathbf{1}_{10}, 3.851\mathbf{1}_{10})$	.121	.487	.371	.588

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 2.3 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.

To investigate the sensitivity of the tests to different types of differences between the two distributions, we also simulated data from $F = MN(\boldsymbol{\mu}_F, \Sigma_F)$ and $G = MN(\boldsymbol{\mu}_G, \Sigma_G)$, where $MN(\boldsymbol{\mu}, \Sigma)$ is the multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance Σ. Here we chose $\boldsymbol{\mu}_F = \mathbf{0}_5$, $\Sigma_F = I_5$, $\boldsymbol{\mu}_G = \mu\mathbf{1}_5$ and $\Sigma_G = \sigma I_5$ with μ and σ being manipulated according to different settings. For this type of distribution, we choose Euclidean distance as the distance measure. Table 2.4 shows that the sim-

Table 2.4: Simulated power for different tests using the samples from $F = MN(\boldsymbol{\mu}_F, \Sigma_F)$ and $G = MN(\boldsymbol{\mu}_G, \Sigma_G)$ where $\boldsymbol{\mu}_F = \mathbf{0}_5$, $\boldsymbol{\mu}_G = \mu\mathbf{1}_5$, and $\Sigma_F = \Sigma_G = I_5$.

	T_{NB}	T_{HT}	T_{KS}	T_{CM}
$\mu = 0$	.059	.047	.058	.051
$\mu = .1$	.072	.062	.082	.071
$\mu = .2$	.093	.120	.163	.161
$\mu = .3$	.154	.276	.324	.319
$\mu = .4$	.258	.493	.557	.574
$\mu = .5$	.403	.726	.770	.794
$\mu = .6$	.570	.906	.935	.949
$\mu = .7$	.752	.979	.983	.993
$\mu = .8$	.878	.997	.999	.999
$\mu = .9$	.946	.999	.999	1
$\mu = 1$	.980	1	1	1

Table 2.5: Simulated power for different tests using the samples from $F = MN(\boldsymbol{\mu}_F, \Sigma_F)$ and $G = MN(\boldsymbol{\mu}_G, \Sigma_G)$ where $\boldsymbol{\mu}_F = \boldsymbol{\mu}_G = \mathbf{0}_5$, $\Sigma_F = I_5$ and $\Sigma_G = \sigma I_5$.

	T_{NB}	T_{HT}	T_{KS}	T_{CM}
$\sigma = 1.1$	.068	.068	.074	.073
$\sigma = 1.2$	.071	.117	.095	.120
$\sigma = 1.3$	.077	.244	.172	.259
$\sigma = 1.4$	.085	.327	.249	.374
$\sigma = 1.5$	.098	.462	.365	.523
$\sigma = 1.6$	.133	.574	.438	.630
$\sigma = 1.7$	.175	.695	.529	.755
$\sigma = 1.8$	.168	.797	.653	.845
$\sigma = 1.9$	.186	.838	.706	.886
$\sigma = 2$	.217	.896	.774	.934

ulated power of the four tests under different choices of μ with σ being fixed at 1, i.e. the two distributions have the same scale. Table 2.5 shows that the simulated power under different choices of σ with μ being fixed at 0, i.e., the two distributions have the same location. From the tables, we can see that our depth-based KS test and CM test perform similarly in detecting location difference, while the CM test is more sensitive to the scale difference than the KS test. In both settings, again, the depth-based CM test performs best among all the tests.

2.7 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 Chapter 1. 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).

First, let us take a look at the raw data. We select 5 species which have the highest combined counts in the two plots. They are Alseis blackiana (Ab), Coussarea curvigemmia (Cc), Faramea occidentalis (Fo), Hybanthus prunifolius (Hp), and Tetragastris panamensis (Tp). We would like to compare the two assemblages to see if there is any difference between them by visually looking at the raw data of the 5 species. Since there is no one-to-one correspondence between the quadrats of the two plots, in order to make the comparison easy, respectively for each species from each assemblage, we rank the quadrats by the counts of that species in the quadrats (from low to high), and denote the ranking by q1, q2,..., q25. Then we can present the counts of the 5 most frequently observed species in 25 quadrats in Table 2.6. Based on the counts of the first 4 species, we suspect there is a location difference between the distributions of the abundance data from the two plots. Furthermore, Table 2.7 shows the comparison of average counts per quadrat of each species from the two plots. To save space, species names are replaced by their alphabetical order. We can see that species 111-159 were not observed in plot 1 but were observed in plot 2, which increases our suspect on the location difference between the distributions of the abundance data from the two plots.

Table 2.6: Counts of the 5 most frequently observed species in 25 quadrats from the two plots in the BCI data.

	Ab		Cc		Fo		Hp		Tp	
plot	1	2	1	2	1	2	1	2	1	2
q1	0	0	0	0	0	0	0	0	0	0
q2	0	0	0	0	0	0	0	0	0	0
q3	1	0	0	0	0	0	0	0	0	0
q4	1	0	0	0	0	0	0	0	0	0
q5	1	0	0	0	0	0	0	0	0	0
q6	1	0	0	0	0	0	0	0	0	0
q7	1	0	0	0	0	0	0	0	0	0
q8	2	0	0	0	0	0	0	0	0	0
q9	2	0	0	0	0	0	0	0	0	0
q10	2	1	0	0	0	0	0	0	1	0
q11	2	1	0	0	0	0	0	0	1	1
q12	2	1	0	0	0	0	0	0	1	1
q13	2	1	0	0	0	0	0	0	1	1
q14	2	1	0	0	0	0	0	0	1	1
q15	2	1	1	0	0	0	0	0	1	1
q16	3	1	1	0	0	0	0	0	2	2
q17	3	1	1	0	1	0	0	0	2	2
q18	3	2	1	0	1	1	0	0	2	2
q19	3	2	2	1	1	1	0	0	2	2
q20	4	2	2	1	1	1	0	0	2	3
q21	5	3	2	2	2	2	0	0	3	3
q22	5	3	14	2	32	4	6	4	4	5
q23	6	4	23	4	38	4	7	5	4	7
q24	7	4	25	4	43	8	8	15	7	9
q25	7	10	56	7	50	10	19	41	16	10
Total	67	38	128	21	169	31	40	65	50	50

Table 2.7: Average counts per quadrat of each observed species from the two plots in the BCI data.

species	1	2	3	4	5	6	7	8	9	10	11	12
plot 1	0.24	0.24	0.08	0.04	2.68	0.08	0.12	1.04	0.16	0.32	1.12	0.04
plot 2	0.04	0	0.04	0	1.52	0	0.08	0.04	0	0.16	0.16	0
species	13	14	15	16	17	18	19	20	21	22	23	24
plot 1	0.04	0.08	0.56	0.04	0.12	0.4	0.04	0.08	0.08	0.04	0.04	0.04
plot 2	0	0	0.16	0	0	0	0.12	0	0.16	0.16	0	0.04
species	25	26	27	28	29	30	31	32	33	34	35	36
plot 1	0.08	0.08	5.12	0.04	0.12	0.28	0.08	0.12	0.04	0.12	0.04	0.76
plot 2	0.12	0.44	0.84	0.28	0.04	0.4	0	0.44	0.08	0.28	0.08	0.2
species	37	38	39	40	41	42	43	44	45	46	47	48
plot 1	0.12	0.04	0.36	6.76	0.08	0.2	0.12	0.04	0.12	0.32	0.04	0.8
plot 2	0.32	0	0.08	1.24	0	0	0.2	0	0.04	0.08	0.04	0.8
species	49	50	51	52	53	54	55	56	57	58	59	60
plot 1	0.04	1.6	0.04	0.08	0.24	0.04	0.12	0.2	0.2	0.08	0.08	0.92
plot 2	0	2.6	0	0.48	0	0.08	0	0	0	0.52	0.08	0.12
species	61	62	63	64	65	66	67	68	69	70	71	72
plot 1	0.12	0.16	0.28	0.04	0.88	0.4	0.24	0.04	0.16	0.08	0.08	0.04
plot 2	0.04	0.12	0.04	0.04	0.28	0.08	0.72	0	0.04	0	0.08	0
species	73	74	75	76	77	78	79	80	81	82	83	84
plot 1	0.08	0.12	0.2	0.16	0.04	0.28	0.16	0.24	0.04	2.88	0.32	0.12
plot 2	0	0.04	0.08	0.12	0.24	0.24	0.88	0.04	0	0.04	0.12	0.04
species	85	86	87	88	89	90	91	92	93	94	95	96
plot 1	0.4	0.16	0.12	0.84	0.04	0.12	0.08	0.16	0.16	0.16	0.08	0.04
plot 2	0.04	0.16	0.04	0.88	0	0	0	0.08	0.28	0.04	0.04	0.04
species	97	98	99	100	101	102	103	104	105	106	107	108
plot 1	2.28	0.24	0.04	0.04	0.04	2	0.04	0.04	0.04	1.6	0.44	0.36
plot 2	0.8	0.08	0	0.16	0	2	0	0.08	0	0.44	0	0.04
species	109	110	111	112	113	114	115	116	117	118	119	120
plot 1	0.04	0.12	0	0	0	0	0	0	0	0	0	0
plot 2	0	0.04	0.08	0.08	0.04	0.08	0.04	0.08	0.08	0.16	0.2	0.04
species	121	122	123	124	125	126	127	128	129	130	131	132
plot 1	0	0	0	0	0	0	0	0	0	0	0	0
plot 2	0.44	0.08	0.04	0.08	0.04	0.2	0.12	0.04	0.08	0.36	0.04	0.08
species	133	134	135	136	137	138	139	140	141	142	143	144
plot 1	0	0	0	0	0	0	0	0	0	0	0	0
plot 2	0.12	0.04	0.6	0.08	0.04	0.04	0.08	0.12	1.8	0.04	0.08	0.08
species	145	146	147	148	149	150	151	152	153	154	155	156
plot 1	0	0	0	0	0	0	0	0	0	0	0	0
plot 2	0.04	0.08	1.64	0.08	0.04	0.12	0.72	0.04	0.04	0.44	0.92	0.12
species	157	158	159									
plot 1	0	0	0									
plot 2	0.16	0.2	0.16									

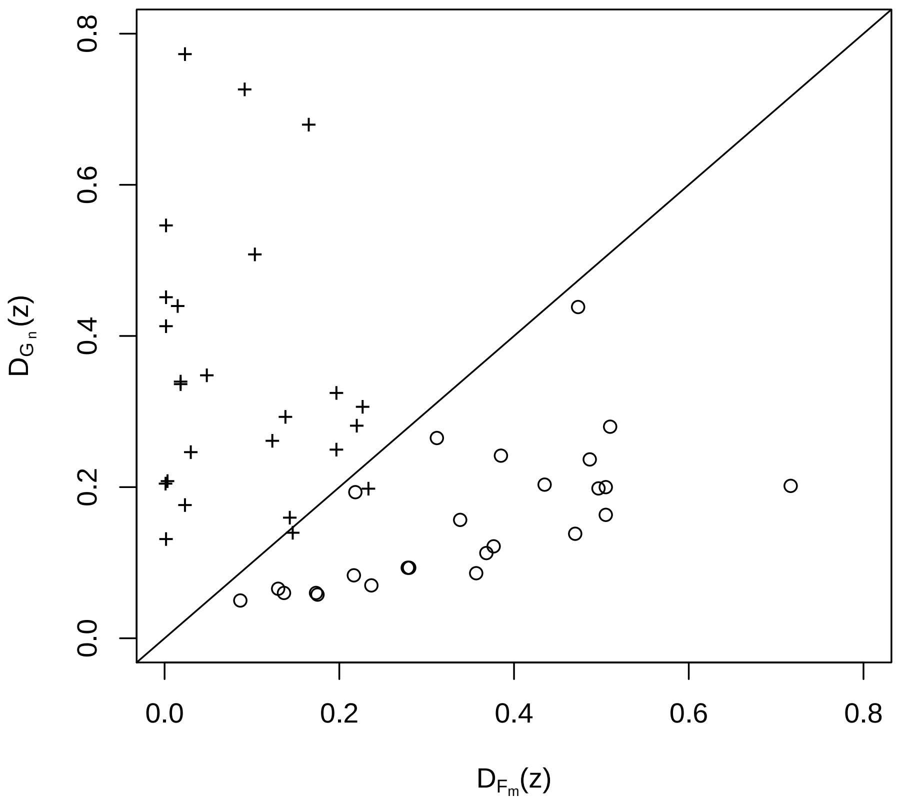

Figure 2.7: *DD*-plot for the samples from the two tropical forest census plots

Before we apply our tests to the data, we first use the *DD*-plot described in Section 2.4 to visualize the difference of these two species assemblages. Figure 2.7 shows the corresponding *DD*-plot based on the distance-based depth by using Bray-Curtis distance. In the plot, the circles represent 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.

2.8 Summary

In this chapter, 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.

The tests we proposed earlier are for comparing two species assemblages. Now we introduce a few tests, which can be used for comparing multiple species assemblages. Kiefer (1959) proposed several multiple sample analogues of the Kolmogorov-Smirnov and the Cramér-von Mises tests. The following multiple sample tests we propose are inspired by Kiefer (1959). Suppose there are K $(K \geq 2)$ samples of species assemblages , denoted by SA_k, $k = 1, \ldots, K$, and let $D_{SA_k}(z)$ denote the sample distance-based depth of observation z w.r.t sample k. Denote the pooled sample by $SA_{pool} = \bigcup_{k=1}^{K} SA_k$. We propose the following test statistics,

- Kolmogorov-Smirnov type test statistics:

$$MT_{KS1} = \sum_{1 \leq i \leq j \leq K} \sup_{z \in SA_{pool}} |D_{SA_i}(z) - D_{SA_j}(z)|$$

$$MT_{KS2} = \sup_{1 \leq i \leq j \leq K, z \in SA_{pool}} |D_{SA_i}(z) - D_{SA_j}(z)|$$

- Cramér-von Mises type test statistics:

$$MT_{CM1} = \sum_{1 \leq i \leq j \leq K} \sum_{z \in SA_{pool}} \Big[D_{SA_i}(z) - D_{SA_j}(z)\Big]^2$$

$$MT_{CM2} = \sup_{1 \leq i \leq j \leq K} \sum_{z \in SA_{pool}} \Big[D_{SA_i}(z) - D_{SA_j}(z)\Big]^2$$

.

Chapter 3

Zero-inflated Poisson mixture model based tests

3.1 Introduction

The nonparametric data depth based tests mentioned in Chapter 2 are good for tracking changes of abundance of individual species. However, sometimes, ecologists are particularly interested in comparison of species diversity across species assemblages. In this chapter, we propose the zero-inflated Poisson mixture model based tests for that purpose.

To facilitate our later on discussion, we first briefly introduce two other ecological data types: incidence data and abundance data from a single quadrat. Incidence data is similar to abundance data from multiple quadrats, but, instead of the count of each individual species, only presence and absence in a sampling quadrat are recorded. If we still use a matrix to represent incidence data as we did to abundance data in Chapter 1, the matrix only has elements 0 and 1, with 1 meaning presence and 0 meaning absence. We refer to abundance data from a sampling procedure where the whole sampling area is treated as one single quadrat as abundance data from a single quadrat.

The mixture models were first used to develop estimators for number of species of a species assemblage due to their capabilities to account for heterogeneity among species (e.g., Ord and Whitmore, 1986; Bunge and Fitzpatrick, 1993; Chao and Bunge, 2002; Böhning and Schön, 2005; Mao and Colwell, 2005; Mao, 2006). More specially, for incidence data, binomial mixture is usually used, and for abundance data from a single quadrat, Poisson mixture is usually used. In Mao and Li (2009), a testing procedure was proposed to compare species assemblages under the binomial mixture model when incidence data is available. Recently Li, Mao, and Wang (2011) developed a testing procedure under the Poisson mixture model when abundance data from a single quadrat is available. The two mixture model based testing procedures can be used for comparing the variations in the temporal or regional patterns of species diversity across species assemblages. But neither of the two mixture models can be directly applied to abundance data from multiple quadrats. Since more and more abundance data from multiple quadrats are collected in practice, there is a need for a unique model which can handle this type of data. In this chapter, we choose to work on the problem of comparison of species diversity across species assemblages under the mixture model framework when the abundance data from multiple quadrats is available. For this purpose, we first introduce the zero-inflated Poisson mixture model for abundance data from multiple quadrats. Based on this mixture model, the comparison of species assemblages amounts to comparing the total number of species and the mixing distributions in the zero-inflated Poisson mixture model. However, neither of them can be estimated well in practice. To circumvent those difficulties, we develop a procedure for comparing some functions of the total number of species and the mixing distributions instead of comparing them directly. Those functions can be readily estimated and at the same time we show that the comparison of those functions is equivalent to the comparison of the total number of species and the mixing distributions, which is ultimately equivalent to the comparison of the species assemblages under our zero-inflated Poisson mixture model.

The rest of the chapter is organized as follows. In Section 3.2, we briefly introduce the binomial mixture model and the incidence based tests, which inspired our research. In Section 3.3, we describe our zero-inflated Poisson mixture model for the abundance data from multiple quadrats. In Section 3.4, we introduce the hypothesis testing problem associated with the species assemblage comparison problem under the zero-inflated Poisson mixture model. In Section 3.5, we describe our testing procedure for comparing species assemblages. In Section 3.6, we demonstrate the procedure of estimating one of the latent distributions in our model. In Section 3.7, we discuss the impact of h_0. In Section 3.8, we discuss the multiple sample version of our testing procedure. In Section 3.9, we report some simulation studies to evaluate the performance of our proposed tests. In Section 3.10, we demonstrate the application of our test to a real ecological data set. Some concluding remarks are given in Section 3.11. All the proofs are collected in Appendix.

3.2 Background: the binomial mixture model based test

In Mao and Li (2009), a testing procedure was proposed to compare species assemblages under the binomial mixture model when incidence data is available. The model we will propose in this thesis is inspired by their model. So first let us take a brief review of their test.

To introduce some necessary notation for the species assemblage comparison problem, we consider two species assemblages. Each assemblage is divided into numerous quadrats. A sample of K_i $(i = 1, 2)$ quadrats is taken from assemblage i. A species is either present or absent in a quadrat, and that is recorded. Define

(i) c_i: the unknown total number of species in assemblage i,

(ii) Z_{ijk}: if species j is observed in quadrat k in assemblage i, then $Z_{ijk} = 1$, otherwise $Z_{ijk} = 0$,

(iii) $n_{i,k}$: the number of species in assemblage i present in exactly k quadrats.

Based on its definition, Z_{ijk} can be modeled by a Bernoulli distribution with rate parameter π_{ij}. Since the π_{ij}'s within the same sample are usually believed to vary across different species, a common practice is to assume that the π_{ij}'s in sample i are drawn from a common latent distribution G_i. We call it the species incidence rate distribution. Therefore the number of quadrats in which a species is present in sample i follows a Binomial mixture $f(\cdot; G_i)$, where

$$f(k; G_i) = \int \binom{K_i}{k} \pi^k (1-\pi)^{K_i - k} dG_i(\pi), \quad k = 0, \quad 1, \ldots, \quad K_i. \tag{3.1}$$

If different species are assumed to be independent of each other, then it is clear that $(n_{i,0}, n_{i,1}, \ldots, n_{i,K_i})'$ follows a multinomial distribution with index c_i and probabilities $f(j; G_i)$. This fact is very helpful in the later on test development. Based on the binomial mixture model described earlier, species assemblage i can be characterized by the total number of species, c_i and the species incidence rate distribution, G_i, $i = 1, \quad 2$ (they together describe patterns of species diversity of assemblage i). Therefore, comparing the two species assemblages can be formulated as a hypothesis testing problem as follows,

$$H_0 : c_1 = c_2 \text{ and } G_1 = G_2$$

versus

$$H_a : c_1 \neq c_2 \text{ or } G_1 \neq G_2.$$

However, neither c_i nor G_i can be estimated well (e.g., Bunge and Fitzpatrick, 1993; Huggins, 2001; Link, 2003; Mao, 2007). Alternative ways need to be considered. Define

$$\tau_i(h) = c_i \int \{1 - (1-\pi)^h\} dG_i(\pi), \quad h = 1, \quad 2, \ldots.$$

$\tau_i(h)$ is a species accumulation function widely used in the ecology literature for the study of species diversity. It calculates the expected number of observed species when h quadrats are randomly chosen from assemblage i. Mao and Li (2009) showed the above hypothesis testing problem is equivalent to

$$H_0 : \tau_1(h) = \tau_2(h) \text{ for } h = 1, 2, \ldots,$$

versus

$$H_a : \tau_1(h) \neq \tau_2(h) \text{ for some } h. \tag{3.2}$$

Since $\tau_i(h)$ only admits a closed-form nonparametric estimator for $h = 1, 2, ..., K_i$, they considered the following hypothesis testing problem implied by that in equation (3.2),

$$H_0 : \tau_1(h) = \tau_2(h) \text{ for } h = 1, 2, \ldots, K$$

versus

$$H_a : \tau_1(h) \neq \tau_2(h) \text{ for some } h, \tag{3.3}$$

where $K = min(K_1, K_2)$. If we define

$$\begin{aligned}
&\boldsymbol{\tau}_i = (\tau_i(1), \tau_i(2), \ldots, \tau_i(K))', \\
&A_i = (a_{i,h,k})_{h=1,k=1}^{K,K_i} \text{ with } a_{i,h,k} = 1 - \binom{K_i - h}{k} / \binom{K_i}{k}, and \\
&\tilde{\mathbf{n}}_i = (n_{i,1}, \ldots, n_{i,K_i})',
\end{aligned}$$

an estimator of the vector $\boldsymbol{\tau}$ can be expressed as $\hat{\boldsymbol{\tau}}_i = A_i \tilde{\mathbf{n}}_i$. It is very easy to find out $\hat{\boldsymbol{\tau}}_i$'s asymptotic normality with the fact $(n_{i,0}, n_{i,1}, \ldots, n_{i,K_i})'$ follows a multinomial distribution. Further, asymptotically, $\hat{\boldsymbol{\tau}}_1 - \hat{\boldsymbol{\tau}}_2$ follows $\mathcal{N}(\boldsymbol{\tau}_1 - \boldsymbol{\tau}_2, \Sigma)$, where Σ is the covariance matrix of $\hat{\boldsymbol{\tau}}_1 - \hat{\boldsymbol{\tau}}_2$ and depends on c_i. Therefore a natural test statistics for the hypothesis testing problem (3.3) is $(\hat{\boldsymbol{\tau}}_1 - \hat{\boldsymbol{\tau}}_2)'\hat{\Sigma}^{-1}(\hat{\boldsymbol{\tau}}_1 - \hat{\boldsymbol{\tau}}_2)$, where $\hat{\Sigma}$ is the sample version of Σ. The test statistics asymptotically follows χ^2 distribution. An adjustment using an eigenvalue decomposition is proposed to overcome computational difficulties of inverting $\hat{\Sigma}$. A bootstrap method is also proposed to approximate the distribution of the test statistics since c_i can not be estimated well. This test procedure performs well, however, given abundance data, the abundance information will not be taken into account, which will lessen its statistical power. So in the next section, we will propose another mixture model which is able to handle abundance data.

3.3 Zero-inflated Poisson mixture model

We continue to use the notation introduced in the previous section. For abundance data, if the species is present, the count of the species is recorded. Therefore define

X_{ijk}: the number of individuals from species j observed in quadrat k in assemblage i.

If the species j is absent in quadrat k in assemblage i, then $X_{ijk} = 0$. Typically, to model the count data X_{ijk}, Poisson distribution can be used. However, in many ecological data sets, a large frequency of zeros for X_{ijk} are common due to the rarity of some species. To account for this zero-inflation, we use the zero-inflated Poisson model. More specifically, the distribution of X_{ijk} is given by

$$Pr(X_{ijk} = x_{ijk}|\pi_{ij}, \lambda_{ij}) = \begin{cases} 1 - \pi_{ij}, & \text{if } x_{ijk} = 0 \\ \pi_{ij}\frac{\exp(-\lambda_{ij})}{1-\exp(-\lambda_{ij})}\frac{\lambda_{ij}^{x_{ijk}}}{x_{ijk}!}, & \text{if } x_{ijk} > 0 \end{cases} \tag{3.4}$$

where π_{ij} is the probability of species j in assemblage i present in a generic quadrat and λ_{ij} is the rate parameter of Poisson distribution for species j. It is clear that $Z_{ijk} = I\{X_{ijk} \neq 0\}$, where $I\{A\}$ is the indicator function and takes 1 if event A is true and 0 otherwise. The above zero-inflated Poisson model can be written as

$$Pr(X_{ijk} = x_{ijk}, Z_{ijk} = z_{ijk} | \pi_{ij}, \lambda_{ij}) = \pi_{ij}^{z_{ijk}} (1 - \pi_{ij})^{1 - z_{ijk}} \left\{ \frac{\exp(-\lambda_{ij})}{1 - \exp(-\lambda_{ij})} \frac{\lambda_{ij}^{x_{ijk}}}{x_{ijk}!} \right\}^{z_{ijk}} .$$

Usually π_{ij} and λ_{ij} may vary among species in one assemblage. To account for this heterogeneity among species, we assume that the π_{ij}'s are drawn from a latent distribution G_i, the λ_{ij}'s are drawn from a latent distribution H_i, and the π_{ij}'s and the λ_{ij}'s are independent. We call H_i the species abundance rate distribution. We further assume that, conditional on π_{ij} and λ_{ij}, the X_{ijk} from each species are independent across all the K_i quadrats.

Therefore, the likelihood function for assemblage i can be written as

$$L(c_i, G_i, H_i) = \prod_{j=1}^{c_i} \int \prod_{k=1}^{K_i} \pi^{z_{ijk}} (1 - \pi)^{1 - z_{ijk}} dG_i(\pi) \int \prod_{k=1}^{K_i} \left\{ \frac{\exp(-\lambda)}{1 - \exp(-\lambda)} \frac{\lambda^{x_{ijk}}}{x_{ijk}!} \right\}^{z_{ijk}} dH_i(\lambda).$$

We referred to the above mixture model as zero-inflated Poisson mixture. By incorporating in the species abundance rate distribution H_i, our model is able to handle abundance data from multiple quadrats.

Remark 1 *The mixture model we proposed here is a very general model in the sense that one can control the density of zero's in the distribution. For example, let* $\pi_{ij} = 1 - \exp(-\lambda_{ij})$, *then*

$$Pr(X_{ijk} = x_{ijk} | \pi_{ij}, \lambda_{ij}) = \begin{cases} \exp(-\lambda_{ij}), & \textit{if } x_{ijk} = 0 \\ \exp(-\lambda_{ij}) \frac{\lambda_{ij}^{x_{ijk}}}{x_{ijk}!}, & \textit{if } x_{ijk} > 0 \end{cases} . \tag{3.5}$$

We see that, given λ_{ij}, X_{ijk} *follows exactly Poisson distribution. The mixture model becomes purely a Poisson mixture. Therefore our model is a very flexible model.*

3.4 Hypothesis testing problem

Following the above zero-inflated Poisson mixture model in Section 3.3, species assemblage i is characterized by the number of species c_i, the species incidence rate distribution G_i, and the species abundance rate distribution H_i. They together describe the species diversity of species assemblage i. Under this mixture model framework, comparing two species assemblages can be formulated as the following hypothesis testing problem,

$$H_0 : c_1 = c_2,\ G_1 = G_2,\ H_1 = H_2$$

$$\text{versus}$$

$$H_a : c_1 \neq c_2 \text{ or } G_1 \neq G_2, \text{ or } H_1 \neq H_2. \tag{3.6}$$

To develop a testing procedure, one may first estimate $\{c_1, c_2, G_1, G_2, H_1, H_2\}$. However, the c_i and G_i can not be estimated well nonparametrically. Therefore, the above testing problem is challenging. To circumvent the difficulties, we search for another hypothesis which is equivalent to the hypothesis in (3.6) and the parameters in this new hypothesis admit closed-form estimators. For this purpose, we define

$$g_i(h,x) = c_i \int (1-(1-\pi))^h dG_i(\pi) \int \frac{\exp(-\lambda)}{1-\exp(-\lambda)} \frac{\lambda^x}{x!} dH_i(\lambda),$$

for $i = 1, 2,\ h = 1, 2, ...$ and $x = 1, 2, ...$ From the above definitions, it is not difficult to see that $\tau_i(h) = \sum_{x=1}^{\infty} g_i(h,x)$. Based on $\tau_i(h)$ and $g_i(h,x)$, we have the following result.

Theorem 2 *Given a positive integer h_0, if H_i's have bounded support, $c_1 = c_2$, $G_1 = G_2$ and $H_1 = H_2$ if and only if $g_1(h_0, x) = g_2(h_0, x)$ for $x = 1, 2, \ldots,$ and $\tau_1(h) = \tau_2(h)$ for $h = 1, 2, \ldots, h_0 - 1, h_0 + 1, \ldots$*

Theorem 2 immediately implies that the problem in (3.6) is equivalent to

$$H_0 : g_1(h_0, x) = g_2(h_0, x) \text{ for } x = 1, 2, ..., \text{ and } \tau_1(h) = \tau_2(h) \text{ for } h = 1, 2, \dots, h_0 - 1, h_0 + 1, ...,$$

versus

$$H_a : g_1(h_0, x) \neq g_2(h_0, x) \text{ for some } x \text{ or } \tau_1(h) \neq \tau_2(h) \text{ for some } h \neq h_0, \tag{3.7}$$

where h_0 can be any chosen positive integer. For simplicity, through the whole thesis, we choose $h_0 = 1$, and define $g_i(x) = g_i(1, x)$, but in Section 3.7, we will discuss tests developed under different h_0's and how different h_0's affect the power of those tests. Therefore, when $h_0 = 1$, we consider the following hypothesis testing problem,

$$H_0 : g_1(x) = g_2(x) \text{ for } x = 1, 2, ..., \text{ and } \tau_1(h) = \tau_2(h) \text{ for } h = 2, 3, ...,$$

versus

$$H_a : g_1(x) \neq g_2(x) \text{ for some } x \text{ or } \tau_1(h) \neq \tau_2(h) \text{ for some } h > 1. \tag{3.8}$$

To develop a procedure for the above hypothesis testing problem, we first need to find estimates for $g_i(x)$ and $\tau_i(h)$. Define $n_{i,k}$ as the number of species in assemblage i that appear in exactly k quadrats. According to Mao et al (2005), a nonparametric estimator of $\tau_i(h)$ is given by

$$\hat{\tau}_i(h) = \sum_{k=1}^{K_i} \left\{ 1 - \frac{\binom{K_i - h}{k}}{\binom{K_i}{k}} \right\} n_{i,k}, \quad h = 1, 2, ..., K_i.$$

To estimate $g_i(x)$, we further define $n^v_{i,k,x}$ as the number of species that appear in exactly k quadrats and appear x times in the v-th ($v = 1, \dots, k$) quadrat among those k quadrats. Denote the event $B_{ij,t_1,\dots,t_k}(x_{t_1}, ..., x_{t_{v-1}}, x, x_{t_{v+1}}, ..., x_{t_k}) = \{X_{ijt_1} = x_{t_1}, \dots, X_{ijt_{v-1}} = x_{t_{v-1}}, X_{ijt_v} = x, X_{ijt_{v+1}} = x_{t_{v+1}} \dots, X_{ijt_k} = x_{t_k},$ and all other $X_{ijk} =$

$0\}$. Based on the definition,

$$n_{i,k,x}^{v} = \sum_{j=1}^{c_i} \sum_{1\le t_1<\ldots<t_k\le K_i} \sum_{x_{t_1}=1}^{\infty} \cdots \sum_{x_{t_{v-1}}=1}^{\infty} \sum_{x_{t_{v+1}}=1}^{\infty} \cdots \sum_{x_{t_k}=1}^{\infty}$$

$$I\{B_{ij,t_1,\ldots,t_k}(x_{t_1},...,x_{t_{v-1}},x,x_{t_{v+1}},...,x_{t_k})\}.$$

Since

$$\begin{aligned}
&E[I\{B_{ij,t_1,\ldots,t_k}(x_{t_1},...,x_{t_{v-1}},x,x_{t_{v+1}},...,x_{t_k})\}] \\
&= \int \pi^k(1-\pi)^{K_i-k} dG_i(\pi) \int \left\{\frac{\exp(-\lambda)}{1-\exp(-\lambda)}\right\}^k \frac{\lambda^{x_{t_1}+\cdots+x_{t_{v-1}}+x+x_{t_{v+1}}+\cdots+x_{t_k}}}{x_{t_1}!\cdots x_{t_{v-1}}!x!x_{t_{v+1}}!\cdots x_{t_k}!} dH_i(\lambda),
\end{aligned}$$

we have, for any $v = 1, ..., k$,

$$E(n_{i,k,x}^{v}) = c_i \int \binom{K_i}{k} \pi^k(1-\pi)^{K_i-k} dG_i(\pi) \int \frac{\exp(-\lambda)}{1-\exp(-\lambda)} \frac{\lambda^x}{x!} dH_i(\lambda). \tag{3.9}$$

Using the result in Mao et al. (2005), $g_i(x)$ can be written as

$$\begin{aligned}
g_i(x) &= \tau_i(1) \int \frac{\exp(-\lambda)}{1-\exp(-\lambda)} \frac{\lambda^x}{x!} dH_i(\lambda) \\
&= \sum_{k=1}^{K_i} \left\{1 - \frac{\binom{K_i-1}{k}}{\binom{K_i}{k}}\right\} c_i \int \binom{K_i}{k} \pi^k(1-\pi)^{K_i-k} dG_i(\pi) \int \frac{\exp(-\lambda)}{1-\exp(-\lambda)} \frac{\lambda^x}{x!} dH_i(\lambda).
\end{aligned} \tag{3.10}$$

Therefore, based on (3.9) and (3.10), we can obtain an unbiased estimate for $g_i(x)$,

$$\hat{g}_i(x) = \sum_{k=1}^{K_i} \left\{1 - \frac{\binom{K_i-1}{k}}{\binom{K_i}{k}}\right\} n_{i,k,x},$$

where $n_{i,k,x} = \sum_{v=1}^{k} n_{i,k,x}^{v}/k$. Using the simple fact that $\sum_{x=1}^{\infty} n_{i,k,x}^{v} = n_{i,k}$ for any $v = 1, \ldots, k$, we have $n_{i,k} = \sum_{x=1}^{\infty} n_{i,k,x}$. Therefore, $\hat{\tau}_i(h)$ can also be written as the function of $n_{i,k,x}$, i.e.,

$$\hat{\tau}_i(h) = \sum_{k=1}^{K_i} \left\{1 - \frac{\binom{K_i-h}{k}}{\binom{K_i}{k}}\right\} \sum_{x=1}^{\infty} n_{i,k,x}, \quad h = 1, 2, ..., K_i.$$

Since $\tau_i(h)$ only admits a closed-form nonparametric estimator for $h = 1, 2, ..., K_i$ and $\hat{g}_i(x)$ is always zero for $x > m$, where m is some arbitrarily large integer, henceforth we consider testing the following hypothesis, which is implied by that in (3.8):

$$H_0 : g_1(x) = g_2(x) \text{ for } x = 1, ..., m, \text{ and } \tau_1(h) = \tau_2(h) \text{ for } h = 2, ..., K,$$

versus

$$H_a : g_1(x) \neq g_2(x) \text{ for some } x \text{ or } \tau_1(h) \neq \tau_2(h) \text{ for some } h, \tag{3.11}$$

where $K = \min(K_1, K_2)$, and m is some arbitrarily large integer. The choice of m will be discussed further in the next section.

Remark 3 *Since H_0 in (3.8) implies H_0 in (3.11), the testing procedures proposed for testing H_0 in (3.8) in the following sections can be also used for testing H_0 in (3.11). When used for testing H_0 in (3.8), the proposed testing procedures can still control the type I error at the nominal level, however, may admit a larger type II error than when used for testing H_0 in (3.11).*

3.5 The proposed test

Denote $\boldsymbol{\eta}_{i,K,m} = (g_i(1), \dots, g_i(m), \tau_i(2), \dots, \tau_i(K))'$. Then the hypothesis testing problem in (3.7) can be written as

$$H_0 : \boldsymbol{\eta}_{1,K,m} = \boldsymbol{\eta}_{2,K,m} \quad \text{versus} \quad H_a : \boldsymbol{\eta}_{1,K,m} \neq \boldsymbol{\eta}_{2,K,m}, \tag{3.12}$$

Let

$$\mathbf{n}_i = (n_{i,1,1}, \dots, n_{i,1,m}, \dots, n_{i,K_i,1}, \dots, n_{i,K_i,m})',$$

$$A1_i = (a_{i,1,1}, \dots, a_{i,1,k}, \dots, a_{i,1,K_i}) \text{ with } a_{i,1,k} = 1 - \binom{K_i - 1}{k} / \binom{K_i}{k},$$

$$A2_i = (a_{i,h,k})_{h=2,k=1}^{K,K_i} \text{ with } a_{i,h,k} = 1 - \binom{K_i - h}{k} / \binom{K_i}{k},$$

$$B1_i = A1_i \bigotimes I_m, \text{ and } B2_i = A2_i \bigotimes \mathbf{1}'_m,$$

where I_m is an m-dimensional identity matrix, $\mathbf{1}_m$ is a vector of m ones, and $\bigotimes$ is the Kronecker product. Define $T_i = \begin{pmatrix} B1_i \\ \\ B2_i \end{pmatrix}$ and $\hat{\boldsymbol{\eta}}_{i,K,m} = T_i \mathbf{n}_i$. It is not difficult to see that $\hat{\boldsymbol{\eta}}_{i,K,m}$ is the estimator of $\boldsymbol{\eta}_{i,K,m}$ developed in the previous section. The following result establishes the asymptotic distribution of $\hat{\boldsymbol{\eta}}_{i,K,m}$.

Theorem 4 *As $c_i \to \infty$, $c_i^{-1/2}(\hat{\boldsymbol{\eta}}_{i,K,m} - \boldsymbol{\eta}_{i,K,m}) \to \mathcal{N}(\mathbf{0}, W_i)$ in distribution, where $W_i = T_i V_i T_i'/c_i$, V_i is the covariance matrix of $\mathbf{n}_i$, and W_i is positive definite.*

Therefore, under the assumption of the independence of the two species assemblages, $\hat{\boldsymbol{\eta}}_{1,K,m} - \hat{\boldsymbol{\eta}}_{2,K,m}$ asymptotically follows $\mathcal{N}(\boldsymbol{\eta}_{1,K,m} - \boldsymbol{\eta}_{2,K,m}, \Sigma_{K,m})$, where $\Sigma_{K,m} = T_1 V_1 T_1' + T_2 V_2 T_2'$.

Based on the above result, a natural test statistic for the hypothesis testing problem in (3.12) is,

$$R_{K,m} = (\hat{\boldsymbol{\eta}}_{1,K,m} - \hat{\boldsymbol{\eta}}_{2,K,m})' \Sigma_{K,m}^{-1} (\hat{\boldsymbol{\eta}}_{1,K,m} - \hat{\boldsymbol{\eta}}_{2,K,m}).$$

It is easy to see that $R_{K,m} \to \chi^2_{m+K-1}$ in distribution under H_0 in (3.12) as $c_i \to \infty$.

In the above test statistic $R_{K,m}$, $\Sigma_{K,m}$ is unknown, and hence should be estimated. Since $\Sigma_{K,m} = T_1 V_1 T_1' + T_2 V_2 T_2'$, in the following we further study the structure of V_i in order to develop an appropriate estimator for $\Sigma_{K,m}$. We first introduce the following notation:

$$\begin{aligned} r_{i,k} &= \int \binom{K_i}{k} \pi^k (1-\pi)^{K_i - k} dG_i(\pi), \\ s_{i,x} &= \int \frac{\exp(-\lambda)}{1-\exp(-\lambda)} \frac{\lambda^x}{x!} dH_i(\lambda), \\ s_{i,x,y} &= \int \left\{ \frac{\exp(-\lambda)}{1-\exp(-\lambda)} \right\}^2 \frac{\lambda^{x+y}}{x!y!} dH_i(\lambda). \end{aligned}$$

Recall that V_i is the covariance matrix of $\mathbf{n}_i = (n_{i,1,1}, \ldots, n_{i,1,m}, \ldots, n_{i,K_i,1}, \ldots, n_{i,K_i,m})'$. The elements of V_i can be specified as follows:

Proposition 5 *(a)* $\mathrm{var}(n_{i,k,x}) = \{c_i r_{i,k} s_{i,x} + (k-1) c_i r_{i,k} s_{i,x,x} - k c_i r_{i,k}^2 s_{i,x}^2\}/k$.

(b) For $x \neq y$*,* $\mathrm{cov}(n_{i,k,x}, n_{i,k,y}) = \{(k-1) c_i r_{i,k} s_{i,x,y} - k c_i r_{i,k}^2 s_{i,x} s_{i,y}\}/k$.

(c) For $k \neq l$*,* $\mathrm{cov}(n_{i,k,x}, n_{i,l,y}) = -c_i r_{i,k} s_{i,x} r_{i,l} s_{i,y}$.

Based on (3.9), $r_{i,k} s_{i,x}$ can be estimated by $n_{i,k,x}/\hat{c}_i$, where $\hat{c}_i$ is some estimator of c_i. Similar to the derivations leading to (3.9), we define $n_{i,k,x,y}^{v_1,v_2}$ as the number of species that appear in exactly k quadrats and appear x times in the v_1-th quadrat and y times in the v_2-th quadrate among those k quadrats. Then, for any $v_1, v_2 = 1, 2, ..., k$, and $v_1 \neq v_2$,

$$E(n_{i,k,x,y}^{v_1,v_2}) = c_i \int \binom{K_i}{k} \pi^k (1-\pi)^{K_i - k} dG_i(\pi) \int \left\{ \frac{\exp(-\lambda)}{1 - \exp(-\lambda)} \right\}^2 \frac{\lambda^{x+y}}{x!y!} dH_i(\lambda).$$

Therefore, $r_{i,k} s_{i,x,y}$ can be estimated by $n_{i,k,x,y}/\hat{c}_i$, where $n_{i,k,x,y} = \sum_{1 \leq v_1 < v_2 \leq K_i} n_{i,k,x,y}^{v_1,v_2} / \binom{k}{2}$.

Plugging the above estimates in V_i, we can obtain an estimator for $\Sigma_{K,m}$. We denote this estimator by $\hat{\Sigma}_{K,m}$.

Proposition 6 $\hat{\Sigma}_{K,m}$ *is positive semi-definite.*

Therefore, our proposed testing procedure is to reject H_0 in (3.12) at a nominal level α if

$$\hat{R}_{K,m} = (\hat{\boldsymbol{\eta}}_{1,K,m} - \hat{\boldsymbol{\eta}}_{2,K,m})' \hat{\Sigma}_{K,m}^{-1} (\hat{\boldsymbol{\eta}}_{1,K,m} - \hat{\boldsymbol{\eta}}_{2,K,m}) > \chi^2_{1-\alpha,m+K-1}, \tag{3.13}$$

where $\chi^2_{1-\alpha,m+K-1}$ is the $(1-\alpha)$ percentile of χ^2_{m+K-1}.

When implementing the above testing procedure, we often encounter the situation that $\hat{\Sigma}_{K,m}$ is singular, and therefore it is impossible to invert $\hat{\Sigma}_{K,m}$. To circumvent this difficulty, we notice that the correlations between the components of $\hat{\boldsymbol{\eta}}_{1,K,m} - \hat{\boldsymbol{\eta}}_{2,K,m}$ are often very large, and the first few principal components of $\hat{\Sigma}_{K,m}$ usually account for the most variability due to its highly correlated components. Therefore, we follow the method proposed in Mao and Li (2009) and only focus on these principal components to test (3.12). To be more specific, consider the eigenvalue decomposition $\hat{\Sigma}_{K,m} = \hat{P}\hat{\Lambda}\hat{P}'$, where $\hat{\Lambda} = \text{diag}\{\hat{\lambda}_1, \hat{\lambda}_2, \ldots, \hat{\lambda}_{m+K-1}\}$, $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_{m+K-1}$ are the eigenvalues of $\hat{\Sigma}_{K,m}$, and $\hat{P}$ is the orthogonal matrix corresponding to the eigenvectors of $\hat{\Sigma}_{K,m}$. Given a constant t in $(0,1)$, say $t = 0.9999$, choose

$$\hat{\nu} = \min\left\{ j : 1 \leq j \leq m+K-1, \sum_{i=1}^{j} \hat{\lambda}_i \geq t \sum_{i=1}^{m+K-1} \hat{\lambda}_i \right\}.$$

Let $\hat{\Lambda}_{\hat{\nu}} = \text{diag}\{\hat{\lambda}_1, \hat{\lambda}_2, \ldots, \hat{\lambda}_{\hat{\nu}}\}$, and $\hat{P}_{\hat{\nu}}$ be the matrix consisting of the first $\hat{\nu}$ columns of $\hat{P}$. Our testing procedure is to reject H_0 in (3.12) at the nominal level α if

$$\hat{R}_{\hat{\nu}} = (\hat{\boldsymbol{\eta}}_{1,K,m} - \hat{\boldsymbol{\eta}}_{2,K,m})' \hat{P}_{\hat{\nu}} \hat{\Lambda}_{\hat{\nu}}^{-1} \hat{P}'_{\hat{\nu}} (\hat{\boldsymbol{\eta}}_{1,K,m} - \hat{\boldsymbol{\eta}}_{2,K,m}) > \chi^2_{1-\alpha,\hat{\nu}}. \tag{3.14}$$

3.5.1 Choice of m

As mentioned earlier, the $n_{i,k,x}$ are always zeros for $x > m$ and so are the $\hat{g}_i(x)$, if we choose m as some arbitrarily large integer. Therefore, they do not contribute to the test statistic $\hat{R}_{K,m}$. In other words, if we choose $m = m_1$ and m_2, and the $n_{i,k,x}$ are all zeros for $x > m_i$ $(i = 1, 2)$, then we always have $\hat{R}_{K,m_1} = \hat{R}_{K,m_2}$. This implies that $\hat{R}_{K,m}$ does not depend on the choice of m as long as it is large enough. However, when implementing the testing procedure (3.13), the threshold $\chi^2_{1-\alpha,m+K-1}$ does depend on the choice of m. Therefore, different choices of m may yield different conclusions. This will not be an issue if we implement the testing procedure (3.14).

Proposition 7 *Testing procedure (3.14) does not depend on the choice of m.*

Due to this invariant property, we choose the testing procedure in (3.14) as our recommended testing procedure, and call it as eigenvalue adjusted (Eva) χ^2 test, similar to the term used in Mao and Li (2009).

3.5.2 Impact of using different $\hat{c}_i$

In the above testing procedure, we need an estimator for c_i. Mao (2007) found that there is no unbiased nonparametric estimate for c_i. However, quite a few lower bound estimators are available in the literature. A popular choice is Chao's lower bound estimator (Chao 1989),

$$\hat{c}_{i,Chao} = n_{i,+} + \frac{(K_i - 1)n_{i,1}^2}{2K_i n_{i,2}},$$

where $n_{i,+} = \sum_{k=1}^{K_i} n_{i,k}$ is the number of species observed in assemblage i. One can also use the trivial upper bound estimator $\hat{c}_i = \infty$ in the calculation of the test statistic in (3.14). Based on our simulation studies, the testing procedure in (3.14) tends to be conservative when the upper bound $\hat{c}_i = \infty$ are used. On the other hand, the testing procedure tends to be liberal if the lower bound estimators are used. The impact of using different $\hat{c}_i$ will be demonstrated further in our simulation studies.

To avoid problems caused by the biased estimates of c_i, alternatively, we can use the bootstrap method to approximate the null distribution of $\hat{R}_{\hat{\nu}}$. More specifically, we first generate the bootstrap resample of $n_{i,+}$, denoted by $n_{i,+}^*$, from Binomial($\hat{c}_{i,Chao}$, $n_{i,+}/\hat{c}_{i,Chao}$). It implies that, in this bootstrap resample, we observe $n_{i,+}^*$ species in assemblage i. Then for species j ($j = 1, ..., n_{i,+}^*$) in assemblage i, randomly choose $k_{i,j}^*$ quadrats out of the K_i quadrats as the quadrats in which species j appears. Here $k_{i,j}^*$ is a random number drawn from a zero-truncated binomial distribution with size K_i and probability $\pi_{i,j}^*$, and $\pi_{i,j}^*$ is drawn from $\hat{Q}_i$, where $\hat{Q}_i$ is the nonparametric maximum likelihood estimator of Q_i with $dQ_i(\pi) = \frac{(1-(1-\pi)^{K_i})dG_i(\pi)}{\int(1-(1-\varpi)^{K_i})dG_i(\varpi)}$ (Mao et al, 2005). Next, for species j ($j = 1, ..., n_{i,+}^*$) in assemblage i, in each one of the $k_{i,j}^*$ quadrats where species j appears, generate the count of species j, i.e., X_{ijk}^*, from a zero-truncated Poisson distribution with the rate parameter $\lambda_{i,j}^*$, where $\lambda_{i,j}^*$ is drawn from $\hat{H}_i$, the nonparametric maximum likelihood estimator of H_i. For the quadrats where species j does not appear, X^*_{ijk} is simply zero. Based on those X^*_{ijk} ($i = 1, 2$, $j = 1, ..., n^*_{i,+}$, $k = 1, ..., K_i$), we can calculate

$$\hat{R}^*_{\hat{\nu}} = (\hat{\boldsymbol{\eta}}^*_{1,K,m} - \hat{\boldsymbol{\eta}}^*_{2,K,m} - (\hat{\boldsymbol{\eta}}_{1,K,m} - \hat{\boldsymbol{\eta}}_{2,K,m}))' \hat{P}^*_{\hat{\nu}} (\hat{\Lambda}^*_{\hat{\nu}})^{-1} (\hat{P}^*_{\hat{\nu}})' (\hat{\boldsymbol{\eta}}^*_{1,K,m} - \hat{\boldsymbol{\eta}}^*_{2,K,m} - (\hat{\boldsymbol{\eta}}_{1,K,m} - \hat{\boldsymbol{\eta}}_{2,K,m})), \tag{3.15}$$

where $\hat{\boldsymbol{\eta}}^*_{1,K,m} - \hat{\boldsymbol{\eta}}^*_{2,K,m}$, $\hat{P}^*_{\hat{\nu}}$ and $\hat{\Lambda}^*_{\hat{\nu}}$ are the counter part of $\hat{\boldsymbol{\eta}}_{1,K,m} - \hat{\boldsymbol{\eta}}_{2,K,m}$, $\hat{P}_{\hat{\nu}}$ and $\hat{\Lambda}_{\hat{\nu}}$, respectively, based on the X^*_{ijk}. We repeat this bootstrap resampling procedure B times and let $\kappa_{1-\alpha}$ be the $(1-\alpha)$ empirical quantile of $\hat{R}^{*1}_{\hat{\nu}}, \ldots, \hat{R}^{*B}_{\hat{\nu}}$, where $\hat{R}^{*j}_{\hat{\nu}}$ is $\hat{R}^*_{\hat{\nu}}$ in (3.15) calculated from the j-th bootstrap resample. Then our Eva-bootstrap testing procedure is to reject H_0 at the level α if $\hat{R}_{\hat{\nu}} > \kappa_{1-\alpha}$.

3.6 The mixture NPMLE of H_i

As we mentioned in the previous section, to use the bootstrap procedure, we need to obtain $\hat{H}_i$, the nonparametric maximum likelihood estimator (NPMLE) of H_i. Therefore in this section, we will discuss how to estimate H_i.

3.6.1 Introduction of the mixture NPMLE

The introduction of the mixture NPMLE is based on Lindsay (1983). First, let us provide some background information about this problem of nonparametric maximum likelihood estimation of the mixing distribution. Suppose $f_\theta(x)$ is a probability density function with parameter θ ($\theta \in \Omega$) and θ is also a random variable with some unknown density function $Q(\theta)$. Then the density function $f_Q(x) = \int f_\theta(x) dQ(\theta)$ is called a mixture density with respect to the mixing distribution Q. If one has observations X_1, X_2, $\ldots$, X_n from the mixture density f_Q (here X_i is not necessarily univariate), the likelihood function of Q will be

$$L(Q) = \prod_{i=1}^{n} f_Q(X_i). \tag{3.16}$$

The problem we want to solve is to find a $\hat{Q}$ which will maximize the likelihood function $L(Q)$. In that sense, $\hat{Q}$ is a good estimation of Q. Let Q be a finite discrete mixing distribution, which takes on values of θ_j with corresponding probability π_j, where $j = 1, \ldots, J$, $\pi_j > 0$, and $\sum_{j=1}^{J} \pi_j = 1$. Therefore we can express Q as $\sum_{j=1}^{J} \pi_j \delta(\theta_j)$.

Suppose X_i's takes on K distinct data values $y_1, y_2, \ldots, y_K$. Given observations $x_1, \ldots, x_n$, let n_k be the number of x's which take on value y_k. Lindsay (1983) defined $\mathbf{f}_\theta = (f_\theta(y_1), \ldots, f_\theta(y_K))$ and $\mathbf{f}_Q = (f_Q(y_1), \ldots, f_Q(y_K))$. Then the log likelihood function becomes

$$\phi(\mathbf{f}_Q) = logL(Q) = \sum_{k=1}^{K} n_k log f_Q(y_k) = \sum_{k=1}^{K} n_k log(\sum_{j=1}^{J} \pi_j f_{\theta_j}(y_k)). \tag{3.17}$$

Let $\Gamma = \{\mathbf{f}_\theta : \theta \in \Omega\}$. Maximizing $L(Q)$ over Q is equivalent to maximizing $\phi(\mathbf{f}_Q)$ over $\mathbf{f}_Q$ in the K-dimensional set conv(Γ), since any convex combination of elements of Γ can be written as $\sum \pi_j \mathbf{f}_{\theta_j}$ (, and therefore conv(Γ)= $\{\mathbf{f}_Q : Q \in \mathcal{M}\}$, where $\mathcal{M}$ is the class of all probability measures on Ω). Lindsay (1983) showed that the above problem is equivalent to a convex optimization problem and there are many similarities between general mixture theory and the theory of optimal design. Furthermore, Lindsay (1983) showed that $\hat{Q}$ maximizes $L(Q)$ if and only if $\sup_\theta D(\theta, \hat{Q}) = 0$, and the support of $\hat{Q}$ is contained in the set of θ for which $D(\theta, \hat{Q}) = 0$, where $D(\theta, Q) = \Phi(\mathbf{f}_\theta; \mathbf{f}_Q) = \sum n_k \{\frac{f_\theta(y_k)}{f_Q(y_k)} - 1\}$.

Remark 8 *Instead of continuous distribution, a finite, discrete distribution is often used as the estimate of the mixing or latent distribution. Lindsay (1995, p.143) mentioned, in most cases, the level of information about the mixing distribution is too small for* *discrimination about the form of the distribution. Further, our goal in this thesis is generating bootstrap samples from an estimate of H instead of obtaining an accurate estimation of H. Therefore, in the next, we will discuss the computing of the finite, discrete MLE of H. But for the information, there are ways to obtain smooth estimators of a mixing distribution (Lindsay, 1995).*

3.6.2 The computing of the Nonparametric MLE of H

To use bootstrap sampling, we need to estimate H_i, the species abundance rate distribution. We want to find the nonparametric maximum likelihood estimator $\hat{H}_i$. The computing of mixture NPMLE is usually based on iterating algorithms such as EM algorithm. The following procedure is developed according to Lindsey (1995). The likelihood function involving H_i is

$$\prod_{j=1}^{c_i} \int (\frac{\exp(-\lambda)}{1-\exp(-\lambda)})^{k_{i,j}} \lambda^{x_{i,j,s_1}+\ldots+x_{i,j,s_t}+\ldots+x_{i,j,s_{k_{i,j}}}} dH_i(\lambda) * C_0 \tag{3.18}$$

,where C_0 is some constant, $k_{i,j}$ is the exact number of quadrats in which the j-th species from i-th assemblage appears and $x_{i,j,s_t}(t = 1, \ldots, k_{i,j})$ is the number of times that the j-th species appears in t-th quadrat while it appears in exactly $k_{i,j}$ quadrats. s_t is used to represent the t-th quadrat among the $k_{i,j}$ quadrats and $1 \le s_t \le K_i$. (3.18) is equivalent to

$$\prod_{k=1}^{K_i} \prod_{x=1}^{\infty} (\int (\frac{\exp(-\lambda)}{1-\exp(-\lambda)})^k \lambda^x dH_i(\lambda))^{N_{k,x}} * C_0$$

where $N_{k,x}$ is the number of species which appear in exactly k quadrats with a total count of x. Further, the log-likelihood is

$$\sum_{k=1}^{K_i} \sum_{x=1}^{\infty} N_{k,x} log\{\int (\frac{\exp(-\lambda)}{1-\exp(-\lambda)})^k \lambda^x dH_i(\lambda)\} + logC_0.$$

3.6.2.1 EM algorithm

Suppose $H_i(\lambda)$ has S support points $\{\lambda_s\}_{s=1}^S$ with corresponding weights $\{w_s\}_{s=1}^S$. Define a latent variable $N_{k,s,x}$, which is the number of species that appear in exactly k quadrats with a total count of x with λ_s as the parameter of the truncated poisson distribution. Now we consider the latent variable $N_{k,s,x}$, then the log-likelihood will be

$$L_H = \sum_{k=1}^{K_i} \sum_{s=1}^{S} \sum_{x=1}^{\infty} N_{k,s,x} log\{(\frac{\exp(-\lambda_s)}{1-\exp(-\lambda_s)})^k \lambda_s^x w_s\} + logC_0 \tag{3.19}$$

When H_i has fixed support points, S. According to Dempster et al. (1977), EM algorithm can be used to find the parameters maximizing the log-likelihood function. Since truncated Poisson distribution belongs to exponential family, the EM algorithm can be implemented easily. In the E-step, given the current estimate of $\{\hat{\lambda}_s\}_{s=1}^S$ and $\{\hat{w}_s\}_{s=1}^S$, one can estimate $\hat{N}_{k,s,x}$ by

$$E(N_{k,s,x}|N_{k,x}) = N_{k,x} \frac{w_s(\frac{\exp(-\lambda_s)}{1-\exp(-\lambda_s)})^j \lambda_s^x}{\sum_{s=1}^S w_s(\frac{\exp(-\lambda_s)}{1-\exp(-\lambda_s)})^j \lambda_s^x}. \tag{3.20}$$

In the M-step, let

$$\frac{\partial L_H}{\partial w_s} = 0 \tag{3.21}$$

$$\frac{\partial L_H}{\partial \lambda_s} = 0 \tag{3.22}$$

$$\sum_{s=1}^{S} w_s = 1 \tag{3.23}$$

Solve for w_s and λ_s from the above equations, we have

$$\hat{w}_s = \frac{\sum_{k=1}^{K_i} \sum_{x=1}^{\infty} \hat{N}_{k,s,x}}{\sum_{s=1}^{S} \sum_{k=1}^{K_i} \sum_{x=1}^{\infty} \hat{N}_{k,s,x}}, \tag{3.24}$$

$\hat{\lambda}_s$ is the solution of the following equation

$$\frac{\lambda_s}{1-\exp(-\lambda_s)} = \frac{\sum_{k=1}^{K_i} \sum_{x=1}^{\infty} x\hat{N}_{k,s,x}}{\sum_{k=1}^{K_i} \sum_{x=1}^{\infty} k\hat{N}_{k,s,x}}. \tag{3.25}$$

It can be verified that $\partial^2 L_H/\partial {w_s}^2 < 0$ and $\partial^2 L_H/\partial {\lambda_s}^2 < 0$, therefore w_s and λ_s maximize (3.19). (3.25) can be solved using Newtonian algorithm. One should iterate between the E step and M step until the absolute difference of L_H from two iterations next to each other is smaller than some predefined constant. In our simulation and real application, that constant is chosen to be 10^{-8}.

The above procedure is estimating the distribution of λ given a fixed number of support points, S. However, when one is estimating the distribution of λ from the abundance data, he does not have any prior information on λ, therefore S should also be estimated from data. In this case, one can use a integrated procedure of the vertex exchange method and the em algorithm. The details are described in Mao (2008).

3.7 Impact of using different h_0

So far we have discussed how to develop the test for comparing species assemblages when h_0 is chosen to be 1. Now let h_0 be any positive integer from 1 to K, denote $\boldsymbol{\eta}_{i,h_0,K,m} = (g_i(h_0,1),\ldots,g_i(h_0,m),\tau_i(1),\tau_i(2),\ldots,\tau(h_0-1),\tau(h_0+1),\ldots,\tau_i(K))'$. For any chosen h_0, through a series of similar arguments, we will reach the conclusion that, to solve the original hypothesis testing problem (3.7) we can consider the following testing problem,

$$H_0 : \boldsymbol{\eta}_{1,h_0,K,m} = \boldsymbol{\eta}_{2,h_0,K,m} \text{ versus } H_1 : \boldsymbol{\eta}_{1,h_0,K,m} \neq \boldsymbol{\eta}_{2,h_0,K,m}. \tag{3.26}$$

Let

$$A1_{i,h_0} = (a_{i,h_0,1},\ldots,a_{i,h_0,k},\ldots,a_{i,h_0,K_i}) \text{ with } a_{i,h_0,k} = 1 - \binom{K_i - h_0}{k} / \binom{K_i}{k},$$

$$A2_{i,h_0} = (a_{i,h,k})_{h=1\&h\neq h_0,k=1}^{K,K_i} \text{ with } a_{i,h,k} = 1 - \binom{K_i - h}{k} / \binom{K_i}{k},$$

$$B1_{i,h_0} = A1_{i,h_0} \bigotimes I_m, \text{ and } B2_{i,h_0} = A2_{i,h_0} \bigotimes \mathbf{1}'_m.$$

Define $T_{i,h_0} = \begin{pmatrix} B1_{i,h_0} \\ \\ B2_{i,h_0} \end{pmatrix}$ and $\hat{\boldsymbol{\eta}}_{i,h_0,K,m} = T_{i,h_0}\boldsymbol{n}_i$, it is easy to see that $\hat{\boldsymbol{\eta}}_{i,h_0,K,m}$ is an unbiased estimator of $\boldsymbol{\eta}_{i,h_0,K,m}$. A natural test statistic for the hypothesis testing problem (3.26) is

$$R_{h_0,K,m} = (\hat{\boldsymbol{\eta}}_{1,h_0,K,m} - \hat{\boldsymbol{\eta}}_{2,h_0,K,m})'\Sigma^{-1}_{h_0,K,m}(\hat{\boldsymbol{\eta}}_{1,h_0,K,m} - \hat{\boldsymbol{\eta}}_{2,h_0,K,m}).$$

It is easy to prove that as c_i goes to ∞, $R_{h_0,K,m}$ follows χ^2 distribution with degree of freedom $m+K-1$ and noncentral parameter $(\boldsymbol{\eta}_{1,h_0,K,m} - \boldsymbol{\eta}_{2,h_0,K,m})'\Sigma^{-1}_{h_0,K,m}(\boldsymbol{\eta}_{1,h_0,K,m} - \boldsymbol{\eta}_{2,h_0,K,m})$, where $\Sigma_{h_0,K,m} = T_{1,h_0}V_1T'_{1,h_0} + T_{2,h_0}V_2T'_{2,h_0}$. So now for the original hypothesis testing problem we have a series of test statistics which are functions of h_0. The natural question to ask is whether the choice of h_0 will have effect on the power of the test. Given H_0 is false, asymptotically, the power of the test is determined by the noncentral parameter of the test statistic. In the following, we will present numerical results to demonstrate the relationship between h_0 and the power of the test.

Let $c = 500$ and $c^* = 400$; G is a Beta distribution with shape parameters 1 and 20, and G^* is a Beta distribution with shape parameters 1 and 15; H is a discrete distribution with support points 1 and 3, and corresponding weights 0.7 and 0.3, and H^* is the same as H but with the second support point taking the value of 2. We consider the comparison of two species assemblages with number of quadrats $K_1 = K_2 = 5$, and $m = 10$. Given the values of c_i, G_i, and H_i ($i = 1$ and 2), the noncentral parameter $(\boldsymbol{\eta}_{1,h_0,K,m} - \boldsymbol{\eta}_{2,h_0,K,m})'\Sigma^{-1}_{h_0,K,m}(\boldsymbol{\eta}_{1,h_0,K,m} - \boldsymbol{\eta}_{2,h_0,K,m})$ can be numerically computed. The results are listed in table 3.1 and table 3.2.

From table 3.1, we can see the noncentral parameters are the same for $h_0 = 1, 2, \ldots, 5$, which means h_0 has no effect on the power of the test when $H_1 = H_2$. From

Table 3.1: The value of the noncentral parameter of χ^2 distribution the test statistic asymptotically follows under different h_0 given H_0 is false and $H_1 = H_2 = H$. One case is represented by (c_1, G_1, H_1) vs (c_2, G_2, H_2).

(c_1, G_1, H_1) vs (c_2, G_2, H_2)	$h_0 = 1$	$h_0 = 2$	$h_0 = 3$	$h_0 = 4$	$h_0 = 5$
(c, G, H) vs (c^*, G, H)	2.777	2.777	2.777	2.777	2.777
(c, G, H) vs (c, G^*, H)	0.733	0.733	0.733	0.733	0.733
(c, G, H) vs (c^*, G^*, H)	4.394	4.394	4.394	4.394	4.394

Table 3.2: The value of the noncentral parameter of χ^2 distribution the test statistic asymptotically follows under different h_0 given H_0 is false and $H_1 \neq H_2$. One case is represented by (c_1, G_1, H_1) vs (c_2, G_2, H_2). The number in the parenthesis is the corresponding power of the test.

(c_1, G_1, H_1) vs (c_2, G_2, H_2)	$h_0 = 1$	$h_0 = 2$	$h_0 = 3$	$h_0 = 4$	$h_0 = 5$
(c, G, H) vs (c, G, H^*)	3.095 (0.1476)	3.127 (0.1489)	3.117 (0.1485)	3.071 (0.1467)	2.996 (0.1438)
(c, G, H) vs (c^*, G, H^*)	5.364 (0.2456)	5.393 (0.2470)	5.386 (0.2466)	5.349 (0.2449)	5.287 (0.2420)
(c, G, H) vs (c, G^*, H^*)	8.052 (0.3767)	8.082 (0.37822)	8.058 (0.37702)	7.991 (0.37368)	7.889 (0.3686)
(c, G, H) vs (c^*, G^*, H^*)	3.893 (0.1801)	3.914 (0.1809)	3.888 (0.1798)	3.826 (0.1772)	3.734 (0.1734)

table 3.2, we see different h_0 yields different noncentral parameter. When $H_1 \neq H_2$, h_0 does have effect on the power of the test. However, the largest power difference for each case is within 0.01. Based on the above simulation results, we tend to believe that if H_i's are the same, asymptotically, the power of the test does not depend on h_0, and even if H_i's are different, the choice of h_0 has very limited effect on the power of the test. Currently, we are not able to provide further evidence for our claim, but in the future we will try to provide theoretical proof or simulation results using the bootstrap procedure.

3.8 Multiple sample versions of the tests

The testing procedure we proposed here is not limited to comparison of two species assemblages. It can be extended to the comparison of L ($L \geq 2$) species assemblages by following the scheme proposed by Mao and Li (2009). Following the notation used in the previous sections, we define, for $i = 1, \ldots, L$,

$$\boldsymbol{\eta}_{i,K,m} = (g_i(1), \ldots, g_i(m), \tau_i(2), \ldots, \tau_i(K))', \tag{3.27}$$

where m is some arbitrarily large integer and $K = min(K_1, \ldots, K_L)$. Then comparing L species assemblages can be formulated as the following hypothesis testing problem:

$$H_0 : \boldsymbol{\eta}_{1,K,m} = \boldsymbol{\eta}_{2,K,m} = \cdots = \boldsymbol{\eta}_{L,K,m}$$

versus

$$H_a : \boldsymbol{\eta}_{i,K,m} \neq \boldsymbol{\eta}_{j,K,m} \text{ for some } i \neq j.$$

Define $\boldsymbol{d} = (\boldsymbol{\eta}'_{1,K,m} - \boldsymbol{\eta}'_{L,K,m}, \boldsymbol{\eta}'_{2,K,m} - \boldsymbol{\eta}'_{L,K,m}, \ldots, \boldsymbol{\eta}'_{L-1,K,m} - \boldsymbol{\eta}'_{L,K,m})'$. The above hypothesis testing problem is equivalent to

$$H_0 : \boldsymbol{d} = \mathbf{0}_{(L-1)(m-1+K)} \text{ versus } H_1 : \boldsymbol{d} \neq \mathbf{0}_{(L-1)(m-1+K)},$$

where $\mathbf{0}_p$ is a vector of p zeros. Denote the estimate of $\boldsymbol{\eta}_{i,K,m}$ by $\hat{\boldsymbol{\eta}}_{i,K,m}$, $i = 1, \ldots, L$. A natural estimate for $\boldsymbol{d}$ can be obtained by $\hat{\boldsymbol{d}} = (\hat{\boldsymbol{\eta}}'_{1,K,m} - \hat{\boldsymbol{\eta}}'_{L,K,m}, \hat{\boldsymbol{\eta}}'_{2,K,m} - \hat{\boldsymbol{\eta}}'_{L,K,m}, \ldots, \hat{\boldsymbol{\eta}}'_{L-1,K,m} - \hat{\boldsymbol{\eta}}'_{L,K,m})'$. One can easily prove that: asymptotically, $\hat{\boldsymbol{d}}$ follows $\mathcal{N}(\boldsymbol{d}, \Sigma_L)$, where

$$\Sigma_L = \bigoplus_{l=1,2,\ldots,L-1} W_l + (\mathbf{1}_{L-1}\mathbf{1}'_{L-1}) \bigotimes W_L, \tag{3.28}$$

W_l is the covariance matrix of $\boldsymbol{\eta}_{l,K,m}$, $\bigoplus$ is the direct sum. We can obtain $\hat{\Sigma}_L$, the estimate of Σ_L, by plugging in (3.28) the estimates of W_i's. Under H_0, as c_i goes to infinity, $\hat{\boldsymbol{d}}'\hat{\Sigma}_L^{-1}\hat{\boldsymbol{d}}$ converges to $\chi^2_{(L-1)(m+K-1)}$, a χ^2 distribution with degree of freedom $(L-1)(m+K-1)$. One should reject H_0 if $\hat{\boldsymbol{d}}'\hat{\Sigma}_L^{-1}\hat{\boldsymbol{d}} > \chi^2_{1-\alpha,(L-1)(m+K-1)}$ at the significance level of α, where $\chi^2_{1-\alpha,(L-1)(m+K-1)}$ is the $1-\alpha$ percentile of $\chi^2_{(L-1)(m+K-1)}$. Based on this χ^2 test, we can also develop its Eva-χ^2 test and Eva-bootstrap test for the L species assemblage comparison problem. Similar to the two species assemblage case, the Eva-χ^2 test does not depend on the choice of m, and the Eva-bootstrap test is not affected by the choice of $\hat{c}_i$ and is able to achieve the desired nominal type I error.

3.9 Simulation study

3.9.1 Type I error study of the two sample tests

In this section, we report a simulation study to assess the type I error of our two sample Eva-χ^2 test. The study consists of 36 simulation settings, which are determined by the following four factors: the total number of species $c_i (c_1 = c_2 = 500$ or 2000), the number of quadrats K_i ($K_1 = K_2 = 50$ or 150), the mixing distribution G_i for π_{ij} ($G_1 = G_2 = \mathscr{B}$, logit$\mathscr{N}$ or $\mathscr{D}_G$), and the mixing distribution H_i for λ_{ij} ($H_1 = H_2 = \mathscr{G}$, log$\mathscr{N}$ or $\mathscr{D}_H$), where $\mathscr{B}$ is the beta distribution with shape parameters 1 and 20, logit$\mathscr{N}$ is obtained by letting $\log(\pi/(1-\pi))$ follow a normal distribution with mean -4 and variance 2, $\mathscr{D}_G$ is discrete with support points 0.01, 0.05, 0.10, and 0.15 and corresponding weights 0.65, 0.20, 0.10, and 0.05, $\mathscr{G}$ is the gamma distribution with shape parameter 1 and scale parameter 2 right truncated at 20, log$\mathscr{N}$ is the lognormal distribution with mean 0 and variance 1 right truncated at 20, and $\mathscr{D}_H$ is discrete with support points 1, 2, 5, and 10 and corresponding weights 0.65, 0.20, 0.10, and 0.05.

To investigate the effect of different estimators for c_i on the type I error of our test, we consider $\hat{c}_i = \hat{c}_{i,Chao}$, and ∞ in the calculation of $\hat{R}_{\hat{\nu}}$ in (3.14). To benchmark the performance, we also include the type I errors of our proposed test when $\hat{c}_i = c_i$ is used in the test. In all the tests, we use $t = 0.9999$ in the eigenvalue decomposition to choose $\hat{\nu}$ and the nominal size of the test α is set at 0.05. Table 3.3-3.4 summarizes the simulated type I errors of the Eva-χ^2 test based on 500 samples. From Table 1, we can see that the estimator of c_i has an important impact on the type I errors of the Eva-χ^2

Table 3.3: The type I error of the proposed tests given different $\hat{c}_i$ being used. One case is represented by (c_i, K_i, G_i, H_i) such that $c_1 = c_2 = 500$ or 2000, $K_1 = K_2 = 50$, $G_1 = G_2 = \mathscr{B}$, logit$\mathscr{N}$ or $\mathscr{D}_G$, and $H_1 = H_2 = \mathscr{G}$, log$\mathscr{N}$ or $\mathscr{D}_H$.

	Eva-χ^2			Eva-bootstrap
(c_i,K_i,G_i,H_i)	$\hat{c}_i = c_i$	$\hat{c}_i = \hat{c}_{Chao}$	$\hat{c}_i = \infty$	$\hat{c}_i = \infty$
$(500,50,\mathscr{B},\mathscr{G})$	0.052	0.090	0.018	0.044
$(500,50,\mathscr{B},\log\mathscr{N})$	0.056	0.100	0.022	0.054
$(500,50,\mathscr{B},\mathscr{D}_H)$	0.046	0.072	0.022	0.046
$(500,50,\text{logit}\mathscr{N},\mathscr{G})$	0.052	0.088	0.032	0.048
$(500,50,\text{logit}\mathscr{N},\log\mathscr{N})$	0.050	0.086	0.028	0.048
$(500,50,\text{logit}\mathscr{N},\mathscr{D}_H)$	0.066	0.098	0.044	0.070
$(500,50,\mathscr{D}_G,\mathscr{G})$	0.040	0.050	0.014	0.040
$(500,50,\mathscr{D}_G,\log\mathscr{N})$	0.066	0.076	0.044	0.062
$(500,50,\mathscr{D}_G,\mathscr{D}_H)$	0.058	0.072	0.028	0.050
$(500,150,\mathscr{B},\mathscr{G})$	0.036	0.094	0.024	0.050
$(500,150,\mathscr{B},\log\mathscr{N})$	0.050	0.110	0.034	0.058
$(500,150,\mathscr{B},\mathscr{D}_H)$	0.064	0.086	0.014	0.048
$(500,150,\text{logit}\mathscr{N},\mathscr{G})$	0.042	0.068	0.026	0.044
$(500,150,\text{logit}\mathscr{N},\log\mathscr{N})$	0.068	0.100	0.038	0.064
$(500,150,\text{logit}\mathscr{N},\mathscr{D}_H)$	0.038	0.080	0.006	0.048
$(500,150,\mathscr{D}_G,\mathscr{G})$	0.048	0.052	0.020	0.040
$(500,150,\mathscr{D}_G,\log\mathscr{N})$	0.050	0.052	0.032	0.048
$(500,150,\mathscr{D}_G,\mathscr{D}_H)$	0.056	0.062	0.034	0.052

test. If we could have high-quality estimators for the c_i, the type I errors of the Eva-χ^2 test would approach its nominal level. If the c_i are underestimated, the Eva-χ^2 test would be liberal. Our Eva-χ^2 test would be conservative if $\hat{c}_i = \infty$ is used. Table 3.3-3.4 also include the type I errors of the Eva-bootstrap test. For computation simplicity, we only consider the Eva-bootstrap test with $\hat{c}_i = \infty$ being used in the calculation of $\hat{R}_{\hat{\nu}}$. The number of bootstrap resamples is set $B = 500$. As we can see from Table 3.3-3.4, the Eva-bootstrap test corrects the conservativeness of its corresponding Eva-χ^2 test and approximately achieves its nominal type I error.

Remark 9 *We have presented the simulation results of the type I error of Eva-bootstrap when using ∞ as $\hat{c}_i$. Alternatively, we can use $\hat{c}_{i,Chao}$ estimated from the observed data (instead of bootstrap sampled data) as $\hat{c}_i$. In the following, we list some simulation*

Table 3.4: The type I error of the proposed tests given different $\hat{c}_i$ being used. One case is represented by (c_i, K_i, G_i, H_i) such that $c_1 = c_2 = 500$ or 2000, $K_1 = K_2 = 150$, $G_1 = G_2 = \mathscr{B}$, logit$\mathscr{N}$ or $\mathscr{D}_G$, and $H_1 = H_2 = \mathscr{G}$, log$\mathscr{N}$ or $\mathscr{D}_H$.

	Eva-χ^2			Eva-bootstrap
(c_i,K_i,G_i,H_i)	$\hat{c}_i = c_i$	$\hat{c}_i = \hat{c}_{Chao}$	$\hat{c}_i = \infty$	$\hat{c}_i = \infty$
(2000,50,$\mathscr{B}$,$\mathscr{G}$)	0.056	0.090	0.020	0.054
(2000,50,$\mathscr{B}$,log$\mathscr{N}$)	0.056	0.092	0.012	0.044
(2000,50,$\mathscr{B}$,$\mathscr{D}_H$)	0.048	0.086	0.026	0.050
(2000,50,logit$\mathscr{N}$,$\mathscr{G}$)	0.058	0.076	0.046	0.042
(2000,50,logit$\mathscr{N}$,log$\mathscr{N}$)	0.036	0.062	0.036	0.058
(2000,50,logit$\mathscr{N}$,$\mathscr{D}_H$)	0.046	0.068	0.026	0.048
(2000,50,$\mathscr{D}_G$,$\mathscr{G}$)	0.07	0.082	0.032	0.072
(2000,50,$\mathscr{D}_G$,log$\mathscr{N}$)	0.042	0.050	0.016	0.046
(2000,50,$\mathscr{D}_G$,$\mathscr{D}_H$)	0.050	0.068	0.026	0.046
(2000,150,$\mathscr{B}$,$\mathscr{G}$)	0.038	0.088	0.012	0.038
(2000,150,$\mathscr{B}$,log$\mathscr{N}$)	0.058	0.100	0.046	0.068
(2000,150,$\mathscr{B}$,$\mathscr{D}_H$)	0.040	0.112	0.016	0.038
(2000,150,logit$\mathscr{N}$,$\mathscr{G}$)	0.042	0.062	0.030	0.058
(2000,150,logit$\mathscr{N}$,log$\mathscr{N}$)	0.044	0.080	0.034	0.050
(2000,150,logit$\mathscr{N}$,$\mathscr{D}_H$)	0.070	0.100	0.032	0.058
(2000,150,$\mathscr{D}_G$,$\mathscr{G}$)	0.058	0.056	0.034	0.052
(2000,150,$\mathscr{D}_G$,log$\mathscr{N}$)	0.048	0.052	0.016	0.040
(2000,150,$\mathscr{D}_G$,$\mathscr{D}_H$)	0.042	0.048	0.022	0.040

Table 3.5: The type I error of the proposed tests given different $\hat{c}_i$ being used. One case is represented by (c_i, K_i, G_i, H_i) such that $c_1 = c_2 = 500$ or 2000, $K_1 = K_2 = 50$, $G_1 = G_2 = \mathscr{B}$, logit$\mathscr{N}$ or $\mathscr{D}_G$, and $H_1 = H_2 = \mathscr{G}$, log$\mathscr{N}$ or $\mathscr{D}_H$.

	Eva-χ^2	Eva-bootstrap
$(c_i$,K_i,G_i,$H_i)$	$\hat{c}_i = c_i$	$\hat{c}_i = \hat{c}_{i,Chao}$
(500,50,$\mathscr{B}$,$\mathscr{G}$)	0.060	0.048
(500,50,$\mathscr{B}$,log$\mathscr{N}$)	0.050	0.040
(500,50,$\mathscr{B}$,$\mathscr{D}_H$)	0.046	0.046
(500,50,logit$\mathscr{N}$,$\mathscr{G}$)	0.034	0.052
(500,50,logit$\mathscr{N}$,log$\mathscr{N}$)	0.048	0.064
(500,50,logit$\mathscr{N}$,$\mathscr{D}_H$)	0.046	0.050
(500,50,$\mathscr{D}_G$,$\mathscr{G}$)	0.046	0.048
(500,50,$\mathscr{D}_G$,log$\mathscr{N}$)	0.062	0.060
(500,50,$\mathscr{D}_G$,$\mathscr{D}_H$)	0.064	0.060
(500,150,$\mathscr{B}$,$\mathscr{G}$)	0.026	0.042
(500,150,$\mathscr{B}$,log$\mathscr{N}$)	0.054	0.066
(500,150,$\mathscr{B}$,$\mathscr{D}_H$)	0.050	0.084
(500,150,logit$\mathscr{N}$,$\mathscr{G}$)	0.052	0.048
(500,150,logit$\mathscr{N}$,log$\mathscr{N}$)	0.062	0.064
(500,150,logit$\mathscr{N}$,$\mathscr{D}_H$)	0.050	0.068
(500,150,$\mathscr{D}_G$,$\mathscr{G}$)	0.040	0.038
(500,150,$\mathscr{D}_G$,log$\mathscr{N}$)	0.048	0.038
(500,150,$\mathscr{D}_G$,$\mathscr{D}_H$)	0.044	0.042

results. For convenience, here we only list the results from Eva-χ^2 when $\hat{c}_i = c_i$ and Eva-bootstrap when $\hat{c}_i = \hat{c}_{i,Chao}$ under the condition when $c_1 = c_2 = 500$ for comparison.

From table 3.5, we can see that the type I error of Eva-bootstrap when $\hat{c}_i = \hat{c}_{i,Chao}$ is also at nominal level. Therefore Eva-bootstrap with $\hat{c}_i = \hat{c}_{i,Chao}$ is another test one can rely on in real practice besides Eva-bootstrap with $\hat{c}_i = \infty$. Next, the question one would naturally ask is that which of the two tests is more powerful. We are not able to answer this question. But we will show some simulation results in the following power comparison study.

Table 3.6: The type I error of the proposed tests when there are 4 species assemblages to compare. One case is represented by (c_i, K_i, G_i, H_i) such that $c_i = 500$, $K_i = 150$, $G_i = \mathscr{B}$, logit$\mathscr{N}$ or $\mathscr{D}_G$, and $H_i = \mathscr{G}$, log$\mathscr{N}$ or $\mathscr{D}_H$.

	Eva-χ^2			Eva-bootstrap
(c_i,K_i,G_i,H_i)	$\hat{c}_i = c_i$	$\hat{c}_i = \hat{c}_{i,Chao}$	$\hat{c}_i = \infty$	$\hat{c}_i = \infty$
(500,150,$\mathscr{B}$,$\mathscr{G}$)	0.064	0.182	0.012	0.060
(500,150,$\mathscr{B}$,log$\mathscr{N}$)	0.066	0.176	0.018	0.050
(500,150,$\mathscr{B}$,$\mathscr{D}_H$)	0.044	0.152	0.010	0.046
(500,150,logit$\mathscr{N}$,$\mathscr{G}$)	0.056	0.138	0.026	0.052
(500,150,logit$\mathscr{N}$,log$\mathscr{N}$)	0.058	0.120	0.012	0.046
(500,150,logit$\mathscr{N}$,$\mathscr{D}_H$)	0.056	0.126	0.024	0.048
(500,150,$\mathscr{D}_G$,$\mathscr{G}$)	0.044	0.050	0.006	0.032
(500,150,$\mathscr{D}_G$,log$\mathscr{N}$)	0.068	0.076	0.012	0.056
(500,150,$\mathscr{D}_G$,$\mathscr{D}_H$)	0.050	0.058	0.014	0.040

3.9.2 Type I error study of the multiple sample tests

In this section, we report a simulation study to assess the type I error of our multiple sample Eva-χ^2 test. We present the case when $L = 4$, which means there are 4 groups of species assemblages. For simplicity, we only present the case when c_i's are 500 and K_i's are 150. The mixing distribution G_i's and H_i's take on distributions we list in the earlier section. We consider the Eva-χ^2 tests with $\hat{c}_i = c_i$, $\hat{c}_{i,Chao}$, and ∞, and the Eva-bootstrap test with $\hat{c}_i = \infty$. In all the tests, we use $t = 0.9999$ in the eigenvalue decomposition to choose $\hat{\nu}$ and the nominal size of the test α is set at 0.05. Table 3.6 summarizes the simulated type I error based on 500 samples. Similar as the two sample test, we can see that the Eva-χ^2 test with $\hat{c}_i = \hat{c}_i$ yields a liberal result, while Eva-χ^2 test with $\hat{c}_i = \infty$ yields a conservative result. The Eva-bootstrap test with $\hat{c}_i = \infty$ yields a type I error at nominal level.

3.9.3 Power comparison between abundance based test and incidence based test

In this section, we refer to the zero-inflated Poisson mixture model based bootstrap test as abundance based test and refer to the binomial mixture model based bootstrap test as incidence based test. From the construction of the two tests, intuitively, one will think that incidence based test will not be able to detect difference on H_i's. Also, it is important to examine how abundance based test performs in detecting difference on c_i's and G_i's. Therefore we perform the following simulation study of power comparison of abundance based test and incidence based test.

In the following study, the number of quadrats for the two samples are $K_1 = K_2 = 50$. When it comes to power simulation, we know that, for two samples, c_i's, G_i's, and H_i's can take on various values or distributions, respectively. There are too many scenarios, and it is impossible to perform an exhaustive comparison . For simplicity, we restrict the values of c_i's, G_i's, and H_i's to be the ones we list in the following. $c = 500$ and $c^* = 450$, and they stand for total number of species; G is a discrete distribution with support points 0.02, 0.1, 0.2, and 0.3 and corresponding weights 0.65, 0.20, 0.10, and 0.05, G^* is similar to G but with first support point equal to 0.025 instead of 0.02; H is a discrete distribution with support points 1, 2, 5, and 10 and corresponding weights 0.65, 0.20, 0.10, and 0.05, H^* is similar to H but with the first support point equal to 4 instead of 1. Therefore there are a total of 13 scenarios. In all the tests, we use $t = 0.9999$ in the eigenvalue decomposition to choose $\hat{\nu}$ and the nominal size of the test α is set at 0.05. The simulation results based on 500 samples are listed in table 3.7.

From table 3.7, we can see that, when only H_i's are different, the incidence based test has a power of 0.054, however, the abundance based test has a power of

Table 3.7: The power comparison of abundance based and incidence based tests under different scenarios. One case is represented by (c_1, G_1, H_1) vs (c_2, G_2, H_2).

	Eva-bootstrap	
(c_1, G_1, H_1) vs (c_2, G_2, H_2)	abundance	incidence
(c, G, H) vs (c^*, G, H)	0.144	0.138
(c, G, H) vs (c, G^*, H)	0.256	0.260
(c, G, H) vs (c, G, H^*)	0.480	0.054
(c, G, H) vs (c^*, G^*, H)	0.202	0.204
(c, G^*, H) vs (c^*, G, H)	0.716	0.706
(c, G, H) vs (c^*, G, H^*)	0.546	0.138
(c, G, H^*) vs (c^*, G, H)	0.524	0.132
(c, G, H) vs (c, G^*, H^*)	0.456	0.234
(c, G, H^*) vs (c, G^*, H)	0.546	0.284
(c, G, H) vs (c^*, G^*, H^*)	0.406	0.200
(c, G, H^*) vs (c^*, G^*, H)	0.496	0.182
(c, G^*, H) vs (c^*, G, H^*)	0.824	0.696
(c, G^*, H^*) vs (c^*, G, H)	0.764	0.702

0.48. The poor performance of incidence based test in detecting difference of H_i's is a result of ignoring the information of H_i's in the construction of the test. For other cases when H_i's are the same but c_i's and G_i's are different, we can see there is no obvious difference between the powers of the abundance based test and incidence based test. In general, we can see that abundance based test performs better than incidence based test by considering the information contained in H_i's.

3.9.4 Power comparison of using different c_i

Previously, we discussed that when one implements the two sample Eva-boostrap test, one can use either $\hat{c}_i = \infty$ or $\hat{c}_i = \hat{c}_{i,Chao}$. Both of the two tests are valid with type I errors at nominal level. In this section, we report a simulation study to assess the power of the two tests. When it comes to power simulation study, we know that, for two samples, c_i's, G_i's, and H_i's can take on various values or distributions, respectively. There are too many scenarios, and it is impossible to perform an exhaustive comparison

Table 3.8: The power comparison of two sample Eva-bootstrap tests with $\hat{c}_i = \infty$ and $\hat{c}_i = \hat{c}_{i,Chao}$. One case is represented by (c_1, G_1, H_1) vs (c_2, G_2, H_2).

	Eva-bootstrap	
(c_1, G_1, H_1) vs (c_2, G_2, H_2)	$\hat{c}_i = \infty$	$\hat{c}_i = \hat{c}_{i,Chao}$
(c, G, H) vs (c^*, G, H)	0.138	0.364
(c, G, H) vs (c, G^*, H)	0.292	0.436
(c, G, H) vs (c, G, H^*)	0.068	0.246
(c, G, H) vs (c^*, G^*, H)	0.250	0.180
(c, G^*, H) vs (c^*, G, H)	0.712	0.934
(c, G, H) vs (c^*, G, H^*)	0.158	0.396
(c, G, H^*) vs (c^*, G, H)	0.170	0.408
(c, G, H) vs (c, G^*, H^*)	0.312	0.474
(c, G, H^*) vs (c, G^*, H)	0.268	0.472
(c, G, H) vs (c^*, G^*, H^*)	0.184	0.166
(c, G, H^*) vs (c^*, G^*, H)	0.240	0.226
(c, G^*, H) vs (c^*, G, H^*)	0.712	0.946
(c, G^*, H^*) vs (c^*, G, H)	0.734	0.932

. For simplicity, we restrict the values of c_i's, G_i's, and H_i's to be the ones we list in the following. The numbers of quadrats for the two samples we consider are $K_1 = K_2 = 50$. $c = 500$ and $c^* = 450$, and they stand for total numbers of species; G is a discrete distribution with support points 0.01, 0.05, 0.1, and 0.15 and corresponding weights 0.65, 0.20, 0.10, and 0.05, G^* is similar to G but with first support point 0.01 taking the density of 0.55, and the second support point 0.05 taking the density of 0.3; H is a discrete distribution with support points 1, 2, 5, and 10 and corresponding weights 0.65, 0.20, 0.10, and 0.05, H^* is similar to H but with the first support point 1 taking the density of 0.55, and the second support point 2 taking the density of 0.3. Therefore there are a total of 13 scenarios. In all the tests, we use $t = 0.9999$ in the eigenvalue decomposition to choose $\hat{\nu}$ and the nominal size of the test α is set at 0.05. The simulation results based on 500 samples are listed in table 3.8.

From table 3.8, we can see that the Eva-bootstrap test with $\hat{c}_i = \hat{c}_{i,Chao}$ is more powerful than the one with $\hat{c}_i = \infty$. The reason for that is most likely that $\hat{c}_{i,Chao}$

is a better approximation to c_i than ∞.

3.10 Real application

The Bosques Project, located in La Selva Biological Station and surrounding areas in the Atlantic lowlands of northeastern Costa Rica, was established in 1997 to study the vegetation dynamics in tropical second-growth rain forests (Chazdon et al. 2005). One of the goals for this project is to provide information about spatial and temporal differences in population of seedlings in tropical second-growth rain forests. Such information can be obtained by comparing the seedling assemblages across different sites and over different years. To demonstrate how our proposed test can be applied to help carry out those comparisons, we are going to study the seedling assemblage data collected from four study sites, Lindero Sur (LSUR), Tirimbina (TIR), Lindero El Peje (LEP), and Cuatro Rios (CR). In all of the four sites, all seedlings were sampled in 144 1m×5m quadrats in 12 strips through the 50m×200m plot. Two groups of data were collected for each site, one in year 1998, and the other in year 2004. The species identity was determined for all seedlings (>20cm in height, but <1cm in diameter at breast height).

First, let us compare species assemblages at LSUR and TIR based on the data collected in year 1998. In LSUR, 132 species with 2287 individuals were observed, and in TIR, 153 species with 3443 individuals were observed.

The lower bound estimates for total numbers of species at LSUR and TIR are 169 and 196, respectively. Furthermore, for each species j in assemblage i, we can estimate the incidence rate π_{ij} by $k_{i,j}/K_i$, number of quadrats species j appears in divided by the total number of quadrats. Given all the estimates of π_{ij}'s, we can form

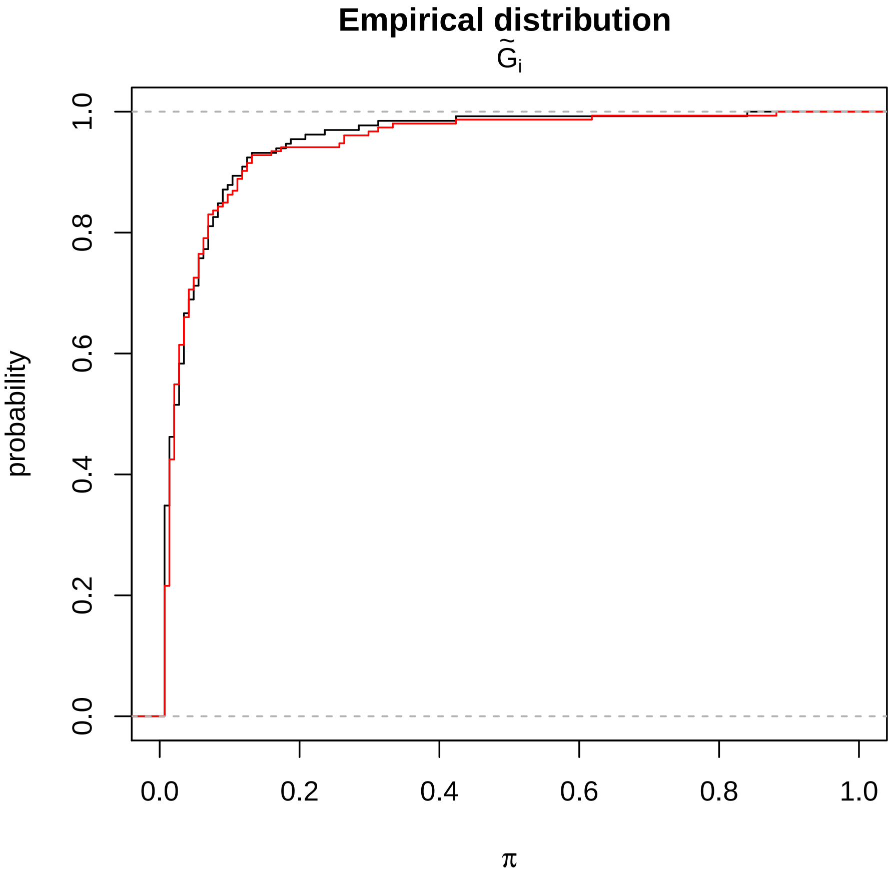

Figure 3.1: Empirical distribution functions of π at LSUR and TIR. The black curve stands for LSUR, and the red one stands for TIR.

the empirical distribution of π, $\tilde{G}_i(\pi)$. We can also estimate the abundance rate λ_{ij} for each species by solving the equation $\frac{\lambda_{ij}}{1-\exp(-\lambda_{ij})} = \frac{\sum_{k=1}^{K_i} X_{ijk}}{k_{i,j}}$. Therefore an empirical distribution of λ, $\tilde{H}_i(\lambda)$, can be constructed as well. Figure 3.1 is a plot of the two empirical distributions of π at LSUR and TIR. Figure 3.2 is a plot of the two empirical distributions of λ at LSUR and TIR. Based on these estimates of c_i's, G_i's, and H_i's, we suspect c_i's and H_i's may be different between these two assemblages. Thus our first impression is that there might be a significant difference between these two species assemblages.

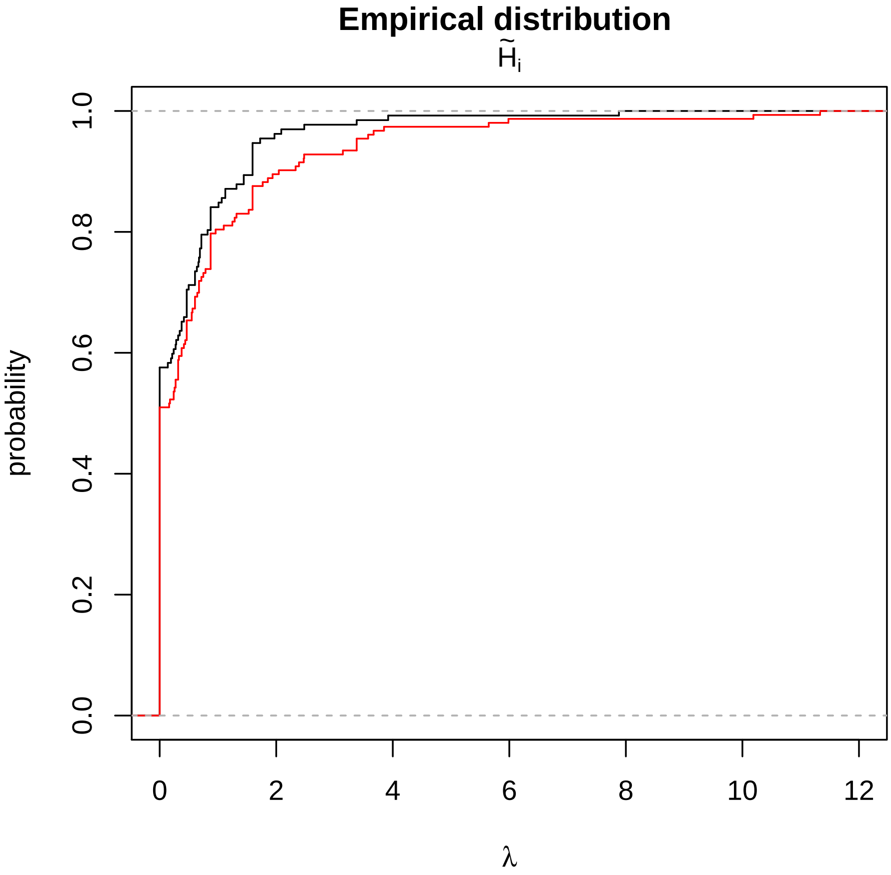

Figure 3.2: Empirical distribution functions of λ at LSUR and TIR. The black curve stands for LSUR, and the red one stands for TIR.

To determine whether there is a significant difference between the two seedling assemblages, we apply our proposed tests. The p-values of the Eva-χ^2 test are 0.0015 and 0.035 given $\hat{c}_i = \hat{c}_{i,Chao}$ and $\hat{c}_i = \infty$, used in $\hat{R}_{\hat{v}}$ in (3.14), respectively. Both p-values are smaller than 0.05. Based on our simulation studies, using $\hat{c}_i = \infty$ often leads to a conservation procedure. The null hypothesis is rejected even when using $\hat{c}_i = \infty$. This implied that there is enough evidence to reject the null hypothesis that there is no difference between these two seedling assemblages. Our Eva-bootstap test yields a p-value 0.004, which further confirms that there is a significant difference between these two seedling assemblages.

Second, let us compare species assemblages across all of the four sites in year 1998 and year 2004 respectively. For year 1998, in LEP, 166 species with 1590 individuals were observed, and in CR, 155 species with 1907 individuals were observed. For year 2004, In LSUR, 147 species with 1659 individuals were observed, in TIR, 145 species with 2030 individuals were observed, in LEP, 166 species with 1357 individuals were observed, and in CR, 158 species with 1810 individuals were observed. To determine whether there is a difference between the four seedling assemblages, we can apply the multiple sample versions of our proposed tests. The testing results are shown in table 3.9. For year 1998, the p-values of the Eva-χ^2 test are 0.004 and 0.070 given $\hat{c}_i = \hat{c}_{i,Chao}$ and $\hat{c}_i = \infty$ respectively. One is smaller than 0.05, and the other is larger than 0.05. Based on our simulation studies, using $\hat{c}_{i,Chao}$ often leads to an aggressive procedure, and using $\hat{c}_i = \infty$ often leads to a conservation procedure. Therefor, only based on those two tests, it is not clear whether we should reject the null hypothesis or not. However, our Eva-bootstap test yields a p-value 0.033, which leads to a rejection of the null hypothesis and implies that there is a significant difference between the four seedling assemblages during year 1998, which is also consistent with hypothesis testing result given in the

Table 3.9: P-values of the tests for comparing species assemblages across all of the four sites of the Bosques Project in year 1998 and year 2004 respectively.

	Eva-χ^2		Eva-bootstrap
Year	$\hat{c}_i = \hat{c}_{Chao}$	$\hat{c}_i = \infty$	$\hat{c}_i = \infty$
1998	0.004	0.070	0.033
2004	0.390	0.718	0.564

previous paragraph.

For year 2004, the p-values of the Eva-χ^2 test are 0.390 and 0.718 given $\hat{c}_i = \hat{c}_{i,Chao}$ and $\hat{c}_i = \infty$ respectively. Both p-values are larger than 0.05. Based on our simulation studies, using $\hat{c}_{i,Chao}$ often leads to an aggressive procedure. The null hypothesis is not rejected even when using $\hat{c}_i = \hat{c}_{i,Chao}$. This implied that there is not enough evidence to reject the null hypothesis that there is no difference between these four seedling assemblages during year 2004. Our Eva-bootstap test yields a p-value 0.564, which further confirms that there is no significant difference between these four seedling assemblages during year 2004.

3.11 Summary

In this chapter, we propose tests for comparison of species diversity across species assemblages given abundance data from multiple quadrats. We employ a zero-inflated Poisson mixture model. We have developed a two sample test and showed the test statistics asymptotically follows a χ^2 distribution. Since the covariance matrix is always singular, we adopt the eigenvalue adjusted procedure proposed by Mao and Li (2009). Because there is no unbiased nonparametric estimation of c_i, using a lower bound estimate of c_i or an upper bound estimate (∞) of c_i will cause our test to be either too liberal or too conservative. Therefore we adopted a bootstrap procedure to approximate the distribution of the test statistic under null hypothesis. The type I error study showed the bootstrap test is a valid test. A generalization of the two sample test to the multiple sample test is presented. In real application, we recommend first using the eigenvalue adjusted test with lower bound and upper bound estimates of c_i. If they yield the same result, then one should take it; otherwise, one should resort to the bootstrap procedure. Two power comparison studies are presented. One is the comparison between the abundance based test we proposed and the incidence based test proposed by Mao and Li (2009). The other comparison is between the bootstrap procedure with $\hat{c}_i = \infty$ and the same procedure with $\hat{c}_i = \hat{c}_{i,Chao}$. The result shows that the bootstrap procedure with $\hat{c}_i = \hat{c}_{i,Chao}$ is more powerful. We think it is because $\hat{c}_{i,Chao}$ is a better estimate of c_i than ∞ is. The result shows the incidence based test is incapable of detecting difference of H_i's, however, abundance based test is as good as incidence based test in detecting difference between c_i's and G_i's. Therefore, in practice, if abundance data is available, we suggest one take advantage of the abundance based test; if only incidence data is available, then one should use incidence based test. In the next chapter, we will summarize the differences between the two types of tests we propose in Chapter 2 and Chapter 3.

Chapter 4

Comparison of the two types of tests

So far we have developed the data depth based tests and the zero-inflated Poisson mixture model based tests separately, and applied them to problems of comparing species assemblages. Now we will discuss the differences between these two types of tests.

First, obviously, the testing objectives are different between the data depth based tests and the mixture model based tests. The data depth based tests are testing whether the two joint distributions of the counts of all the observed species from two species assemblages differ significantly. They are good for tracking changes in abundance of individual species. The mixture model based tests are testing whether numbers of species, species incidence rate distributions, and species abundance rate distributions are the same for two species assemblages. They can be used to compared species diversity across species assemblages.

Second, the data depth based tests are generally used when two species assemblages have many species in common. Because a species identity is attached to the abundance of each species, if too few species are in common between two species assemblages, there is no need to perform the tests, since the data depth based tests will always yield rejection of the null hypothesis. However, for the mixture model based tests, the testing objective of comparing species diversity across species assemblages determines that species need not to be labeled by having a species identity attached to the abundance. This loss of the label allows the mixture model based tests for comparison of species assemblages that have no species in common.

Third, the data depth based tests and the mixture model based tests work under different assumptions. The former requires quadrats are independent, while the latter requires species are independent. More precisely, if we present a species assemblage by an abundance data matrix as we did in Chapter 1, then the data depth based tests require that each row of the matrix is a random sample from certain multivariate distribution, while the mixture model based tests require each column is a random sample from certain multivariate distribution. Actually, these two types of tests are very sensitive to these independence assumptions, as we can see from the following simulation studies. Similarly to how we simulated data in Section 3.9 for the mixture model based tests, we generate abundance data by letting $c_1 = c_2 = 10$, $K_1 = K_2 = 150$, $G_1 = G_2 = \mathscr{D}_G$, and $H_1 = H_2 = \mathscr{D}_H$. We know the generated abundance data are independent across different species but correlated across quadrats. Treating abundance data from each quadrat as an observation, and applying the data depth based tests, we get huge type I errors, 0.730 and 0.576, for our KS and CM tests respectively. Again, similarly to how we simulated data in Section 2.6 for the data depth based tests, we generate $m = n = 30$ random observations from $F = PL(\boldsymbol{\mu}_F, \Sigma_F)$ and $G = PL(\boldsymbol{\mu}_G, \Sigma_G)$, where $\boldsymbol{\mu}_F = \boldsymbol{\mu}_G = \mathbf{1}_{500}$, $\Sigma_F = \Sigma_G = 0.5\mathbf{1}_{500}\mathbf{1}_{500}' + 0.5I_{500}$. We know the generated abundance data are independent across different quadrats but correlated across species. Treating the abundance data for each species across all the quadrats as an observation, and applying the zero-inflated Poisson mixture model, we get a type I error of 0.230 for the Eva-bootstrap test. Therefore, the data depth based tests are more suitable for the situations when sampling quadrats are geographically far away from each other (thus more likely to be independent of each other), while the mixture model based tests are more suitable for species assemblages in which species rarely interact with each other.

Additionally, applying the two types of tests to the same real data will render different conclusions. Take the BCI data mentioned in Chapter 1 and Chapter 2 as an example. At the significance level of $\alpha = 0.05$, both of our depth-based KS and CM tests yield p-values 0.001, meaning there is significant difference between the two species assemblages. However, the p-values of the Eva-χ^2 test are 0.088 and 0.167 given $\hat{c}_i = \hat{c}_{i,Chao}$ and $\hat{c}_i = \infty$, and the p-value of the Eva-bootstrap test is 0.171, meaning there is no significant difference between the two species assemblages. The different conclusions of the two types of tests are not surprising. First, that may be because the independence assumption for one type of the tests is not met. Second, since the testing objectives of the two types of tests are different, it is possible that the counts of certain species are distributed differently in these two species assemblages, however, in terms of species diversity there is no obvious difference between the two species assemblages.

In all, given these differences between the two types of tests, readers should choose their test accordingly.

Chapter 5

Concluding Remarks

This thesis presents two types of statistical tests for comparison of species assemblages given abundance data from multiple quadrats. They are the data depth based nonparametric tests and the zero-inflated Poisson mixture model based tests.

As we mentioned earlier, the data depth based tests and the mixture model based tests work under different assumptions. The former requires quadrats are independent, while the latter requires species are independent. Therefore given abundance data, the first question one should ask is whether the quadrats or the species are independent. The underlying statistical problem is how to test whether a group of multivariate vectors are i.i.d samples from certain underlying multivariate distribution. There are solutions under some special constraints, for example, Paindaveine (2009) proposed multivariate runs tests for the elliptical distribution family. But so far we have not seen in the literature a general solution which can be applied to any distribution. Thus we think more research can be done in this area.

The new distance-based data depth we proposed in this thesis is compatible with any distance measure, which gives our tests great flexibility. Although the proposed data depth based tests were motivated by the species assemblages comparison problem and were demonstrated by examples of count data, they can be applied with other data types as well. Future research may be conducted on applying this new notion of data depth to other data types or areas such as comparing samples of functional data, or samples of image data, because properly defined distance measures are already available. Also some of the properties of this new data depth are not clear at this moment, such as affine invariance and maximality at center (Zuo and Serfling, 2000); the asymptotical properties of the two depth based tests are also interesting to study. Theoretical research might be conducted in those areas.

When it comes to the study of species assemblages, there are two general categories of methods : static sampling based methods and dynamic theory based methods. The former focuses on analysis of species assemblages observed at certain time points, while the latter focuses on fundamental birth, death, migration processes, etc. Both of the species assemblage comparison methods we proposed in this thesis fall into the category of static sampling analysis, which is based on data collected at one time point. They ignore the internal birth, death, immigration processes of the species assemblages. We think a very interesting research problem would be how to incorporate the dynamic theory into the statistical testing procedure or how to combine the static sampling analysis and the dynamic theory when performing statistical tests. There are researchers trying to achieve that, for example, Alonso et al. (2008) presented an alternative formulation of statistical sampling theory that incorporates species asymmetries in sampling and dynamics, and illustrated the theory on a stochastic community model. That will be the direction of our future research on comparison of species assemblages.

Bibliography

Aitchison, J. and Ho, C. H. (1989). The multivariate poisson lognormal distribution. *Biometrika*, 76:643–653.

Alonso, D., Ostling, A., and Etienne, R. S. (2008). The implicit assumption of symmetry and the species abundance distribution. *Ecology Letters*, 11(2):93–105.

Bartoszy'nski, R., Pearl, D. K., and Lawrence, J. (1997). A multidimensional goodnessof-fit test based on interpoint distances. *Journal of the American Statistical Association*, 92.

Böhning, D. and Schön, D. (2005). Nonparametric maximum likelihood estimation fo population size based on the counting distribution. *Applied Statistics*, 54.

Bray, J. R. and Curtis, J. T. (1957). An ordination of the upland forest communities of southern wisconsin. *Ecological Monographs*, 27:325–459.

Bunge, J. and Fitzpatrick, M. (1993). Estimating the number of species: a review. *Journal of the American Statistical Association*, 88:364–373.

Chao, A. (1989). Estimating population size for sparse data in capture-recapture experiments. *Biometrics*, 45:427–438.

Chao, A. and Bunge, J. (2002). Estimating the number of species in a stochastic abundance model. *Biometrics*, 58:531–539.

Chao, A., Chazdon, R. L., Colwell, R. K., and Shen, T.-J. (2006). Abundance-based similarity indices and their estimation when there are unseen species in samples. *Biometrics*, 62(2):361–371.

Chazdon, R. L., Redondo Brenes, A., and Vilchez Alvarado, B. (2005). Effects of climate and stand age on annual tree dynamics in tropical second-growth rain forests. *Ecology*, 86:1808–1815.

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. *Springer-Verlag and R. G. Landes Company, Berlin, Germany, and Georgetown, Texas.*

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society Series B Methodological*, 39(1):1–38.

Donoho, D. L. (1982). Breakdown properties of multivariate location estimators. *Ph.D. qualifying paper, 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. *Annals of Statistics*, 69:57–68.

Fay, M. P.and 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.

Ghosh, A. K. and Chaudhuri, P. (2005). On maximum depth and related classifiers. *Scandinavian Journal of Statistics*, 32(2):327–350.

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. (2011). Generalized mahalanobis depth in the reproducing kernel hilbert space. *Statistical Papers*, 52:511–522.

Hubbell, S., Condit, R., and Foster, R. (2005). Barro colorado forest census plot data. *URL http://ctfs.arnarb.harvard.edu/webatlas/datasets/bci*, 26:523–527.

Hubbell, S., Foster, R. B., O'Brien, S., Harms, K., Condit, R., Wechsler, B., Wright, S., and Loo de Lao, S. (1999). Light gap disturbances, recruitment limitation, and tree diversity in a neotropical forest. *Science*, 283:554–557.

Huggins, R. (2001). A note on the difficulties associated with the analysis of capturerecapture experiments with heterogeneous capture probabilities. *Statistics and Probability Letters*, 54:147–152.

Kiefer, J. (1959). K-sample analogues of the kolmogorov-smirnov and cramér-v. mises tests. *The Annals of Mathematical Statistics*, 30(2):420–447.

Li, J., Ban, J., and Santiago, L. S. (2011a). Nonparametric tests for homogeneity of species assemblages: A data depth approach. *Biometrics*, 67(4):1481–1488.

Li, J., Cuesta-Albertos, J. A., and Liu, R. Y. (2012). *DD*-classifier: Nonparametric classification procedure based on *DD*-plot. *Journal of the American Statistical Association*. To appear.

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.

Li, J., Mao, C. X., and Wang, S. (2011b). Species assemblage comparison with abundance data. *Submitted*.

Lindsay, B. G. (1983). The geometry of mixture likelihoods: A general theory. *The Annals of Statistics*, 11(1):pp. 86–94.

Lindsay, B. G. (1995). *Mixture models: theory, geometry and applications*. Institute of Mathematical Statistics.

Link, W. A. (2003). Nonidentifiability of population size from capture-recapture data with heterogeneous detection probabilities. *Biometrics*, 59:1123–1130.

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

Liu, R. Y. (1995). Control charts for multivariate processes. *Journal of the American Statistical Association*, 90(432):1380–1387.

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

Liu, R. Y. and Singh, K. (1997). Notions of limiting p-values based on data depth and bootstrap. *Journal of the American Statistical Association*, 92(437):266–277.

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. C. (1936). On the generalized distance in statistics. *Proc. Nat. Acad. Sci. India*, 12:49–55.

Mao, C. (2008). Computing an npmle for a mixing distribution in two closed heterogeneous population size models. *Biometrical Journal*, 50(6):983–992.

Mao, C. X. (2006). Inference of the number of species via geometric lower bounds. *Journal of American Statistical Association*, 101:1663–1670.

Mao, C. X. (2007). Estimating population sizes for capture-recapture sampling with binomial mixtures. *Computational Statistics & Data Analysis*, 51(11):5211 – 5219.

Mao, C. X. and Colwell, R. K. (2005). Estimation of the species richness: mixture models, the role of rare species, and inferential challenges. *Ecology*, 86:1143 – 1153.

Mao, C. X., Colwell, R. K., and Chang, J. (2005). Estimating the species accumulation curve using mixtures. *Biometrics*, 61(2):433 – 441.

Mao, C. X. and Li, J. (2009). Comparing species assemblages via species accumulation curves. *Biometrics*, 65(4):1063–1067.

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.

McGill, B. J., Etienne, R. S., Gray, J. S., Alonso, D., Anderson, M. J., Benecha, H. K., Dornelas, M., Enquist, B. J., Green, J. L., He, F., Hurlbert, A. H., Magurran, A. E., Marquet, P. A., Maurer, B. A., Ostling, A., Soykan, C. U., Ugland, K. I., and White, E. P. (2007). Species abundance distributions: moving beyond single prediction theories to integration within an ecological framework. *Ecology Letters*, 10(10):995–1015.

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 & Data Analysis*, 37:195–208.

Numa, C., Verdú, J. R., Rueda, C., and Galante, E. (2012). Comparing dung beetle species assemblages between protected areas and adjacent pasturelands in a mediterranean savanna landscape. *Rangeland Ecology & Management*, 65(2):137–143.

Ord, J. K. and Whitmore, G. (1986). The poisson-inverse gaussian distribution as a model for species abundance. *Communications in Statistics-Theory and Methods*, 15:853–871.

Paindaveine, D. (2009). On multivariate runs tests for randomness. *Journal of the American Statistical Association*, 104(488):1525–1538.

Pipan, T. and Culver, D. (2007). Regional species richness in an obligate subterranean dwelling faunaepikarst copepods. *Journal of Biogeography*, 34:854–861.

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

Stahel, W. (1981). Robust schaetzungen: Infinitesmale optimalitaet und schaetzungen von kovarianzmatrizen. *Ph.D. thesis, ETH Zurich.*

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

Warwick, R. M. and Clarke, K. R. (1993). Increased variability as a symptom of stress in marine communities. *Journal of Experimental Marine Biology and Ecology*, 172:215–226.

Warwick, R. M., Clarke, K. R., and Suharsono (1990). A statistical analysis of coral community responses to the 1982c-83 el niño in the thousand islands, indonesia. *Coral Reefs*, 8:171–179. 10.1007/BF00265008.

Wells, K., Kalko, E. K. V., and Lakim, M. B. (2007). Effects of rain forest logging on species richness and assemblage composition of small mammals in Southeast Asia. *Journal of Biogeography*, pages 1087–1099.

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

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

Appendix A

Proof in Chapter 3

Proof of Theorem 2. It is straightforward that, if $c_1 = c_2$, $G_1 = G_2$ and $H_1 = H_2$, then $g_1(h_0, x) = g_2(h_0, x)$ for $x = 1, 2, ...$ and $\tau_1(h) = \tau_2(h)$ for $h = 1, 2, \dots, h_0 - 1, h_0 + 1, \dots$. When $g_1(h_0, x) = g_2(h_0, x)$ for $x = 1, 2, \dots$, $\tau_1(h_0) = \tau_2(h_0)$ since $\tau_i(h_0) = \sum_{x=1}^{\infty} g_i(h_0, x)$. Together with $\tau_1(h) = \tau_2(h)$ for $h = 1, 2, \dots, h_0 - 1, h_0 + 1, \dots$, we can have $c_1 = c_2$ and $G_1 = G_2$, following Theorem 2 of Mao and Li (2009). Therefore, $g_1(h_0, x) = g_2(h_0, x)$ implies that

$$\int \frac{\exp(-\lambda)}{1 - \exp(-\lambda)} \frac{\lambda^x}{x!} dH_1(\lambda) = \int \frac{\exp(-\lambda)}{1 - \exp(-\lambda)} \frac{\lambda^x}{x!} dH_2(\lambda), \quad x = 1, 2, \dots, \tag{A.1}$$

thus

$$\int \frac{\exp(-\lambda)}{1 - \exp(-\lambda)} \frac{\lambda^{x+y}}{(x+y)!} dH_1(\lambda) = \int \frac{\exp(-\lambda)}{1 - \exp(-\lambda)} \frac{\lambda^{x+y}}{(x+y)!} dH_2(\lambda), \tag{A.2}$$

for $x = 1, 2, \dots$ and $y = 1, 2, \dots$. Multiplying both sides of (A.2) with $(x + y)!/x!$, we will have

$$\int \frac{\exp(-\lambda)\lambda^{(x+y)}/x!}{1 - \exp(-\lambda)} dH_1(\lambda) = \int \frac{\exp(-\lambda)\lambda^{(x+y)}/x!}{1 - \exp(-\lambda)} dH_2(\lambda), \tag{A.3}$$

for $x = 1, 2, \dots$ and $y = 1, 2, \dots$. Because

$$\sum_{x=1}^{\infty} \int \frac{\exp(-\lambda)\lambda^{(x+y)}/x!}{1 - \exp(-\lambda)} dH_i(\lambda) = \int \sum_{x=1}^{\infty} \frac{\exp(-\lambda)\lambda^{(x+y)}/x!}{1 - \exp(-\lambda)} dH_i(\lambda) = \int \lambda^y dH_i(\lambda),$$

for any positive integer y,

$$\int \lambda^y dH_1(\lambda) = \int \lambda^y dH_2(\lambda).$$

Given that both $H_1(\lambda)$ and $H_2(\lambda)$ have bounded support, the moment generation functions of H_1 and H_2 are identical, therefore, $H_1 = H_2$.

■

Proof of Theorem 4. We define $I^v_{i,j,k,x} = I\{$In assemblage i, species j appears exactly in k quadrats and appears x times in the v-th quadrat among those k quadrats$\}$. It is easy to see that $n^v_{i,k,x} = \sum_{j=1}^{c_i} I^v_{i,j,k,x}$. Since $n_{i,k,x} = \sum_{v=1}^{k} n^v_{i,k,x}/k$, $n_{i,k,x} = \sum_{j=1}^{c_i}\sum_{v=1}^{k} I^v_{i,j,k,x}/k$. Therefore, $\mathbf{n}_i = \sum_{j=1}^{c_i} \mathbf{I}_{i,j}$, where

$$\mathbf{I}_{i,j} = (\sum_{v=1}^{1} I^v_{i,j,1,1}/1, ..., \sum_{v=1}^{1} I^v_{i,j,1,m}/1, ..., \sum_{v=1}^{K_i} I^v_{i,j,K_i,1}/K_i, ..., \sum_{v=1}^{K_i} I^v_{i,j,K_i,m}/K_i)'.$$

Based on our assumptions, the $\mathbf{I}_{i,j}$ are i.i.d., therefore, $c_i^{-1/2}\{\mathbf{n}_i - E(\mathbf{n}_i)\}$ converges to $\mathcal{N}(\mathbf{0}, V_i/c_i)$ in distribution as c_i goes to ∞.

Since $W_i = T_i V_i T_i'$ and V_i is positive definite, to prove that W_i is positive definite, we need to prove that for any nonzero vector $\boldsymbol{a} \in \mathbb{R}^{m+K-1}$, $\boldsymbol{a}'T_i \neq 0$, which is equivalent to that the linear equations $\boldsymbol{a}'T_i = 0$ does not have nonzero solution. Using the knowledge of linear algebra, that is equivalent to that $Rank(T_i) = m + K - 1$, since T_i is $(m + K - 1) \times mK_i$ and $m + K - 1 \leq mK_i$.

Let us look at a sub-matrix of T_i, R_i, which is made of all the rows of T_i, column $m(K_i - K) + 1$, $m(K_i - K) + 2$, $\ldots$, $m(K_i - K) + m$, and column $m(K_i + 1 - K) + 1$, $m(K_i + 2 - K) + 1$, $\ldots$, $mK_i + 1$. Thus R_i is a $(m + K - 1) \times (m + K - 1)$ matrix like this,

$$R_i = \begin{pmatrix} & 1-\frac{\binom{K_i-1}{K_i+2-K}}{\binom{K_i}{K_i+2-K}} & 1-\frac{\binom{K_i-1}{K_i+3-K}}{\binom{K_i}{K_i+3-K}} & \cdots & 1-\frac{\binom{K_i-1}{K_i}}{\binom{K_i}{K_i}} \\ \left(1-\frac{\binom{K_i-1}{K_i+1-K}}{\binom{K_i}{K_i+1-K}}\right)\bigotimes \mathbf{I}_{m\times m} & 0 & 0 & \cdots & 0 \\ & \vdots & \vdots & \cdots & \vdots \\ & 0 & 0 & \cdots & 0 \\ \left(1-\frac{\binom{K_i-2}{K_i+1-K}}{\binom{K_i}{K_i+1-K}}\right)\bigotimes \mathbf{1}'_m & 1-\frac{\binom{K_i-2}{K_i+2-K}}{\binom{K_i}{K_i+2-K}} & 1-\frac{\binom{K_i-2}{K_i+3-K}}{\binom{K_i}{K_i+3-K}} & \cdots & 1-\frac{\binom{K_i-2}{K_i}}{\binom{K_i}{K_i}} \\ \left(1-\frac{\binom{K_i-3}{K_i+1-K}}{\binom{K_i}{K_i+1-K}}\right)\bigotimes \mathbf{1}'_m & 1-\frac{\binom{K_i-3}{K_i+2-K}}{\binom{K_i}{K_i+2-K}} & 1-\frac{\binom{K_i-3}{K_i+3-K}}{\binom{K_i}{K_i+3-K}} & \cdots & 1-\frac{\binom{K_i-3}{K_i}}{\binom{K_i}{K_i}} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ \left(1-\frac{\binom{K_i-K}{K_i+1-K}}{\binom{K_i}{K_i+1-K}}\right)\bigotimes \mathbf{1}'_m & 1-\frac{\binom{K_i-K}{K_i+2-K}}{\binom{K_i}{K_i+2-K}} & 1-\frac{\binom{K_i-K}{K_i+3-K}}{\binom{K_i}{K_i+3-K}} & \cdots & 1-\frac{\binom{K_i-K}{K_i}}{\binom{K_i}{K_i}} \end{pmatrix}.$$

It is easy to see that for the element of row x and column y of R_i, $R_{i_{xy}}$, where $x > m$ and $y > m$, if $K_i < x + y - 2m$, then $R_{i_{xy}} = 1$. Also $R_{i_{xy}} = 1$ for $x = m + K - 1$ and $y = 1, \ldots, m$, and $x = 1$ and $y = m + K - 1$. Therefore

$$R_i = \begin{pmatrix} & 1-\frac{\binom{K_i-1}{K_i+2-K}}{\binom{K_i}{K_i+2-K}} & \cdots & 1-\frac{\binom{K_i-1}{K_i-1}}{\binom{K_i}{K_i-1}} & 1 \\ \left(1-\frac{\binom{K_i-1}{K_i+1-K}}{\binom{K_i}{K_i+1-K}}\right)\bigotimes \mathbf{I}_{m\times m} & 0 & \cdots & 0 & 0 \\ & \vdots & \cdots & \vdots & \vdots \\ & 0 & \cdots & 0 & 0 \\ \left(1-\frac{\binom{K_i-2}{K_i+1-K}}{\binom{K_i}{K_i+1-K}}\right)\bigotimes \mathbf{1}'_m & 1-\frac{\binom{K_i-2}{K_i+2-K}}{\binom{K_i}{K_i+2-K}} & \cdots & 1 & 1 \\ \vdots & \vdots & \iddots & \vdots & \vdots \\ \left(1-\frac{\binom{K_i+1-K}{K_i+1-K}}{\binom{K_i}{K_i+1-K}}\right)\bigotimes \mathbf{1}'_m & 1 & \cdots & 1 & 1 \\ \mathbf{1}'_m & 1 & \cdots & 1 & 1 \end{pmatrix}.$$

For $x_1 > m$, $x_2 > m$, and $y > m$, where $K_i \geq x_1+y-2m$ and $K_i \geq x_2+y-2m$, one can easily prove that if $x_1 > x_2 > m \geq x$, then $1 > R_{i_{x_1y}} > R_{i_{x_2y}} > R_{i_{xy}} > 0$; for $m + K - 1 > x_1 > m$, $m + K - 1 > x_2 > m$, and $y \leq m$, if $x_1 > x_2 > m \geq x$, then $1 > R_{i_{x_1y}} > R_{i_{x_2y}} > R_{i_{xy}} > 0$. Similarly, for $y_1 > m$, $y_2 > m$ and $x > m$, where

$K_i \geq x + y_1 - 2m$ and $K_i \geq x + y_2 - 2m$, if $y_1 > y_2 > m \geq y$, then $1 > R_{i_{xy_1}} > R_{i_{xy_2}} > R_{i_{xy}} > 0$. Based on the above results, one can perform a series of elementary row transformations, and R_i can be transformed into

$$R_i^* = \left(\begin{array}{ccccc} & 1 - \frac{\binom{K_i-1}{K_i+2-K}}{\binom{K_i}{K_i+2-K}} & \cdots & 1 - \frac{\binom{K_i-1}{K_i-1}}{\binom{K_i}{K_i-1}} & 1 \\ \left(1 - \frac{\binom{K_i-1}{K_i+1-K}}{\binom{K_i}{K_i+1-K}}\right) \bigotimes \mathbf{I}_{m\times m} & 0 & \cdots & 0 & 0 \\ & \vdots & \cdots & \vdots & \vdots \\ & 0 & \cdots & 0 & 0 \\ \mathbf{0}'_m & 0 & \cdots & 1 & 1 \\ \vdots & \vdots & \iddots & 0 & 0 \\ \mathbf{0}'_m & 1 & \cdots & 0 & 0 \\ \mathbf{1}'_m & 0 & \cdots & 0 & 0 \end{array} \right),$$

where $\mathbf{0}'_m$ is a m dimension vector with all its elements equal to zero.

By far, one can easily prove by definition that the column vectors of R_i^* are linearly independent. Thus $Rank(R_i) = Rank(R_i^*) = m + K - 1$. Since R_i is a submatrix of T_i, $Rank(T_i) \geq Rank(R_i) = m + K - 1$. Considering T_i is $(m+K-1) \times mK_i$ and $m+K-1 \leq mK_i$, we have $Rank(T_i) \leq m+K-1$. Therefore $Rank(T_i) = m+K-1$ and W_i is positive definite. Since $\hat{\boldsymbol{\eta}}_{i,K,m} = T_i\mathbf{n}_i$, Theorem 4 follows by using the delta method.

■

Proof of Proposition 5. (a) Since $n_{i,k,x} = \sum_{v=1}^{k} n_{i,k,x}^{v}/k$,

$$\mathrm{var}(n_{i,k,x}) = \{\mathrm{var}(n_{i,k,x}^1) + (k-1)\,\mathrm{cov}(n_{i,k,x}^1, n_{i,k,x}^2)\}/k.$$

Based on the definitions of $n_{i,k}$ and $n_{i,k,x}^1$, $\{n_{i,0}$, and $n_{i,k,x}^1, k = 1, ..., K_i, x = 1, 2, ..., \}$ follows a multinomial distribution with size c_i and probabilities $r_{i,0}$ and $r_{i,k}s_{i,x}$. Therefore, based on the properties of multinomial distribution, $\mathrm{var}(n_{i,k,x}^1) = c_i r_{i,k} s_{i,x}(1 -$

$r_{i,k}s_{i,x})$. We also have $\text{cov}(n^1_{i,k,x}, n^1_{i,k,y}) = -c_i r_{i,k}{}^2 s_{i,x}s_{i,y}$, and $\text{cov}(n^1_{i,k,x}, n^1_{i,l,y}) = -c_i r_{i,k} r_{i,l} s_{i,x} s_{i,y}$, which will be used in the proof for (b) and (c).

To find $\text{cov}(n^1_{i,k,x}, n^2_{i,k,x})$, recall $n^v_{i,k,x} = \sum_{j=1}^{c_i} I^v_{i,j,k,x}$, where $I^v_{i,j,k,x} = I\{$In assemblage i, species j appears exactly in k quadrats and appears x times in the v-th quadrat among those k quadrats$\}$. Since the species are independent of each other,

$$\begin{aligned}\text{cov}(n^1_{i,k,x}, n^2_{i,k,x}) &= \sum_{j=1}^{c_i} \text{cov}(I^1_{i,j,k,x}, I^2_{i,j,k,x}) \\ &= \sum_{j=1}^{c_i} E(I^1_{i,j,k,x} I^2_{i,j,k,x}) - \sum_{j=1}^{c_i} E(I^1_{i,j,k,x}) E(I^2_{i,j,k,x})\end{aligned}$$

Clearly, $E(I^1_{i,j,k,x}) = E(I^2_{i,j,k,x}) = r_{i,k}s_{i,x}$, and

$$\begin{aligned}E(I^1_{i,j,k,x} I^2_{i,j,k,x}) &= E[\sum_{1\le t_1<...<t_k\le K_i} \sum_{x_{t_3}=1}^{\infty} \cdots \sum_{x_{t_k}=1}^{\infty} I\{B_{ij,t_1,\dots,t_k}(x, x, x_{t_3}, ..., x_{t_k})\}] \\ &= r_{i,k}s_{i,x,x}\end{aligned}$$

Thus, $\text{cov}(n^1_{i,k,x}, n^2_{i,k,x}) = c_i r_{i,k} s_{i,x,x} - c_i (r_{i,k}s_{i,x})^2$.

Therefore,

$$\begin{aligned}\text{var}(n_{i,k,x}) &= \{c_i r_{i,k} s_{i,x}(1 - r_{i,k}s_{i,x}) + (k-1)[c_i r_{i,k} s_{i,x,x} - c_i(r_{i,j}s_{i,x})^2]\}/k \\ &= \{c_i r_{i,k} s_{i,x} + (k-1) c_i r_{i,k} s_{i,x,x} - k c_i r^2_{i,j} s^2_{i,x}\}/k.\end{aligned}$$

(b) Again based on the definition of $n_{i,k,x}$,

$$\text{cov}(n_{i,k,x}, n_{i,k,y}) = \{\text{cov}(n^1_{i,k,x}, n^1_{i,k,y}) + (k-1)\,\text{cov}(n^1_{i,k,x}, n^2_{i,k,y})\}/k.$$

Similar to the proof in (a), we can obtain

$$\begin{aligned}\text{cov}(n^1_{i,k,x}, n^1_{i,k,y}) &= -c_i r_{i,k}{}^2 s_{i,x} s_{i,y} \\ \text{cov}(n^1_{i,k,x}, n^2_{i,k,y}) &= c_i r_{i,k} s_{i,x,y} - c_i r_{i,k}{}^2 s_{i,x} s_{i,y}.\end{aligned}$$

Therefore,

$$\begin{aligned}\text{cov}(n_{i,k,x}, n_{i,k,y}) &= \{-c_i {r_{i,k}}^2 s_{i,x}s_{i,y} + (k-1)(c_i r_{i,k}s_{i,x,y} - c_i {r_{i,k}}^2 s_{i,x}s_{i,y})\}/k \\ &= \{(k-1)c_i r_{i,k}s_{i,x,y} - kc_i r_{i,k}^2 s_{i,x}s_{i,y}\}/k.\end{aligned}$$

(c) Since $\text{cov}(n_{i,k,x}, n_{i,l,y}) = \text{cov}(n^1_{i,k,x}, n^1_{i,l,y}) = -c_i r_{i,k}s_{i,x}r_{i,l}s_{i,y}$, the proof completes.

■

Proof of Proposition 6. First, define $n_{i,+} = \sum_{k=1}^{K_i} n_{i,k}$, which is the number of species observed in assemblage i. W.l.o.g, we assume species j, $j = 1, \quad \ldots, \quad n_{i,+}$, are observed among the c_i species. Following the notation used in the proof of Theorem 4, we denote the sample covariance matrix of $\mathbf{I}_{i,1}$,...,$\mathbf{I}_{i,n_{i,+}}$ by S_i, and its element on the j-th row and k-th column by $S_{i,j,k}$. We denote the plug-in estimate of V_i by $\hat{V}_i$, and its element on the j-th row and k-th column by $\hat{V}_{i,j,k}$. In the following, we first prove that $\hat{V}_i = n_{i,+} \cdot S_i + \boldsymbol{n_i}\boldsymbol{n_i}'(1/n_{i,+} - 1/\hat{c}_i)$.

First of all, for any diagonal element of S_i, $S_{i,x+m(k-1),x+m(k-1)}$, when $0 < x \leq m$ and $1 \leq k \leq K_i$, we have,

$$\begin{aligned}S_{i,x+m(k-1),x+m(k-1)} &= 1/n_{i,+} \sum_{j=1}^{n_{i,+}} \left(\sum_{v=1}^{k} I^v_{i,j,k,x}/k - 1/n_{i,+} \sum_{j=1}^{n_{i,+}} \sum_{v=1}^{k} I^v_{i,j,k,x}/k \right)^2 \\ &= 1/n_{i,+} \sum_{j=1}^{n_{i,+}} \left(\sum_{v=1}^{k} I^v_{i,j,k,x}/k \right)^2 - \left(\sum_{j=1}^{n_{i,+}} \sum_{v=1}^{k} I^v_{i,j,k,x}/k \right)^2 /n^2_{i,+}.\end{aligned}$$

Since $I^v_{i,j,k,x}$ is either 0 or 1, $I^v_{i,j,k,x} = 0$ for $j = (n_{i,+}{+}1), \ldots, c_i$, and $n_{i,k,x} = \sum_{j=1}^{n_{i,+}} \sum_{v=1}^{k} I^v_{i,j,k,x}/k$,

$$\begin{aligned}S_{i,x+m(k-1),x+m(k-1)} &= 1/n_{i,+} \sum_{j=1}^{n_{i,+}} \left(1/k^2 \left(\sum_{v=1}^{k} I^v_{i,j,k,x} + 2 \sum_{1 \leq v_1 < v_2 \leq k} I^{v_1}_{i,j,k,x} I^{v_2}_{i,j,k,x} \right)\right) - n^2_{i,k,x}/n^2_{i,+} \\ &= 1/n_{i,+}\{1/k^2(k \cdot n_{i,k,x} + k(k-1)n_{i,k,x,x})\} - n^2_{i,k,x}/n^2_{i,+} \\ &= 1/k\{n_{i,k,x}/n_{i,+} + (k-1)n_{i,k,x,x}/n_{i,+}\} - n^2_{i,k,x}/n^2_{i,+}.\end{aligned}$$

Since $\hat{V}_{i,x+m(k-1),x+m(k-1)}$ is the estimate of $var(n_{i,k,x})$, based on our estimating procedure, we have,

$$\hat{V}_{i,x+m(k-1),x+m(k-1)} = 1/k\{n_{i,k,x} + (k-1)n_{i,k,x,x} - kn^2_{i,k,x}/\hat{c}_i\}.$$

Therefore, $\hat{V}_{i,x+m(k-1),x+m(k-1)} = n_{i,+} \cdot S_{i,x+m(k-1),x+m(k-1)} + n^2_{i,k,x}(1/n_{i,+} - 1/\hat{c}_i)$.

Similarly, for the off-diagonal elements of S_i when $0 < x, y \leq m$ and $1 \leq k \leq K_i$,

$$\begin{aligned} S_{i,x+m(k-1),y+m(k-1)} &= 1/n_{i,+} \sum_{j=1}^{n_{i,+}} \left(\sum_{v=1}^{k} I^v_{i,j,k,x}/k - 1/n_{i,+} \sum_{j=1}^{n_{i,+}} \sum_{v=1}^{k} I^v_{i,j,k,x}/k \right) \\ &\left(\sum_{v=1}^{k} I^v_{i,j,k,y}/k - 1/n_{i,+} \sum_{j=1}^{n_{i,+}} \sum_{v=1}^{k} I^v_{i,j,k,y}/k \right) \\ &= 1/n_{i,+} \sum_{j=1}^{n_{i,+}} \left(\sum_{v_1=1}^{k} I^{v_1}_{i,j,k,x}/k \cdot \sum_{v_2=1}^{k} I^{v_2}_{i,j,k,y}/k \right) - n_{i,k,x}n_{i,k,y}/n^2_{i,+} \\ &= 1/n_{i,+} \sum_{j=1}^{n_{i,+}} 1/k^2 \cdot 2 \sum_{1 \leq v_1 < v_2 \leq k} \left(I^{v_1}_{i,j,k,x} I^{v_2}_{i,j,k,y} \right) - n_{i,k,x}n_{i,k,y}/n^2_{i,+} \\ &= 1/k\{(k-1)n_{i,k,x,y}/n_{i,+} - k \cdot n_{i,k,x}/n_{i,+} \cdot n_{i,k,y}/n_{i,+}\}. \end{aligned}$$

Since $\hat{V}_{i,x+m(k-1),y+m(k-1)}$ is the estimate of $\text{cov}(n_{i,k,x}, n_{i,k,y})$,we can see that

$$\begin{aligned} \hat{V}_{i,x+m(k-1),y+m(k-1)} &= 1/k\{(k-1)n_{i,k,x,y} - kn_{i,k,x}n_{i,k,y}/\hat{c}_i\} \\ &= n_{i,+} \cdot S_{i,x+m(k-1),y+m(k-1)} + n_{i,k,x}n_{i,k,y}(1/n_{i,+} - 1/\hat{c}_i). \end{aligned}$$

Last, let us consider the off-diagonal elements of S_i when $0 < x, y \leq m$ and $1 \leq k_1 \neq k_2 \leq K_i$,

$$\begin{aligned} S_{i,x+m(k_1-1),y+m(k_2-1)} &= 1/n_{i,+} \sum_{j=1}^{n_{i,+}} \left(\sum_{v=1}^{k_1} I^v_{i,j,k_1,x}/k_1 - 1/n_{i,+} \sum_{j=1}^{n_{i,+}} \sum_{v=1}^{k_1} I^v_{i,j,k_1,x}/k_1 \right) (\sum_{v=1}^{k_2} I^v_{i,j,k_2,y}/k_2 \\ &- 1/n_{i,+} \sum_{j=1}^{n_{i,+}} \sum_{v=1}^{k_2} I^v_{i,j,k_2,y}/k_2) \\ &= 1/n_{i,+} \sum_{j=1}^{n_{i,+}} \left(\sum_{v_1=1}^{k_1} \sum_{v_2=1}^{k_2} I^{v_1}_{i,j,k_1,x} I^{v_2}_{i,j,k_2,y}/(k_1k_2) \right) - n_{i,k,x}n_{i,k,y}/n^2_{i,+}. \end{aligned}$$

No species can appear in exactly k_1 and k_2 quadrats at the same time, therefore $I^{v_1}_{i,j,k_1,x} I^{v_2}_{i,j,k_2,y} = 0$. Thus

$$S_{i,x+m(k_1-1),y+m(k_2-1)} = -n_{i,k,x}/n_{i,+} \cdot n_{i,k,y}/n_{i,+}.$$

Since $\hat{V}_{i,x+m(k_1-1),y+m(k_2-1)}$ is the estimate of $\mathrm{cov}(n_{i,k,x}, n_{i,l,y})$, we can see that,

$$\begin{aligned} \hat{V}_{i,x+m(k_1-1),y+m(k_2-1)} &= -n_{i,k_1,x} n_{i,k_2,y}/\hat{c}_i \\ &= n_{i,+} \cdot S_{i,x+m(k_1-1),y+m(k_2-1)} + n_{i,k_1,x} n_{i,k_2,y}(1/n_{i,+} - 1/\hat{c}_i). \end{aligned}$$

So far, based on the above results, we have proved that $\hat{V}_i = n_{i,+} \cdot S_i + \boldsymbol{n_i}\boldsymbol{n_i}'(1/n_{i,+} - 1/\hat{c}_i)$. Because S_i is a sample covariance matrix and $n_{i,+} > 0$, $n_{i,+} \cdot S_i$ is positive semi-definite; because $1/n_{i,+} - 1/\hat{c}_i \geq 0$ and $\boldsymbol{n_i}\boldsymbol{n_i}'$ is positive semi-definite, $\boldsymbol{n_i}\boldsymbol{n_i}'(1/n_{i,+} - 1/\hat{c}_i)$ is positive semi-definite. Therefore $\hat{V}_i$ is positive semi-definite. We know that $\hat{\Sigma}_{K,m} = T_1\hat{V}_1T_1' + T_2\hat{V}_2T_2'$. For any non-zero vector $\boldsymbol{x}$ of dimension $m+K-1$, $\boldsymbol{x}'\hat{\Sigma}_{K,m}\boldsymbol{x} = (\boldsymbol{x}'T_1)\hat{V}_1(\boldsymbol{x}'T_1)' + (\boldsymbol{x}'T_2)\hat{V}_2(\boldsymbol{x}'T_2)'$. Because $\hat{V}_i \geq 0$, $\boldsymbol{x}'\hat{\Sigma}_{K,m}\boldsymbol{x} \geq 0$. Therefore $\hat{\Sigma}_{K,m}$ is positive semi-definite.

■

Proof of Proposition 7. Assume m_1 is a large integer such that $n_{i,k,x} = 0$ for $x \geq m_1$ and m_2 is an integer greater than m_1. It is obvious that, for $i = 1, 2$, $\hat{\eta}_{i,K,m_1} = (\hat{g}_i(1), \ldots, \hat{g}_i(m_1), \hat{\tau}_i(2), \ldots, \hat{\tau}_i(K))'$ and $\hat{\eta}_{i,K,m_2} = (\hat{g}_i(1), \ldots, \hat{g}_i(m_1), \mathbf{0}_{m_2-m_1}, \hat{\tau}_i(2), \ldots, \hat{\tau}_i(K))'$, where $\mathbf{0}_{m_2-m_1}$ is a $m_2 - m_1$ dimension vector with all its elements equal to zero.

Moreover, if we write $\hat{\Sigma}_{K,m_1}$ as a block matrix, $\begin{array}{cc} & \begin{array}{cc} m_1-1 & K \end{array} \\ \begin{array}{c} m_1-1 \\ K \end{array} & \left(\begin{array}{cc} A & B \\ C & D \end{array}\right) \end{array}$, then

$$\hat{\Sigma}_{K,m_2} = \begin{array}{cc} & \begin{array}{ccc} m_1-1 & m_2-m_1 & K \end{array} \\ \begin{array}{c} m_1-1 \\ m_2-m_1 \\ K \end{array} & \left(\begin{array}{ccc} A & O_{(m_1-1)\times(m_2-m_1)} & B \\ O_{(m_2-m_1)\times(m_1-1)} & O_{(m_2-m_1)\times(m_2-m_1)} & O_{(m_2-m_1)\times K} \\ C & O_{K\times(m_2-m_1)} & D \end{array}\right) \end{array},$$

where $O_{x\times y}$ stands for a $x \times y$ matrix with all its elements equal to zero.

To find the eigenvalues of $\hat{\Sigma}_{K,m_2}$, we solve the following equation of λ,

$$|\hat{\Sigma}_{K,m_2} - \lambda I| = 0.$$

$$\begin{vmatrix} A-\lambda I_{m_1-1} & O_{(m_1-1)\times(m_2-m_1)} & B \\ O_{(m_2-m_1)\times(m_1-1)} & -\lambda I_{m_2-m_1} & O_{(m_2-m_1)\times K} \\ C & O_{K\times(m_2-m_1)} & D-\lambda I_K \end{vmatrix} = 0,$$

from which one immediately derives

$$\begin{vmatrix} A-\lambda I_{m_1-1} & B & O_{(m_1-1)\times(m_2-m_1)} \\ C & D-\lambda I_K & O_{K\times(m_2-m_1)} \\ O_{(m_2-m_1)\times(m_1-1)} & O_{(m_2-m_1)\times K} & -\lambda I_{m_2-m_1} \end{vmatrix} =$$

$$\begin{vmatrix} \hat{\Sigma}_{K,m_1} - \lambda I_{m_1-1+K} & O_{(m_1-1+K)\times(m_2-m_1)} \\ O_{(m_2-m_1)\times(m_1-1+K)} & -\lambda I_{m_2-m_1} \end{vmatrix} = 0,$$

which is $|\hat{\Sigma}_{K,m_1} - \lambda I_{m_1-1+K}| \cdot | - \lambda I_{m_2-m_1}| = 0$. Therefore $\hat{\Sigma}_{K,m_2}$ and $\hat{\Sigma}_{K,m_1}$ have the same non-zero eigenvalues. Thus $\hat{v}$ does not depend on m.

Assume $\begin{array}{cc} m_1-1 & \\ K & \end{array}\!\!\begin{pmatrix} \mathbf{y}_{i1} \\ \mathbf{y}_{i2} \end{pmatrix}$ is the standardized eigenvector of $\hat{\Sigma}_{K,m_1}$ corresponding to $\hat{\lambda}_i$ $(i = 1, \ldots, \hat{v})$. Then we have

$$\hat{\Sigma}_{K,m_1} \times \begin{pmatrix} \mathbf{y}_{i1} \\ \mathbf{y}_{i2} \end{pmatrix} = \begin{pmatrix} A & B \\ C & D \end{pmatrix} \times \begin{pmatrix} \mathbf{y}_{i1} \\ \mathbf{y}_{i2} \end{pmatrix} = \begin{pmatrix} A\mathbf{y}_{i1} + B\mathbf{y}_{i2} \\ C\mathbf{y}_{i1} + D\mathbf{y}_{i2} \end{pmatrix} =$$

$$\hat{\lambda}_i \begin{pmatrix} \mathbf{y}_{i1} \\ \mathbf{y}_{i2} \end{pmatrix}.$$

Therefore $\hat{\Sigma}_{K,m_2} \times \begin{pmatrix} \mathbf{y}_{i1} \\ \mathbf{0}_{m_2-m_1} \\ \mathbf{y}_{i2} \end{pmatrix} =$

$$\begin{pmatrix} A & O_{(m_1-1)\times(m_2-m_1)} & B \\ O_{(m_2-m_1)\times(m_1-1)} & O_{(m_2-m_1)\times(m_2-m_1)} & O_{(m_2-m_1)\times K} \\ C & O_{K\times(m_2-m_1)} & D \end{pmatrix} \times \begin{pmatrix} \mathbf{y}_{i1} \\ \mathbf{0}_{m_2-m_1} \\ \mathbf{y}_{i2} \end{pmatrix} = \begin{pmatrix} A\mathbf{y}_{i1} + B\mathbf{y}_{i2} \\ \mathbf{0}_{m_2-m_1} \\ C\mathbf{y}_{i1} + D\mathbf{y}_{i2} \end{pmatrix}$$

$= \hat{\lambda}_i \begin{pmatrix} \mathbf{y}_{i1} \\ \mathbf{0}_{m_2-m_1} \\ \mathbf{y}_{i2} \end{pmatrix}$, and $\begin{pmatrix} \mathbf{y}_{i1} \\ \mathbf{0}_{m_2-m_1} \\ \mathbf{y}_{i2} \end{pmatrix}$ is the eigenvector of $\hat{\Sigma}_{K,m_2}$, obviously standardized. By far one can easily verify that $\hat{R}_{\hat{v}}$ does not depend on m. At the beginning, we proved $\hat{v}$ does not depend on m. So in all, the test procedure does not depend on m.

■