Statistica Sinica **23** (2013), 1215-1233
doi:http://dx.doi.org/10.5705/ss.2011.270

SPECIES ASSEMBLAGE COMPARISON WITH ABUNDANCE-BASED DATA USING ZERO-INFLATED POISSON MIXTURES

Jun Li and Jifei Ban

University of California, Riverside

Abstract: Species assemblage comparison is important in many ecological studies. In this paper, we develop a novel test for comparing species assemblages when abundance data from multiple quadrats are available. The test is based on the zero-inflated Poisson mixture model which we introduce to characterize the species assemblage given abundance data from multiple quadrats. We present a simulation study to evaluate the performance of our proposed test. The application of our test is further demonstrated on an ecological dataset.

Key words and phrases: Mixture model, species richness, zero-inflated Poisson.

1. Introduction

Comparison of species assemblages has important applications in ecology, since it provides crucial information about the spatial and temporal variations of ecosystems. There are two typical types of data collected in ecological studies: abundance-based data that contains the information of counts of each observed species in each sampling unit, and incidence-based data that only notes whether a species is present or absent in each sampling unit. Depending on the sampling procedure, the abundance-based data can be further divided into two categories. In one, the whole sampling area is treated as a single sampling unit, and the count information of each observed species is summarized for the whole area. The other has the sampling area divided into numerous plots, a sample of plots is randomly taken, and the count information of each observed species is recorded for each of the sampled plots. Following the terminology commonly used in ecology, we call those plots quadrats. We refer to the first type as abundance data from a single quadrat, and refer to the second type as abundance data from multiple quadrats.

In the literature, mixture models are popular choices to model the ecological data due to their capabilities to account for heterogeneity among species (see, for example, Ord and Whitmore (1986), Bunge and Fitzpatrick (1993), Chao and Bunge (2002), Böhning and Schön (2005), Mao and Colwell (2005), Mao (2006)). More specifically, for incidence-based data, the binomial mixture model is usually used, and for abundance data from a single quadrat, the Poisson mixture model is used. In Mao and Li (2009), a testing procedure was proposed to compare species assemblages under the binomial mixture model when the incidence-based data are available. Recently Li, Mao, and Wang (2012) developed a testing procedure under the Poisson mixture model when the abundance data from a single quadrat are available. In this paper, we focus on the species assemblage comparison problem when the abundance data from multiple quadrats are available. We also choose to work on the comparison problem under the mixture model framework. 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 numbers of species and the mixing distributions in the zero-inflated Poisson mixture model. However, neither of them can be well estimated nonparametrically in practice. To circumvent these difficulties, we develop a procedure for comparing some functions of the total numbers 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 numbers of species and the mixing distributions, which is ultimately equivalent to the comparison of species assemblages under our zero-inflated Poisson mixture model.

The rest of the paper is organized as follows. In Section 2, we describe our zero-inflated Poisson mixture model for abundance data from multiple quadrats. In Section 3, we introduce the hypothesis testing problem associated with the species assemblage comparison problem under the zero-inflated Poisson mixture model. In Section 4, we describe our testing procedure for comparing species assemblages. In Section 5, we report some simulation studies to evaluate the performance of our proposed test. In Section 6, we demonstrate the application of our test to an ecological data set. All proofs are collected in the Appendix.

2. Zero-inflated Poisson Mixture Model

To introduce some necessary notation, 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. If the species is present, the count of the species is recorded. Define

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

(ii) 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}, the Poisson distribution can be used. However, in many ecological data sets, some species may be present only in a small number of quadrats, which leads to a large frequency of zeros in the data. To account for this, we use the zero-inflated Poisson model. More specifically, the distribution of X_{ijk} is given by

$$Pr(X_{ijk} = x_{ijk}|\pi_{ij}, \lambda_{ij}) = \begin{cases} 1 - \pi_{ij}, & \text{if } x_{ijk}, = 0, \\ \pi_{ij} \frac{\exp(-\lambda_{ij})}{1-\exp(-\lambda_{ij})} \frac{\lambda_{ij}^{x_{ijk}}}{x_{ijk}!}, & \text{if } x_{ijk} > 0, \end{cases} \tag{2.1}$$

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 easy to see that this model includes the regular Poisson as a special case with $\pi_{ij} = 1 - \exp(-\lambda_{ij})$. Define $Z_{ijk} = I\{X_{ijk} \neq 0\}$, where $I\{A\}$ is the indicator function. The zero-inflated Poisson can be written as

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

Usually π_{ij} and λ_{ij} vary among species in one assemblage. To account for this heterogeneity among species, we assume that the π_{ij} are drawn from a latent distribution G_i, the λ_{ij} are drawn from a latent distribution H_i, and the π_{ij} and the λ_{ij} are independent. 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

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

We refer to this as the zero-inflated Poisson mixture.

3. Hypothesis Testing Problem

In the zero-inflated Poisson mixture model, species assemblage i is characterized by the number of species c_i and the mixing distributions G_i and H_i. Then comparing two species assemblages can be formulated as the hypothesis testing problem

$$H_0 : c_1 = c_2,\ G_1 = G_2,\ H_1 = H_2 \quad \text{versus} \quad H_a : c_1 \neq c_2 \text{ or } G_1 \neq G_2, \text{ or } H_1 \neq H_2. \tag{3.1}$$

Since it is difficult to verify parametric distribution assumptions for the G_i and H_i in practice, we take a nonparametric approach. To develop a testing procedure, one might first estimate $\{c_1, c_2, G_1, G_2, H_1, H_2\}$ nonparametrically. However, the c_i and G_i cannot be estimated well (e.g., Bunge and Fitzpatrick (1993), Huggins (2001), Link (2003), Mao (2006)). To circumvent such difficulties, we search for another hypothesis that is equivalent to (3.1), for which the parameters in the hypothesis admit close-form estimators. Toward this end, take

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

for $i = 1, 2$, $h = 1, 2, \ldots$ and $x = 1, 2, \ldots$. It is not difficult to see that $\tau_i(1) = \sum_{x=1}^{\infty} g_i(x)$, and that $\tau_i(h)$ is the species accumulation function, used widely in the ecology literature, and the desired result follows.

Theorem 1. *Given that the H_i have bounded support, $c_1 = c_2$, $G_1 = G_2$, and $H_1 = H_2$ if and only if $g_1(x) = g_2(x)$ for $x = 1, 2, \ldots,$ and $\tau_1(h) = \tau_2(h)$ for $h = 2, 3, \ldots$.*

Accordingly, the testing problem at (3.1) is equivalent to

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

versus

$$H_a : g_1(x) \neq g_2(x) \text{ for some } x \text{ or } \tau_1(h) \neq \tau_2(h) \text{ for some } h. \tag{3.2}$$

To go further, we need to find estimates for $g_i(x)$ and $\tau_i(h)$. Let $n_{i,k}$ be the number of species in assemblage i that appear in exactly k quadrats. According to Mao, Colwell, and Chang (2005), a nonparametric estimator of $\tau_i(h)$ is

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

To estimate $g_i(x)$, we take $n_{i,k,x}^v$ as the number of species that appear in exactly k quadrats and appear x times in the v-th ($v = 1, \ldots, k$) quadrat among those k quadrats. Let $B_{ij,t_1,\ldots,t_k}(x_{t_1}, \ldots, x_{t_{v-1}}, x, x_{t_{v+1}}, \ldots, x_{t_k}) = \{X_{ijt_1} = x_{t_1}, \ldots, X_{ijt_{v-1}} = x_{t_{v-1}}, X_{ijt_v} = x, X_{ijt_{v+1}} = x_{t_{v+1}} \ldots, X_{ijt_k} = x_{t_k}$, and all other $X_{ijk} = 0\}$. Then

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

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

Since

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

we have, for any $v = 1,\ldots,k$,

$$E(n_{i,k,x}^v) = c_i \int \binom{K_i}{k} \pi^k (1-\pi)^{K_i-k} dG_i(\pi) \int \frac{\exp(-\lambda)}{1-\exp(-\lambda)} \frac{\lambda^x}{x!} dH_i(\lambda). \tag{3.3}$$

Using the result in Mao, Colwell, and Chang (2005), $g_i(x)$ can be written as

$$\begin{aligned}
g_i(x) &= \tau_i(1) \int \frac{\exp(-\lambda)}{1-\exp(-\lambda)} \frac{\lambda^x}{x!} dH_i(\lambda)\\
&= \sum_{k=1}^{K_i} \left\{1 - \frac{\binom{K_i-1}{k}}{\binom{K_i}{k}}\right\} c_i \int \binom{K_i}{k} \pi^k (1-\pi)^{K_i-k} dG_i(\pi)\\
&\quad\times \int \frac{\exp(-\lambda)}{1-\exp(-\lambda)} \frac{\lambda^x}{x!} dH_i(\lambda). \qquad (3.4)
\end{aligned}$$

Therefore, based on (3.3) and (3.4), we have an unbiased estimate of $g_i(x)$,

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

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

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

Since $\tau_i(h)$ only admits a close-form nonparametric estimator for $h = 1,\ldots,$ K_i and $\hat{g}_i(x)$ is always zero for $x > m$, m is an arbitrarily large integer, henceforth we consider testing the hypothesis, implied by that at (3.2),

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

versus

$$H_a:\ g_1(x)\neq g_2(x) \text{ for some } x \text{ or } \tau_1(h)\neq\tau_2(h) \text{ for some } h, \tag{3.5}$$

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

Remark 1. Since H_0 at (3.2) implies H_0 at (3.5), the testing procedures proposed for testing H_0 at (3.5) can be also used for testing H_0 at (3.2). When used for testing H_0 at (3.2), the testing procedures still control the type I error at the nominal level but may admit a larger type II error.

4. The Proposed Test

If $\boldsymbol{\eta}_{i,K,m}=(g_i(1),\ldots,g_i(m),\tau_i(2),\ldots,\tau_i(K))'$, the hypothesis testing problem at (3.5) can be written as

$$H_0:\boldsymbol{\eta}_{1,K,m}=\boldsymbol{\eta}_{2,K,m} \quad \text{versus} \quad H_1:\boldsymbol{\eta}_{1,K,m}\neq\boldsymbol{\eta}_{2,K,m}, \tag{4.1}$$

Let

$$\begin{aligned}
\mathbf{n}_i &= (n_{i,1,1},\ldots,n_{i,1,m},\ldots,n_{i,K_i,1},\ldots,n_{i,K_i,m})',\\
A1_i &= (a_{i,1},\ldots,a_{i,k},\ldots,a_{i,K_i}) \text{ with } a_{i,k}=1-\binom{K_i-1}{k}\Big/\binom{K_i}{k},\\
A2_i &= (a_{i,h,k})_{h=2,k=1}^{K,K_i} \text{ with } a_{i,h,k}=1-\binom{K_i-h}{k}\Big/\binom{K_i}{k},\\
B1_i &= A1_i\bigotimes I_m, \text{ and } B2_i=A2_i\bigotimes \mathbf{1}_m',
\end{aligned}$$

where I_m is an m-dimensional identity matrix, $\mathbf{1}_m$ is a vector of m ones, and $\bigotimes$ is the Kronecker product. Take $T_i=\begin{pmatrix}B1_i\\B2_i\end{pmatrix}$ 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. Let $\mathcal{N}(\boldsymbol{\mu},\Omega)$ denote the multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix Ω.

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

Therefore, given the independence of the two species assemblages, $\hat{\boldsymbol{\eta}}_{1,K,m}-\hat{\boldsymbol{\eta}}_{2,K,m}$ is asymptotically $\mathcal{N}(\boldsymbol{\eta}_{1,K,m}-\boldsymbol{\eta}_{2,K,m},\Sigma_{K,m})$, where $\Sigma_{K,m}=T_1V_1T_1'+T_2V_2T_2'$, and a natural test statistic for the hypothesis testing problem at (4.1) 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 at (4.1) as $c_i\to\infty$.

In $R_{K,m}$, $\Sigma_{K,m}$ is unknown, and must be estimated. Since $\Sigma_{K,m} = T_1 V_1 T_1' + T_2 V_2 T_2'$, we study the structure of V_i in order to develop an appropriate estimator for it. For this, let

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

to get the following.

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

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

(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 (3.3), $r_{i,k} s_{i,x}$ 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 (3.3), we take $n_{i,k,x,y}^{v_1,v_2}$ as the number of species that appear in exactly k quadrats and that 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 $v_1, v_2 = 1, 2, \ldots, k$, 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 $n_{i,k,x,y} = \sum_{1 \le v_1 < v_2 \le K_i} n_{i,k,x,y}^{v_1,v_2} / \binom{k}{2}$.

Plugging these estimates into V_i, we obtain an estimator for $\Sigma_{K,m}$, denoted $\hat{\Sigma}_{K,m}$, and note the following.

Proposition 2. *$\hat{\Sigma}_{K,m}$ is a positive semi-definite matrix.*

Plugging $\hat{\Sigma}_{K,m}$ into $R_{K,m}$, we can reject H_0 in (4.1) at a nominal level α if

$$\hat{R}_{K,m} = (\hat{\boldsymbol{\eta}}_{1,K,m} - \hat{\boldsymbol{\eta}}_{2,K,m})' \hat{\Sigma}_{K,m}^{-1} (\hat{\boldsymbol{\eta}}_{1,K,m} - \hat{\boldsymbol{\eta}}_{2,K,m}) > \chi^2_{1-\alpha,m+K-1}, \tag{4.2}$$

where $\chi^2_{1-\alpha,m+K-1}$ is the $(1-\alpha)$ quantile of χ^2_{m+K-1}. When implementing this, we often encounter singular $\hat{\Sigma}_{K,m}$. To circumvent this, we note that the correlations between the components of $\hat{\boldsymbol{\eta}}_{1,K,m} - \hat{\boldsymbol{\eta}}_{2,K,m}$ are often very large, and that the first few principal components of $\hat{\Sigma}_{K,m}$ usually account for the most variability. Thus we follow Mao and Li (2009) and focus on these principal components to test (4.1). Specifically, consider the eigenvalue decomposition $\hat{\Sigma}_{K,m} = \hat{P}\hat{\Lambda}\hat{P}'$, where $\hat{\Lambda} = \text{diag}\{\hat{\lambda}_1, \ldots, \hat{\lambda}_{m+K-1}\}$, $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_{m+K-1}$ 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 $t = 0.9999$, take

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

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 at (4.1) at the nominal level α if

$$\hat{R}_{\hat{\nu}} = (\hat{\boldsymbol{\eta}}_{1,K,m} - \hat{\boldsymbol{\eta}}_{2,K,m})' \hat{P}_{\hat{\nu}} \hat{\Lambda}_{\hat{\nu}}^{-1} \hat{P}_{\hat{\nu}}' (\hat{\boldsymbol{\eta}}_{1,K,m} - \hat{\boldsymbol{\eta}}_{2,K,m}) > \chi^2_{1-\alpha,\hat{\nu}}. \tag{4.3}$$

4.1. Choice of m

Recall that the $n_{i,k,x}$ are zero for $x > m$ as are the $\hat{g}_i(x)$, if we choose m as some arbitrarily large integer. Thus, if we choose $m = m_1$ and m_2, and the $n_{i,k,x}$ are zeros for $x > m_i$ $(i = 1,2)$, then $\hat{R}_{K,m_1} = \hat{R}_{K,m_2}$. Therefore $\hat{R}_{K,m}$ does not depend on the choice of m as long as it is large enough. However, when implementing (4.2) the threshold $\chi^2_{1-\alpha,m+K-1}$ does depend on the choice of m, and different choices of m may yield different conclusions. This is not an issue if we implement (4.3), as follows. In (4.3) the threshold $\chi^2_{1-\alpha,\hat{\nu}}$ only depends on the degree of freedom $\hat{\nu}$. Given t, $\hat{\nu}$ is determined by the nonzero eigenvalues of $\hat{\Sigma}_{K,m}$, which can be shown not to depend on the choice of m as long as it is large enough. It can be also shown that $\hat{R}_{\hat{\nu}}$ does not depend on those choices of m. Therefore, the test at (4.3) yields the same conclusion no matter the choice of m, as long as it is large enough that the $n_{i,k,x}$ are zero for $x > m$. Accordingly, the test at (4.3) is recommended. We call it the eigenvalue adjusted (Eva) χ^2 test, similar to the term used in Mao and Li (2009).

4.2. Impact of using different $\hat{c}_i$

In our testing procedure, we need an estimator for c_i. Mao (2007) found that there is no unbiased 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 (4.3). Based on our simulations, the test tends to be conservative when the upper bound $\hat{c}_i = \infty$ is used, while it tends to be liberal if the lower bound estimators are used. The impact of using different $\hat{c}_i$ can be seen in our simulation studies.

Alternatively, to avoid problems caused by the biased estimates of c_i, we can use the bootstrap method to approximate the null distribution of $\hat{R}_{\hat{\nu}}$. Here first generate the bootstrap resample of $n_{i,+}$, denoted $n^*_{i,+}$, from Binomial ($\hat{c}_{i,Chao}$,$n_{i,+}/\hat{c}_{i,Chao}$). With $n^*_{i,+}$ species in assemblage i, for species j ($j = 1, \ldots, n^*_{i,+}$), randomly choose $k^*_{i,j}$ 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 $\pi^*_{i,j}$, where $\pi^*_{i,j}$ is drawn from $\hat{Q}_i$ with $\hat{Q}_i$ the nonparametric maximum likelihood estimator of Q_i with $dQ_i(\pi) = [(1-(1-\pi)^{K_i})dG_i(\pi)]/[\int(1-(1-\varpi)^{K_i})dG_i(\varpi)]$ (Mao, Colwell, and Chang (2005)). Next, for species j ($j = 1, \ldots, n^*_{i,+}$) in assemblage i, in each one of the $k^*_{i,j}$ quadrats where species j appears, generate the count of species j, X^*_{ijk}, from a zero-truncated Poisson distribution with 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} ($i = 1, 2,\ j = 1, \ldots, n^*_{i,+},\ k = 1, \ldots, K_i$), we can calculate

$$\begin{aligned}\hat{R}^*_{\hat{\nu}} = (\hat{\boldsymbol{\eta}}^*_{1,K,m} - \hat{\boldsymbol{\eta}}^*_{2,K,m} - (\hat{\boldsymbol{\eta}}_{1,K,m} - \hat{\boldsymbol{\eta}}_{2,K,m}))'\hat{P}^*_{\hat{\nu}}(\hat{\Lambda}^*_{\hat{\nu}})^{-1}(\hat{P}^*_{\hat{\nu}})' \\ \times(\hat{\boldsymbol{\eta}}^*_{1,K,m} - \hat{\boldsymbol{\eta}}^*_{2,K,m} - (\hat{\boldsymbol{\eta}}_{1,K,m} - \hat{\boldsymbol{\eta}}_{2,K,m})),\end{aligned} \tag{4.4}$$

where $\hat{\boldsymbol{\eta}}^*_{1,K,m} - \hat{\boldsymbol{\eta}}^*_{2,K,m}$, $\hat{P}^*_{\hat{\nu}}$ and $\hat{\Lambda}^*_{\hat{\nu}}$ are the counterparts 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}. Repeat the 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 (4.4) calculated from the j-th bootstrap resample. The Eva-bootstrap testing procedure is to reject H_0 in (4.1) at the level α if $\hat{R}_{\hat{\nu}} > \kappa_{1-\alpha}$.

Remark 2. The proposed testing procedures can be easily extended to the comparison of L ($L \geq 2$) species assemblages. Following earlier notation, we take, for $l = 1, \ldots, L$, $\boldsymbol{\eta}_{l,K,m} = (g_l(1), \ldots, g_l(m), \tau_l(2), \ldots, \tau_l(K))'$, where m is some arbitrarily large number and $K = min(K_1, \ldots, K_L)$. Then comparing L species assemblages can be formulated as the hypothesis testing problem

$$H_0 : \boldsymbol{\eta}_{1,K,m} = \cdots = \boldsymbol{\eta}_{L,K,m} \text{ versus } H_1 : \boldsymbol{\eta}_{i,K,m} \neq \boldsymbol{\eta}_{j,K,m} \text{ for some } i \neq j.$$

Let $\boldsymbol{d} = (\boldsymbol{\eta}'_{1,K,m} - \boldsymbol{\eta}'_{L,K,m}, \ldots, \boldsymbol{\eta}'_{L-1,K,m} - \boldsymbol{\eta}'_{L,K,m})'$. This 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 $\boldsymbol{\eta}_{l,K,m}$ by $\hat{\boldsymbol{\eta}}_{l,K,m}$, $l = 1, \ldots, L$. A natural estimate for $\boldsymbol{d}$ is $\hat{\boldsymbol{d}} = (\hat{\boldsymbol{\eta}}'_{1,K,m} - \hat{\boldsymbol{\eta}}'_{L,K,m}, \hat{\boldsymbol{\eta}}'_{2,K,m} - \hat{\boldsymbol{\eta}}'_{L,K,m}, \ldots,$

$\hat{\boldsymbol{\eta}}'_{L-1,K,m} - \hat{\boldsymbol{\eta}}'_{L,K,m})'$. One can easily prove that, asymptotically, $\hat{\boldsymbol{d}}$ is $\mathcal{N}(\boldsymbol{d}, \Sigma_L)$, where

$$\Sigma_L = \bigoplus_{l=1,2,\ldots,L-1} W_l + (\mathbf{1}_{L-1}\mathbf{1}'_{L-1}) \bigotimes W_L, \tag{4.5}$$

W_l is the covariance matrix of $\hat{\boldsymbol{\eta}}_{l,K,m}$, and $\bigoplus$ is the direct sum. Therefore, we can reject H_0 at the significance level α if $\hat{\boldsymbol{d}}'\hat{\Sigma}_L^{-1}\hat{\boldsymbol{d}} > \chi^2_{1-\alpha,(L-1)(m+K-1)}$, where $\hat{\Sigma}_L$ is the estimator of Σ_L by plugging the estimators of the W_l in (4.5). There are then Eva-χ^2 and Eva-bootstrap tests for the L species assemblage comparison problem. As in 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 $\hat{c}_i$.

5. Simulations

We first report on a simulation study to assess the type I error of the Eva-χ^2 test. The study consisted of 36 simulation settings, determined by the total number of species $c_i(c_1 = c_2 = 500$ or 2,000), 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 $\mathscr{D}_H$). Here $\mathscr{B}$ is the beta distribution with shape parameters 1 and 20, logit$\mathscr{N}$ is obtained by letting $\log\{\pi/(1-\pi)\}$ be normal 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 type I error, we took $\hat{c}_i = \hat{c}_{i,Chao}$ and $\hat{c}_i = \infty$ in the calculation of $\hat{R}_{\hat{\nu}}$ at (4.3). To benchmark the performance, we also include the type I errors of the test when $\hat{c}_i = c_i$ was used in the test. In all tests, we used $t = 0.9999$ in the eigenvalue decomposition to choose $\hat{\nu}$, and the nominal size of the test α was set at 0.05. Tables 1 and 2 summarize the simulated type I errors of the Eva-χ^2 test based on 500 replications. From Tables 1 and 2, we can see that the estimator of c_i has a significant impact on the type I errors. With high-quality estimators for the c_i, the type I errors would approach the nominal level. If the c_i are underestimated, the Eva-χ^2 test is liberal; if $\hat{c}_i = \infty$ is used, the Eva-χ^2 test is conservative. Tables 1 and 2 also include the type I errors of the Eva-bootstrap test. For computation simplicity, we only considered the Eva-bootstrap test with $\hat{c}_i = \infty$ used in the calculation of $\hat{R}_{\hat{\nu}}$. The number of bootstrap resamples, B, was 500. As we can see from Tables 1 and 2, the Eva-bootstrap test corrects the conservativeness of the corresponding Eva-χ^2 test and approximately achieves its nominal type I error. Based on the

Table 1. 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 2,000, $K_1 = K_2 = 50$, $G_1 = G_2 = \mathscr{B}$, 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
$(500,50,\mathscr{B},\text{log}\mathscr{N})$	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
$(500,50,\text{logit}\mathscr{N},\text{log}\mathscr{N})$	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
$(500,50,\mathscr{D}_G,\text{log}\mathscr{N})$	0.066	0.076	0.044	0.062
$(500,50,\mathscr{D}_G,\mathscr{D}_H)$	0.058	0.072	0.028	0.050
$(2000,50,\mathscr{B},\mathscr{G})$	0.056	0.090	0.020	0.054
$(2000,50,\mathscr{B},\text{log}\mathscr{N})$	0.056	0.092	0.012	0.044
$(2000,50,\mathscr{B},\mathscr{D}_H)$	0.048	0.086	0.026	0.050
$(2000,50,\text{logit}\mathscr{N},\mathscr{G})$	0.058	0.076	0.046	0.042
$(2000,50,\text{logit}\mathscr{N},\text{log}\mathscr{N})$	0.036	0.062	0.036	0.058
$(2000,50,\text{logit}\mathscr{N},\mathscr{D}_H)$	0.046	0.068	0.026	0.048
$(2000,50,\mathscr{D}_G,\mathscr{G})$	0.07	0.082	0.032	0.072
$(2000,50,\mathscr{D}_G,\text{log}\mathscr{N})$	0.042	0.050	0.016	0.046
$(2000,50,\mathscr{D}_G,\mathscr{D}_H)$	0.050	0.068	0.026	0.046

simulation results, we suggest using the Eva-χ^2 test with both $\hat{c}_i = \hat{c}_{i,Chao}$ and $\hat{c}_i = \infty$. If they yield the same conclusion, take it; otherwise, resort to the Eva-bootstrap test.

As suggested by the referees, we also conducted simulations with a smaller $c_i (c_1 = c_2 = 100)$. The simulation results were similar to the ones reported above, indicating our procedure can work well with relatively small populations. To investigate the effect of other estimators of c_i on type I errors, we also used $\hat{c}_i =$ Chao's abundance coverage estimator and Chao's incidence coverage estimator (Chao and Lee (1992), Chao, Ma, and Yang (1993), Lee and Chao (1994)) in the calculation of $\hat{R}_{\hat{\nu}}$ at (4.3). As in the $\hat{c}_i = \hat{c}_{i,Chao}$ case, the test based on these estimators also tended to be liberal, rejecting more often than desired.

We also carried out a simulation study to assess the power of the Eva-bootstrap test for detecting differences of two species assemblages, with $K_i = 50$ $(i = 1, 2)$. Then each of the species assemblages can be represented by (c_i, G_i, H_i) $(i = 1, 2)$. The exact simulation settings are listed in Table 3, where $\mathscr{D}_G$, $\mathscr{D}_G^*$, $\mathscr{D}_H$, and $\mathscr{D}_H^*$ are discrete distributions with support points $(0.02, 0.1, 0.2, 0.3)$, $(0.025, 0.1, 0.2, 0.3)$, $(1, 2, 5, 10)$, and $(4, 2, 5, 10)$, respectively, and common weights of $(0.65, 0.2, 0.1, 0.05)$. In Settings 1-3, the two species

Table 2. 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 2,000, $K_1 = K_2 = 150$, $G_1 = G_2 = \mathscr{B}$, 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)	$\hat{c}_i = c_i$	$\hat{c}_i = \hat{c}_{Chao}$	$\hat{c}_i = \infty$	$\hat{c}_i = \infty$
$(500,150,\mathscr{B},\mathscr{G})$	0.036	0.094	0.024	0.050
$(500,150,\mathscr{B},\text{log}\mathscr{N})$	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
$(500,150,\text{logit}\mathscr{N},\text{log}\mathscr{N})$	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
$(500,150,\mathscr{D}_G,\text{log}\mathscr{N})$	0.050	0.052	0.032	0.048
$(500,150,\mathscr{D}_G,\mathscr{D}_H)$	0.056	0.062	0.034	0.052
$(2000,150,\mathscr{B},\mathscr{G})$	0.038	0.088	0.012	0.038
$(2000,150,\mathscr{B},\text{log}\mathscr{N})$	0.058	0.100	0.046	0.068
$(2000,150,\mathscr{B},\mathscr{D}_H)$	0.040	0.112	0.016	0.038
$(2000,150,\text{logit}\mathscr{N},\mathscr{G})$	0.042	0.062	0.030	0.058
$(2000,150,\text{logit}\mathscr{N},\text{log}\mathscr{N})$	0.044	0.080	0.034	0.050
$(2000,150,\text{logit}\mathscr{N},\mathscr{D}_H)$	0.070	0.100	0.032	0.058
$(2000,150,\mathscr{D}_G,\mathscr{G})$	0.058	0.056	0.034	0.052
$(2000,150,\mathscr{D}_G,\text{log}\mathscr{N})$	0.048	0.052	0.016	0.040
$(2000,150,\mathscr{D}_G,\mathscr{D}_H)$	0.042	0.048	0.022	0.040

assemblages differ in only one out of the three species assemblage characteristics c, G and H; in Settings 4−9, the two assemblages differ in two out of the three characteristics; in Settings 10−13, the two assemblages differ in all three characteristics.

The simulated power of the Eva-bootstrap test, from 500 simulations for each setting, is reported in column "abundance-Eva" of Table 3. In this simulation study, ignoring the count information of the observed species in each of the sampled quadrats leads to incidence-based data. Given such data, the Eva-bootstrap test in Mao and Li (2009) can be used to detect species assemblage differences. We also applied the Eva-bootstrap test to the data. The simulated powers are reported in column "incidence-Eva" of Table 3. In both cases, the nominal level of the test was set to 0.05. To distinguish the two Eva-bootstrap tests, we refer to the one proposed here as the abundance Eva-bootstrap test, and the Mao and Li (2009) test as the incidence Eva-bootstrap test.

As we can see from Table 3, our abundance Eva-bootstrap test has good power detecting a variety of species assemblage differences. Furthermore, in Settings 1, 2, 4, 5 where the H_i are the same for both species assemblages, our abundance Eva-bootstrap test performs similarly to the incidence Eva-bootstrap

Table 3. The simulated power of the bootstrap adjusted χ^2 tests for abundance-based data and incidence-based data.

Setting	(c_1, G_1, H_1)	(c_2, G_2, H_2)	abundance-Eva	incidence-Eva
1	$(500, \mathscr{D}_G, \mathscr{D}_H)$	$(450, \mathscr{D}_G, \mathscr{D}_H)$	0.144	0.138
2	$(500, \mathscr{D}_G, \mathscr{D}_H)$	$(500, \mathscr{D}_G^*, \mathscr{D}_H)$	0.256	0.260
3	$(500, \mathscr{D}_G, \mathscr{D}_H)$	$(500, \mathscr{D}_G, \mathscr{D}_H^*)$	0.480	0.054
4	$(500, \mathscr{D}_G, \mathscr{D}_H)$	$(450, \mathscr{D}_G^*, \mathscr{D}_H)$	0.202	0.204
5	$(500, \mathscr{D}_G^*, \mathscr{D}_H)$	$(450, \mathscr{D}_G, \mathscr{D}_H)$	0.716	0.706
6	$(500, \mathscr{D}_G, \mathscr{D}_H)$	$(450, \mathscr{D}_G, \mathscr{D}_H^*)$	0.546	0.138
7	$(500, \mathscr{D}_G, \mathscr{D}_H^*)$	$(450, \mathscr{D}_G, \mathscr{D}_H)$	0.524	0.132
8	$(500, \mathscr{D}_G, \mathscr{D}_H)$	$(500, \mathscr{D}_G^*, \mathscr{D}_H^*)$	0.456	0.234
9	$(500, \mathscr{D}_G, \mathscr{D}_H^*)$	$(500, \mathscr{D}_G^*, \mathscr{D}_H)$	0.546	0.284
10	$(500, \mathscr{D}_G, \mathscr{D}_H)$	$(450, \mathscr{D}_G^*, \mathscr{D}_H^*)$	0.406	0.200
11	$(500, \mathscr{D}_G, \mathscr{D}_H^*)$	$(450, \mathscr{D}_G^*, \mathscr{D}_H)$	0.496	0.182
12	$(500, \mathscr{D}_G^*, \mathscr{D}_H)$	$(450, \mathscr{D}_G, \mathscr{D}_H^*)$	0.824	0.696
13	$(500, \mathscr{D}_G^*, \mathscr{D}_H^*)$	$(450, \mathscr{D}_G, \mathscr{D}_H)$	0.764	0.702

test, indicating that our abundance Eva-bootstrap test which is more general does not lose efficiency for testing simpler hypotheses. In Setting 3, where the c_i and G_i are the same in both species assemblages and the assemblages differ only in H, the incidence Eva-bootstrap test has no power. This is expected since the incidence Eva-bootstrap test can only test $H_0 : c_1 = c_2, G_1 = G_2$. In the remaining settings where the assemblages differ in both H and (c, G), the abundance Eva-bootstrap test significantly outperforms the incidence Eva-bootstrap test. Thus, overall, our abundance Eva-bootstrap test outperforms the incidence Eva-bootstrap test. This is not surprising since our abundance Eva-bootstrap test uses all the information in the data, while the incidence Eva-bootstrap test does not. Ignoring the abundance information in the data and resorting to the simpler incidence Eva-bootstrap test can lead to significant loss of power for detecting differences of species assemblages.

6. An 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, Redondo Brenes, and Vilchez Alvarado (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 choose the seedling assemblage data collected from the study sites Lindero Sur (LSUR) and Tirimbina (TIR). In both sites, all seedlings were sampled in 144 1m × 5m quadrats in 12 strips through the 50m × 200m plot. The species identity was determined for all seedlings (> 20 cm in height, but < 1 cm in diameter at breast height). In LSUR, 132 species with 2,287 individuals were observed, and in TIR, 153 species with 3,443 were observed. Chao's lower bound estimators $\hat{c}_{i,Chao}$ for these two sites are 169 and 196, respectively.

To determine whether there is a difference between these two seedling assemblages, we can apply our test. We choose $m = 118$ and the number of principal components $\hat{\nu} = 4$ according to $t = 0.9999$. 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 $\hat{R}_{\hat{\nu}}$ in (4.3), respectively. Both p-values are smaller than 0.05. Based on our simulation studies, using $\hat{c}_i = \infty$ often leads to a conservative procedure. The null hypothesis is rejected even when using $\hat{c}_i = \infty$. This implies that there is enough evidence to reject the null hypothesis that there is no difference between these two seedling assemblages. Our Eva-bootstrap test yields a p-value 0.004, which further confirms that there is a significant difference between them.

With abundance data from multiple quadrats, we can always treat them as incidence data by ignoring the count information of the observed species in each sampling quadrat and then apply the incidence Eva-bootstrap test proposed in Mao and Li (2009) to test whether the two species assemblages are the same. Applying the incidence Eva-bootstrap test to the same assemblages, the p-value is 0.018. The fact that the p-value of our abundance Eva-bootstrap test is smaller than the p-value of the incidence Eva-bootstrap test further confirms that our abundance Eva-bootstrap test is, in general, more powerful than the incidence Eva-bootstrap test.

For each of these seedling assemblages, one can also pool all the count information for each observed species across all the sampling quadrats and treat the data as abundance data from a single quadrat. Then the test proposed in Li, Mao, and Wang (2012) can be applied. The p-value from that test is 0.308, thus finds no significant difference. The opposing results can be explained as follows. Pooling the count information and applying the test in Li, Mao, and Wang (2012) can only test whether or not the overall abundances of the species are the same, but it cannot tell whether or not the distributions of the abundances of each species across the sampling quadrats are the same. The disparity between the distributions of the abundances of each species across the sampling quadrates can be of interest to ecologists.

7. Concluding Remarks

In this paper, we propose a testing procedure for comparing species assemblages when abundance data from multiple quadrats are available. The testing procedure is based on the zero-inflated Poisson mixture model we introduce to characterize the abundance data from multiple quadrats. Since we use the nonparametric approach for estimating G and H, to verify our zero-inflated Poisson mixture model assumption, we need only check whether the zero-truncated Poisson distribution is a reasonable distribution for the non-zero abundances of each observed species. Some existing goodness-of-fit procedures, for example Chi-squared goodness-of-fit test, can be used for this purpose.

The abundance data we deal with in this paper are taken from a single snapshot of the species assemblages. Due to the nature of such data, our method does not take into account the dynamics of species assemblages, i.e., how birth, death, immigration and other factors affect the species abundance. Recently, stochastic models have been proposed to study such dynamics when such information as abundance data from sequential sampling or species traits is available (for example, Alonso, Ostling, and Etienne (2008), Jabot (2010)). It is of interest to incorporate those models into our comparison procedures in our future research.

Acknowledgement

The authors thank Professors Robert K. Colwell and Robin L. Chazdon for generously sharing data. They also thank the Editor, an Associate Editor, and the referees for their thoughtful comments, which have helped improve our article. Jun Li would like to thank National Science Foundation for supporting her research through Grant DMS-0907655.

Appendix

Proof of Theorem 1. It is straightforward that, if $c_1 = c_2$, $G_1 = G_2$, and $H_1 = H_2$, then $g_1(x) = g_2(x)$ for $x = 1, 2, \ldots$ and $\tau_1(h) = \tau_2(h)$ for $h = 2, 3, \ldots$. When $g_1(x) = g_2(x)$ for $x = 1, 2, \ldots$, $\tau_1(1) = \tau_2(1)$ since $\tau_i(1) = \sum_{x=1}^{\infty} g_i(x)$. Together with $\tau_1(h) = \tau_2(h)$ for $h = 2, \ldots, \infty$, we have $c_1 = c_2$ and $G_1 = G_2$, following Theorem 2 of Mao and Li (2009). Therefore, $g_1(x) = g_2(x)$ implies that

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

thus

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

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

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

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

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

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, $H_1 = H_2$.

Proof of Theorem 2. We put $I^v_{i,j,k,x} = I\{$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

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

Based on our assumptions, the $\mathbf{I}_{i,j}$ are i.i.d., and so $\{\mathbf{n}_i - E(\mathbf{n}_i)\}$ converges to $\mathcal{N}(\mathbf{0}, V_i)$ in distribution as c_i goes to ∞. Since $\hat{\boldsymbol{\eta}}_{i,K,m} = T_i\mathbf{n}_i$, the result follows by using the delta method.

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

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

Based on the definitions of $n_{i,k}$ and $n^1_{i,k,x}$, $\{n_{i,0},$ and $n^1_{i,k,x}, k = 1, \ldots, K_i, x = 1, 2, \ldots\}$ follows a multinomial distribution with size c_i and probabilities $r_{i,0}$ and $r_{i,k}s_{i,x}$. Therefore, $\mathrm{var}(n^1_{i,k,x}) = c_i r_{i,k}s_{i,x}(1 - r_{i,k}s_{i,x})$. We also have $\mathrm{Cov}\,(n^1_{i,k,x}, n^1_{i,k,y}) = -c_i {r_{i,k}}^2 s_{i,x}s_{i,y}$, and $\mathrm{Cov}\,(n^1_{i,k,x}, n^1_{i,l,y}) = -c_i r_{i,k}r_{i,l}s_{i,x}s_{i,y}$.

To find $\mathrm{Cov}\,(n^1_{i,k,x}, n^2_{i,k,x})$, recall that $n^v_{i,k,x} = \sum_{j=1}^{c_i} I^v_{i,j,k,x}$. Since the species are independent of each other,

$$\mathrm{Cov}\,(n^1_{i,k,x}, n^2_{i,k,x}) = \sum_{j=1}^{c_i} \mathrm{Cov}\,(I^1_{i,j,k,x}, I^2_{i,j,k,x})$$

$$= \sum_{j=1}^{c_i} E(I^1_{i,j,k,x} I^2_{i,j,k,x}) - \sum_{j=1}^{c_i} E(I^1_{i,j,k,x}) E(I^2_{i,j,k,x})$$

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

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

Thus, $\operatorname{Cov}(n^1_{i,k,x}, n^2_{i,k,x}) = c_i r_{i,k}s_{i,x,x} - c_i(r_{i,k}s_{i,x})^2$ and

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

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

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

Similar to the proof in (a), we can obtain $\operatorname{Cov}(n^1_{i,k,x}, n^1_{i,k,y}) = -c_i {r_{i,k}}^2 s_{i,x}s_{i,y}$ and $\operatorname{Cov}(n^1_{i,k,x}, n^2_{i,k,y}) = c_i r_{i,k}s_{i,x,y} - c_i {r_{i,k}}^2 s_{i,x}s_{i,y}$. Therefore,

$$\begin{aligned} \operatorname{Cov}(n_{i,k,x}, n_{i,k,y}) &= \left\{ -c_i r_{i,k}s_{i,x}r_{i,k}s_{i,y} + (k-1)(c_i r_{i,k}s_{i,x,y} - c_i r^2_{i,k}s_{i,x}s_{i,y}) \right\} k^{-1} \\ &= \left\{ (k-1)c_i r_{i,k}s_{i,x,y} - kc_i r^2_{i,k}s_{i,x}s_{i,y} \right\} k^{-1}. \end{aligned}$$

(c) Since $\operatorname{Cov}(n_{i,k,x}, n_{i,l,y}) = \operatorname{Cov}(n^1_{i,k,x}, n^1_{i,l,y}) = -c_i r_{i,k}s_{i,x}r_{i,l}s_{i,y}$.

Proof of Proposition 2. Take $n_{i,+} = \sum_{k=1}^{K_i} n_{i,k}$, the number of species observed in assemblage i. W.l.o.g, we assume that species j, $j = 1, \ldots, n_{i,+}$, is observed in the sample. Following the notation in the proof of Theorem 2, we denote the sample covariance matrix of $\mathbf{I}_{i,1}, \ldots, \mathbf{I}_{i,n_{i,+}}$ by S_i, and its element in 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 show 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, $S_{i,x+m(k-1),x+m(k-1)}$ we have,

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

$$= \frac{1}{n_{i,+}} \sum_{j=1}^{n_{i,+}} \Big\{ \frac{1}{k^2} \Big(\sum_{v=1}^{k} I_{i,j,k,x}^{v} + 2 \sum_{1 \le v_1 < v_2 \le k} I_{i,j,k,x}^{v_1} I_{i,j,k,x}^{v_2} \Big) \Big\} - \frac{n_{i,k,x}^2}{n_{i,+}^2}$$
$$= \frac{1}{n_{i,+}} \left[\frac{1}{k^2} \left\{ k \cdot n_{i,k,x} + k(k-1) n_{i,k,x,x} \right\} \right] - \frac{n_{i,k,x}^2}{n_{i,+}^2}$$
$$= \frac{1}{k} \left\{ \frac{n_{i,k,x}}{n_{i,+}} + \frac{(k-1) n_{i,k,x,x}}{n_{i,+}} \right\} - \frac{n_{i,k,x}^2}{n_{i,+}^2}.$$

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

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

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

Similarly, for any off-diagonal elements in S_i, when $0 < x, y \le m$ and $1 \le k \le K_i$,

$$S_{i,x+m(k-1),y+m(k-1)} = \frac{1}{k} \Big\{ \frac{(k-1) n_{i,k,x,y}}{n_{i,+}} - \frac{k n_{i,k,x} n_{i,k,y}}{n_{i,+}^2} \Big\}.$$

For $\hat{V}_{i,x+m(k-1),y+m(k-1)}$, the estimate of $\mathrm{Cov}\,(n_{i,k,x}, n_{i,k,y})$, we have

$$\hat{V}_{i,x+m(k-1),y+m(k-1)} = \frac{1}{k} \Big\{ (k-1) n_{i,k,x,y} - \frac{k n_{i,k,x} n_{i,k,y}}{\hat{c}_i} \Big\}.$$

Therefore,

$$\hat{V}_{i,x+m(k-1),y+m(k-1)} = n_{i,+} \cdot S_{i,x+m(k-1),y+m(k-1)} + n_{i,k,x} n_{i,k,y} \Big(\frac{1}{n_{i,+}} - \frac{1}{\hat{c}_i} \Big).$$

When $0 < x, y \le m$ and $1 \le k \ne l \le K_i$,

$$S_{i,x+m(k-1),y+m(l-1)} = -\frac{n_{i,k,x} n_{i,l,y}}{n_{i,+}^2}.$$

For $\hat{V}_{i,x+m(k-1),y+m(l-1)}$, which is the estimate of $\mathrm{Cov}\,(n_{i,k,x}, n_{i,l,y})$, we have $\hat{V}_{i,x+m(k-1),y+m(l-1)} = -n_{i,k,x} n_{i,l,y} / \hat{c}_i$. Therefore, $\hat{V}_{i,x+m(k-1),y+m(l-1)} = n_{i,+} \cdot S_{i,x+m(k-1),y+m(l-1)} + n_{i,k,x} n_{i,l,y} (1/n_{i,+} - 1/\hat{c}_i)$.

Thus we have shown that $\hat{V}_i = n_{i,+} \cdot S_i + \boldsymbol{n_i} \boldsymbol{n_i}' (1/n_{i,+} - 1/\hat{c}_i)$. From this it is not difficult to see that $\hat{V}_i$ is positive semi-definite, since S_i is the sample covariance matrix, $\boldsymbol{n_i} \boldsymbol{n_i}'$ is positive semi-definite, $n_{i,+} > 0$, and $\hat{c}_i \ge n_{i,+}$. Since $\hat{\Sigma}_{K,m} = T_1 \hat{V}_1 T_1' + T_2 \hat{V}_2 T_2'$, the result follows.

References

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

Böhning, D. and Schön, D. (2005). Nonparametric maximum likelihood estimation of population size based on the counting distribution. *Appl. Statist.* **54**, 721-737.

Bunge, J. and Fitzpatrick, M. (1993). Estimating the number of species: a review. *J. Amer. Statist. Assoc.* **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. and Lee, S. M. (1992). Estimating the number of classes via sample coverage. *J. Amer. Statist. Assoc.* **87**, 210-217.

Chao, A., Ma, M. C., and Yang, M. C. K. (1993). Stopping rules and estimation for recapture debugging with unequal failure rates. *Biometrika* **80**, 193-201.

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.

Huggins, R. (2001). A note on the difficulties associated with the analysis of capture-recapture experiments with heterogeneous capture probabilities. *Statist. Probab. Lett.* **54**, 147-152.

Jabot, F. (2010). A stochastic dispersal-limited trait-based model of community dynamics. Journal of Theoretical Biology 262, 650-661.

Lee, S. M. and Chao, A. (1994). Estimating population size via sample coverage for closed capture-recapture models. Biometrics 50, 88-97.

Li, J., Mao, C. X. and Wang, S. (2012). Comparing species assemblages with abundance data. *Scand. J. Statist.*, under review.

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

Mao, C. X. (2006). Inference of the number of species via geometric lower bounds. *J. Amer. Statist. Assoc.* **101**, 1663-1670.

Mao, C. X. (2007). Estimating population sizes for capture-recapture sampling with binomial mixtures. *Comput. Statist. Data Anal.* **51**, 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**, 433-441.

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

Ord, J. K. and Whitmore, G. (1986). The Poisson-inverse Gaussian distribution as a model for species abundance. *Comm. Statist. Theory Methods* **15**, 853-871.

Department of Statistics, University of California - Riverside, Riverside, CA 92521, U.S.A.
E-mail: jun.li@ucr.edu

Department of Statistics, University of California - Riverside, Riverside, CA 92521, U.S.A.
E-mail: jifei.ban@gmail.com

(Received October 2011; accepted August 2012))