Biometrics 61, 433–441
June 2005

DOI: 10.1111/j.1541-0420.2005.00316.x

Estimating the Species Accumulation Curve Using Mixtures

Chang Xuan Mao,[1,*] Robert K. Colwell,[2] and Jing Chang[3]

[1]Department of Statistics, University of California, Riverside, California 92521, U.S.A.
[2]Department of Ecology and Evolutionary Biology, University of Connecticut, Storrs, Connecticut 06269, U.S.A.
[3]Department of Preventive Medicine, School of Medicine, University of Southern California, Los Angeles, California 90089, U.S.A.
[*]*email:* cmao@stat.ucr.edu

Summary. As a significant tool in ecological studies, the species accumulation curve or the collector's curve is the graph of the expected number of detected species as a function of sampling effort. The problem of estimating the species accumulation curve based on an empirical data set arising from quadrat sampling is studied in a nonparametric binomial mixture model. It will be shown that estimating the species accumulation curve not only is independent of the unknown number of species but also includes estimating the number of species as a limiting case. For the purpose of interpolation, moment-based estimators, associated with asymptotic confidence intervals, are developed from several points of view. A likelihood-based procedure is developed for the purpose of extrapolation, associated with bootstrap confidence intervals. The proposed methods are illustrated by ecological data sets.

Key words: Conditional likelihood; Rarefaction; Species richness.

1. Introduction

The number of species or *species richness* in a species assemblage is a significant measure of biodiversity at the habitat level (Bunge and Fitzpatrick, 1993; Colwell and Coddington, 1994; Mao and Colwell, 2005). Because a complete census is feasible only under a few special situations, it is necessary to estimate species richness by sampling the target species assemblage. The *species accumulation curve*, that is, the plot of the expected number of detected species as a function of sampling effort, arises as a graphical representation of the sampling process (Sanders, 1968; Palmer, 1990). Species accumulation curves have also been used by ecologists to perform quantitative comparison among species assemblages (Sanders, 1968; Colwell and Coddington, 1994; Gotelli and Colwell, 2002), and used to estimate the expected number of new species to be detected given a level of additional sampling effort, which can lead to efficient planning and sampling protocols (Soberón and Llorente, 1993; Colwell and Coddington, 1994; Moreno and Halffter, 2000; Shen, Chao, and Lin, 2003). One approach to sampling a species assemblage is to take a random sample of individuals. The other approach is to record whether or not each species is detected in a set of randomly selected sampling units. The term "*quadrats*" will refer to such sampling units, including areas of various shapes and sizes, traps, lures, mist nets, etc.

Although species accumulation curves are routinely used by ecologists, statistical treatments are sparse (Mingoti and Meeden, 1992; Ugland, Gray, and Ellingsen, 2003). In this article, species accumulation curves based on quadrat sampling will be studied in a nonparametric binomial mixture model. Both moment-based and plug-in estimation procedures will be presented with asymptotic or bootstrap confidence intervals. Two ecological data sets are used as an illustration. A set of functions written in `R` is available from the first author.

2. Mixture Models

To estimate a species accumulation curve, we must obtain an empirical sample of K randomly selected quadrats from a species assemblage with information on the species detected in each quadrat. To describe the target species assemblage and the empirical sample, we define

- c as the unknown total number of species or species richness of the species assemblage;
- π_i as the probability that the ith species is detected in a single quadrat, $i = 1, 2, \ldots, c$;
- Y_i as the number of quadrats in which the ith species is detected in the empirical sample;
- n_j as the number of species with $Y_i = j$ in the empirical sample, $j = 0, 1, 2, \ldots, K$; and
- n_+ as the number of detected species (with $Y_i > 0$) in the empirical sample.

The detection probabilities π_i are assumed to be from a discrete latent distribution Θ. The c species are classified into disjoint homogeneous species groups in the sense that the species within the same group have the same detection probability. The species assemblage is homogeneous if all species have the same detection probability,

$$\pi_i = \pi_0, \quad i = 1, 2, \ldots, c, \tag{1}$$

that is, Θ is degenerate at the common detection probability π_0. The latent distribution Θ will put all mass over the interval (0, 1], which means that all species must be detectable.

Let $g_\pi(y)$ be a binomial density and let $g_\Theta(y)$ be a mixture of binomial densities,

$$g_\pi(y) = C_y^K \pi^y (1-\pi)^{K-y} \quad \text{and} \quad g_\Theta(y) = \int g_\pi(y)\, d\Theta(\pi),$$

where the constant C_b^a, with $a \in (-\infty, \infty)$ and $b \in [0, \infty)$ being integers, is defined by

$$C_b^a = \begin{cases} 0, & 0 \le a < b, \\ a(a-1)\cdots(a-b+1)/b!, & \text{otherwise.} \end{cases}$$

Because the Y_i arise as a random sample from the binomial mixture g_Θ, the counts $n_0, n_1, \ldots, n_K$ arise from a multinomial density, which, with the unknown n_0 replaced by $c - n_+$, is the full data likelihood for c and Θ, and written as (Norris and Pollock, 1996)

$$\begin{aligned} L(c, \Theta) &= \frac{c!}{\prod_{j=0}^{K} n_j!} \prod_{j=0}^{K} g_\Theta^{n_j}(j) \\ &= \frac{c!}{(c-n_+)! \prod_{j=1}^{K} n_j!} g_\Theta^{c-n_+}(0) \prod_{j=1}^{K} g_\Theta^{n_j}(j). \end{aligned} \tag{2}$$

Note that if $Y_i = 0$, then the ith species is not detected in the empirical sample. Those $Y_i > 0$, as a sample from the zero-truncated density $g_\Theta(x)\{1 - g_\Theta(0)\}^{-1}$, can be treated as a sample from a mixture $f_Q(x)$ of zero-truncated binomial densities $f_\pi(x)$, where, for $x = 1, 2, \ldots, K$,

$$f_\pi(x) = C_x^K \frac{\pi^x (1-\pi)^{K-x}}{1-(1-\pi)^K} \quad \text{and}$$

$$f_Q(x) = \int f_\pi(x)\, dQ(\pi) = \frac{g_\Theta(x)}{1 - g_\Theta(0)},$$

and the latent distribution Q is derived from Θ (Mao and Lindsay, 2002),

$$dQ(\pi) = \frac{\{1-(1-\pi)^K\}\, d\Theta(\pi)}{\int \{1-(1-\pi)^K\}\, d\Theta(\pi)}. \tag{3}$$

Note that in the homogeneous case, the derived latent distribution Q is also degenerate at the common detection probability π_0 in (1). The observed counts $n_1, n_2, \ldots, n_K$ given the number of detected species n_+ also arise from a multinomial density, which is the conditional likelihood for the derived latent distribution Q and written as

$$L(Q; n_+) = \frac{n_+!}{\prod_{j=1}^{K} n_j!} \prod_{j=1}^{K} f_Q^{n_j}(j). \tag{4}$$

3. Methods

3.1 *Estimating the Species Accumulation Function*

Let $\tau(h)$ be the expected number of species detected in a set of h randomly selected quadrats. Because the probability that a species is detected in a set of h randomly selected quadrats is $1 - (1-\pi)^h$ if it has a detection probability π, we can write

$$\tau(h) = c \int \{1 - (1-\pi)^h\}\, d\Theta(\pi). \tag{5}$$

There are nonparametric approaches (Norris and Pollock, 1996; Pledger, 2000) which provide likelihood-based estimators for species richness c and the latent distribution Θ of the detection probabilities. Any pair of estimators for (c, Θ) can yield an estimator for the species accumulation function $\tau(h)$ in (5) by plug-in.

However, because the problem of estimating c and Θ is difficult (Huggins, 2001; Dorazio and Royle, 2003; Link, 2003) and the quality of an estimator for c and that of the estimator for Θ will affect the quality of the estimator for $\tau(h)$, we will develop alternative estimators for the species accumulation function $\tau(h)$ in (5), with the help of two representations of $\tau(h)$.

LEMMA 1: *For $h = 1, 2, \ldots, K$, we have*

$$\tau(h) = \sum_{j=1}^{K} \left\{ 1 - \frac{C_j^{K-h}}{C_j^K} \right\} \cdot c g_\Theta(j), \tag{6}$$

and for $h \ge 1$, we have

$$\tau(h) = \tau(K) \cdot \{1 + \gamma(h; Q)\}, \tag{7}$$

where $\gamma(h;\ Q)$ is a functional of the derived latent distribution Q,

$$\gamma(h; Q) = \int \frac{(1-\pi)^K - (1-\pi)^h}{1-(1-\pi)^K}\, dQ(\pi). \tag{8}$$

It is clear that an "estimator" for $g_\Theta(j)$ is the "sample proportion" n_j/c. An estimator for the species accumulation function $\tau(h)$ in (6) is obtained by replacing $g_\Theta(j)$ in (6) by n_j/c,

$$\tilde{\tau}(h) = \sum_{j=1}^{K} \left\{ 1 - \frac{C_j^{K-h}}{C_j^K} \right\} n_j, \quad h = 1, 2, \ldots, K. \tag{9}$$

The estimator $\tilde{\tau}(h)$ is nonparametric in the sense that no restriction is put on the latent distribution Θ and species richness c, called a moment-based estimator because $\{g_\Theta(y)/g_{\pi_0}(y)\}_{y=0}^K$ is the moment sequence of a measure over $(0, \infty]$ (Lindsay, 1995, p. 52). By similar arguments to those in Smith and Grassle (1977), we can show Theorem 1, as follows.

THEOREM 1: *The estimator $\tilde{\tau}(h)$ is the minimum variance unbiased estimator for $\tau(h)$.*

Note that n_+, the number of detected species in the empirical sample, is an estimator for $\tau(K)$, the expected number of detected species in K quadrats. Estimation of the species accumulation function $\tau(h)$ in (7) can be reduced to estimating $\gamma(h;\ Q)$ in (8). Note that from (6) and (7), as $\tau(K) = c\{1 - g_\Theta(0)\}$, we can write

$$\gamma(h;Q) = -\sum_{j=1}^{K} \frac{C_j^{K-h}}{C_j^K} \frac{g_\Theta(j)}{1-g_\Theta(0)}$$
$$= -\sum_{j=1}^{K} \frac{C_j^{K-h}}{C_j^K} f_Q(j), \quad h = 1, 2, \ldots, K.$$

If one replaces $f_Q(j)$ by the sample proportion n_j/n_+, then one obtains an estimator

$$\tilde{\gamma}(h;Q) = -\sum_{j=1}^{K} \frac{C_j^{K-h}}{C_j^K} \frac{n_j}{n_+} = \frac{\tilde{\tau}(h)}{n_+} - 1. \tag{10}$$

We will perform further assessment on the moment-based estimator $\tilde{\tau}(h)$ in (9) and develop estimators for the derived latent distribution Q in (3) so that $\gamma(h;Q)$ in (8) can be estimated.

3.2 *Moment-Based Estimation*

The asymptotic normality of the moment-based estimator $\tilde{\tau}(h)$ in (9) or $\log\tilde{\tau}(h)$ can be shown with species richness c going to infinity and the number of quadrats h being fixed.

THEOREM 2: *For each $h \le K$, as c goes to infinity, we have*

$$c^{-1/2}\{\tilde{\tau}(h) - \tau(h)\} \to \mathcal{N}(0, \sigma^2(h)),$$
$$c^{1/2}\{\log\tilde{\tau}(h) - \log\tau(h)\} \to \mathcal{N}(0, c^2\sigma^2(h)/\tau^2(h))$$

in distribution, where $\mathcal{N}(\mu, \sigma^2)$ stands for a normal distribution with mean μ and variance σ^2, and the variance function $\sigma^2(h)$ is defined by

$$\sigma^2(h) = \sum_{j=1}^{K} \left\{1 - \frac{C_j^{K-h}}{C_j^K}\right\}^2 g_\Theta(j) - \frac{\tau^2(h)}{c^2}. \tag{11}$$

From (5), $\tau(h)/c$ depends only on Θ. For a fixed latent distribution Θ, the asymptotic variance of the moment-based estimator $\tilde{\tau}(h)$ in (9), which is also the variance of $\tilde{\tau}(h)$,

$$v(h) = c\sigma^2(h),$$

increases in species richness c while the asymptotic variance of $\log\tilde{\tau}(h)$ decreases in c.

One can construct asymptotic confidence intervals for $\tau(h)$ in (6) with an "estimator" for the asymptotic variance $v(h)$ given by

$$\tilde{v}(h) = \sum_{j=1}^{K} \left\{1 - \frac{C_j^{K-h}}{C_j^K}\right\}^2 n_j - \frac{\tilde{\tau}^2(h)}{c}, \quad h = 1, 2, \ldots, K. \tag{12}$$

One needs an estimator for species richness c in (12) in order to use $\tilde{v}(h)$. In particular, one can set $c = \infty$ in (12) to obtain conservative confidence intervals. For various estimators for species richness c, see Bunge and Fitzpatrick (1993) and Mao and Colwell (2005). For example, Mao and Colwell (2005) presented an estimator for c (also see Chao, 1989),

$$\tilde{c} = n_+ + \frac{(K-1)n_1^2}{2Kn_2}. \tag{13}$$

A well-known approach in ecology to estimation of species accumulation function $\tau(h)$ in (5) with $h \le K$ is the *randomization procedure* (Colwell and Coddington, 1994). This is a simulation-based approximation to the *enumeration procedure* in which one enumerates all subsamples of h quadrats of the empirical sample, counts the number of detected species in each subsample, and calculates their mean $\bar{\tau}(h)$ as an estimator for $\tau(h)$.

THEOREM 3: *The estimator $\bar{\tau}(h)$ is a U-statistic and identical to $\tilde{\tau}(h)$ in (9), where*

$$\bar{\tau}(h) = \sum_{j=1}^{K} \left\{1 - \frac{C_h^{K-j}}{C_h^K}\right\} n_j. \tag{14}$$

An analytic formula for the expectation of the number of detected species in a subsample of h quadrats, conditioning upon the empirical sample, has been found by Ugland et al. (2003) independently, which is identical to $\bar{\tau}(h)$ in (14) after simplification. Both Ugland et al. (2003) and the randomization procedure provided a conditional variance, which, unfortunately, should not be used to construct confidence intervals for $\tau(h)$.

Neither the randomization procedure nor the method in Ugland et al. (2003) can produce an estimator for species accumulation function $\tau(h)$ in (5) with $h > K$. The representation in (6) holds only for $h \le K$. In the proof of Lemma 1, we show that $\tau(h)$ for $h > K$ can be represented by an infinite series, under the restriction that $\Theta(\pi)$ puts all its mass in (0, 1/2). While the first K terms in the infinite series that involve $\{g_\Theta(j)\}_{j=1}^K$ can be estimated, no simple estimators are available for the other terms that are not nonparametrically identifiable.

Ecologists often assume that species accumulation function $\tau(h)$ in (5) is determined by a few unknown parameters. Several well-known parametric functions can be written as

$$\tau_1(h;\alpha_1,\beta_1) = \alpha_1(1 - e^{-\beta_1 h}), \quad \tau_2(h;\alpha_2,\beta_2) = \alpha_2 h/(\beta_2 + h),$$
$$\tau_3(h;\alpha_3,\beta_3) = \alpha_3 + \beta_3 \log h, \quad \tau_4(h;\alpha_3,\beta_3) = \alpha_4 h^{\beta_4}.$$

The negative exponential model $\tau_1(h;\ \alpha_1,\ \beta_1)$ (Holdridge et al., 1975) is a special case of $\tau(h)$ in (5), with the latent distribution Θ being degenerate at the common detection probability π_0 in (1). The hyperbola model $\tau_2(h;\ \alpha_2,\ \beta_2)$ (de Caprariis, Lindemann, and Collins, 1976) stands for the Michaelis–Menten equation of enzyme-catalyzed reaction kinetics (Raaijmakers, 1987). The third model $\tau_3(h;\ \alpha_3,\ \beta_3)$ is called the log-linear model (Gleason, 1922) and the fourth model is called the log–log model because $\log\ \tau_4(h;\ \alpha_4,\ \beta_4) = \log\alpha_4 + \beta_4 \log h$ (Arrhenius, 1923).

For each $\tau_i(h;\ \alpha_i,\ \beta_i)$, the parameters α_i and β_i can be estimated, for example, by minimizing $\sum_{h=1}^K\{\tau_i(h;\alpha_i,\beta_i) - \bar{\tau}(h)\}^2$ (Colwell and Coddington, 1994). Let $\hat{\alpha}_i$ and $\hat{\beta}_i$ be estimators for α_i and β_i, respectively. Then, one can estimate $\tau(h)$ by $\tau_i(h;\hat{\alpha}_i,\hat{\beta}_i)$ for any $h > K$. Because $\tau_i(h;\ \alpha_i,\ \beta_i)$ is only an approximation to $\tau(h)$ in (5), estimating $\tau(h)$ via parametric functions will have a bias. The dependence structure of the $\tilde{\tau}(h)$ is also an obstacle to a further investigation on statistical properties of the estimators of α_i and β_i.

THEOREM 4: *For $h \neq k$, $1 \leq h, k < K$, the correlation of $\tilde{\tau}(h)$ and $\tilde{\tau}(k)$ is given by*

$$\frac{1}{\sigma(h)\sigma(k)}\left\{\sum_{j=1}^{K}\left(1-\frac{C_j^{K-h}}{C_j^K}\right)\left(1-\frac{C_j^{K-k}}{C_j^K}\right)g_\Theta(j) - \frac{\tau(h)\tau(k)}{c^2}\right\}.$$

3.3 *Likelihood-Based Plug-In Estimation*

The latent distribution Q in (3) is not nonparametrically identifiable. To consider an identifiable set of latent distributions, we define index(Q) as the number of support points of Q, with a support point at zero or one being counted as 1/2 (Lindsay, 1995, p. 49),

$$\text{index}(Q) = \sum_{\pi\in(0,1)} I(dQ(\pi) > 0) + 1/2\{I(dQ(1) > 0) + I(dQ(0) > 0)\}. \quad (15)$$

The set of identifiable latent distributions consists of those Q that satisfy

$$\text{index}(Q) \leq (K-1)/2. \quad (16)$$

Because the case $\pi = 0$ is excluded, if there is no species with a perfect detection probability $\pi = 1$, then index(Q) is simply the number of support points of Q. The relationship between the number of quadrats K in the empirical sample and the number of support points of Q is complicated. A species assemblage can be as diverse as possible in the sense that all species have distinct detection probabilities. The properties of quadrats such as the quadrat size will also affect the detection probabilities. In practice, when K is sufficiently large, the number of quadrats K in the empirical sample might not affect the number of support points of Q and the identifiability constraint in (16) might be satisfied.

If the species assemblage is homogeneous, then the maximum likelihood estimator $\hat{\pi}_0$ for the common detection probability π_0 in (1) solves

$$\frac{\hat{\pi}_0}{1-(1-\hat{\pi}_0)^K} = \frac{\sum_{j=1}^{K} jn_j}{K\sum_{j=1}^{K} n_j}.$$

Let $\hat{Q}_0$ be the degenerate distribution at $\hat{\pi}_0$. The plug-in estimator $\gamma(h;\hat{Q}_0)$ for $\gamma(h;\ Q)$ in (8) and the plug-in estimator $\hat{\tau}_0(h)$ for $\tau(h)$ in (7) are given by

$$\gamma(h;\hat{Q}_0) = \frac{(1-\hat{\pi}_0)^K - (1-\hat{\pi}_0)^h}{1-(1-\hat{\pi}_0)^K} \quad \text{and}$$
$$\hat{\tau}_0(h) = n_+ + n_+\gamma(h;\hat{Q}_0). \quad (17)$$

There exists a discrete distribution $\hat{Q}$ over the closed interval [0, 1], called the *nonparametric maximum likelihood estimator* (*NPMLE*) (Lindsay, 1983a,b), which maximizes the conditional likelihood $L(Q;\ n_+)$ in (4), treating the number of support points, the support points, and the mixing weights as unknown parameters. If $\hat{Q}$ satisfies the identifiability constraint in (16) and the gradient function $D(\pi;\hat{Q})$ is not identically zero, then it will be unique, where

$$D(\pi;Q) = \sum_{x=1}^{K}\frac{n_x f_\pi(x)}{n_+ f_Q(x)} - 1.$$

The plug-in estimator $\gamma(h;\hat{Q})$ for $\gamma(h;\ Q)$ in (8) and the plug-in estimator $\hat{\tau}(h)$ for $\tau(h)$ in (7) are given by

$$\gamma(h;\hat{Q}) = \int \frac{(1-\pi)^K - (1-\pi)^h}{1-(1-\pi)^K}\, d\hat{Q}(\pi) \quad \text{and}$$
$$\hat{\tau}(h) = n_+ + n_+\gamma(h;\hat{Q}). \quad (18)$$

If $n_K = 0$, then the NPMLE $\hat{Q}$ will not put any mass on $\pi = 1$. Although the latent distribution Q in (3) is only allowed to put mass over (0, 1], theoretically the NPMLE $\hat{Q}$ can take zero as a support point. Let $\hat{\pi}$ be the smallest support point for $\hat{Q}$ with a corresponding mixing weight $\hat{m}$. Let $\tilde{Q}$ be a distribution obtained from $\hat{Q}$ by eliminating $\hat{\pi}$ and adjusting mixing weights accordingly. Because

$$\lim_{\pi\to 0}\frac{(1-\pi)^K - (1-\pi)^h}{1-(1-\pi)^K} = \frac{-K+h}{K} = \frac{h}{K} - 1,$$

if $\hat{\pi} = 0$, then we have

$$\gamma(h;\hat{Q}) = (1-\hat{m})\gamma(h;\tilde{Q}) - \hat{m} + K^{-1}\hat{m}\cdot h.$$

The right-hand side of this equality becomes an approximation of $\gamma(h;\hat{Q})$ when $\hat{\pi} \approx 0$. If h is sufficiently large, then $\gamma(h;\tilde{Q}) \approx \gamma(\infty;\tilde{Q})$ and

$$\gamma(h;\hat{Q}) \approx (1-\hat{m})\gamma(\infty;\tilde{Q}) - \hat{m} + K^{-1}\hat{m}\cdot h.$$

This means that $\gamma(h;\hat{Q})$ eventually will look like a linear function in h and approach $\gamma(\infty;\hat{Q})$.

Instead of maximizing the conditional likelihood $L(Q;\ n_+)$ in (4), one might minimize

$$AIC = -2\log L(Q;n_+) + 2\{2\cdot\text{index}(Q) - 1\}.$$

The minimum Akaike Information Criterion (AIC) estimator, denoted by $\check{Q}$, balances the goodness-of-fit and the number of parameters used to fit the data. The minimum AIC estimator can be found by selecting an estimator among the estimators with different numbers of support points. Pledger (2000) started the selection process for the latent distribution Θ from the homogeneous case and added support points sequentially, with the likelihood ratio test used to decide whether an additional support point is necessary. This is a forward selection procedure. We will consider a backward selection procedure. We start from the NPMLE $\hat{Q}$, merge the pair of the closest support points, run the EM algorithm with the merged distribution as the initial distribution to obtain a new estimator for Q, repeat the merging step and the EM step to produce a sequence of estimators, and find the one with the smallest AIC as the minimum AIC estimator $\tilde{Q}$. Note that this backward selection procedure is computationally efficient because for each fixed number of support points, the initial distribution supplied to the EM algorithm often yields a large likelihood and the EM algorithm converges rapidly. The plug-in estimator

$\gamma(h;\check{Q})$ for $\gamma(h;\,Q)$ in (8) and the plug-in estimator $\check{\tau}(h)$ for $\tau(h)$ in (7) are given by

$$\gamma(h;\check{Q}) = \int \frac{(1-\pi)^K - (1-\pi)^h}{1-(1-\pi)^K}\, d\check{Q}(\pi) \quad \text{and}$$

$$\check{\tau}(h) = n_+ + n_+\gamma(h;\check{Q}). \tag{19}$$

It is much less likely for $\check{Q}$ to put some mass on zero even if the NPMLE $\hat{Q}$ does.

Finally, construction of confidence intervals for the species accumulation function $\tau(h)$ in (5) can be achieved by generating bootstrap resamples $(n_0^*,\, n_1^*, \ldots, n_K^*)$ from the multinomial density $L(c,\,\Theta)$ in (2), with species richness c and the mixture density g_Θ being estimated. For example, if $d\check{Q}(0) = 0$, then c can be estimated by $\check{\tau}(\infty)$ and $g_\Theta(y)$ is estimated by $g_{\check{\Theta}}(y)$, where an estimator for Θ is given by

$$d\check{\Theta}(\pi) = \frac{\{1-(1-\pi)^K\}^{-1}\, d\check{Q}(\pi)}{\displaystyle\int \{1-(1-\pi)^K\}^{-1}\, d\check{Q}(\pi)}.$$

The observed counts $(n_1^*,\, n_2^*, \ldots, n_K^*)$ are used to calculate, for example, the minimum AIC estimator $\check{Q}^*$ for the derived latent distribution Q in (3) and the moment-based estimator $\tilde{\tau}^*(h)$ for the species accumulation function $\tau(h)$ in (6).

A resample of the observed counts $(n_1^*,\, n_2^*, \ldots, n_K^*)$ can also be obtained by a two-step procedure, in which one first generates n_+^* from a binomial density $\mathrm{Bin}(c,\,\tau(K)/c)$ with index c and probability $\tau(K)/c$, and then generates $(n_1^*,\, n_2^*, \ldots, n_K^*)$ from the multinomial density $L(Q;\,n_+)$ in (4) with n_+ replaced by n_+^* in each resample, provided that one has estimates for c, $\tau(K)$ and f_Q, respectively. However, as c goes to infinity, the density $\mathrm{Bin}(c, \tau(K)/c)$ goes to a Poisson density $\mathrm{Poi}(\tau(K))$ with mean $\tau(K)$. One can generate n_+^* from $\mathrm{Poi}(n_+)$, which means that one only needs to estimate $f_Q(x)$, for example, by $f_{\hat{Q}}(x)$. For any estimator $\hat{c} < \infty$ for species richness c, because the variance of $\mathrm{Bin}(\hat{c}, n_+/\hat{c})$ is $n_+(1 - n_+/\hat{c})$, smaller than n_+, the variance of $\mathrm{Poi}(n_+)$, a confidence interval obtained by sampling n_+^* from $\mathrm{Poi}(n_+)$ tends to be wider than the corresponding confidence interval with the same confidence level obtained by sampling n_+^* from $\mathrm{Bin}(\hat{c}, n_+/\hat{c})$.

4. Numeric Studies

4.1 *Simulation*

A simulation study is performed to assess estimators for $\gamma(h;\,Q)$ in (8). Each simulation scenario consists of three parameters: the number of detected species n_+, the derived latent distribution Q, and the number of quadrats K in the empirical sample. We summarize our findings here while the detailed results are available in a technical report.

For $h \le K$, the estimator $\gamma(h;\hat{Q})$ in (18), the estimator $\gamma(h;\,\check{Q})$ in (19), and the estimator $\tilde{\gamma}(h;Q)$ in (10) have little difference and all behave like a normal random variable. This suggests that $\tilde{\gamma}(h;Q)$ in (10) should be used as it is the simplest one in calculation. As h becomes larger and larger, the density of $\gamma(h;\,\check{Q})$ becomes less and less symmetric. The median of the density of $\gamma(h;\,\check{Q})$ is close to the true value $\gamma(h;\,Q)$ for a not-too-large h, say, $h \le 3K$. For the limiting case $h = \infty$, the density of $\gamma(\infty;\check{Q})$ is highly skewed with a long right tail. The 97.5% quantile of the density of $\gamma(\infty;\check{Q})$ can be huge, for example, larger than 10^6. For each $h \ge 1$, the estimator $\gamma(h;\,\check{Q})$ can be positively or negatively median biased. The median bias tends to be positive and increase when h increases.

4.2 *Examples*

The first data set is taken from a rain forest ant study at the La Selva Biological Station in Costa Rica (Longino, Coddington, and Colwell, 2002), in which $n_+ = 197$ ant species were detected at $K = 41$ sites. The second data set is taken from the North American Breeding Bird Survey, where $n_+ = 67$ bird species were detected along a survey route with $K = 50$ equidistant locations in 1997 (Dorazio and Royle, 2003). The observed counts n_x are given in Table 1.

We calculate the degenerate distribution $\hat{Q}_0$, assuming that the species assemblage is homogeneous, the NPMLE $\hat{Q}$, and the minimum AIC estimate $\check{Q}$ for each data set. In order to calculate bootstrap confidence intervals, we generate 400 resamples, with n_+^* from $\mathrm{Poi}(n_+)$ and $(n_1^*,\, n_2^*, \ldots, n_K^*)$ from the multinomial density $L(Q;\,n_+)$ in (4), with the latent distribution Q replaced by $\check{Q}$, and in each resample, n_+ replaced by n_+^*.

For the ant data, the sequence of maximized likelihoods are $-543.427, -543.789, -544.669, -550.642, -566.740, -641.052$, and -1261.409, omitting the logarithm of a multinomial coefficient. The estimate $\hat{Q}_0$ is degenerate at the estimated common detection probability $\hat{\pi}_0 = 0.175$, with $\log L(\hat{Q}_0; n_+) = -1261.409$. The NPMLE $\hat{Q}$, with $\log L(\hat{Q}; n_+) = -543.427$, has seven support points 0, 0.045, 0.107, 0.280, 0.473, 0.602, 0.860, with mixing weights 0.103, 0.409, 0.165, 0.171, 0.077, 0.060, 0.015. The minimum AIC estimate $\check{Q}$, with $\log L(\check{Q}; n_+) = -544.669$, has five support points 0.024,

Table 1

The count n_x is the number of species detected exactly in x quadrats. The observed counts n_x at $x \le 37$ are shown and the observed counts n_x at $x > 37$ are zero. The total numbers of detected species are $n_+ = 197$ (the ant data) and $n_+ = 67$ (the bird data).

	x	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Ant	n_x	50	29	24	13	6	9	3	4	1	7	6	2	6	5	1	2
Bird	n_x	14	10	6	3	3	6	2	3	2	1	1	1	0	3	2	0
	x	17	18	19	20	21	22	23	24	25	26	27	28	29	34	35	37
Ant	n_x	2	2	3	5	1	1	3	1	1	4	1	1	1	1	1	1
Bird	n_x	3	1	1	1	0	0	1	2	0	0	1	0	0	0	0	0

0.089, 0.289, 0.546, 0.858 with mixing weights 0.380, 0.297, 0.185, 0.122, 0.016.

For the bird data, the sequence of maximized likelihoods are −188.191, −190.183, −205.875, and −349.359, omitting the logarithm of a multinomial coefficient. The estimate $\hat{Q}_0$ is degenerate at the estimated common detection probability $\hat{\pi}_0 = 0.144$, with $\log L(\hat{Q}_0; n_+) = -349.359$. The NPMLE $\hat{Q}$, with $\log L(\hat{Q}; n_+) = -188.191$, has four support points 0.026, 0.123, 0.310, 0.462, with mixing weights 0.426, 0.322, 0.182, 0.071. The minimum AIC estimate $\check{Q}$, with $\log L(\check{Q}; n_+) = -190.183$, has three support points 0.027, 0.132, 0.364, with mixing weights 0.440, 0.327, 0.233.

For each $h \le K$, we compare the three estimates $\tilde{\tau}(h)$ in (9), $\hat{\tau}(h)$ in (18), and $\check{\tau}(h)$ in (19). They are close to each other in both data sets because we have, for the ant data,

$$\max_{1\le h\le K} |\tilde{\tau}(h) - \hat{\tau}(h)| = 0.058, \quad \max_{1\le h\le K} |\tilde{\tau}(h) - \check{\tau}(h)| = 0.127,$$

$$\max_{1\le h\le K} |\hat{\tau}(h) - \check{\tau}(h)| = 0.170,$$

and for the bird data,

$$\max_{1\le h\le K} |\tilde{\tau}(h) - \hat{\tau}(h)| = 0.069, \quad \max_{1\le h\le K} |\tilde{\tau}(h) - \check{\tau}(h)| = 0.096,$$

$$\max_{1\le h\le K} |\hat{\tau}(h) - \check{\tau}(h)| = 0.142.$$

The moment-based estimates $\tilde{\tau}(h)$ in (9), the 95% asymptotic confidence intervals, and the 95% bootstrap confidence intervals for $\tau(h) \le K$ are presented in Figure 1. The 95% bootstrap confidence interval for the species accumulation function $\tau(h)$ in (6) consists of the 2.5% quantile and the 97.5% quantile of the resample estimates $\tilde{\tau}^*(h)$. Note that the 95% bootstrap confidence interval for $\tau(h)$ and the 95% asymptotic confidence interval for $\tau(h)$ with $c = \infty$ in (12) provide similar lower and upper confidence limits.

Figure 2 presents the estimate $\check{\tau}(h)$ in (19) with a 95% bootstrap confidence interval. We also calculate $\hat{\tau}(h)$ (18) and $\hat{\tau}_0(h)$ (17). Three curves $(h, \hat{\tau}_0(h))$, $(h, \hat{\tau}(h))$, and $(h, \check{\tau}(h))$ cross at $(K,\ n_+)$ in each data set. For the ant data, the curve $(h, \hat{\tau}(h))$ finally tends to be a straight line because the NPMLE $\hat{Q}$ has a zero support point and the curves $(h, \hat{\tau}(h))$ and $(h, \check{\tau}(h))$ agree with each other for h smaller or not much larger than K. For the bird data, the curves $(h, \hat{\tau}(h))$ and $(h, \check{\tau}(h))$ are almost identical. For each data set, the curve $(h, \hat{\tau}_0(h))$ grows up and tends to its asymptote much faster than the curves $(h, \hat{\tau}(h))$ and $(h, \check{\tau}(h))$.

Finally, consider the limiting case $\tau(\infty)$, identical to species richness c. For the ant data, $\hat{\tau}(\infty) = \infty$ and $\check{\tau}(\infty) = 240.9$, with a 95% confidence interval (215.9, ∞). For the bird data, $\hat{\tau}(\infty) = 77.6$ and $\check{\tau}(\infty) = 77.1$, with a 95% confidence interval (71.5, ∞), where the lower confidence limit is the 5% quantile of the resample estimates $\check{\tau}^*(\infty)$. An alternative 95% confidence interval is given by the 2.5% quantile and the 97.5% quantile of the $\check{\tau}^*(\infty)$. For the ant data, we have (214.4, 2×10^7) and for the bird data, we have (70.8, 118.7). The upper confidence limit 2×10^7 is noninformative for the ant data. Note that $\tilde{c} = 239.1$ in (13) for the ant data and $\tilde{c} = 76.6$

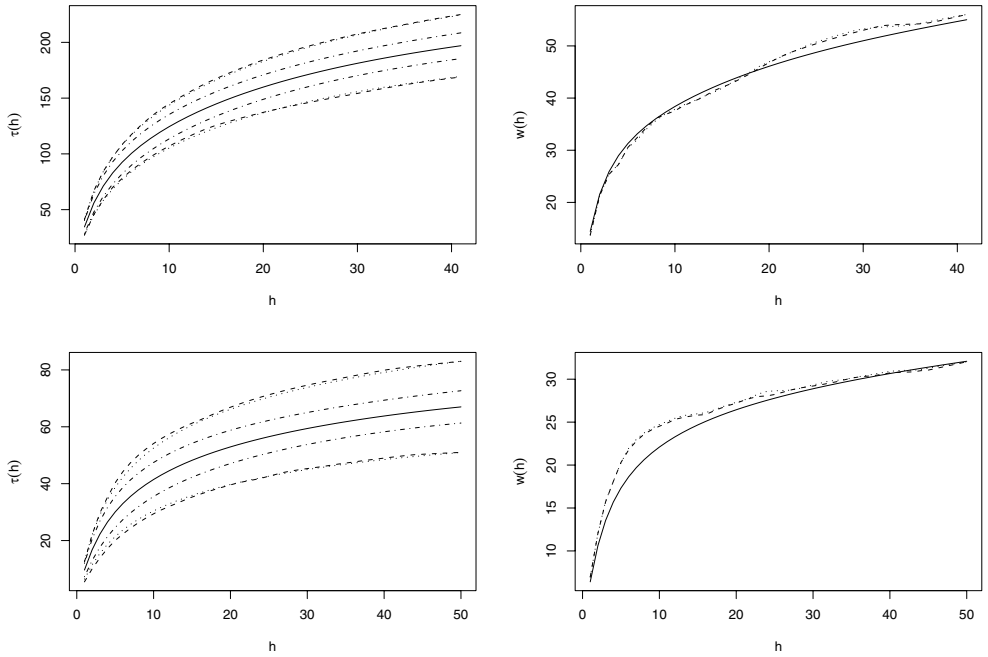

Figure 1. The moment-based estimate $\tilde{\tau}(h)$ in (9) with $h \le K$, the 95% confidence intervals, and the width $w(h)$ of a confidence interval. The left panels present, at each h, the estimate $\tilde{\tau}(h)$ (—), a 95% bootstrap confidence interval $(\tilde{\tau}^*_l(h), \tilde{\tau}^*_u(h))$ (- -), a 95% asymptotic confidence interval $\tilde{\tau}(h) \pm 1.96\tilde{v}^{1/2}(h)$ $(\cdots)$ with $c = \infty$ in $\tilde{v}(h)$, and a 95% asymptotic confidence interval $\tilde{\tau}(h) \pm 1.96\tilde{v}^{1/2}(h)$ $(\cdot - \cdot)$ with $c = \tilde{c}$ in $\tilde{v}(h)$, where $\tilde{\tau}^*_u(h)/\tilde{\tau}^*_l(h)$ is the 97.5%/2.5% quantile from the resample estimates $\tilde{\tau}^*(h)$. The right panels present $\tilde{w}(h) = 3.92\tilde{v}^{1/2}(h)$ (—) with $c = \infty$ in $\tilde{v}(h)$, $\tilde{w}^*(h) = \tilde{\tau}^*_u(h) - \tilde{\tau}^*_l(h)$ $(\cdots)$, and $\tilde{w}^*(h) = \tilde{\tau}^*_u(h) - \check{\tau}_l(h)$ (- -), where $\check{\tau}^*_u(h)/\check{\tau}^*_l(h)$ is the 97.5%/2.5% quantile from the resample estimates $\check{\tau}^*(h)$. The top panels are for the ant data and the bottom panels are for the bird data.

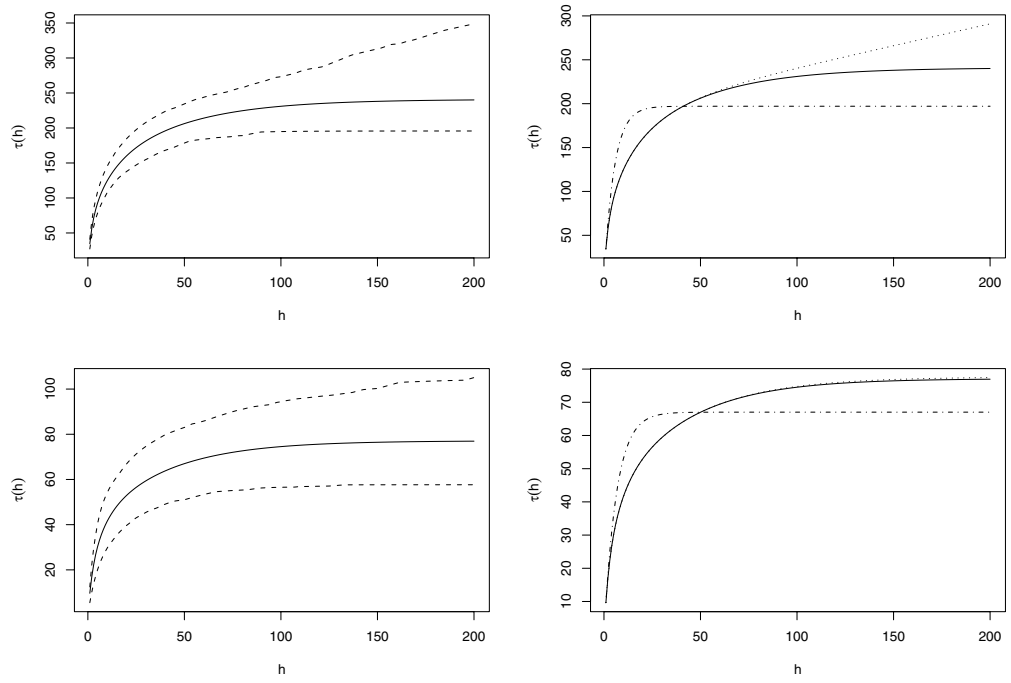

Figure 2. The likelihood-based plug-in estimates $\check{\tau}(h)$ in (19), $\hat{\tau}(h)$ in (18), and $\hat{\tau}_0(h)$ in (17). The left panels present, at each h, the estimate $\check{\tau}(h)$ (—) with a 95% bootstrap confidence interval (- -). The right panels present $\check{\tau}(h)$ (—), $\hat{\tau}(h)$ ($\cdots$), and $\hat{\tau}_0(h)$ ($\cdot-\cdot$). The top panels are for the ant data and the bottom panels are for the bird data.

for the bird data so that $\check{\tau}(\infty)$ and $\tilde{c}$ are close to each other for both data sets. Dorazio and Royle (2003) analyzed the bird data, where they obtained an estimate 76.1 for species richness c based on a mixture of two components and a 95% profile likelihood-based confidence interval (68.7, 95.3).

5. Conclusion

The nonparametric binomial mixture model can be applied to real species assemblages, in which the detection probabilities vary arbitrarily across species. Only one part of the species accumulation function $\tau(h)$ in (5), namely, the part with $h \leq K$, is nonparametrically identifiable, and for this part, the moment-based estimator $\tilde{\tau}(h)$ in (9) arises as an optimal choice. To estimate the entire species accumulation function $\tau(h)$ in (5), we propose to estimate the derived latent distribution Q in (3) under the identifiability constraint in (16), which might be satisfied as the number of quadrats in the empirical sample is usually large. The minimum AIC estimator $\check{Q}$ is recommended. Besides the likelihood-based estimators, there are other estimators for Q, for example, the minimum distance estimators.

A nice property associated with both the moment-based and plug-in methods is that it is neither necessary to estimate species richness c nor necessary to estimate the latent distribution Θ of detection probabilities. Asymptotic confidence intervals are readily available for $h \leq K$. Bootstrap confidence intervals can also be calculated with more computational effort. An estimator for species richness c is required only when one does not wish to use the conservative asymptotic or bootstrap confidence intervals. Because estimation of the species accumulation function $\tau(h)$ in (5) becomes more and more difficult as h increases, we recommend that our estimation method can be used for relatively small h, say $h \leq 3K$, which is significantly useful for practical purposes in ecology and conservation biology. The problem of estimating $\tau(h)$ in (5) at a large h, including the limiting case $c = \tau(\infty)$, deserves further investigation.

Acknowledgements

The authors thank the editor, the associate editor, and referees for their careful reviews and insistence on improvements of the presentation.

References

Arrhenius, O. (1923). Statistical investigations in the constitution of plant associations. *Ecology* **4,** 68–73.

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.

Colwell, R. K. and Coddington, J. A. (1994). Estimating terrestrial biodiversity through extrapolation. *Philosophical Transactions of the Royal Society: Biological Sciences* **345,** 101–118.

de Caprariis, P., Lindemann, R. H., and Collins, C. M. (1976). A method for determining optimum sample size in species diversity studies. *Mathematical Geology* **8,** 575–581.

Dorazio, R. M. and Royle, J. A. (2003). Mixture models for estimating the size of a closed population when

capture rates vary among individuals. *Biometrics* **59,** 351–364.

Gleason, H. A. (1922). On the relation between species and area. *Ecology* **3,** 158–162.

Gotelli, N. and Colwell, R. K. (2001). Quantifying biodiversity: Procedures and pitfalls in the measurement and comparison of species richness. *Ecology Letters* **4,** 379–391.

Holdridge, L. R., Grenke, W. G., Haheway, W. H., Liang, T., and Tosi, J. A. (1975). *Forest Environments in Tropical Life Zones*. Oxford: Pergamon Press.

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

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

Lindsay, B. G. (1983b). The geometry of mixture likelihoods: Part II: The exponential family. *Annals of Statistics* **11,** 783–792.

Lindsay, B. G. (1995). *Mixture Models: Theory, Geometry, and Applications*. Hayward, California: Institute of Mathematical Statistics.

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

Longino, J. T., Coddington, J., and Colwell, R. K. (2002). The ant fauna of a tropical rainforest: Estimating species richness three different ways. *Ecology* **83,** 689–702.

Mao, C. X. and Colwell, R. K. (2005). Estimation of species richness: Mixture models, the role of rare species, and inferential challenges. *Ecology*, in press.

Mao, C. X. and Lindsay, B. G. (2002). Diagnostics for the homogeneity of capture probabilities in a Bernoulli census. *Indian Journal of Statistics* **64,** 626–639.

Mingoti, S. A. and Meeden, G. (1992). Estimating the total number of distinct species using presence and absence data. *Biometrics* **48,** 863–875.

Moreno, C. E. and Halffter, G. (2000). Assessing the completeness of bat biodiversity inventories using species accumulation curves. *Journal of Applied Ecology* **37,** 149–158.

Norris, J. L. I. and Pollock, K. H. (1996). Nonparametric MLE under two closed capture–recapture models with heterogeneity. *Biometrics* **52,** 639–649.

Palmer, M. W. (1990). The estimation of species richness by extrapolation. *Ecology* **71,** 1195–1198.

Pledger, S. (2000). Unified maximum likelihood estimates for closed capture–recapture models using mixtures. *Biometrics* **56,** 434–442.

Raaijmakers, J. G. W. (1987). Statistical analysis of the Michaelis–Menten equations. *Biometrics* **43,** 793–803.

Sanders, H. L. (1968). Marine benthic diversity: A comparative study. *American Naturalist* **102,** 243–282.

Shen, T.-J., Chao, A., and Lin, C.-F. (2003). Predicting the number of new species in further taxonomic sampling. *Ecology* **84,** 798–804.

Smith, W. and Grassle, J. F. (1977). Sampling properties of a family of diversity measures. *Biometrics* **33,** 283–292.

Sobeŕon, J. M. and Llorente, J. B. (1993). The use of species accumulation functions for the prediction of species richness. *Conservation Biology* **7,** 480–488.

Ugland, K. I., Gray, J. S., and Ellingsen, K. E. (2003). The species-accumulation curve and estimation of species richness. *Journal of Animal Ecology* **72,** 888–897.

Received March 2003. *Revised September* 2004. *Accepted September* 2004.

Appendix

Proofs

To prove Lemma 1, we first use a latent distribution Ω on $\lambda = \pi/(1-\pi)$ obtained from the latent distribution Θ on π by changing of variable, and write $\tau(h)/c$ as

$$\frac{\tau(h)}{c} = \int \frac{(1+\lambda)^K - (1+\lambda)^{K-h}}{(1+\lambda)^K}\, d\Omega(\lambda) = \int \sum_{j=1}^{\infty} \frac{\left(C_j^K - C_j^{K-h}\right)\lambda^j}{(1+\lambda)^K}\, d\Omega(\lambda),$$

where, for $h > K$, we assume that Ω puts all its mass in $(0, 1)$. As $C_j^K = 0$ for $j > K$, we write the last equality in terms of the latent distribution Θ on π as

$$\frac{\tau(h)}{c} = \sum_{j=1}^{K} \left(1 - \frac{C_j^{K-h}}{C_j^K}\right) \int C_j^K \pi^j (1-\pi)^{K-j}\, d\Theta(\pi) - \sum_{j=K+1}^{\infty} C_j^{K-h} \int \frac{\pi^j}{(1-\pi)^{j-K}}\, d\Theta(\pi),$$

where, for $h > K$, Θ puts all its mass in $(0, 1/2)$. For $h \le K$, we can write $\tau(h)/c$ as

$$\frac{\tau(h)}{c} = \sum_{j=1}^{K} \left(1 - \frac{C_j^{K-h}}{C_j^K}\right) g_\Theta(j).$$

To obtain the representation of $\gamma(h;\, Q)$, we write $\tau(h)/\tau(K)$ as

$$\frac{\tau(h)}{\tau(K)} = \frac{\int \{1-(1-\pi)^h\}\, d\Theta(\pi)}{\int \{1-(1-\pi)^K\}\, d\Theta(\pi)} = \int \frac{1-(1-\pi)^h}{1-(1-\pi)^K}\, dQ(\pi).$$

To prove Theorems 2 and 4, we write $\tilde{\tau}(h) = \sum_{i=1}^{c} W_i(h)$, where

$$W_i(h) = \sum_{j=1}^{K} \left(1 - \frac{C_j^{K-h}}{C_j^K}\right) I(Y_i = j).$$

By simple calculation, we obtain, for each $i = 1, 2, \ldots, c$,

$$EW_i(h) = \sum_{j=1}^{K} \left(1 - \frac{C_j^{K-h}}{C_j^K}\right) g_\Theta(j),$$

$$EW_i(h)W_i(k) = \sum_{j=1}^{K} \left(1 - \frac{C_j^{K-h}}{C_j^K}\right)\left(1 - \frac{C_j^{K-k}}{C_j^K}\right) g_\Theta(j).$$

Theorem 2 holds by the central limit theorem and the delta method. Theorem 4 holds as

$$E\tilde{\tau}(h)\tilde{\tau}(k) = \sum_{i=1}^{c}\sum_{m=1}^{c} EW_i(h)W_m(k) = \sum_{i=1}^{c} EW_i(h)W_i(k).$$

To prove Theorem 3, we define $Z(i,j) = I$ (the ith species detected in the jth quadrat). Let $\mathcal{T}$ be the set of all combinations $\{l_1,\ l_2, \ldots, l_h\}$ drawn from $\{1,\ 2, \ldots, K\}$. It is clear that

$$C_h^K \bar{\tau}(h) = \sum_{j=1}^{K} \sum_{\left\{i:\sum_{k=1}^{K} Z(i,k)=j\right\}} \sum_{\{l_1,l_2\ldots,l_h\}\in\mathcal{T}} I\left(\sum_{m=1}^{h} Z(i,l_m) > 0\right)$$

$$= \sum_{j=1}^{K} \sum_{\left\{i:\sum_{k=1}^{K} Z(i,k)=j\right\}} \left(C_h^K - C_h^{K-j}\right)$$

$$= \sum_{j=1}^{K} \left(C_h^K - C_h^{K-j}\right) n_j.$$

The U-statistic is simply another representation of $\bar{\tau}(h)$. Finally, $\tilde{\tau}(h) = \bar{\tau}(h)$ because

$$\frac{C_h^{K-j}}{C_h^K} = \frac{C_j^{K-h}}{C_j^K} = \begin{cases} \dfrac{(K-j)!(K-h)!}{K!(K-j-h)!}, & h+j \le K, \\ 0, & \text{otherwise.} \end{cases}$$