Estimating species richness by a Poisson-compound gamma model

BY JI-PING WANG

*Department of Statistics, Northwestern University, 2006 Sheridan Road, Evanston,
Illinois 60208, U.S.A.*

jzwang@northwestern.edu

SUMMARY

We propose a Poisson-compound gamma approach for species richness estimation. Based on the denseness and nesting properties of the gamma mixture, we fix the shape parameter of each gamma component at a unified value, and estimate the mixture using nonparametric maximum likelihood. A least-squares crossvalidation procedure is proposed for the choice of the common shape parameter. The performance of the resulting estimator of N is assessed using numerical studies and genomic data.

Some key words: Crossvalidation; Nesting property of gamma mixtures; Nonparametric maximum likelihood estimation; Poisson-compound gamma model; Species richness estimation.

1. INTRODUCTION

A typical dataset in the species problem comprises a series of counts $\{x_i : i = 1, \dots, D\}$, recording the numbers of individuals from a total of D observed distinct species. These counts are often summarized into the sufficient statistic $n = (n_1, \dots, n_t)$, where $n_j = \sum_i I\{x_i = j\}$, $t = \max\{x_i\}$. The parameter for estimation is the total number of the distinct species N in the underlying population. Research on this old but challenging problem is motivated by several novel genomic applications, and the following are two examples.

Serial analysis of gene expression data and expressed sequence tag data are two similar types of genomic data widely used for surveying transcribed genes. Both are collections of tags or signatures of genes, representing random samples from the entire gene expression population or mRNA pool (Thygesen & Zwinderman, 2006; Morris et al., 2003; Wang et al., 2005). We let n_j stand for the number of genes that have j tags in a sample; but owing to the limited sampling or sequencing effort, some expressed genes may not be observed. By knowing the total number of expressed genes, biologists can gain insights into the diversity of active genes in the given tissue.

The diversity of microbial species present can provide an interesting profile of a microbial community (Acinas et al., 2004; Hong et al., 2006). For example, using the 16S ribosomal RNA gene as the signature, researchers can identify distinct microbial species present in a soil sample. However, sampling is often shallow in practice, i.e. only a small fraction of the species is obtained, and the total diversity needs to be estimated.

In both examples, the gene expression level or the species abundance varies substantially. The first challenge in the estimation of N is how to model the heterogeneity. Ignoring it may result in a large downward bias (Chao & Lee, 1992; Bunge & Fitzpatrick, 1993; Norris & Pollock, 1998; Böhning & Schön, 2005). In a likelihood-based approach, one first specifies a probability

distribution function $f(x_i; \lambda_i)$ ($\lambda_i > 0$), and then imposes a nondegenerate distribution $Q(\lambda; \theta)$, $\theta \in \Theta \subset \mathbb{R}^d$, typically continuous, for the species abundance index λ_i . Thus, the marginal distribution of X becomes a mixture distribution, i.e. $f(x; Q) = \int f(x; \lambda) dQ(\lambda)$. However, an inappropriate choice of Q can lead to systematic bias of the resulting estimator.

A more robust approach is to estimate Q by nonparametric maximum likelihood (Norris & Pollock, 1998). The nonparametric maximum likelihood estimator, $\hat{Q}$, is in the form of a discrete distribution (Lindsay, 1995), and in the case of a Poisson mixture it is unique with a finite number of support points for any dataset. Wang & Lindsay (2005) proposed a penalized nonparametric maximum likelihood estimator, which is more stable than that of Norris & Pollock (1998). Nevertheless, severe underestimation can occur due to the interplay of factors including inadequate sampling effort, heterogeneity and skewness of the abundance curve. This weakness is also observed in other non-likelihood-based nonparametric approaches. In the motivating genomic applications, shallow sampling is typical. Thus, underestimation is the second major challenge.

In the motivating applications, a continuous Q can be appealing for interpretability of species abundance. More importantly, it could better capture the information of species abundance near zero, and hence improve the estimation of N when underestimation is a major concern. The objective of this study is to model Q using a smooth curve while estimating it using nonparametric maximum likelihood based on a Poisson-compound gamma model.

2. THE MODEL AND MAXIMUM LIKELIHOOD ESTIMATION

2.1. A general model

Consider that we have a mixture distribution $f(x; H)$ defined based on a density kernel $f(x; \lambda)$, $\lambda \in \mathbb{R}$ and a mixing distribution $H(\lambda; \mu, \nu)$, $\mu, \nu \in \mathbb{R}$. Here, we assume that H is continuous and has location and dispersion parameters μ and ν , respectively. Furthermore, as $\nu \rightarrow 0$, H becomes degenerate, and thus, $f(x; H)$ converges to a unicomponent distribution. If we fix ν and allow a mixing distribution G for μ , then the overall mixing distribution for λ becomes continuous, having a form $Q(\lambda) = \int H(\lambda; \mu, \nu) dG(\mu)$. The mixture density of X becomes $f_\nu(x; G) = \int_\mu f_\nu^*(x; \mu) dG(\mu)$, where $f_\nu^*(x; \mu) = \int_\lambda f(x; \lambda) dH(\lambda; \mu, \nu)$. If H is chosen to be conjugate for f , then f_ν^* can often be obtained in a simple closed form. For each given ν , the global maximum likelihood estimator of G , say $\hat{G}_\nu$, can be found by nonparametric maximum likelihood estimation (Lindsay, 1983, 1995). The estimator $\hat{G}_\nu$ is a discrete distribution, while the resulting mixing distribution Q of λ , denoted by $\hat{Q}_\nu(\lambda)$, is continuous. In this paper, we term $\hat{Q}_\nu(\lambda)$ the smooth nonparametric maximum likelihood estimator (Magder & Zeger, 1996).

Alternatively, one can ignore the H structure and directly estimate $Q(\lambda)$ in the defined mixture model $f(x; Q) = \int f(x; \lambda) dQ$ by nonparametric maximum likelihood. The resulting discrete $\hat{Q}(\lambda)$ maximizes the likelihood over all Q ; we call it the discrete nonparametric maximum likelihood estimator.

2.2. Properties of gamma mixtures

Let $h(\lambda; \alpha, \beta) = \beta^\alpha \{\Gamma(\alpha)\}^{-1} \lambda^{\alpha-1} e^{-\beta\lambda}$, $\alpha, \beta > 0$ be the gamma density. We consider two reparameterizations: $(\alpha, \beta) \mapsto (\alpha, \mu)$ and $(\alpha, \beta) \mapsto (\mu, \beta)$, where $\mu = \alpha/\beta$. The following denseness property of the gamma mixture is cited from Fang & Chlamtac (1999, p. 1064):

LEMMA 1. *Let $Q(\lambda)$ be the cumulative distribution function of a nonnegative random variable. Then it is possible to choose a sequence of distribution functions $Q_m(\lambda)$, each of which corresponds to a gamma mixture, such that $Q_m(\lambda) \rightarrow Q(\lambda)$ as $m \rightarrow \infty$, at any point where $Q(\lambda)$ is*

continuous. For example, we may choose

$$Q_m(\lambda) = \sum_{k=1}^{\infty} [Q(k/m) - Q((k-1)/m)] Q_m^k(\lambda),$$

where $Q_m^k(\lambda)$ is a gamma distribution with mean k/m and variance k/m^2 .

Intuitively, when m becomes large, each individual gamma component Q_m^k approximates a delta function at k/m with a weight equal to the increase in the cumulative distribution function of Q from $(k-1)/m$ to k/m . Therefore, a gamma mixture model can be used to approximate any distribution defined in $\mathbb{R}_+$. In practice, Q is unknown and the sample size is limited, and one can fit a gamma mixture by the method of maximum likelihood.

The second property is the nesting property of gamma mixtures.

LEMMA 2. Let $h(\lambda; \theta, \mu_0)$ be the reparameterized gamma density where $\theta = \alpha$ or β and μ_0 is the mean parameter. Then for any given $\delta > 0$, there exists a unique probability measure G for μ such that for all $\lambda > 0$, $\int h(\lambda; \theta + \delta, \mu) dG(\mu) = h(\lambda; \theta, \mu_0)$.

Proof. See the Appendix. □

The existence part of Lemma 2 for $\theta = \alpha$ is a result from Lindsay (1995, p. 54), given in the new parameterization. The existence result for $\theta = \beta$ case and the uniqueness for both α and β cases are new. The uniqueness is a consequence of the identifiability presented in Theorem 2 below. Subsequently, we shall use $h\{\theta, G(\mu)\}$ to denote a gamma mixture defined as $\int h(\lambda; \theta, \mu) dG(\mu)$. The following nesting property is an immediate consequence of Lemma 2.

THEOREM 1. Let $\delta > 0, \theta > 0$ and $h(\theta, \mu)$ be the reparameterized gamma density, where $\theta = \alpha$ or β . Let $\mathcal{G} = \{G(\mu) : G \text{ is a probability measure on } \mathbb{R}_+\}$. Define the mixture sets $\mathcal{H}_\theta = [h\{\theta, G(\mu)\} : G \in \mathcal{G}]$ and $\mathcal{H}_{\theta+\delta} = [h\{\theta + \delta, G(\mu)\} : G \in \mathcal{G}]$. Then $\mathcal{H}_\theta \subset \mathcal{H}_{\theta+\delta}$.

Lemma 2 implies that an arbitrary gamma distribution $h(\theta, \mu_0)$ can be uniquely expressed as a gamma mixture $h\{\theta + \delta, G(\mu)\}$ for any given $\delta > 0$. Hence, a gamma mixture model, $h\{\theta, G(\mu)\} \in \mathcal{H}_\theta$, can be rewritten as a model in $\mathcal{H}_{\theta+\delta}$, but with a different $G(\mu)$. This implies that $G(\mu)$ is not identifiable in the entire distribution space $\mathcal{G}$ if θ is unknown. Nevertheless, if θ is given, G is identifiable in $\mathcal{G}$.

THEOREM 2. Let $h\{\theta, G(\mu)\}$ be a gamma mixture model as defined earlier with a given $\theta > 0$ and an unknown $G(\mu)$ defined in $\mathbb{R}_+$. Then $G(\mu)$ is identifiable in $\mathcal{G}$.

Proof. See the Appendix. □

The identifiability of the gamma mixture model under a given unified θ implies the uniqueness of $G(\mu)$, as stated in Lemma 2. The gamma mixture with different θ s is not identifiable in the entire distribution space $\mathcal{G}$. If we assume that the mixture is finite, i.e. the mixing distribution G is discrete with a finite number of support points, then the mixture becomes identifiable (Teicher, 1963). In practice, as we fit a finite mixture of gammas, the identifiability among the finite mixture set is adequate to justify the estimation.

2.3. Poisson-compound gamma model

The denseness property suggests that there exists a sequence of finite gamma mixtures that can approximate any Q in $\mathbb{R}_+$. As Q is unknown, we need to estimate Q based on the data. If $\tilde{Q}$ is such an estimator in a finite gamma mixture form, then by the nesting property $\tilde{Q}$ can be rewritten as a new gamma mixture, say $\tilde{Q}_\theta$, with a unified θ parameter. However, the likelihood for such

$\tilde{Q}_\theta$ is always dominated by that of the smooth nonparametric maximum likelihood estimator $\hat{Q}_\theta$ obtained with the given θ . In other words, the smooth nonparametric maximum likelihood estimator with an appropriate choice of θ can also approximate Q arbitrarily well regardless of the form of true Q . Hence, robustness of this method is anticipated.

We consider a Poisson-compound gamma model for the species problem. If the compound gamma has a fixed θ and a mixing distribution G for μ , then $f_\theta(x; G) = \int \int e^{-\lambda} \lambda^x (x!)^{-1} h(\lambda; \theta, \mu) d\lambda dG(\mu)$. We can define $f_\theta^*(x; \mu) = \int e^{-\lambda} \lambda^x (x!)^{-1} h(\lambda; \theta, \mu) d\lambda$. For example, under the (α, μ) parameterization, $f_\alpha^*(x; \mu) = \Gamma(x + \alpha)(\alpha/\mu)^\alpha / \{x! \Gamma(\alpha)(1 + \alpha/\mu)^{x+\alpha}\}$. Then $f_\alpha(x; G) = \int f_\alpha^*(x; \mu) dG(\mu)$. The likelihood function is

$$L_\theta(N, G; n) \propto \binom{N}{D} f_\theta(0; G)^{N-D} \{1 - f_\theta(0; G)\}^D \times \prod_{j>0} \left\{ \frac{f_\theta(j; G)}{1 - f_\theta(0; G)} \right\}^{n_j}, \quad (1)$$

where the expressions before and after the multiplication sign are referred to as the marginal and conditional likelihood, respectively, and are denoted by L_θ^m and L_θ^c henceforth.

Based on the likelihood in (1), there are two possible likelihood approaches, conditional or unconditional (Sanathanan, 1972, 1977). In the conditional approach, the mixing distribution G , say, $\hat{G}_c$, is first estimated from L_θ^c alone, and then the estimator of N by $\hat{N}_c = \lfloor D / \{1 - f_\theta(0, \hat{G}_c)\} \rfloor$, where $\lfloor x \rfloor$ stands for the integer no greater than x . The unconditional approach seeks the pair $(\hat{N}_u, \hat{G}_u)$ to maximize the full likelihood, which must also satisfy $\hat{N}_u = \lfloor D / \{1 - f_\theta(0, \hat{G}_u)\} \rfloor$ (Wang & Lindsay, 2005).

In this paper, we consider the conditional approach for its computational simplicity. Below we shall use $\hat{G}$ to denote the conditional nonparametric maximum likelihood estimator of G . A frequently occurring issue in the conditional discrete nonparametric maximum likelihood approach is the boundary problem, where a component $\lambda \approx 0^+$ is fitted. The same problem arises in the unconditional discrete (Wang & Lindsay, 2005) and the finite mixture approaches. Unfortunately, the continuity brought into Q in the Poisson-compound gamma model does not help overcome this issue. As the gamma distribution has $\mu = \alpha/\beta$, $\sigma^2 = \mu/\beta$, with the given α , fitting a $\mu \approx 0^+$ implies $\beta \rightarrow \infty$, and hence, $\sigma^2 \rightarrow 0$. Therefore, the fitted gamma component corresponds to a spike of λ at 0^+ . Such a tiny component is problematic because it blows up the resulting point estimate of N . In Wang & Lindsay (2008), it was further shown that this boundary problem is structural in the sense that the fitted probability at $X = 1$ for the zero-truncated sample must be no less than the empirical probability $n_1 / \sum n_i$. It can be shown that this is also true for smooth nonparametric maximum likelihood estimation.

An effective solution to this problem is to penalize the likelihood. Define the odds parameter ψ as $\psi(G) = f_\theta(0; G) / \{1 - f_\theta(0; G)\}$. Then $\hat{N}_c$ can be written as $\hat{N}_c = \lfloor D \{1 + \psi(\hat{G})\} \rfloor$. Thus, penalizing large ψ values can stabilize the conditional estimator. Such a penalty function of ψ in the likelihood has a natural interpretation as a Bayesian prior for ψ , in which, however, the hyperparameter needs to be subjectively prespecified (Wang & Lindsay, 2005). To overcome this weakness, here we use a partial prior approach similar to that proposed in Wang & Lindsay (2008) for the binomial mixture. We impose an exponential prior for the odds parameter in a form of $p(\psi; \eta) = \eta e^{-\eta\psi}$ ($\psi, \eta > 0$). As we are estimating the entire mixing distribution in $\mathbb{R}_+^\infty$ while the prior is specified for a functional of G in $\mathbb{R}_+$, such a prior is termed a partial prior. We use an empirical Bayes method to find $(\hat{\eta}, \hat{G})$ that maximizes the joint likelihood

$$\ell_\theta^*(G, \eta; n) = \log(L_\theta^c) + \log(\eta) - \eta\psi, \quad (2)$$

where L_θ^c is the conditional likelihood defined in (1). It was shown in Wang & Lindsay (2008) that, $\psi(\hat{G}) = 1/\hat{\eta}$, and $\hat{G}$ maximizes ℓ_θ^* if and only if it maximizes the alternative likelihood

$$\ell_\theta^{**}(G; n) = \log(L_\theta^c) - \log\{\psi(G)\}. \quad (3)$$

Clearly, as $\psi \rightarrow \infty$, we have $\ell_\theta^{**} \rightarrow -\infty$. Thus, the boundary problem is prevented. Compared to the quadratic penalty in Wang & Lindsay (2005), the exponential partial prior approach is simpler and less ad hoc. The proposed estimator of N in this paper is based on the likelihood ℓ_θ^* or its equivalent ℓ_θ^{**} . The penalized nonparametric maximum likelihood estimator of G from ℓ_θ^{**} can be reliably found using an EM algorithm (Wang, 2007).

3. CHOOSING θ BY LEAST-SQUARES CROSSVALIDATION

The nesting property of the gamma mixture directly results in the following monotonicity property of the likelihood function defined in (1), (2) and (3), whose proof is omitted.

PROPOSITION 1. *Let $\hat{\ell}_\theta$ be the maximized loglikelihood value from the smooth nonparametric maximum likelihood estimation based on conditional or unconditional likelihood from (1), or based on the likelihood function with the partial prior as defined in (2) or (3). Then $\hat{\ell}_\theta$ is a monotone nondecreasing function of θ .*

Proposition 1 implies that the global maximum likelihood is achieved at the discrete nonparametric maximum likelihood estimator, namely $\theta = \infty$. When $\theta = \alpha$, as $\sigma^2 = \mu^2/\alpha$, a larger α for given μ implies a more spiky gamma density. As $\alpha \rightarrow \infty$, each gamma component will converge to a degenerate distribution. The same conclusion holds when $\theta = \beta$. By design, the tuning parameter θ brings better smoothness to $Q(\lambda)$, and should reduce the underestimation. The likelihood value under smooth nonparametric maximum likelihood estimation can be regarded as the profile likelihood for θ . Unfortunately, the corresponding likelihood ratio test does not have a known distribution (Lindsay, 1995). Hence we propose a crossvalidation procedure for model selection.

In our estimation procedure, the parameter θ plays the same role as a bandwidth parameter in kernel density estimation; it tunes the smoothness of the resulting gamma mixture density. However, our problem differs from the latter in two senses: the species abundance value from the gamma mixture is not directly observed, and the density estimator is defined as a gamma mixture rather than a convolution of Gaussian densities. We investigate a crossvalidation procedure similar to that in density estimation (Silverman, 1992, p. 48) to determine the value of the tuning parameter. Our goal is to find a θ that minimizes the summed square error

$$\text{SSE} = \sum_{j=1}^{\infty} \{f(j; \hat{G}) - f(j; G)\}^2 = \sum_{j=1}^{\infty} \{f(j; \hat{G})\}^2 - 2 \sum_{j=1}^{\infty} \{f(j; \hat{G})f(j; G)\} + \sum_{j=1}^{\infty} \{f(j; G)\}^2, \quad (4)$$

where $f(j; \hat{G})$ is the estimated zero-truncated Poisson mixture probability at j , and $f(j; G)$ is its true value. The last term in (4), $\sum_{j=1}^{\infty} \{f(j; G)\}^2$, does not involve $\hat{G}$, so a good choice of $\hat{G}$ is one that minimizes $\sum_{j=1}^{\infty} \{f(j; \hat{G})\}^2 - 2 \sum_{j=1}^{\infty} \{f(j; \hat{G})f(j; G)\}$. Because $f(j; G)$ is unknown, we replace it with an empirical version, n_j/D . In addition we define $\hat{f}_{-j}$ as the leave-one-out crossvalidated probability density after excluding one species with count j from the sample.

Fig. 1. Plots of (a) maximized loglikelihood relative to that achieved at discrete nonparametric maximum likelihood, (b) M_0 , and (c) $\hat{N}$ versus $\log(\alpha)$ at $\alpha = 1, \dots, 10, 20, 40, 60, 100$ for the example data. The dashed line in (c) corresponds to the discrete nonparametric maximum likelihood estimator, $\hat{N} = 886$.

Further replacing $f(j; \hat{G})$ with $\hat{f}_{-j}(j)$ in the second term leads us to minimize the function,

$$M_0(\theta) = \sum_{j=1}^{\infty} \{f(j; \hat{G})\}^2 - 2 \sum_{j \in \{j: n_j > 0\}} \frac{n_j}{D} \hat{f}_{-j}(j), \quad (5)$$

with respect to θ . As $\hat{f}_{-j}(j) \approx f(j; \hat{G})$,

$$M_0(\theta) \approx \sum_{j=1}^{\infty} \{f(j; \hat{G}) - n_j/D\}^2 - \sum_j (n_j/D)^2.$$

Given the data, minimizing $M_0(\theta)$ tends to minimize $\sum_{j=1}^{\infty} \{f(j; \hat{G}) - n_j/D\}^2$, to provide good fit to the data. Incorporating the leave-one-out crossvalidated probability $\hat{f}_{-j}(j)$ in $M_0(\theta)$ in (5) introduces a variance component into the objective function $M_0(\theta)$, owing to a small perturbation of the data. In the Poisson-compound gamma model, an over-small θ results in a large summed square error due to lack of fit or bias, whereas an over-large θ , though always beating any smaller θ in terms of the likelihood by the nesting property, may yield a relatively larger variance in $f(j; \hat{G})$. Minimizing $M_0(\theta)$ over θ balances the two competing components in this model.

The summed square error defined earlier is analogous to the integrated square error used in the kernel density estimation. Although we are unable to establish similar asymptotics because of the complexity of nonparametric maximum likelihood estimation of mixtures, this procedure works surprisingly well in our application. We illustrate this using simulated species data from a Poisson-compound gamma model, where $N = 1000$, and the species abundance is a two-component gamma mixture $0.5\text{Ga}(\alpha = 2, \mu = 1) + 0.5\text{Ga}(\alpha = 2, \mu = 2)$. The data are: $n = (n_1, \dots, n_{14}) = (269, 183, 88, 56, 34, 8, 9, 6, 0, 1, 1, 1, 1, 1)$. We consider the $\theta = \alpha$ setting by treating α as a tuning parameter. In Fig. 1, we plot the maximized log conditional likelihood under the exponential partial prior on a grid of $\alpha \in \mathcal{A} = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 40, 60, 100\}$ together with $M_0(\alpha)$ and the resulting $\hat{N}$.

The loglikelihood plot in Fig. 1(a) confirms the monotonicity. The crossvalidation procedure identified $\alpha = 2$ on the grid that minimizes $M_0(\alpha)$, yielding $\hat{N} = 981$. The estimator $\hat{N}$ decreases in α , until reaching 886 at $\alpha = \infty$. This overall decreasing pattern of $\hat{N}$ is typical, though strict monotonicity seems not to hold based on our experience. The discrepancy between $\hat{N}$ at $\alpha = 2$ and at $\alpha = \infty$ suggests that an appropriate choice of α model could lead to a less biased estimator than the discrete nonparametric maximum likelihood estimator.

Table 1. *Species abundance distribution and sampling depth in 12 simulation settings*

Setting	Distribution type (Q)	d
	gamma	
1	Ga(4,3·125)	0.90
2	Ga(4,1)	0.59
3	Ga(1,0.25)	0.20
	gamma mixture	
4	0.5Ga(2,1) + 0.5Ga(2,2)	0.65
5	0.5Ga(2,1) + 0.5Ga(4,1)	0.57
	Lognormal	
6	LN (0.75,0.75)	0.82
7	LN (−.5,2)	0.50
8	LN (−1,1)	0.36
	Lognormal mixture	
9	0.5LN(−0.5,1) + 0.5LN(0.5,1)	0.61
	Discrete	
10	0.8 $\Delta_{1.2}$ + 0.2 $\Delta_{6.7}$	0.76
11	0.89 $\Delta_{0.5}$ + 0.11 $\Delta_{6.7}$	0.46
12	0.8 $\Delta_{0.2}$ + 0.2 $\Delta_{1.3}$	0.29

Ga, gamma (α, μ); LN, lognormal(μ, σ); Δ_x , a degenerate distribution at x ; d is defined as the sampling depth, which is the expected fraction of species to be observed in the sample, i.e. $d = 1 - f(0; Q)$.

4. SIMULATION STUDY

We now investigate whether this new approach could preserve the desired robustness as a nonparametric maximum likelihood estimator, and whether it could significantly reduce the underestimation. Table 1 lists 12 settings, where the species abundance distribution $Q(\lambda)$ varied from gamma, gamma mixture, lognormal and lognormal mixture, to discrete distributions, with the expected fraction of the sampled species, defined as the sampling depth, d , ranging from 0.20 to 0.90. For each setting, we simulated 200 samples from a population containing $N = 1000$ species. Crossvalidation was applied to each sample on a grid $\alpha \in \mathcal{A} = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 40, 60, 100, \infty\}$. One could carry out the crossvalidation for $\alpha < 1$, but the gamma density goes to ∞ at 0^+ for $\alpha < 1$. Consequently, the resulting estimator can be extremely variable or biased. Hence, in this simulation, we took only $\alpha \geq 1$.

Overall, the crossvalidation procedure works satisfactorily in selecting appropriate models for different settings. For the single gamma models including settings 1–3, and the gamma mixture setting 4 where the components had the same α parameter, the medians of the selected α values, denoted as $\hat{\alpha}$, were 4, 4, 1 and 2, respectively, which were consistent with the true values. For setting 5, where the gamma mixture components had differing α values, the crossvalidation tended to select some intermediate $\hat{\alpha}$ values with a median of 3. For the three discrete Q models, settings 10–12, the medians of $\hat{\alpha}$ were 100, 80 and 1, respectively. The difference probably arises because in the first two, the sampling was deeper, and the two discrete components were farther apart than setting 12. As a result, the crossvalidation procedure can better identify the latent structure in the former two. For the four lognormal models including settings 6–9, the medians were 2, 1, 1 and 1, respectively, suggesting that the crossvalidation procedure tended to select gamma components with a larger coefficient of variation to approximate the lognormal distributions. In Fig. 2, we present the histograms of $\hat{\alpha}$ from settings 2, 4, 5, 8, 11 and 12.

Fig. 2. Histogram of selected α values in simulation settings (a) 2, (b) 4, (c) 5, (d) 8, (e) 11, and (f) 12. Each bar in the histogram represents the frequency of selected α models for $\alpha \in \mathcal{A} = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 40, 60, 100, \infty\}$. The bar at $\alpha = 101$ represents the frequency at $\hat{\alpha} = \infty$.

Table 2 summarizes the Monte Carlo estimates from the proposed estimator, denoted by $\hat{N}_{\text{PCG}}$, in comparison with other popular estimators, including the penalized discrete nonparametric maximum likelihood estimator $\hat{N}_{\text{WL}}$ of Wang & Lindsay (2005); two sample coverage-based estimators $\hat{N}_{c_1}$ and $\hat{N}_{c_2}$, corresponding to $\hat{N}_2$ and $\hat{N}_3$, respectively from Chao & Lee (1992), where $\hat{N}_{c_2}$ provides further bias correction based on $\hat{N}_{c_1}$; a coverage-duplication estimator $\hat{N}_{c_3}$ based on a Poisson-gamma model from Chao & Bunge (2002); and a jackknife estimator $\hat{N}_J$ from Burnham & Overton (1978, 1979). The jackknife order was selected up to five based on a procedure proposed in the original papers. A confidence interval for $\hat{N}_{\text{PCG}}$ was obtained using a standard bootstrap procedure as follows. Let $g(j; \hat{G}_{\hat{\alpha}})(j = 1, 2, \dots)$ be the fitted zero-truncated Poisson mixture probabilities from a Monte Carlo sample with D observed distinct species. We generated 200 bootstrap samples, each containing D observations sampled from $g(j; \hat{G}_{\hat{\alpha}})$. The crossvalidation procedure was applied to each bootstrap sample and $\hat{N}_{\text{PCG}}$ was calculated. The interval spanned by the 0.025 and 0.975 quantiles of the N estimates was regarded as the bootstrap confidence interval. We also computed three additional likelihood-based estimators without a partial prior, including $\hat{N}_{\text{FM}}$, $\hat{N}_{\text{PCG}_2}$ and $\hat{N}_{\text{PCG}_3}$. Estimator $\hat{N}_{\text{FM}}$ was obtained by fitting a finite mixture of zero-truncated Poisson distributions to the data. Estimators $\hat{N}_{\text{PCG}_2}$ and $\hat{N}_{\text{PCG}_3}$ were the conditional and unconditional maximum likelihood estimators, respectively, based on the Poisson-compound gamma model without fixing α or β . In each case, the number of components was selected using the Bayesian information criterion; and the confidence interval coverage was based on the same bootstrap procedure as for $\hat{N}_{\text{PCG}}$.

For each setting, the summary statistics based on 200 Monte Carlo estimates presented in Table 2 include the median, standard deviation, mean absolute error and the coverage of a nominal 95% confidence interval. For $\hat{N}_{c_1}$, $\hat{N}_{c_2}$ and $\hat{N}_{c_3}$, we used a log-transformation procedure from Chao (1987) for the 95% confidence interval. For computational reasons the number of Monte Carlo replicates is too small for precise estimation of coverages, but despite this major differences between the procedures are clear.

Overall, $\hat{N}_{\text{PCG}}$ exhibited the desirable robustness similar to $\hat{N}_{\text{WL}}$, $\hat{N}_{c_1}$, $\hat{N}_J$ and $\hat{N}_{\text{FM}}$, having no extreme behaviour in any setting. In contrast, $\hat{N}_{\text{PCG}_2}$, $\hat{N}_{\text{PCG}_3}$ and $\hat{N}_{c_3}$, which make gamma

Table 2. Comparing nine different estimators with respect to median bias, mean absolute error and 95% confidence interval coverage in 12 simulation settings listed in Table 1

Setting	$\hat{N}$	$\hat{M}$	s	MAE	Cov.	Setting	$\hat{M}$	s	MAE	Cov.	Setting	$\hat{M}$	s	MAE	Cov.
1	$\hat{N}_{PCG}$	1003	25	19	96	2	1021	135	95	95	3	925	234	195	98
	$\hat{N}_{WL}$	1004	46	33	90		998	112	86	92		674	148	314	82
	$\hat{N}_{c_1}$	982	14	20	75		925	44	76	70		615	101	376	18
	$\hat{N}_{c_2}$	988	16	16	88		953	57	60	89		654	150	326	46
	$\hat{N}_{c_3}$	989	16	16	90		993	89	67	96		—	—	—	—
	$\hat{N}_J$	1075	17	74	100		1134	108	140	43		590	85	406	1
	$\hat{N}_{FM}$	974	16	27	38		868	87	131	14		541	177	448	22
	$\hat{N}_{PCG_2}$	1001	17	13	92		991	90	68	95		—	—	—	—
4	$\hat{N}_{PCG_3}$	999	17	13	87	5	981	85	66	95	6	—	—	—	—
	$\hat{N}_{PCG}$	1097	127	118	99		1012	140	104	96		1022	65	50	98
	$\hat{N}_{WL}$	955	98	84	99		965	110	88	99		1008	73	58	95
	$\hat{N}_{c_1}$	900	34	101	23		895	44	102	45		965	21	34	69
	$\hat{N}_{c_2}$	951	45	55	87		934	59	72	83		989	25	22	93
	$\hat{N}_{c_3}$	1014	71	55	96		999	105	78	95		1001	28	22	95
	$\hat{N}_J$	1036	80	60	84		1094	106	107	33		1064	65	74	33
	$\hat{N}_{FM}$	874	170	145	27		886	189	148	22		885	31	114	14
7	$\hat{N}_{PCG_2}$	1038	83	72	92	8	1002	107	79	96	9	1100	207	120	26
	$\hat{N}_{PCG_3}$	1028	81	66	95		991	101	76	94		1094	62	99	100
	$\hat{N}_{PCG}$	847	130	156	78		1033	191	147	100		1041	139	108	100
	$\hat{N}_{WL}$	755	91	229	57		794	115	199	97		920	96	99	100
	$\hat{N}_{c_1}$	674	29	325	0		769	72	226	23		856	35	142	7
	$\hat{N}_{c_2}$	739	43	257	2		940	140	124	87		932	50	71	76
	$\hat{N}_{c_3}$	846	98	153	73		—	—	—	—		1036	98	82	95
	$\hat{N}_J$	802	112	193	19		892	126	131	55		984	112	76	75
10	$\hat{N}_{FM}$	628	50	363	80	11	699	106	285	32	12	821	66	176	28
	$\hat{N}_{PCG_2}$	—	—	—	—		—	—	—	—		—	—	—	—
	$\hat{N}_{PCG_3}$	581	97	396	25		—	—	—	—		—	—	—	—
	$\hat{N}_{PCG}$	1013	36	31	83		1035	133	99	92		915	168	148	94
	$\hat{N}_{WL}$	1026	60	49	67		1048	142	116	66		742	129	247	89
	$\hat{N}_{c_1}$	1060	36	65	52		1003	71	55	96		625	64	373	2
	$\hat{N}_{c_2}$	1171	54	180	2		1487	172	490	2		676	93	314	21
	$\hat{N}_{c_3}$	1343	115	365	0		—	—	—	—		—	—	—	—
10	$\hat{N}_J$	1160	45	164	0	11	1183	144	215	46	12	797	104	205	32
	$\hat{N}_{FM}$	1002	31	25	86		1006	461	386	51		534	244	459	7
	$\hat{N}_{PCG_2}$	1018	39	34	80		—	—	—	—		—	—	—	—
	$\hat{N}_{PCG_3}$	1025	59	41	87		—	—	—	—		—	—	—	—

$\hat{M}$, s and MAE are the sample median, standard deviation and mean of absolute error calculated based on 200 Monte Carlo estimates for each setting. Cov. is the average of 95% confidence interval coverage evaluated from 200 samples. $\hat{N}_{c_1}$, $\hat{N}_{c_2}$ and $\hat{N}_{c_3}$ were estimated based on less abundant subset of species defined as $n_j \leq 10$. For $\hat{N}_{PCG}$, $\hat{N}_{WL}$, $\hat{N}_{FM}$, $\hat{N}_{PCG_2}$ and $\hat{N}_{PCG_3}$, the coverage of confidence interval was estimated based on a similar bootstrap procedure. ‘—’ stands for situations where extreme bias/variance occurred.

or gamma mixture assumptions for Q , are susceptible to model misspecification. For example, they all had large bias in multiple settings. In particular, for $\hat{N}_{PCG_2}$ and $\hat{N}_{PCG_3}$, the Bayesian information criterion tended to eliminate the boundary component if its weight was small, but not otherwise. Consequently, large bias was still occasionally observed. In most other cases, the positive bias appeared to be systematic, which occurred when using gamma models to approximate

Table 3. *Microbial species, Arabidopsis thaliana root expressed sequence tag and Traffic data*

j	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17 ⁺
Microbial (n_j)	381	65	23	18	4	5	3	0	1	0	4	0	3	2	0	1	3
Root(n_j)	2187	490	133	121	37	51	22	19	7	8	6	7	6	4	5	5	18
Traffic(n_j)	1317	239	42	14	4	4	1	0	0	0	0	0	0	0	0	0	0

lognormal and discrete distributions in settings 7, 8, 9, 11 and 12. A large bias was frequently observed when a fitted gamma component had $\hat{\alpha} < 1$. Such results were excluded from Table 2. Although $\hat{N}_{\text{PCG}}$ was built on the same model, its estimation strategy, including using the partial prior, crossvalidation and controlling $\alpha \geq 1$, appeared to be effective in alleviating such issues regardless of the true form of Q . For $\hat{N}_{\text{FM}}$, we observed a few extremely large estimates owing to the boundary problem in several settings. In the summary in Table 2, such extreme estimates, if greater than 2000, were replaced with 2000.

We further calculated the median relative bias, defined as bias/ N for each setting. The average of the absolute median relative bias across different settings for $\hat{N}_{\text{PCG}}$, $\hat{N}_{\text{WL}}$, $\hat{N}_{\text{c}_1}$, $\hat{N}_{\text{c}_2}$, $\hat{N}_{\text{c}_3}$, $\hat{N}_{\text{J}}$ and $\hat{N}_{\text{FM}}$ was 4.6%, 10%, 14.6%, 14.8%, 6.3%, 13.9% and 19.2%, respectively. The average bias of 6.3% for $\hat{N}_{\text{c}_3}$ was obtained after excluding four settings where it had large bias. In particular, when the sampling was shallow and underestimation was a severe issue, $\hat{N}_{\text{PCG}}$ significantly improved over other estimators regarding bias.

The proposed estimator $\hat{N}_{\text{PCG}}$ pays a price of relatively larger variance for the achieved robustness, improved bias and confidence interval coverage when the true Q is continuous. For example, when compared to the coverage-based estimator $\hat{N}_{\text{c}_2}$, $\hat{N}_{\text{PCG}}$ had relatively larger mean absolute error in seven settings, though the coverage of confidence interval of $\hat{N}_{\text{PCG}}$ was uniformly better in such settings.

5. DATA ANALYSIS

In the motivating genomic applications, each dataset often presents a shallow survey of the entire population of genes or microbial species. Hence, underestimation is probably the primary concern. In Table 3, we present three datasets including the microbial species data from [Acinas et al. \(2004\)](#), the *Arabidopsis thaliana* expressed sequence tag data from [Wang & Lindsay \(2005\)](#) and the traffic data from [Simar \(1976\)](#) and [Böhning & Schön \(2005\)](#).

In total, 513 species were observed in the microbial sample. The point estimate from $\hat{N}_{\text{PCG}}$ was 3005 at $\hat{\alpha} = 3$ with 95% confidence interval (1691, 4530). The point estimate and confidence bounds from other estimators including $\hat{N}_{\text{WL}}$, $\hat{N}_{\text{c}_1}$, $\hat{N}_{\text{c}_2}$, $\hat{N}_{\text{J}}$, $\hat{N}_{\text{PCG}_2}$ and $\hat{N}_{\text{PCG}_3}$ are also presented in Table 4 for comparison. The estimator $\hat{N}_{\text{c}_3}$ gave all negative estimates, and was therefore excluded. Estimates from other estimators varied substantially. The point estimates from $\hat{N}_{\text{c}_1}$ and the order 5 jackknife estimator $\hat{N}_{\text{J}}$ were 1674 and 1913, respectively, and both are likely to be underestimates of N based on our experience gained from the simulation. It is possible that there are actually more than 3000 species, as indicated by the confidence limits from $\hat{N}_{\text{PCG}}$.

For the expressed sequence tag data and traffic data, the estimates from $\hat{N}_{\text{PCG}}$ were 8980 (8383, 18 771) at $\hat{\alpha} = \infty$ and 6935 (5198, 12 409) at $\hat{\alpha} = 3$, respectively. In both the cases, the bootstrap estimates were highly skewed to the right. The large skewness should not be simply understood as the negative side of variability, but instead it may reflect an appealing feature in coverage. For example, the true N in the traffic data was known to be 9461, and $\hat{N}_{\text{PCG}}$ is the only estimator that provides a confidence interval that covers the true N . Similarly, for the root data,

Table 4. Results for data in Table 3

Estimators	$\hat{N}_{PCG}$	$\hat{N}_{WL}$	$\hat{N}_{c_1}$	$\hat{N}_{c_2}$	$\hat{N}_j$	$\hat{N}_{PCG_2}$	$\hat{N}_{PCG_3}$
Microbial	3005 (1691, 4530)	2053 (1519, 2585)	1674 (1388, 2052)	2510 (1878, 3434)	1913 (1634, 2191)	— (521, 2272)	1098 (855, 2712)
Root	8980 (8383, 18 771)	8919 (8211, 10 918)	9090 (8470, 9783)	13 948 (12 459, 15 674)	9891 (9192, 10 589)	9483 —	9478 (6530, 18 033)
Traffic	6935 (5198, 12 409)	5496 (4991, 6918)	5684 (5031, 6461)	6788 (5665, 8223)	6588 (5892, 7284)	— —	— —

The point estimate and 95% confidence interval (parentheses) are presented for each dataset. N_{WL} was computed by using ESTstatCF program from <http://bioinfo.stats.northwestern.edu/~jzwang>. ‘—’ indicates that the estimates were extremely large due to boundary problems or possible bias.

8980 could be a conservative estimate of the total number of expressed genes in the root tissue. As the *Arabidopsis thaliana* genome contains approximately 27 000 protein coding genes, the $\hat{N}_{PCG}$ confidence limits suggest that there can be up to 2/3 of the entire transcriptome expressed in the root tissue.

6. DISCUSSION

This estimation strategy can be potentially applied to many applications that involve mixture models. An immediate application is to the capture–recapture problem. Instead of using a finite mixture of binomials (Pledger, 2000), one could use a binomial-compound beta model (Morgan & Ridout, 2008). The crossvalidation procedure can be investigated further to select the number of components in finite mixture models. Detailed discussion can be found in the online Supplementary Material. The R package (R Development Core Team, 2010) species, which automates different estimators for species richness estimation, can be found at <http://bioinfo.stats.northwestern.edu/~jzwang>.

ACKNOWLEDGEMENT

The author would like to thank Professor Bruce Lindsay of the Department of Statistics at Pennsylvania State University, two referees, the associate editor and the editor. The author is also grateful to Professor Bruce Spencer, Professor Thomas Severini and Professor Wenxin Jiang of the Department of Statistics at Northwestern University. This work was supported by the National Institutes of Health, U.S.A., and the National Center for Supercomputing Applications.

SUPPLEMENTARY MATERIAL

Supplementary material is available at *Biometrika* online.

APPENDIX

Proof of Lemma 2. We show the existence of G for $\theta = \alpha$ or β . The uniqueness is guaranteed by the identifiability of gamma mixtures under a unified α or β parameter, stated in Theorem 2.

When $\theta = \alpha$, given any $\delta > 0$ and $\mu_0 > 0$, the following G satisfies $\int h(\lambda; \alpha + \delta, \mu) dG(\mu) = h(\lambda; \alpha, \mu_0)$, for all $\lambda > 0$:

$$dG(\mu) = \frac{\alpha + \delta}{\mu^2} \frac{\Gamma(\alpha + \delta)(\alpha/\mu_0)^\alpha}{\Gamma(\alpha)\Gamma(\delta)} \left(\frac{\alpha + \delta}{\mu}\right)^{-\alpha-\delta} \left(\frac{\alpha + \delta}{\mu} - \frac{\alpha}{\mu_0}\right)^{\delta-1} I\left(\frac{\alpha + \delta}{\mu} > \frac{\alpha}{\mu_0}\right) d\mu.$$

It is straightforward to verify that $G(\mu)$ is a well-defined probability distribution function on $\{0, (\alpha + \delta)\mu_0/\alpha\}$.

When $\theta = \beta$, if such a $G(\mu)$ exists, it has to satisfy that for all $\lambda > 0$,

$$\int \frac{(\beta + \delta)^{(\beta + \delta)\mu}}{\Gamma\{(\beta + \delta)\mu\}} \lambda^{(\beta + \delta)\mu - \beta\mu_0} dG(\mu) = \frac{\beta^{\beta\mu_0}}{\Gamma(\beta\mu_0)} e^{\delta\lambda}. \quad (\text{A1})$$

We can define $G(\mu)$ as a discrete probability measure as such that

$$P_G\{(\beta + \delta)\mu = \beta\mu_0 + i\} = \frac{\beta^{\beta\mu_0}}{\Gamma(\beta\mu_0)} \frac{\Gamma(\beta\mu_0 + i)}{(\beta + \delta)^{\beta\mu_0 + i}} \frac{\delta^i}{i!} \quad (i = 0, 1, \dots). \quad (\text{A2})$$

To verify whether (A2) is a probability measure, consider a $\text{Po}(x; \delta\lambda)$, where λ has a prior distribution of $\text{Ga}(\beta\mu_0, \beta)$. Then

$$\begin{aligned} \text{pr}(X = i) &= \int \frac{e^{-\delta\lambda} (\delta\lambda)^i}{i!} \frac{\beta^{\beta\mu_0}}{\Gamma(\beta\mu_0)} \lambda^{\beta\mu_0 - 1} e^{-\beta\lambda} d\lambda \\ &= \frac{\beta^{\beta\mu_0}}{\Gamma(\beta\mu_0)} \frac{\Gamma(\beta\mu_0 + i)}{(\beta + \delta)^{\beta\mu_0 + i}} \frac{\delta^i}{i!} \quad (i = 0, 1, \dots). \end{aligned}$$

This is exactly in the same form as P_G . Plugging G into the left-hand side of the equality (A1), we have

$$\frac{\beta^{\beta\mu_0}}{\Gamma(\beta\mu_0)} \sum_{i \geq 0} \frac{(\delta\lambda)^i}{i!} = \frac{\beta^{\beta\mu_0}}{\Gamma(\beta\mu_0)} e^{\delta\lambda}. \quad \square$$

Proof of Theorem 2. Suppose that under the reparameterization (α, μ) setting, the mixing distribution in μ is $G(\mu)$. As α is fixed, $G(\mu)$ corresponds to a unique mixing distribution in β , e.g. $G^*(\beta)$, in the original (α, β) parameterization and $G^*(\beta) = 1 - G\{(\alpha/\beta)^-\}$, where $G\{(\alpha/\beta)^-\} = P(\mu < \alpha/\beta)$. Conversely, we have $G(\mu) = 1 - G^*\{(\alpha/\beta)^-\}$. As $G(\mu)$ and $G^*(\beta)$ are uniquely determined by each other, it suffices to show that under the original parameterization (α, β) , $G^*(\beta)$ is identifiable. Suppose there is another G^{**} such that for all $\lambda > 0$,

$$\int \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda} dG^*(\beta) \equiv \int \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda} dG^{**}(\beta). \quad (\text{A3})$$

This equality holds at $\lambda = 1$, and hence

$$\int \frac{\beta^\alpha}{\Gamma(\alpha)} e^{-\beta} dG^*(\beta) \equiv \int \frac{\beta^\alpha}{\Gamma(\alpha)} e^{-\beta} dG^{**}(\beta). \quad (\text{A4})$$

The left- and right-hand sides of equation (A3) divided by those of (A4) in parallel gives

$$\int e^{-\beta(\lambda-1)} dP^*(\beta) \equiv \int e^{-\beta(\lambda-1)} dP^{**}(\beta), \quad (\text{A5})$$

where

$$dP^*(\beta) = \frac{\beta^\alpha e^{-\beta} dG^*(\beta)}{\int \beta^\alpha e^{-\beta} dG^*(\beta)}, \quad dP^{**}(\beta) = \frac{\beta^\alpha e^{-\beta} dG^{**}(\beta)}{\int \beta^\alpha e^{-\beta} dG^{**}(\beta)}.$$

If we let $t = 1 - \lambda$, then the left- and right-hand sides of (A5) are the moment generating functions for P^* and P^{**} , respectively, which are finite for t in a neighbourhood around zero, and hence $P^* = P^{**}$. As $dG^*(\beta) = \beta^{-\alpha} e^\beta dP^*(\beta) / \{\int \beta^{-\alpha} e^\beta dP^*(\beta)\}$, $dG^{**}(\beta) = \beta^{-\alpha} e^\beta dP^{**}(\beta) / \{\int \beta^{-\alpha} e^\beta dP^{**}(\beta)\}$, and therefore $G^* = G^{**}$. Hence $G(\mu)$ is identifiable in $\mathcal{G}$.

Now consider the (β, μ) setting. Suppose the mixing distribution has a cumulative distribution function $G(\mu)$, and there exists another G^* such that for all $\lambda > 0$

$$\int \frac{\beta^{\beta\mu}}{\Gamma(\beta\mu)} \lambda^{\beta\mu-1} e^{-\beta\lambda} dG(\mu) \equiv \int \frac{\beta^{\beta\mu}}{\Gamma(\beta\mu)} \lambda^{\beta\mu-1} e^{-\beta\lambda} dG^*(\mu). \quad (\text{A6})$$

This equality holds at $\lambda = 1$, and hence

$$\int \frac{\beta^{\beta\mu}}{\Gamma(\beta\mu)} dG(\mu) \equiv \int \frac{\beta^{\beta\mu}}{\Gamma(\beta\mu)} dG^*(\mu). \quad (\text{A7})$$

The left- and right-hand sides of equation (A6) divided by those of (A7) in parallel gives

$$\int \lambda^{\beta\mu} dP(\mu) \equiv \int \lambda^{\beta\mu} dP^*(\mu), \quad (\text{A8})$$

where

$$dP(\mu) = \frac{\beta^{\beta\mu} \{\Gamma(\beta\mu)\}^{-1} dG(\mu)}{\int \beta^{\beta\mu} \{\Gamma(\beta\mu)\}^{-1} dG(\mu)} \quad \text{and} \quad dP^*(\mu) = \frac{\beta^{\beta\mu} \{\Gamma(\beta\mu)\}^{-1} dG^*(\mu)}{\int \beta^{\beta\mu} \{\Gamma(\beta\mu)\}^{-1} dG^*(\mu)}.$$

Let $t = \beta \log(\lambda)$. The two sides of equation (A8) become the moment generating functions of P and P^* , respectively, and is bounded when λ is around 1, or t around 0. Thus, based on by the same argument as above, $G(\mu)$ is identifiable in $\mathcal{G}$. $\square$

REFERENCES

- ACINAS, S., KLEPAC-CERAJ, V., HUNT, D., PHARINO, C., CERAJ, I., DISTEL, D. & POLZ, M. (2004). Fine-scale phylogenetic architecture of a complex bacterial community. *Nature* **430**, 551–4.
- BÖHNING, D. & SCHÖN, D. (2005). Nonparametric maximum likelihood estimation of population size based on the counting distribution. *J. R. Statist. Soc. C* **54**, 721–37.
- BUNGE, J. & FITZPATRICK, M. (1993). Estimating the number of species: a review. *J. Am. Statist. Assoc.* **88**, 364–73.
- BURNHAM, K. P. & OVERTON, W. S. (1978). Estimation of the size of a closed population when capture probabilities vary among animals. *Biometrika* **65**, 625–33.
- BURNHAM, K. P. & OVERTON, W. S. (1979). Robust estimation of population size when capture probabilities vary among animals. *Ecology* **60**, 927–36.
- CHAO, A. (1987). Estimating the population size for capture-recapture data with unequal catchability. *Biometrics* **43**, 783–91.
- CHAO, A. & BUNGE, J. (2002). Estimating the number of species in a stochastic abundance model. *Biometrics* **58**, 531–9.
- CHAO, A. & LEE, S.-M. (1992). Estimating the number of classes via sample coverage. *J. Am. Statist. Assoc.* **87**, 210–7.
- FANG, Y. & CHLAMTAC, I. (1999). Teletraffic analysis and mobility modeling of PCS networks. *IEEE Trans. Commun.* **47**, 1062–72.
- HONG, S.-H., BUNGE, J., JEON, S.-O. & EPSTEIN, S. (2006). Predicting microbial species richness. *Proc. Nat. Acad. Sci.* **103**, 117–22.
- LINDSAY, B. G. (1983). The geometry of mixture likelihoods: a general theory. *Ann. Statist.* **11**, 86–94.
- LINDSAY, B. G. (1995). *Mixture Models: Theory, Geometry and Applications*, NSF-CBMS Regional Conference Series in Probability and Statistics, vol. 5. Hayward, CA: Institute of Mathematical Statistics.
- MAGDER, L. S. & ZEGER, S. Z. (1996). A smooth nonparametric estimate of a mixing distribution using mixtures of gaussians. *J. Am. Statist. Assoc.* **91**, 1141–51.
- MORGAN, B. J. T. & RIDOUT, M. S. (2008). A new mixture model for capture heterogeneity. *Appl. Statist.* **57**, 433–46.
- MORRIS, J. S., BAGGERLY, K. A. & COOMBES, K. R. (2003). Bayesian shrinkage estimation of the relative abundance of m-rna transcripts using sage. *Biometrics* **59**, 476–86.
- NORRIS, J. L. & POLLOCK, K. H. (1998). Non-parametric mle for poisson species abundance models allowing for heterogeneity between species. *Environ. Ecol. Statist.* **5**, 391–402.
- PLEDGER, S. (2000). Unified maximum likelihood estimates for closed capture-recapture models using mixtures. *Biometrics* **56**, 434–42.
- R DEVELOPMENT CORE TEAM (2010). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0. URL: <http://www.R-project.org>.
- SANATHANAN, L. (1972). Estimating the size of a multinomial population. *Ann. Math. Statist.* **43**, 142–52.
- SANATHANAN, L. (1977). Estimating the size of a truncated sample. *J. Am. Statist. Assoc.* **72**, 669–72.
- SILVERMAN, B. W. (1992). *Density Estimation for Statistics and Data Analysis*. New York: Chapman & Hall.
- SIMAR, L. (1976). Maximum likelihood estimation of a compound poisson process. *Ann. Statist.* **4**, 1200–9.
- TEICHER, H. (1963). Identifiability of finite mixtures. *Ann. Math. Statist.* **34**, 1265–9.

- THYGESEN, H. H. & ZWINDERMAN, A. H. (2006). Modeling sage data with a truncated gamma-poisson model. *BMC Bioinformatics* **7**, doi:10.1186/1471-2105-7-157.
- WANG, J.-P. (2007). A linearization procedure and a VDM/ECM algorithm for penalized and constrained nonparametric maximum likelihood estimation for mixture models. *Comp. Statist. Data Anal.* **51**, 2946–57.
- WANG, J.-P. & LINDSAY, B. G. (2008). An exponential partial prior for improving nonparametric maximum likelihood estimation in mixture models. *Statist. Methodol.* **5**, 30–45.
- WANG, J.-P. Z., LINDSAY, B., CUI, L., WALL, P., MARION, J., ZHANG, J. & DEPAMPHILIS, C. W. (2005). Gene capture prediction and overlap estimation in EST sequencing from one or multiple libraries. *BMC Bioinformatics* **6**, 300.
- WANG, J.-P. Z. & LINDSAY, B. G. (2005). A penalized nonparametric maximum likelihood approach to species richness estimation. *J. Am. Statist. Assoc.* **100**, 942–59.

[Received March 2009. Revised January 2010]