A Penalized Nonparametric Maximum Likelihood Approach to Species Richness Estimation

Ji-Ping Z. WANG and Bruce G. LINDSAY

We propose a class of penalized nonparametric maximum likelihood estimators (NPMLEs) for the species richness problem. We use a penalty term on the likelihood because likelihood estimators that lack it have an extreme instability problem. The estimators are constructed using a conditional likelihood that is simpler than the full likelihood. We show that the full-likelihood NPMLE solution given by Norris and Pollock can be found (with great accuracy) by using an appropriate penalty term on the conditional likelihood, so it is an element of our class of estimators. A simple and fast algorithm for the penalized NPMLE is developed; it can be used to greatly speed up computation of the unconditional NPMLE. It can also be used to find profile mixture likelihoods. Based on our goal of attaining high stability while retaining sensitivity, we propose an adaptive quadratic penalty function. A systematic simulation study, using a wide range of scenarios, establishes the success of this method relative to its competitors. Finally, we discuss an application in the gene number estimation using expressed sequence tag (EST) data from genomics.

KEY WORDS: Mixture model; NPMLE computing; Number of classes; Penalized NPMLE; Species richness.

1. INTRODUCTION

A sample of individuals is collected from a population consisting of N distinct species (or classes), and their species is identified. If we suppose that N is unknown, then estimation of N based on such a sample is often termed the "species problem" in statistics nomenclature. In this article we consider the setting in which the sampling probabilities vary over species.

This article presents a penalized nonparametric maximum likelihood estimator (NPMLE) of N with desirable properties in bias, variability, and robustness. A penalized form is used to account for an inherent instability in the likelihood. Although the scope of the article is limited to the species problem, the penalized NPMLE and the computing algorithm developed here can be easily extended to the capture-mark-recapture problem as well as to other nonparametric mixture problems. The algorithm can also be used to construct profile likelihoods.

The species problem has a wide variety of important applications covering multiple disciplines including ecology, linguistics, and numismatics (see Bunge and Fitzpatrick 1993 for an extensive review). Existing methods in this area can be loosely classified into parametric or nonparametric, depending on whether the species abundance pattern is modeled by a parametric form. We believe that the nonparametric approaches are generally more desirable because they have competitive performance while adding robustness. Popular nonparametric approaches under comparison with the proposed approach in this article include the lower-bound estimator, $\hat{N}_{c_0}$, of Chao (1984); two coverage coefficient of variation (CV)-based estimators, $\hat{N}_{c_1}$ and $\hat{N}_{c_2}$, of Chao and Lee (1992) (denoted by $\hat{N}_2$ and $\hat{N}_3$ in their article); the coverage-duplication based solution $\hat{N}_{c_3}$ of Chao and Bunge (2002), the jackknife solution, $\hat{N}_J$ of Burnham and Overton (1978, 1979), and the unconditional NPMLE solution, $\hat{N}_{UNP}$, of Norris and Pollock (1996, 1998).

Bunge and Fitzpatrick (1993, p. 364) commented that "there is not as yet a globally preferable estimator." Simulation studies that have appeared in the literature are generally not conclusive. The weakness of the aforementioned approaches have also been recognized in the literature. As a lower-bound estimator, $\hat{N}_{c_0}$ is stable but usually biased downward. The coverage estimators $\hat{N}_{c_1}$, $\hat{N}_{c_2}$, and $\hat{N}_{c_3}$ depend on the choice of a tuning parameter τ, where one splits the sample into "rare" (observed no more than τ times) and "abundant" (seen more than τ times) groups. One then estimates N based on the rare species frequencies (Chao, Ma, and Yang 1993; Chao, Huang, Chen, and Kuo 2000). However, it was noticed that $\hat{N}_{c_1}$, $\hat{N}_{c_2}$, and $\hat{N}_{c_3}$ are all sensitive to τ (Chao and Bunge 2002). As shown in our simulation study, a bad choice of τ can bias the estimate substantially. The coverage-duplication estimator $\hat{N}_{c_3}$ can fail due to a negative estimate of the duplication fraction (Chao and Bunge 2002). The jackknife estimator $\hat{N}_J$ is robust, but often has a positive bias (Smith and van Belle 1984). The unconditional NPMLE $\hat{N}_{UNP}$ achieves remarkable insensitivity to τ but requires extremely long computing times, especially when N is large. We also demonstrate that it has an instability problem.

Two nonparametric likelihood approaches exist, the unconditional and conditional, as identified by Sanathanan (1972). In this article we unify these approaches and extend them by considering the addition of a penalty term to the conditional likelihood. The unconditional NPMLE can then be obtained, to a high degree of approximation, by maximizing the conditional likelihood under an appropriate penalty. This device provides a great reduction in computing time for the unconditional estimator.

However, both the conditional and unconditional NPMLEs have a severe instability problem. We show that by modifying the penalty, one can stabilize the resulting estimators. We discuss the statistical interpretation of the penalty in terms of Bayesian estimation, and demonstrate that the addition of an adaptive quadratic penalty term greatly improves the performance of the conditional NPMLE over a wide range of simulation settings.

2. NONPARAMETRIC MAXIMUM LIKELIHOOD ESTIMATOR SOLUTIONS

Let $\mathbf{X} = \{X_1, \ldots, X_N\}$ be the number of observed individuals from each distinct species, where the unobserved (and therefore

Ji-Ping Z. Wang is Assistant Professor, Department of Statistics, Northwestern University, Evanston, IL 60208 (E-mail: *jzwang@northwestern.edu*). Bruce G. Lindsay is Professor, Department of Statistics, Pennsylvania State University, University Park, PA 16802 (E-mail: *bgl@psu.edu*). This research was supported in part by National Science Foundation grant DMS-01-04443 to Lindsay. The authors thank Annie Chao, two referees, the associate editor, and the editor for helpful comments and suggestions.

Journal of the American Statistical Association
September 2005, Vol. 100, No. 471, Theory and Methods
DOI 10.1198/016214504000002005

unknown) species have $X_i = 0$. If so, then $n_j = \sum_{i=1}^{N} I(X_i = j)$, for $j = 1, \ldots$ becomes the number of those species that had j individuals in the sample. The variable n_0 then represents the unknown number of species that were present in the population but never observed.

Assume that the X_i's, for $i = 1, \ldots, N$, are iid observations from $f(x; M)$, where M represents a set of unknown parameters (or an unknown mixing distribution in the nonparametric mixture case). Let $t = \max_i(x_i)$, and let $D = n_1 + \cdots + n_t$ be the total of observed distinct species. As shown by Sanathanan (1972), the likelihood function for this model can be factored into two parts,

$$\begin{aligned} L(N, M; \mathbf{X}) &= \binom{N}{n_1, \ldots, n_t} \prod_{j=0}^{t} [f(j; M)]^{n_j} \\ &\propto \binom{N}{D} f(0; M)^{N-D} [1 - f(0; M)]^{D} \\ &\quad \times \prod_{j>0} \left[\frac{f(j; M)}{1 - f(0; M)} \right]^{n_j} \\ &\equiv L_m(N, M) \times L_c(M). \end{aligned} \tag{1}$$

Here the likelihood $L_m(N, M)$ is from the binomial marginal distribution of D, which depends on both N and M. The conditional distribution of $\mathbf{X}$ given D generates $L_c(M)$, which depends on M alone.

Sanathanan (1972, 1977) discussed two likelihood-based estimation methods for this problem. The unconditional method finds the pair $(\hat{N}, \hat{M})$ that maximize $L(N, M; \mathbf{X})$ globally. In the second method, one finds $\hat{M}$ by first maximizing $L_c(M)$, then calculating the MLE of N from $L_m(N, M)$ treating $M = \hat{M}$. If we define the *odds parameter* as

$$\theta(M) = \frac{f(0; M)}{1 - f(0; M)}, \tag{2}$$

then, for a given θ, the conditional estimator $\hat{N}_c$ equals $\langle D(1 + \theta) \rangle = \langle \frac{D}{1 - f(0; M)} \rangle$, where "$\langle a \rangle$" means the largest integer no greater than a (see also Lindsay and Roeder 1987). In other words, we obtain the conditional MLE given $f(0; M)$ by simply rescaling D by the non-0 probability. [Because the penalized approaches under consideration in this article all concern θ directly rather than $f(0; M)$, we write the conditional estimator $\hat{N}_c$ in terms of θ in the following context.] Sanathanan (1972, 1977) also estabilished that the unconditional and conditional MLEs of N have the same asymptotic distribution when the regularity conditions hold (e.g., in the Poisson–gamma model).

2.1 Poisson Mixture Model and Conditional NPMLE

We assume the model where f in (1) is a Q-mixture of Poisson variables, that is,

$$f(x; Q) = \int \frac{e^{-\lambda} \lambda^x}{x!} \, dQ(\lambda),$$

where Q is an arbitrary unknown distribution on the parameter space Ω. This arises from the following hierarchical model: Suppose that the sampling count for species i is a Poisson random variable with abundance parameter λ_i, and that the distribution of the abundance parameters among the species is Q. Mao and Lindsay (2003) identified that the conditional log-likelihood in (1) can be reparameterized into a P-mixture of 0-truncated Poisson densities as

$$\ell_c(P) = \sum_{j=1}^{t} n_j \log[g(j; P)], \tag{3}$$

where $g(j; P) = \int \frac{e^{-\lambda} \lambda^j}{j!(1 - e^{-\lambda})} \, dP(\lambda)$ and

$$dP(\lambda) = \frac{(1 - e^{-\lambda}) \, dQ(\lambda)}{\int (1 - e^{-\lambda}) \, dQ(\lambda)}. \tag{4}$$

If $\Omega = (0, \infty)$, then $Q \to P$ is a one-to-one transformation with inverse $dQ(\lambda) = \frac{(1 - e^{-\lambda})^{-1} \, dP(\lambda)}{\int (1 - e^{-\lambda})^{-1} \, dP(\lambda)}$.

However, if we were to include the boundary point 0 in Ω, then the map is not invertible. This corresponds to the fact that because n_0 is not observed, any mass of Q placed at 0 is not identifiable. It is natural to exclude 0 from Ω, because any species with $\lambda = 0$ is not actually present and should not count in N. (Looking ahead, we will show that the near nonidentifiability of the mass in Q for λ near 0 still presents a severe problem.)

In what follows, we reserve the notation Q for the mixing distribution of the original Poisson mixture and P for its 0-truncated transformation. Due to the invertible relationship between Q and P on $(0, \infty)$, for convenience we write ℓ_m and ℓ_c in P or Q interchangeably, that is, $\ell_m(N, Q) \equiv \ell_m(N, P)$ and $\ell_c(Q) \equiv \ell_c(P)$.

The advantage of the form (3) is that it is a standard nonparametric mixture likelihood (Lindsay 1995) of iid observations from a P-mixture of 0-truncated Poisson variables. The absence of the parameter N in $\ell_c(P)$ makes this a simpler optimization problem. One can first find $\hat{P}$ from $\ell_c(P)$, a conditional MLE. Notice that because $\theta(Q)$ from (2) has the form $\theta(Q) = \frac{f(0; Q)}{1 - f(0; Q)}$, we can write θ as a function of P in the simple linear form,

$$\theta = \int (e^{\lambda} - 1)^{-1} \, dP. \tag{5}$$

We can then maximize the marginal likelihood with P fixed at $\hat{P}$ to obtain an N-estimator of the form $\langle D(1 + \theta(\hat{P})) \rangle$. We call the resulting estimator the *conditional NPMLE* and denote it by $\hat{N}_{CNP}$, although technically it is a two-step "conditional-then-marginal" MLE.

The conditional NPMLE of N is simple to implement and fast to compute. However, our numerical experience shows that it has a severe instability problem related to the boundary of the parameter space. In these cases, the conditional NPMLE of $\hat{P}$ from (3) contains λ at or near 0 as a support point. In particular, if one were to allow Q to have mass at $\lambda = 0$, then the NPMLE would often put positive mass there. In such cases, if one restricted the parameter space to $\Omega = (0, \infty)$, then the maximum over Q would not be attained (and an algorithm will keep trying to put mass near 0), because the solution is on the excluded boundary.

To solve this, we could try allowing $\lambda = 0$ to be a mass point in the algorithm. But this creates two new problems. First, we

cannot invert the Q to P map to find a unique $\hat{Q}$, because the mass at $\lambda = 0$ is not well defined. Second, this situation has a severe impact on N estimation. Recall that the estimator of N is $\langle D(1 + \theta(\hat{P}))\rangle$, and see that $\theta(\hat{P})$ goes to ∞ as any support point in $\hat{P}$ approaches 0; see (5). That is, if we allow mass at $\lambda = 0$, then the estimator will be ∞, and if we allow mass arbitrarily near 0, then $\hat{N}_{CNP}$ will "blow up." In practice, a tiny λ component is frequently fit in $\hat{P}$, which blows up $\hat{N}_{CNP}$.

2.2 A Special Challenge

This numerical instability is closely related to the results of Mao and Lindsay (2002), who identified some challenging theoretical defects in the model. In particular, they showed that it is theoretically impossible to create an upper confidence limit for N that would hold a target confidence level across all possible abundance distributions Q (unless one uses the trivial value of ∞ as the upper limit).

There is a simple logic behind the mathematics. Suppose that there were M species with abundance $\lambda = 0$. Of course, they are never seen, and so we could not hope to determine their number. Even though we have excluded $\lambda = 0$ from Ω, this does not exclude the possibility that there are many species with vanishingly small λ, small enough that none of them would ever appear in a reasonable sample. To put this in another way, there is less and less statistical information about the distribution Q as $\lambda \to 0$, and it is these small λ's that can have a large effect on N estimation.

There exist a substantial number of nonparametric estimators of N, and they are often accompanied by a method for constructing standard errors. How can this be if upper confidence limits are impossible? It is because these estimators might better be called *proxy* estimators, because they consistently estimate proxy functionals of the model that equal N only for some subset of all distributions Q. Outside this subset, the estimators display nonvanishing bias at all sample sizes.

The ramifications of this can be seen in our simulations. For many simulation settings, the 95% central portion of the estimators' sampling distributions will not even cover the true value of N. On the other hand, the NPMLE has a sampling distribution that typically covers the true value, but it does so by frequently creating very large values due to its aforementioned instability.

In the face of this, our goal in this research was to see whether one could penalize the nonparametric likelihood in such a way that the resulting NPMLE retained much of its flexibility relative to other methods, but also behaved more stably.

2.3 Penalized Likelihood

Let $\ell(P)$ be a log-likelihood function, where in our case P is an unknown probability measure. The *penalized likelihood* corresponding to penalty parameter γ and penalty function $h(P)$ is defined as

$$\ell^{\gamma}(P) = \ell(P) - \gamma h(P). \tag{6}$$

The penalized conditional likelihood corresponding to the conditional likelihood ℓ_c in (3) is then

$$\ell_c^{\gamma}(P) = \ell_c(P) - \gamma h(P). \tag{7}$$

The resulting penalized full log-likelihood is

$$\ell^{\gamma}(P) = \ell_m(N, P) + \ell_c^{\gamma}(P).$$

We write h in terms of P rather than Q, because ℓ_c is more naturally parameterized by P in the conditional likelihood form (3).

Remark 1. For $\gamma > 0$ and $h(P) > 0$, clearly a maximizer of (7) tends to avoid P with large values of $h(P)$. Note, however, that the form (7) also arises when one maximizes ℓ_c subject to the constraint $h(P) = h_0$ using the method of Lagrange multipliers, with the γ being the Lagrange parameter. For this reason, our methods are also relevant for profile likelihood construction.

To estimate N, we first find the penalized NPMLE of P, $\hat{P}^{\gamma}$, from $\ell_c^{\gamma}(P)$, then define the *penalized conditional NPMLE* of N to be

$$\hat{N}^{\gamma} = \langle D(1 + \theta(\hat{P}^{\gamma}))\rangle. \tag{8}$$

The penalized likelihood form in (7) provides rich possibilities for controlling the behavior of estimators of N. However, one important issue in construction of the penalty is that computing the NPMLE from $\ell_c^{\gamma}(P)$ may not be trivial. It is known that if the functional $h(P)$ is linear [i.e., of the form $h(P) = \int h(\lambda)\, dP(\lambda)$], then the NPMLE can be found using a simple extension of standard NPMLE methods (Lindsay 1995). For example, the odds parameter $\theta(P)$ would be such a linear functional, with $h(\lambda) = (e^{\lambda} - 1)^{-1}$. In this article we consider three choices of penalty function $h(P)$ all involving the odds parameter θ. One of the innovations of this article is a fast extension of the standard computing algorithm to the situation where the penalizing function is a differentiable function of a linear functional like $\theta(P)$, thereby expanding our possibilities.

2.4 Unconditional NPMLE via Penalized Likelihood

The first choice of the penalizing function is

$$h_1(P) = \log\left[\frac{\theta(P)}{1 + \theta(P)}\right].$$

Note that $\log[\frac{\theta(P)}{1+\theta(P)}] = \log[f(0, P)] \le 0$. A larger $f(0, P)$ causes more penalty (or, strictly, less reward) if $\gamma > 0$ in (7). We actually include this penalty because it can be used to generate a close approximant to the unconditional NPMLE. The following proposition provides a way to simplify the unconditional solution.

Proposition 1. (a) Suppose that $(\hat{N}, \hat{Q})$ is the unconditional MLE (parametric or nonparametric), and let $\hat{\theta} = \theta(\hat{Q})$. Then $\hat{N} = \langle D(1 + \hat{\theta})\rangle$ and $\ell_c(\hat{Q}) = \sup_Q\{\ell_c(Q) : \theta(Q) = \hat{\theta}\}$.

(b) Consider the full log-likelihood function $\ell(N, Q) = \ell_m(N, Q) + \ell_c(Q)$ and the modified objective function

$$\ell^*(Q) = \ell_m(N, Q)|_{N=\langle D(1+\theta(Q))\rangle} + \ell_c(Q).$$

Then $\hat{Q}$ is the unconditional MLE (parametric or nonparametric) from $\ell(N, Q)$ if and only if $\hat{Q}$ also maximizes $\ell^*(Q)$.

(c) For each $\theta \equiv \theta(Q) \in (0, \infty)$,

$$\begin{aligned}\ell_m(N, Q)|_{N=\langle D(1+\theta(Q))\rangle} &= \sup_N(\ell_m|\theta) \\ &\approx \log\left[\frac{e^{-D}D^D}{D!}\right] - .5\log\left[\frac{\theta}{1+\theta}\right], \end{aligned} \tag{9}$$

where "$\approx$" means that the difference of the two sides goes to 0 as $D \to \infty$.

For the proof see Appendix A.

Remark 2. The approximation in part (c) of Proposition 1 corresponds to approximating the binomial marginal likelihood with a Poisson likelihood modified by a θ function. As $\theta \to \infty$, the modifier vanishes. The approximation error, defined as the difference of the two sides of (9), decreases as D or θ increases. For example, at $D = 100$, as θ varies from .2 to 4, the error changes from 3.5E–3 to 4.2E–5. At $D = 500$, the error decreases from 6.9E–4 to 8.3E–6 as θ varies in the same range.

By part (a) of Proposition 1, the unconditional NPMLE, denoted by $\hat{N}_{UNP}$, is determined by $\hat{Q}$. Moreover, an objective function for finding $\hat{Q}$ is given by the profile likelihood ℓ^* in part (b) of Proposition 1. If we substitute the approximation in part (c) into $\ell^*(P)$, then we arrive at a penalized likelihood of the form

$$\ell_1 = \ell_c(P) - \gamma_1 \log\left[\frac{\theta}{1+\theta}\right], \qquad \gamma_1 = .5. \tag{10}$$

The NPMLE $\hat{Q}$ from ℓ_1 is therefore an approximation to the unconditional NPMLE of Q. Due to the excellence of Stirling's approximation (see Remark 2), the resulting point estimator, denoted by $\hat{N}_u$, can be practically regarded as equivalent to $\hat{N}_{UNP}$ (as we demonstrate later).

From this result, we can directly derive a relationship between $\hat{N}_u$ and $\hat{N}_{CNP}$ based on the monotonicity property of the penalized likelihood in Theorem 1, and extend it to a relationship between exact unconditional and conditional MLEs in both nonparametric and parametric situations in Corollary 1.

Theorem 1. Let $\hat{P}^\gamma$ be the NPMLE for the penalized likelihood form (6), where $h(P) = h[\theta(P)]$ is an increasing function of θ. Let $\hat{N}^\gamma$ be the corresponding N estimator based on (8). If $\gamma_1 \le \gamma_2$, then $\theta(\hat{P}^{\gamma_1}) \ge \theta(\hat{P}^{\gamma_2})$, which implies that $\hat{N}^{\gamma_1} \ge \hat{N}^{\gamma_2}$.

For the proof see Appendix B.

Corollary 1. (a) $\hat{N}_u \le \hat{N}_{CNP}$.

(b) The unconditional MLE of N is always no greater than the conditional MLE in both the parametric and nonparametric cases.

For the proof see Appendix C.

The monotonicity stated in Theorem 1 explains how the variability can be controlled by penalty in the NPMLE solutions. In particular, the exact unconditional MLE, whether parametric or nonparametric, can be regarded as a penalized version of conditional MLE and thus is always no greater than the latter. The degree of shrinkage is jointly determined by the penalty and penalizing function. For example, increasing the penalty parameter γ brings a more stable estimator, but at the possible cost of negative bias due to overpenalization.

Our goal is to build a penalty that balances sensitivity and stability. Note that $\log[\frac{\theta}{1+\theta}] = \log[f(0; Q)]$; therefore, in ℓ_1 we are actually penalizing the zero probability. Unfortunately, although this penalty does improve stability over the conditional method, it turns out to be inadequate to eliminate the boundary problem (see an example in simulation section). Note that $\log[\frac{\theta}{1+\theta}]$ increases in θ but becomes flat rapidly for $\theta > .5$ [Fig. 1(b)]. This near-constancy implies that this penalty function becomes ignorable as θ grows large. As a result, it does not guarantee the elimination of a boundary support point, and so it retains the instability of the conditional NPMLE. For this reason, we consider the following, more extreme penalties.

2.5 More Extreme Penalties

A natural choice for the penalty function is the odds functional $\theta(P)$ itself. Recall that $\hat{N} = \langle D + D\hat{\theta}\rangle$, so imposing a penalty on θ reduces the magnitude of $\hat{N}$. In fact, it is clear that the penalized likelihood

$$\ell_2(P) = \ell_c(P) - \gamma_2\theta(P), \qquad \gamma_2 > 0 \tag{11}$$

cannot have its maximum at $\theta(P) = \infty$, so extreme estimates due to the boundary problem cannot occur. By Theorem 1, we can tune the penalty term γ_2 to control the variability of $\hat{N}$. Despite our initial optimism about $\ell_2(P)$, it was shown by simulation in Wang (2003) that the optimal choice of γ_2 depends strongly on the true value of θ. A larger true θ value requires more bias correction, and vice versa.

This consideration led us to consider instead the following, more severe penalty:

$$\ell_3(P) = \ell_c(P) - \gamma_3(\theta - \mu)^2 I(\theta > \mu), \qquad \gamma_3, \mu > 0. \tag{12}$$

Under the penalty function $(\theta - \mu)^2 I(\theta > \mu)$ in (12), we penalize the quadratic distance between θ and μ when θ exceeds threshold μ. This offers no penalty for small θ, but the penalty becomes much larger than those in $\ell_1(P)$ and $\ell_2(P)$ as θ becomes large. This penalty function requires a choice of two parameters, γ_3 and μ. If a chosen μ is too small, then the penalty can be too strong, and the resulting penalized estimator of N can be too conservative. But choosing μ too large can be ineffective at variability control. The parameter γ_3 affects the estimator in the opposite way.

To avoid underpenalizing and overpenalizing, we used adaptive values for the penalty parameters. We let $\mu = \hat{N}_{c_0}/D - 1$ and $\gamma_3 = \frac{1}{2\mu}$, where $\hat{N}_{c_0}$ is the lower-bound estimator of Chao (1984). The resulting estimator is denoted by $\hat{N}_{WL}$. The logic behind these choices is that the parameter μ, as defined based on $\hat{N}_{c_0}$, is a lower-bound estimator of θ. The estimator $\hat{N}_{c_0}$ is very stable but usually negatively biased. In our experience, $\hat{\theta}_{CNP}$ is never smaller than μ, so the penalty term always shrinks the NPMLE toward $\hat{N}_{c_0}$. The term γ_3 was chosen so that the penalty term corresponds to a normal prior with variance μ (see the next section for further discussion), which could be anticipated to bound the estimator above by μ plus about $3\,\sigma = 3\sqrt{\mu}$.

Of course, this shrinkage of the NPMLE toward μ is likely to cause bias for some P. However, in our experience the estimation bias of the NPMLE estimators is less damaging to inference than their excessive variability. In particular, we found that this variability is reflected in sampling distributions for the estimators that are right-skewed with a very long right tail (even to ∞, due to the boundary problem).

2.6 Interpreting the Penalties Using Bayes Risk

Our goal is to control the variability of NPMLE-based estimators through penalizing the likelihood. A Bayesian interpretation of the penalized log-likelihood function is that it equals the logarithm of the posterior distribution, given that the penalty term is the logarithm of the prior distribution on the latent distribution P, and so penalized estimation corresponds to maximum a posteriori estimation.

We can give an interpretation of this through Bayes risk. If our goal were to improve the risk in estimating N at smaller values of $\theta(P)$ (at the price of increased risk for larger values of $\theta(P)$), then we would use a Bayes prior that downweights large $\theta(P)$. This seems to be a sensible strategy given that estimation is already quite bad when $\theta(P)$ is large (which corresponds to many species of vanishingly small observability).

This leads us to interpret our penalties in terms of the prior information implicit in them. The implicit prior measures for the odds functional $\theta(P)$ corresponding to the penalties in ℓ_1, ℓ_2, and ℓ_3 are plotted in Figure 1 and denoted by

$$\ell_1 : m_1(\theta) \propto \left[1 + \frac{1}{\theta}\right]^{.5},$$

$$\ell_2 : m_2(\theta) \propto e^{-\gamma_2 \theta}, \qquad \gamma_2 > 0,$$

and

$$\ell_3 : m_3(\theta) \propto e^{-\gamma_3(\theta-\mu)^2} I\{\theta > \mu\} + cI(\theta \le \mu), \qquad \mu, \gamma_3 > 0, c > 0.$$

The density m_1 is proportional to an approximation to the profile marginal likelihood $\sup_N(L_m|\theta)$ (Prop. 1). This is very much an improper prior, because m_1 is everywhere greater than 1; it also does not integrate on intervals neighboring 0. The prior m_2 is an exponential distribution with mean $1/\gamma_2$. The prior m_3 is a hybrid of a uniform distribution on $(0, \mu)$ and the right half of a normal distribution with mean μ and variance $\sigma^2 = \frac{1}{2\gamma_3}$ on (μ, ∞). From this, it is clear that very little mass occurs to the right of $\mu + 3\sigma$.

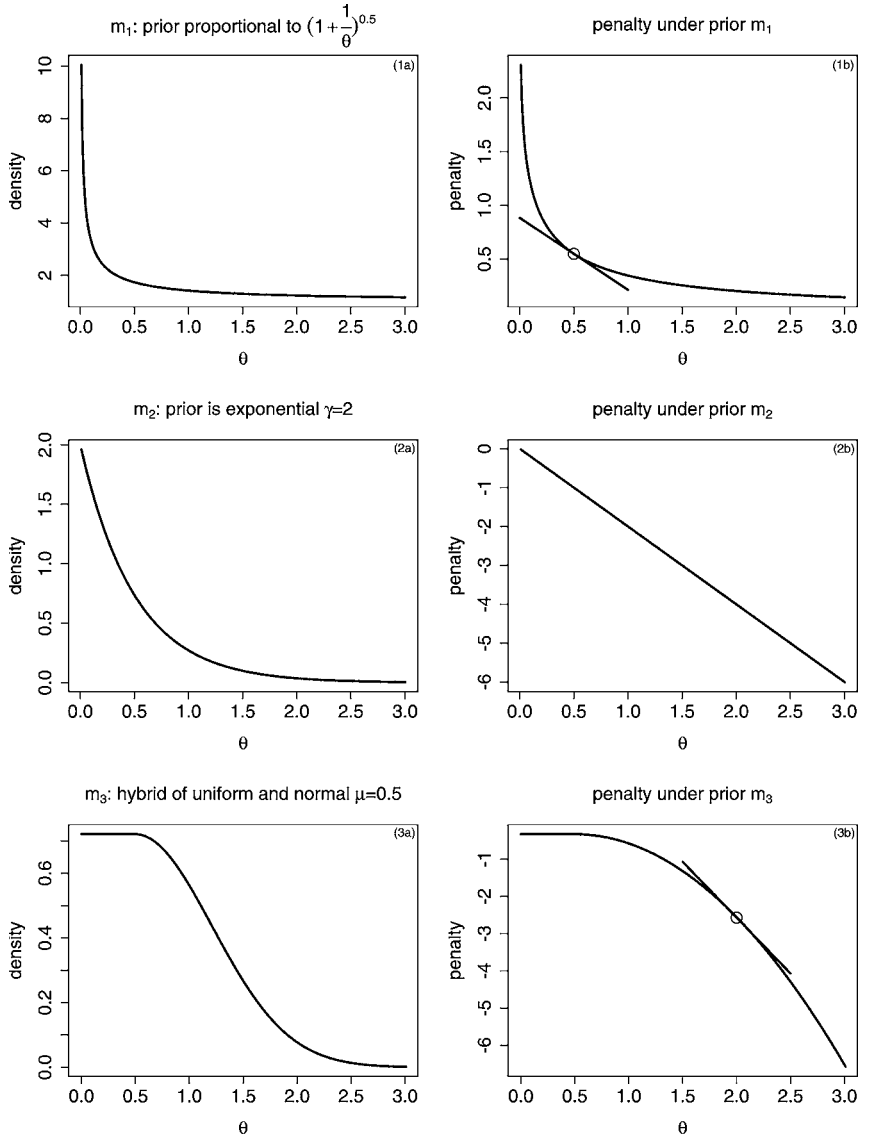

Figure 1. Priors and Corresponding Penalizing Functions for ℓ_1, ℓ_2, and ℓ_3.

The foregoing three priors all shrink the N estimates in the sense that the penalized NPMLE $\hat{N}_{\ell_i} \le \hat{N}_{CNP}$ (see Thm. 1). The choice of γ (and μ in ℓ_3), however, directly determines the shape of the posterior distribution of θ. If the true θ is large, then we would like the prior to have larger mean value or to be flatter, and vice versa. This suggests that if no strong prior information is available, then we might wish to generate an adaptive prior based on the sample itself.

The prior m_3 is such a creation. Because $\hat{N}_{CNP}$ rarely (if ever) goes below the lower-bound estimator $\hat{N}_{c_0}$, the part of the prior that affects the estimator is the right tail, the normal distribution for θ with $\mu = \hat{N}_{c_0}/D - 1$ and $\sigma^2 = \mu$. This choice of σ^2 is natural, because larger values of μ imply more unsampled species and thus more uncertainty.

Remark 3. As an alternative to construction a penalty via prior specification, one can think of the additional likelihood term as being the likelihood contribution from some additional (and fictional) data. This can be helpful in construction and interpretation, because then one can get some feeling of how much prior information one is adding to the problem.

For example, suppose that in addition to the other data, one had an independent observation x from a binomial distribution with parameters m and $p = 1 - f(0; Q) = 1/(1+\theta)$. This fictional data would correspond to adding m additional species to the population, with each species generated from the same abundance distribution Q as the others, and finding that t were observed and $m - x$ were not. These data would have the log-likelihood

$$x \log \frac{1}{1+\theta} + (m-x) \log \frac{\theta}{1+\theta}.$$

One could then interpret x/m as the estimated likelihood that a randomly drawn species would be observed, and the parameter m would represent the weight that one is giving these "prior data." If $x > 0$, then this augmentation term would become negative infinity as $\theta \to \infty$, and so the maximum of the penalized likelihood could not occur there. For example, if we used $x = \gamma$ and $m = \gamma$, then the augmented likelihood term would be $-\gamma \log(1+\theta)$, corresponding to the monotonic-increasing penalty function $h(P) = \log(1+\theta)$. The penalized likelihood ℓ_1 is another example where x is known as D and m is set approximately as $\langle D(1+\theta) \rangle$.

2.7 Optimization Results and an Algorithm

The theory of optimization for a standard mixture log-likelihood such as $\ell_c(P)$ in (3), or the penalized modifications such as found here, is based on principles of convex optimization theory (Lindsay 1995). In such problems, there is a natural extension of the likelihood equations used in ordinary parametric likelihoods. Let Δ_λ denote the degenerate distribution at λ. One starts with the gradient function $D(P, \lambda)$, defined as the derivative in ε of the objective function $L(P)$ (log-likelihood or the penalized version) along the path $(1-\varepsilon)P + \varepsilon\Delta_\lambda$, evaluated at $\varepsilon = 0$. We say that $L(P)$ is maximized *locally* at $\hat{P}$ if and only if

$$D(\hat{P}, \lambda) \le 0 \quad \text{for all } \lambda \in \Omega.$$

This condition is necessary and sufficient for $\hat{P}$ to be a global maximum if $L(P)$ is a concave function of P.

This gradient criterion naturally leads to algorithms that can be used to optimize L. We have developed an extension of VDM/EM algorithm (Lindsay and Roeder 1992) to find the NPMLE from ℓ_2 very reliably. Critical in its development is the linearity in P of the penalty $\theta(P)$, because this enables one to extend the EM algorithm to include the penalty term. For $\ell_1(P)$, the penalty $-\gamma_1 \log[\frac{\theta}{1+\theta}]$ is a convex function of the linear functional $\theta(P)$, so the concavity of $\ell_1(P)$ need not hold, and so we are not guaranteed a global solution at convergence. The likelihood $\ell_3(P)$ is concave due to the concavity of the penalty $-\gamma_3(\theta - \mu)^2 I(\theta > \mu)$; thus a global maximum is guaranteed. However, linearity in P does not hold for either ℓ_1 or ℓ_3, and we have found that direct maximization of ℓ_1 or ℓ_3 is difficult. The following optimization theorem, however, can be used as the basis for a novel strategy for finding the NPMLE in ℓ_1 or ℓ_3 by iteratively maximizing the same likelihood form as in ℓ_2.

Before giving our result, we offer some background on the existence of NPMLEs for mixture likelihoods. In particular, suppose that $\ell(P)$ has the standard mixture likelihood form $\sum_{nj} \log \int L_j(\lambda)\, dP(\lambda)$, where $L_j(\lambda)$ is the mixture density at the ith distinct value for $i = 1, \ldots, n$ (truncated Poisson with mean parameter λ in our case). Suppose that we are interested in the penalized likelihood $\ell(P) - \gamma h(P)$, where $h(P) = \int h(\lambda)\, dP(\lambda)$ is a linear function of P. Then, according to Lindsay (1995, p. 142), there exists an NPMLE provided that the extended likelihood curve $\{(L_1(\lambda), \ldots, L_n(\lambda), h(\lambda)) : \lambda \in \Omega\}$ is a closed and bounded set.

In cases like ours, boundedness in L_j is automatic because we have discrete densities, so each likelihood component is bounded by 1. The closed condition is a little more difficult, because this requires that we include the limit point $\lambda = 0$ of the parameter space. Indeed, this is the source of the NPMLE instability. However, once we move to a penalized likelihood optimization problem, this is not a problem, as long as the penalty function becomes infinite for P containing such limit points. Therefore, even if we include these points to ensure the closedness of the image set, we know that they cannot be used in the solution.

But Lindsay's results do not cover nonlinear penalties, so we must extend the relevant existence results for penalized mixture likelihoods.

Theorem 2. Consider a penalized mixture likelihood of the form $\ell(P) - \gamma h[\theta(P)]$, where $\ell(P)$ is a standard mixture likelihood, where $\theta(P) = \int \theta(\lambda)\, dP(\lambda)$ is a linear functional of P, and where $h(\theta)$ is a differentiable function of θ. Suppose also that the image set of the function $(L_1(\lambda), \ldots, L_n(\lambda), \theta(\lambda))$ is closed and bounded. Then the following apply:

(a) There exists a distribution $\hat{P}_\omega$ that maximizes $\ell^\omega(P) = \ell(P) - \omega\theta(P)$.

(b) A necessary and sufficient condition for $\ell(P) - \gamma h[\theta(P)]$ to be locally maximized at $\hat{P}$ (in the sense of satisfying the gradient requirement) is that $\hat{P}$ is also the global maximizer of $\ell^\omega(P)$ when ω is set equal to the constant $\gamma h'[\theta(\hat{P})]$; that is, each solution to the h penalty problem can be found as a solution to the linear problem by using an appropriate penalty weight.

(c) If $h(\theta)$ is convex in θ, then any local maxima to the penalized likelihood is the global maximum.

For the proof see Appendix D.

Applying Theorem 2 to our particular penalized likelihoods gives the following result.

Corollary 2. (a) If there exists $\hat{P}$ that globally maximizes ℓ_1 such that $\theta(\hat{P}_1) < \infty$, then $\hat{P}_1$ must be the unique global maximum to ℓ_2 under the penalty $\gamma_2 = \gamma_1 * h_1'(\hat{\theta}) = \gamma_1 \frac{1}{\hat{\theta}(1+\hat{\theta})}$, where $\hat{\theta} = \theta(\hat{P}_1)$.

(b) There is a unique mixing distribution $\hat{P}_3$ that globally maximizes ℓ_3. Let $\hat{P}_{CNP}$ be the conditional NPMLE without penalty. If $\theta(\hat{P}_{CNP}) \le \mu$, then $\hat{P}_{CNP}$ is the unique global maximizer of ℓ_3. Otherwise, $\hat{P}_3$ is the maximizer of ℓ_3 if and only if $\hat{P}_3$ maximizes ℓ_2 at $\gamma_2 = \gamma_3 h_3'[\theta(\hat{P}_3)] = 2\gamma_3[\theta(\hat{P}_3) - \mu]$.

For the proof see Appendix E.

Remark 4. For ℓ_1, the penalty term h_1 is unbounded below as $\theta \to 0$ (and hence $-\gamma_1 h_1$ can go to ∞). However, in our experience this does not present a practical problem with finding a solution. Forcing θ toward 0 can be done only by letting the λ supports of P become infinite; but then the rest of the log-likelihood goes to minus infinity at a faster rate.

Remark 5. For the penalized likelihood with linear or convex penalty in the truncated Poisson case, the global NPMLE solution exists and is unique. See a proof in Appendix D for the uniqueness.

Geometrically, the idea of the algorithm is exceedingly simple. One finds the NPMLE for the likelihood ℓ_2 iteratively, with γ_2 updated in each iteration as the product of γ_1 (or γ_3 for ℓ_3) times the derivative of $h_1(\theta)$ [or $h_3(\theta)$ for ℓ_3], evaluated at $\theta(\hat{P})$ from the last iteration.

In the ℓ_1 case, the algorithm comprises the following steps:

1. Initialize $\theta^{(0)} = (\hat{N}_{c_0} - D)/D$, and find $\hat{P}^{(1)} = \arg_P \sup\{\ell_2|_{\gamma_2=\gamma_1 h_1'(\theta^{(0)})}\}$.
2. At the tth iteration, let $\theta^{(t-1)} = \theta[\hat{P}^{(t-1)}]$, and find $\hat{P}^{(t)} = \arg_P \sup\{\ell_2|_{\gamma_2=\gamma_1 h_1'(\theta^{(t-1)})}\}$.
3. Repeat step 2 until $|\theta^{(t)} - \theta^{(t-1)}| < tol$.

We used $tol = 1/D$, because $\hat{N} = \langle D + D\hat{\theta}\rangle$, and so the N-estimate is changing by less than one unit per step. Despite the fact that part (a) of Corollary 2 guarantees only a local condition in ℓ_1, we have never encountered multiple solutions in our simulation study and real data analysis. Our examination of the profile likelihood of $\theta(P)$, which can be constructed from ℓ_2, suggests that it is almost always unimodal, which confirms similar findings by Norris and Pollock (1998).

To maximize ℓ_3, we followed this procedure: (1) obtain $\hat{P}_{CNP}$; and (2) if $\theta(\hat{P}_{CNP}) \le \hat{N}_{c_0}$, then the algorithm stops. In our experience, $\theta(\hat{P}_{CNP})$ smaller than $\hat{N}_{c_0}$ was never found, which is a lower-bound estimator. Otherwise, we need only replace $\gamma_1 h_1'(\theta)$ by $\gamma_3 h_3'(\theta)$ in the foregoing algorithm. A unique global maximizer is guaranteed by Theorem 2.

3. CONFIDENCE INFERENCE

As we have noted, in a nonparametric inference it might be most honest to think of determining lower confidence limits for N and using ∞ as the upper bound. However, in a practical sense we might wish to have an a finite upper bound that reflects some reasonable optimism about the underlying abundance distribution Q having not too much mass at very small λ. This thinking is implicit in the finite confidence limits that have been given for the coverage and jackknife estimators.

For our nonparametric likelihood methods, we cannot appeal to the asymptotic results of Sanathanan (1977), because they were for finite-dimensional parametric models. We have therefore investigated bootstrap approaches for confidence interval construction.

In a *Poisson-based bootstrap*, we would plug our fitted parameter values $\hat{N}$ and $\hat{Q}$ into our Poisson mixture model and use it to simulate $\hat{N}$ new X values. The values of $X_i = 0$ would be truncated away, giving us a new dataset. However, this seems a bit artificial, because in typical species datasets it is individuals that are sampled from the population, and then the species count comes from aggregating individuals.

This suggests an alternative *multinomial-based bootstrap*; that is, we create an estimated population, then simulate draws of individuals from it. We propose doing this as follows. Suppose that $\hat{N}$ is the point estimate of N and that the support points of $\hat{Q}$ are $(\hat{\lambda}_1, \hat{\lambda}_2, \ldots, \hat{\lambda}_k)$ with weights $(\hat{\pi}_1, \hat{\pi}_2, \ldots, \hat{\pi}_k)$. We create a multinomial sampling model by creating $\hat{N}$ cells, with each cell corresponding to a species. We divide the cells into k groups, with each cell group corresponding to species from a particular abundance level $\hat{\lambda}_j$, $j = 1, \ldots, k$. The number of cells in the groups are $\hat{N}\hat{\pi}_1, \hat{N}\hat{\pi}_2, \ldots, \hat{N}\hat{\pi}_k$. The multinomial parameter for each cell in the jth group is then equal to

$$p_j = \frac{\hat{\lambda}_j}{\hat{N}\sum_i \hat{\lambda}_i \hat{\pi}_i}$$

for $i = 1, \ldots, k$.

Bootstrap samples of fixed size S are then generated by repeatedly drawing individuals from this multinomial population. For every bootstrap sample, the penalized NPMLE of N is computed. A confidence interval is then constructed using Efron's percentile method (Efron 1981). Note that in this sampling the sample size S is fixed; in Poisson sampling it would be random. This scheme therefore better matches data that were collected by sampling a fixed number of individuals rather than using Poisson sampling.

We note that we are able to do this semiparametric bootstrap because we have fit a model to the full sampling mechanism, not just developed an estimator for N. We illustrate this method in Section 5 and briefly compare it with other bootstrap approaches in Section 6.

4. REAL DATA ANALYSIS AND SIMULATION STUDY

In this section we investigate the behavior of the two penalized NPMLE estimators, $\hat{N}_u$ from ℓ_1 and $\hat{N}_{WL}$ from ℓ_3 under the adaptive penalizing function using $\mu = \hat{N}_{c_0}/D - 1$, and compare them with the lower-bound estimator $\hat{N}_{c_0}$ (Chao 1984), two coverage-based estimators $\hat{N}_{c_1}$ and $\hat{N}_{c_2}$ (Chao and Lee 1992), the coverage-duplication estimator $\hat{N}_{c_3}$ (Chao and Bunge 2002), the jackknife estimator $\hat{N}_J$ (Burnham and Overton 1979), and the unconditional NPMLE $\hat{N}_{UNP}$ (Norris and Pollock 1998).

Our primary goal is to find which estimators are the most widely reliable for inference in this difficult model. We start

with a case study demonstrating that some of the estimators are quite sensitive to the choice of τ and others are not. This information then provides some guidance for the selection of τ in our second study, in which we compare the estimators in simulations from a wide range of models. From this study, we draw conclusions about which methods are most reliable.

4.1 Real Data 1: Fisher's Butterfly Data

We first investigate the choice of τ. In the process, we also illustrate the relationship between $\hat{N}_u$ and $\hat{N}_{UNP}$. The famous Malayan butterfly data in Table 1 was originally published by Fisher, Corbet, and Williams (1943) and was recently analyzed by Chao and Bunge (2002). As suggested by Chao et al. (1993) and Chao et al. (2000), we split the sample into "rare" (n_1–n_τ) and "abundant" (beyond n_τ) species. In practice, one can first obtain an estimate, $\hat{N}_\tau$, based completely on the "rare" species data, then calculate the final estimate by $\hat{N} = \hat{N}_\tau + \sum_{j=\tau+1}^{t} n_j$. In the following analyses, the estimates are obtained in this way if τ is specified. Table 2 presents the results from various estimators.

For the unpenalized conditional NPMLE solution, the boundary problem (i.e., $\hat{P}$ contains a 0 or tiny component) occurred at $\tau = 10, 11, 12, 13, 14, 20, 24$, and, consequently, $\hat{N}_{CNP}$ gave extremely large numbers at these τ's (not reported here). As a remedial measure, we fitted a model with one less component in this situation, which often shifts the smallest components rightward and yields more conservative (smaller) estimates. Here we denote the modified version as $\hat{N}_{mCNP}$. By Corollary 1, $\hat{N}_{CNP}$ must be no less than $\hat{N}_u$, $\hat{N}_{UNP}$, and $\hat{N}_{WL}$. Even after modification, $\hat{N}_{mCNP}$ still ended up greater than $\hat{N}_u$ or $\hat{N}_{UNP}$, except for $\tau = 11$.

We found that the approximation of $\hat{N}_u$ to $\hat{N}_{UNP}$ is very satisfactory with a great saving of computing time. As predicted earlier, $\hat{N}_u$ is identical to $\hat{N}_{UNP}$ for all τ's except for $\tau = 12$, where it differs by 1 (723 vs. 722). Chao and Bunge (2002) reported that the approach of Norris and Pollock (1998) took up to an hour for a single point estimate, whereas with our codes written in JAVA, on average it took a few seconds to finish computing such that $|\theta^{(t)} - \theta^{(t-1)}| < 1/D_\tau$, where $D_\tau = \sum_{j=1}^{\tau} n_j$.

As for $\hat{N}_{WL}$, the penalty appears to be an effective way to solve the boundary problem and stabilize the estimator. The estimates at different τ's are fairly consistent except for $\tau = 11$, where the estimate was pulled up to 739. This is probably related to the failure of $\hat{N}_{CNP}$ at $\tau = 11$ due to the boundary problem. In contrast, $\hat{N}_{mCNP}$ gave 711 at $\tau = 11$, tending to be relatively conservative, even smaller than the lower-bound estimator, $\hat{N}_{c_0} = 714$.

Relative insensitivity to the cutoff τ is a feature of all of the NPMLE-based estimators listed in Table 2. This allows one to use the relatively "rare" species to obtain estimates without concern about the choice of τ. In contrast, the coverage-based estimators $\hat{N}_{c_1}$, $\hat{N}_{c_2}$, and $\hat{N}_{c_3}$ all increased systematically with τ, and so its choice plays an important role in estimation properties. The estimator of N can change dramatically in τ, as we show in the genomic example of next section.

Table 2. Comparing Estimates for Fisher's Malayan Butterfly Data (Fisher et al. 1943)

τ	$\hat{N}_{mCNP}$	$\hat{N}_u$	$\hat{N}_{UNP}$	$\hat{N}_{WL}$	$\hat{N}_{c_1}$	$\hat{N}_{c_2}$	$\hat{N}_{c_3}$
10	716	715	715	716	712	737	757
11	711	715	715	739	714	740	761
12	729	723	722	730	716	744	765
13	731	724	724	728	717	746	768
14	726	723	723	724	719	750	774
15	724	722	722	724	721	753	777
20	721	718	718	725	723	772	802
24	721	719	719	722	737	779	810

NOTE: τ is the cutoff defining "rare" and "abundant" species. $\hat{N}_{mCNP}$ is the modified conditional NPMLE estimator without penalty. $\hat{N}_u$ is the approximatant to $\hat{N}_{UNP}$ from ℓ_2 at $\gamma_2 = .5$. $\hat{N}_{UNP}$, $\hat{N}_{c_1}$, $\hat{N}_{c_2}$, and $\hat{N}_{c_3}$ were reported by Chao and Bunge (2002, table 2, p. 535). $\hat{N}_{WL}$ is the penalized NPMLE estimator from ℓ_3 under the adaptive quadratic penalizing function.

4.2 Simulation Study

Under the Poisson mixture model, we assume that the mean parameter Λ of each Poisson observation is generated from a latent distribution $Q(\lambda)$. In the simulation, we can first generate $\lambda_1, \lambda_2, \ldots, \lambda_N$ from $Q(\lambda)$, then generate corresponding Poisson observations using these λ's as the mean parameter values. To compare the performance of these estimators, we construct a factorial design concerning three factors that characterize Q, namely *sampling depth* (d), *coefficient of variation* (CV), and *skewness* (ρ). The first factor, sampling depth, is defined as the probability of capturing a randomly selected species, that is, $d = 1 - f(0; Q) = \frac{1}{1+\theta(Q)}$. This factor measures the sampling intensity relative to the total species number N. The CV factor, introduced by Chao and Lee (1992), defined as $CV = \frac{\sqrt{\text{var}(\lambda)}}{E(\lambda)}$, accounts for the heterogeneity in the species abundance distribution. Skewness, defined as $\rho = \frac{E(\lambda - E(\lambda))^3}{[E(\lambda - E(\lambda))^2]^{3/2}}$, is another factor that affects the performance of an estimator (Haas and Stokes 1998) but has not attracted much attention in the literature. Because the species data usually is right-skewed, we only simulate data with positive skewness. The goal of the following simulations is to compare the relative performance of these estimators with respect to coverage, bias, variability, and robustness in different situations.

Simulation Design I. We first simulated data from a population with $N = 1{,}000$ and $Q = \text{gamma}(\alpha, \beta)$ according to a two-way factorial design in Table 3. Because a gamma distribution has two parameters, only the first two factors, sampling depth and CV, were controlled for in this design. Note that in the Poisson–gamma model, $d = 1 - \frac{\beta^\alpha}{(1+\beta)^\alpha}$, $CV = \sqrt{\text{var}(\Lambda)}/E(\Lambda) = 1/\sqrt{\alpha}$, and $\rho = 2/\sqrt{\alpha}$. The scale parameter β is not involved in CV or ρ. The α and β values corresponding to each crossover of $d \in \{.2, .3, .4, .5, .7, .9\}$ by $CV \in \{1.41, 1.00, .71, .50\}$ are tabulated in Table 3. The expected number of sampled individuals $E(S) = E(\sum_{i=1}^{N} X_i)$ at each situation is also calculated in Table 3, to indicate how the actual sampling effort varies with heterogeneity at each targeted sampling depth d.

Table 1. Malayan Butterfly Data of Fisher et al. (1943)

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	>24	D
118	74	44	24	29	22	20	19	20	15	12	14	6	12	6	9	9	6	10	10	11	5	3	3	119	620

Table 3. Design of Simulation I

	$d = .2$ ($\theta = 4$)		$d = .3$ ($\theta = 3.3$)		$d = .4$ ($\theta = 1.5$)		$d = .5$ ($\theta = 1$)		$d = .75$ ($\theta = .33$)		$d = .9$ ($\theta = .11$)	
α, CV	β	$E(S)$	β	$E(S)$	β	$E(S)$	β	$E(S)$	β	$E(S)$	β	$E(S)$
$\alpha = 4$ ($CV = .50$)	17.43	229	10.72	373	7.34	545	5.29	756	2.41	1,660	1.28	3,125
$\alpha = 2$ ($CV = .71$)	8.47	236	5.12	391	3.44	581	2.41	830	1	2,000	.46	4,348
$\alpha = 1$ ($CV = 1.00$)	4	250	2.33	429	1.5	667	1	1,000	.33	3,030	.11	9,091
$\alpha = 1/2$ ($CV = 1.41$)	1.78	281	.96	521	.56	893	.33	1,515	.067	7,463	.01	50,000

Simulation Design II. To investigate how the skewness ρ affects estimation, we consider the simplest case, a two-point mixture, that is, $Q = \pi\Delta_{\lambda_1} + (1 - \pi)\Delta_{\lambda_2}$. Even in this simple model, however, it is still difficult to obtain the exact Q determined by the specified values of CV, d, and ρ, because three nonlinear equations must be solved simultaneously. We thus obtained an approximate Q by evaluating these parameters on a three-way grid of $\lambda_1 \times \lambda_2 \times \pi$ for each crossing of $CV \times d \times \rho = \{1, 1.5\} \times \{.3, .5, .75\} \times \{1.5, 2.5, 4.69\}$. Table 4 presents the three-way factorial design with both the targeted and actual obtained values of CV, d, and ρ.

For every simulation setting, we generated 200 datasets, using the cutoff $\tau = 10$ for $\hat{N}_u, \hat{N}_{WL}, \hat{N}_{c_1}$, and $\hat{N}_{c_2}$. We excluded the estimator $\hat{N}_{c_3}$ from this study for two reasons. First, it could fail due to a negative or wildly large estimate of coverage-duplication coefficient when d is small. For example, in Simulation I we observed 31 cases out of 200 that had negative estimates at $\alpha = .5$ and $d = .2$, and 4 out of 200 cases that had negative estimates at $\alpha = .5$ and $d = .3$. Second, the estimator $\hat{N}_{c_3}$ was built on the Poisson–gamma assumption. In Simulation II, where the true mixing distribution was discrete, we found that this estimator behaved wildly bad (results not shown here) even when d is relatively large.

We determined the order of the jackknife estimator $\hat{N}_J$ by the sequential testing procedure of Burnham and Overton (1978). If the determined order exceeded 5 (which frequently occurred), then the fifth-order estimate was taken to avoid outliers. For $\hat{N}_u$, extreme values were often observed, especially when the true θ was relatively large (e.g., $\theta > .33$ corresponding to $d < .75$). If $\hat{\theta}$ kept increasing and exceeded 5 during the iterations, then we terminated the algorithm and reported the estimate based on current $\hat{\theta}$. Otherwise, the variability of $\hat{N}_u$ could have been even larger than that reported here. (The same problem and a similar remedial measure were also reported in Norris and Pollock 1998.)

The summary statistics of the sampling distributions of these estimators based on 200 Monte Carlo samples are presented in Tables 7 and 8 for Simulation I and in Table 9 for Simulation II. These statistics include sample mean ($\hat{N}$), median ($\hat{M}$), standard deviation (s), root mean squared error (RMSE), mean absolute error (MAE), and central 95% interval of the sampling distribution. One sign of the difficulty of this estimation problem is that in many cases this central interval of the sampling distribution did not cover the true value of N used in the simulation. The only estimator to do this reliably, $\hat{N}_u$, did so because of its wide variability.

In general, the nonparametric estimators all improved as the sampling depth d increased. This is not a surprise, because a larger d implies that more information about Q is present in the data and there are fewer unseen species to estimate. In the Poisson–gamma model, when d was fixed, bias increased with CV, whereas variation remained relatively stable. On the other hand, when CV was fixed, both bias and variation decreased as d increased, but the bias decreased at a much faster pace and overwhelmingly determined the behavior of these estimators except for $\hat{N}_u$, where variability, not bias, was more important.

For the discrete Q case, the results in Table 9 show considerable complication in how the three factors impacted these estimators. With control of skewness ρ and sampling depth d, it appears that both $\hat{N}_{c_1}$ and $\hat{N}_{c_2}$ behaved worse for larger CV,

Table 4. Design of Simulation II

		[a]$d \approx .3$			$d \approx .5$			$d \approx .75$		
		[a]$\rho \approx 1.5$	$\rho \approx 2.5$	$\rho \approx 4.69$	$\rho \approx 1.5$	$\rho = 2.5$	$\rho = 4.69$	$\rho = 1.5$	$\rho = 2.5$	$\rho = 4.69$
[a]$CV \approx 1$	[c]λ_1	.2	.3	.3	.5	.6	.6	1.2	1.3	1.3
	[c]λ_2	1.3	1.8	2.2	3.0	3.6	4.4	6.7	7.7	10
	[c]π	.8	.89	.96	.8	.89	.96	.8	.89	.96
	[b]d	.29	.32	.28	.5	.51	.47	.76	.76	.74
	[b]CV	1.05	1.01	.99	1	1	.99	.96	1	1
	[b]ρ	1.5	2.49	4.69	1.5	2.49	4.69	1.5	2.49	4.69
$CV \approx 1.5$	λ_1	.2	.2	.3	.5	.5	.6	1.2	1.3	1.3
	λ_2	3.1	2.2	3.5	7.9	6.7	7.2	19.2	14.5	15.7
	π	.8	.89	.96	.8	.89	.96	.8	.89	.96
	d	.34	.26	.29	.51	.51	.47	.76	.76	.74
	CV	1.49	1.49	1.5	1.5	1.5	1.5	1.5	1.5	1.5
	ρ	1.5	2.49	4.69	1.5	2.49	4.69	1.5	2.49	4.69

[a]The targeted values of CV, d, and ρ.

[b]The actual values of CV, d, and ρ used in the simulation.

[c]The parameters in Q, that is, $Q = \pi\Delta_{\lambda_1} + (1 - \pi)\Delta_{\lambda_2}$.

Table 5. An Example

j	1	2	3	4	5	6	7	8
n_j	196	83	59	30	18	2	7	7

whereas the other four estimators, $\hat{N}_u, \hat{N}_{WL}, \hat{N}_{c_0}$, and $\hat{N}_J$, often showed opposite behavior. When d and CV were controlled, larger skewness pushed more mass onto the smallest component. At $\rho = 4.69$, when the mixing distribution was close to being degenerate at the smallest component, $\hat{N}_{c_1}, \hat{N}_{c_2}$, and $\hat{N}_J$ all produced wildly positive bias, whereas $\hat{N}_u, \hat{N}_{WL}$, and $\hat{N}_{c_0}$ were much less vulnerable to this change.

To compare the relative performance of these estimators, we ranked the six estimators in terms of MAE, median bias, and coverage in Table 10 for Simulation I and Table 11 for Simulation II. We used the median here because the sampling distributions of $\hat{N}_{c_0}, \hat{N}_u, \hat{N}_{WL}$, and $\hat{N}_J$ are all skewed strongly to the right. For the MAE or median-bias criterion, a "+" sign indicates that the estimator was one of the best two among the six, whereas a "−" indicates the worst two. In the "95% coverage column," a "+" indicates that the central 95% sampling interval covered the true $N = 1{,}000$, and "−" indicates that it did not.

One obvious conclusion is that there is no uniform winner over all criteria and all models, and so an overall conclusion requires setting some priorities. Here is how we set ours. The goal is to find estimators that would work most reliably across the widest range of situations. Due to the severe bias–variance trade-off in this model, we believe that statistical reliability is best measured by the estimator's rate of success using the 95% coverage criterion. Our reasoning is that: if this property fails, then it is clear that "point estimator plus/minus two standard errors" gives a wildly misleading sense of confidence about the location of the true value of N.

This leads us to the conclusion that the jackknife estimator worked well in Simulation I but not in Simulation II, and the coverage estimator $\hat{N}_{c_0}$ worked well in Simulation II but not in Simulation I. We conclude that only $\hat{N}_{WL}$ and $\hat{N}_u$ succeeded well in both simulations.

Comparing $\hat{N}_{WL}$ and $\hat{N}_u$, we see that the penalized estimator did succeed in reducing the MAE of the former, although $\hat{N}_u$ had better bias properties. We note, however, that in our implementation of $\hat{N}_u$ we prevented extreme large values by terminating the algorithm if $\hat{\theta}$ kept increasing and exceeded 5 in our simulations, so in all honesty, our version was actually a penalized estimator. Otherwise, its MAE would have been a very large number. In our judgment, then, $\hat{N}_{WL}$ is the overall winner.

Our foregoing conclusion is further supported by the rankings under the MAE or RMSE criterion. We note that $\hat{N}_{WL}$, $\hat{N}_J$, and $\hat{N}_{c_2}$ all performed competitively well under the MAE criterion in Simulation I (Table 10). However, as $\hat{N}_{WL}$ continued to be a winner in Simulation II (Table 11), $\hat{N}_{c_2}$ and $\hat{N}_J$ became the two least recommendable among the six estimators. These results suggest that $\hat{N}_{WL}$ tends to be more robust than those other two estimators.

To illustrate the instability of the unconditional NPMLE, Table 5 presents an example that was simulated from the Poisson–gamma model at $\alpha = .5$ and $\beta = .56$. The estimate from $\hat{N}_u$ was 2,417, corresponding to $\hat{\theta} = 5.01$. (If we had not set the upper limit on θ, then $\hat{N}_u$ would have been 2,706.) The true value of θ was 1.5. We can compare this with $\hat{N}_{UNP}$ by computing the profile (maximized over Q for fixed N) of the full likelihood. The results are presented in Table 6 (where the likelihood values have a common constant removed).

Table 6 suggests that the likelihood is unimodal, with the true unconditional NPMLE $\hat{N}_{UNP}$ somewhere between 2,700 and 2,730. However, it is extremely flat, with the likelihood at the true N of 1,000 being nearly the same as the likelihood at 3,000. In such flat cases, the penalty has a major impact, and, correspondingly, the estimate from $\hat{N}_{WL}$ was 847.

5. REAL DATA 2: EXPRESSED SEQUENCE TAG DATA

Our research on this topic was motivated by an application in genomics. An expressed sequence tag (EST) is a fragment of a gene sequence that can be considered a label of the corresponding gene in the expressed form of mRNA transcript (Adams et al. 1991). An EST dataset constitutes a random sample from a population, the mRNA transcript pool containing a finite but unknown number, N, of genes. The sampled ESTs can be classified by genes of origin, where here "gene" plays the role of "species." We wish to estimate the number of genes represented in the mRNA transcript pool. The sequence data can be summarized as the frequency, n_j (Table 12), where n_j represents the number of genes with j ESTs existing in the sample. We next use an EST dataset to demonstrate how various methods behave in a real problem with very large counts. We also examine the role of τ and the sampling behavior of the estimators in this important biological problem.

The *Arabidopsis thaliana* root EST data was obtained from the NCBI dbEST at *http://www.ncbi.nih.gov/dbEST* (Asamizu, Nakamura, Sato, and Tabata 2000). (Interested readers can find more relevant results in Wang, Lindsay, Leebens-Mack, Cui, Wall, Webb, and de Pamphilis 2004.) The 6,043 total EST tags were classified into 3,126 distinct genes. The most abundant gene had 49 ESTs; in contrast, 2,187 of the 3,126 genes sampled had only one EST (Table 12).

Table 13 compares $\hat{N}_u$ and $\hat{N}_{WL}$ with $\hat{N}_{c_1}$ and $\hat{N}_{c_2}$ at different τ's. The two NPMLE estimators $\hat{N}_u$ and $\hat{N}_{WL}$ suggest, consistently at different τ's, that about 9,000 genes are expressed in the root tissue of *Arabidopsis thaliana*. In contrast, $\hat{N}_{c_2}$ is doubled as τ changes from 10 to 24, whereas $\hat{N}_{c_1}$ increases by 3,574. Clearly, $\hat{N}_{c_2}$ is not recommended due to instability in the highly skewed scenario. For comparison, the jackknife estimator $\hat{N}_J$ and lower-bound estimator $\hat{N}_{c_0}$ give 9,923 and 8,006.

To assess the accuracy and variability of the estimators in this setting, we used the bootstrap simulation method of Section 3. We generated 200 bootstrap samples based on our estimated

Table 6. Profile Nonparametric Likelihood for the Data in Table 5

n	1,000	1,500	2,000	2,500	2,700	2,720	2,730	3,000
$\log L(N, \hat{Q}_N)$	−.33	−.22	−.21	−.20239	−.202234	−.2022336	−.2022342	−.2024

Table 7. Result of Simulation I, N = 1,000

		$d = .2$						$d = .3$							$d = .4$						
CV	$\hat{N}$	[a]$\bar{\hat{N}}$	[b]$\hat{M}$	[c]s	[d]*RMSE*	[e]*MAE*	[f]*95%*	$\hat{N}$	$\bar{\hat{N}}$	$\hat{M}$	s	*RMSE*	*MAE*	*95%*	$\hat{N}$	$\bar{\hat{N}}$	$\hat{M}$	s	*RMSE*	*MAE*	*95%*
.5	$\hat{N}_u$	931	808	347	354	280	(604, 1,932)	$\hat{N}_u$	1,153	927	512	535	337	(708, 2,506)	$\hat{N}_u$	1,164	969	549	573	307	(757, 1,967)
	$\underline{\hat{N}_{WL}}$	937	893	217	226	182	(638, 1,441)	$\underline{\hat{N}_{WL}}$	945	915	160	169	141	(708, 1,294)	$\underline{\hat{N}_{WL}}$	978	958	150	151	122	(758, 1,308)
	$\hat{N}_{c_0}$	878	843	198	232	197	(592, 1,341)	$\hat{N}_{c_0}$	869	862	113	173	150	(688, 1,077)	$\hat{N}_{c_0}$	888	883	89	143	124	(721, 1,085)
	$\hat{N}_{c_1}$	886	852	190	221	184	(602, 1,343)	$\hat{N}_{c_1}$	872	864	107	167	145	(700, 1,089)	$\hat{N}_{c_1}$	890	886	82	138	121	(743, 1,071)
	$\hat{N}_{c_2}$	929	876	241	252	196	(618, 1,447)	$\hat{N}_{c_2}$	904	881	136	167	140	(704, 1,172)	$\hat{N}_{c_2}$	922	909	105	131	109	(752, 1,167)
	$\hat{N}_J$	675	667	83	336	325	(513, 855)	$\hat{N}_J$	963	958	102	109	87	(768, 1,174)	$\hat{N}_J$	1,113	1,112	137	178	139	(879, 1,419)
.71	$\hat{N}_u$	943	813	376	380	301	(571, 1,770)	$\hat{N}_u$	1,046	823	536	538	355	(627, 3,469)	$\hat{N}_u$	1,115	911	505	518	308	(733, 2,760)
	$\underline{\hat{N}_{WL}}$	812	790	166	252	222	(572, 1,214)	$\underline{\hat{N}_{WL}}$	845	816	151	217	192	(627, 1,190)	$\underline{\hat{N}_{WL}}$	917	886	123	149	127	(735, 1,170)
	$\hat{N}_{c_0}$	755	735	138	252	257	(546, 1,128)	$\hat{N}_{c_0}$	768	761	94	251	236	(610, 975)	$\hat{N}_{c_0}$	809	799	65	202	191	(697, 961)
	$\hat{N}_{c_1}$	762	745	135	274	250	(552, 1,124)	$\hat{N}_{c_1}$	771	761	85	244	230	(620, 981)	$\hat{N}_{c_1}$	814	809	63	197	187	(706, 946)
	$\hat{N}_{c_2}$	806	779	177	263	232	(559, 1,244)	$\hat{N}_{c_2}$	810	797	116	223	203	(620, 1,108)	$\hat{N}_{c_2}$	861	846	88	165	148	(720, 1,076)
	$\hat{N}_J$	674	662	78	336	326	(538, 825)	$\hat{N}_J$	921	915	109	134	110	(714, 1,150)	$\hat{N}_J$	1,032	1,008	122	127	97	(861, 1,314)
1	$\hat{N}_u$	817	693	362	406	352	(446, 1,872)	$\hat{N}_u$	997	802	479	479	367	(549, 2,316)	$\hat{N}_u$	1,052	815	583	585	366	(647, 2,858)
	$\underline{\hat{N}_{WL}}$	665	665	128	359	336	(452, 906)	$\underline{\hat{N}_{WL}}$	687	689	84	325	313	(543, 890)	$\underline{\hat{N}_{WL}}$	825	806	122	214	192	(655, 1,062)
	$\hat{N}_{c_0}$	607	604	98	406	393	(431, 813)	$\hat{N}_{c_0}$	668	668	78	347	333	(538, 866)	$\hat{N}_{c_0}$	706	703	54	300	297	(618, 809)
	$\hat{N}_{c_1}$	613	608	97	400	387	(435, 828)	$\hat{N}_{c_1}$	674	671	74	335	326	(552, 851)	$\hat{N}_{c_1}$	720	713	55	286	280	(636, 839)
	$\hat{N}_{c_2}$	658	650	130	366	344	(448, 948)	$\hat{N}_{c_2}$	738	725	108	284	265	(570, 988)	$\hat{N}_{c_2}$	793	781	86	224	209	(671, 983)
	$\hat{N}_J$	658	655	87	354	342	(498, 840)	$\hat{N}_J$	849	840	130	199	175	(656, 1,116)	$\hat{N}_J$	918	885	132	156	134	(736, 1,275)
1.41	[g]$\hat{N}_u$	746	591	360	441	401	(399, 1,603)	$\hat{N}_u$	761	634	370	441	388	(483, 1,922)	$\hat{N}_u$	850	700	430	456	360	(546, 2,417)
	[h]$\underline{\hat{N}_{WL}}$	546	537	104	467	454	(403, 767)	$\underline{\hat{N}_{WL}}$	588	585	63	417	411	(480, 731)	$\underline{\hat{N}_{WL}}$	702	688	94	313	298	(550, 917)
	[i]$\hat{N}_{c_0}$	478	465	71	530	524	(373, 623)	$\hat{N}_{c_0}$	538	533	46	466	462	(453, 632)	$\hat{N}_{c_0}$	605	605	43	398	395	(527, 693)
	[j]$\hat{N}_{c_1}$	484	474	73	523	516	(386, 665)	$\hat{N}_{c_1}$	561	560	50	443	439	(476, 683)	$\hat{N}_{c_1}$	624	623	40	379	376	(551, 712)
	[k]$\hat{N}_{c_2}$	552	524	124	467	448	(403, 918)	$\hat{N}_{c_2}$	648	642	83	363	352	(523, 838)	$\hat{N}_{c_2}$	703	701	61	304	297	(600, 839)
	[l]$\hat{N}_J$	594	582	100	419	406	(449, 849)	$\hat{N}_J$	689	670	98	327	312	(545, 934)	$\hat{N}_J$	763	752	107	261	244	(624, 1,077)

[a]Sample mean.
[b]Median.
[c]Standard deviation.
[d]Root mean square.
[e]Mean absolute error ($\sum_{i=1}^{200} |\hat{N}_i - 1{,}000|/200$).
[f]Central 95% percentile from 200 point estimates of Monte Carlo samples.
[g]$\hat{N}_u$ is the proposed approximant to $\hat{N}_{UNP}$.
[h]$\underline{\hat{N}_{WL}}$ is the proposed estimator based on quadratic penalizing function.
[i]$\hat{N}_{c_0}$ is the lower bound estimator of Chao (1984).
[j]$\hat{N}_{c_1}$.
[k]$\hat{N}_{c_2}$ are two coverage CV-based estimators of Chao and Lee (1992).
[l]$\hat{N}_J$ is the jackknife estimator of Burnham and Overton (1978, 1979).

Table 8. Result of Simulation I, N = 1,000 (continued)

		$d = .5$							$d = .75$							$d = .9$						
CV	$\hat{N}$	$\bar{\hat{N}}$	$\hat{M}$	*s*	*RMSE*	*MAE*	*95%*	$\hat{N}$	$\bar{\hat{N}}$	$\hat{M}$	*s*	*RMSE*	*MAE*	*95%*	$\hat{N}$	$\bar{\hat{N}}$	$\hat{M}$	*s*	*RMSE*	*MAE*	*95%*
.5	$\hat{N}_u$	1,113	979	441	454	217	(810, 2,458)	$\hat{N}_u$	1,041	992	191	195	89	(912, 1,323)	$\hat{N}_u$	1,005	991	69	69	31	(954, 1,132)
	$\underline{\hat{N}_{WL}}$	1,001	983	120	120	97	(811, 1,277)	$\underline{\hat{N}_{WL}}$	1,029	1,008	85	90	67	(912, 1,214)	$\underline{\hat{N}_{WL}}$	1,015	1,000	49	52	32	(963, 1,148)
	$\hat{N}_{c_0}$	904	903	63	115	99	(783, 1,030)	$\hat{N}_{c_0}$	953	951	33	58	50	(898, 1,019)	$\hat{N}_{c_0}$	980	980	16	26	22	(951, 1,016)
	$\hat{N}_{c_1}$	911	905	62	109	94	(802, 1,025)	$\hat{N}_{c_1}$	961	958	32	50	42	(908, 1,020)	$\hat{N}_{c_1}$	981	981	14	23	22	(957, 1,008)
	$\hat{N}_{c_2}$	944	936	80	98	81	(803, 1,098)	$\hat{N}_{c_2}$	984	981	39	42	35	(923, 1,060)	$\hat{N}_{c_2}$	987	986	15	20	16	(961, 1,017)
	$\hat{N}_J$	1,142	1,125	119	186	150	(949, 1,429)	$\hat{N}_J$	1,125	1,122	59	139	126	(1,036, 1,204)	$\hat{N}_J$	1,075	1,074	17	77	75	(1,046, 1,105)
.71	$\hat{N}_u$	1,027	911	404	404	208	(761, 2,165)	$\hat{N}_u$	1,028	962	201	203	111	(886, 1,589)	$\hat{N}_u$	987	975	48	50	36	(939, 1,114)
	$\underline{\hat{N}_{WL}}$	931	921	107	127	106	(963, 1,149)	$\underline{\hat{N}_{WL}}$	1,006	987	94	94	73	(890, 1,224)	$\underline{\hat{N}_{WL}}$	1,004	992	48	48	37	(944, 1,125)
	$\hat{N}_{c_0}$	831	831	54	178	170	(731, 939)	$\hat{N}_{c_0}$	919	920	30	86	81	(865, 982)	$\hat{N}_{c_0}$	966	966	16	38	34	(933, 998)
	$\hat{N}_{c_1}$	843	841	54	167	157	(735, 943)	$\hat{N}_{c_1}$	928	929	26	76	72	(879, 979)	$\hat{N}_{c_1}$	959	959	13	43	41	(932, 981)
	$\hat{N}_{c_2}$	891	881	73	131	114	(754, 1,039)	$\hat{N}_{c_2}$	957	957	31	53	45	(899, 1,019)	$\hat{N}_{c_2}$	964	963	13	39	36	(937, 990)
	$\hat{N}_J$	1,040	1,035	104	111	80	(868, 1,247)	$\hat{N}_J$	1,063	1,065	60	87	69	(969, 1,056)	$\hat{N}_J$	1,035	1,036	17	39	38	(1,001, 1,064)
1	$\hat{N}_u$	939	822	405	408	285	(695, 2,421)	$\hat{N}_u$	958	922	130	136	108	(834, 1,331)	$\hat{N}_u$	970	959	41	52	45	(929, 1,087)
	$\underline{\hat{N}_{WL}}$	876	841	133	182	161	(699, 1,189)	$\underline{\hat{N}_{WL}}$	958	953	79	90	75	(842, 1,141)	$\underline{\hat{N}_{WL}}$	994	979	51	51	41	(935, 1,130)
	$\hat{N}_{c_0}$	751	750	44	253	249	(674, 835)	$\hat{N}_{c_0}$	876	875	30	127	124	(821, 939)	$\hat{N}_{c_0}$	952	952	14	50	48	(927, 980)
	$\hat{N}_{c_1}$	777	774	44	228	223	(697, 869)	$\hat{N}_{c_1}$	871	870	23	131	129	(823, 911)	$\hat{N}_{c_1}$	939	939	10	62	61	(918, 957)
	$\hat{N}_{c_2}$	851	846	64	163	150	(735, 985)	$\hat{N}_{c_2}$	892	891	27	112	108	(839, 938)	$\hat{N}_{c_2}$	942	942	11	59	58	(920, 960)
	$\hat{N}_J$	933	932	100	120	100	(799, 1,217)	$\hat{N}_J$	984	982	71	73	47	(890, 1,127)	$\hat{N}_J$	995	992	33	33	16	(967, 1,033)
1.41	$\hat{N}_u$	868	752	324	349	291	(653, 1,981)	$\hat{N}_u$	900	863	110	148	132	(789, 1,226)	$\hat{N}_u$	952	937	46	66	61	(911, 1,085)
	$\underline{\hat{N}_{WL}}$	788	771	103	236	214	(639, 1,016)	$\underline{\hat{N}_{WL}}$	917	904	83	118	102	(804, 1,084)	$\underline{\hat{N}_{WL}}$	964	954	41	55	48	(914, 1,064)
	$\hat{N}_{c_0}$	670	666	38	333	330	(604, 754)	$\hat{N}_{c_0}$	837	837	27	166	163	(784, 891)	$\hat{N}_{c_0}$	935	934	16	67	65	(908, 974)
	$\hat{N}_{c_1}$	674	669	31	329	326	(623, 741)	$\hat{N}_{c_1}$	818	818	20	183	181	(779, 854)	$\hat{N}_{c_1}$	924	924	11	77	76	(902, 947)
	$\hat{N}_{c_2}$	724	723	42	280	276	(657, 813)	$\hat{N}_{c_2}$	830	829	22	172	170	(787, 871)	$\hat{N}_{c_2}$	927	927	12	74	73	(904, 954)
	$\hat{N}_J$	815	779	110	216	205	(698, 1,174)	$\hat{N}_J$	916	907	83	118	105	(826, 1,206)	$\hat{N}_J$	967	951	52	62	51	(926, 1,172)

NOTE: Footnotes are the same as in Table 7.

Table 9. Result of Simulation II, N = 1,000

		$\rho = 1.5$							$\rho = 2.5$							$\rho = 4.7$						
$d \approx$	$CV \approx$	$\hat{N}$	$\bar{\hat{N}}$	$\hat{M}$	s	RMSE	MAE	95%	$\hat{N}$	$\bar{\hat{N}}$	$\hat{M}$	s	RMSE	MAE	95%	$\hat{N}$	$\bar{\hat{N}}$	$\hat{M}$	s	RMSE	MAE	95%
.3	1	$\hat{N}_u$	1,277	936	693	746	545	(582, 2,734)	$\hat{N}_u$	1,454	1,169	699	834	569	(706, 3,008)	$\hat{N}_u$	1,312	1,115	547	630	407	(757, 2,769)
		$\hat{N}_{WL}$	790	764	141	253	229	(597, 1,144)	$\hat{N}_{WL}$	968	953	162	166	136	(704, 1,297)	$\hat{N}_{WL}$	1,027	1,018	176	178	140	(745, 1,399)
		$\hat{N}_{c_0}$	651	639	78	359	349	(528, 830)	$\hat{N}_{c_0}$	788	774	93	232	214	(641, 1,016)	$\hat{N}_{c_0}$	879	880	118	169	141	(671, 1,140)
		$\hat{N}_{c_1}$	636	628	70	372	364	(523, 805)	$\hat{N}_{c_1}$	786	777	90	233	215	(647, 974)	$\hat{N}_{c_1}$	915	912	126	152	123	(681, 1,149)
		$\hat{N}_{c_2}$	694	683	101	322	308	(544, 925)	$\hat{N}_{c_2}$	918	892	151	172	148	(693, 1,249)	$\hat{N}_{c_2}$	1,150	1,128	243	285	217	(732, 1,658)
		$\hat{N}_J$	855	829	131	196	168	(658, 1,104)	$\hat{N}_J$	1,034	1,031	139	143	118	(788, 1,292)	$\hat{N}_J$	1,037	1,061	121	127	106	(820, 1,243)
	1.5	$\hat{N}_u$	1,337	1,021	719	794	569	(574, 2,810)	$\hat{N}_u$	1,313	1,102	680	748	549	(554, 2,744)	$\hat{N}_u$	1,216	1,077	433	484	298	(755, 2,307)
		$\hat{N}_{WL}$	779	767	121	253	227	(570, 1,003)	$\hat{N}_{WL}$	787	769	159	266	231	(534, 1,088)	$\hat{N}_{WL}$	1,034	1,009	171	175	137	(750, 1,416)
		$\hat{N}_{c_0}$	567	563	52	437	433	(478, 673)	$\hat{N}_{c_0}$	607	603	89	404	394	(471, 794)	$\hat{N}_{c_0}$	896	884	126	163	139	(673, 1,170)
		$\hat{N}_{c_1}$	495	495	33	507	505	(440, 563)	$\hat{N}_{c_1}$	569	564	66	437	431	(449, 721)	$\hat{N}_{c_1}$	963	945	130	135	112	(740, 1,229)
		$\hat{N}_{c_2}$	542	539	45	462	458	(470, 639)	$\hat{N}_{c_2}$	690	678	111	330	312	(515, 952)	$\hat{N}_{c_2}$	1,490	1,441	310	581	494	(1,008, 2,148)
		$\hat{N}_J$	819	792	144	232	199	(599, 1,116)	$\hat{N}_J$	818	805	131	225	190	(600, 1,059)	$\hat{N}_J$	1,045	1,051	117	126	102	(822, 1,263)
.5	1	$\hat{N}_u$	1,195	1,051	482	521	261	(831, 2,753)	$\hat{N}_u$	1,136	1,037	329	356	183	(870, 2,124)	$\hat{N}_u$	1,155	1,047	416	444	202	(875, 2,600)
		$\hat{N}_{WL}$	1,037	1,027	128	133	105	(834, 1,318)	$\hat{N}_{WL}$	1,061	1,039	128	141	104	(875, 1,342)	$\hat{N}_{WL}$	1,066	1,049	132	148	110	(877, 1,383)
		$\hat{N}_{c_0}$	863	865	59	150	138	(744, 973)	$\hat{N}_{c_0}$	945	940	73	91	76	(819, 1,095)	$\hat{N}_{c_0}$	984	984	90	92	73	(838, 1,153)
		$\hat{N}_{c_1}$	843	844	48	164	157	(749, 934)	$\hat{N}_{c_1}$	996	991	71	71	56	(876, 1,155)	$\hat{N}_{c_1}$	1,157	1,152	110	193	161	(978, 1,366)
		$\hat{N}_{c_2}$	957	955	75	87	70	(808, 1,099)	$\hat{N}_{c_2}$	1,217	1,208	129	253	218	(1,012, 1,517)	$\hat{N}_{c_2}$	1,530	1,513	231	579	530	(1,184, 2,008)
		$\hat{N}_J$	1,148	1,124	154	214	161	(926, 1,483)	$\hat{N}_J$	1,217	1,196	144	261	218	(1,007, 1,552)	$\hat{N}_J$	1,232	1,226	151	277	232	(1,016, 1,555)
	1.5	$\hat{N}_u$	1,069	1,063	109	129	92	(925, 1,264)	$\hat{N}_u$	1,069	1,037	215	226	100	(915, 1,388)	$\hat{N}_u$	1,107	1,038	302	320	142	(895, 1,824)
		$\hat{N}_{WL}$	1,069	1,069	90	113	90	(927, 1,253)	$\hat{N}_{WL}$	1,049	1,042	94	106	78	(915, 1,284)	$\hat{N}_{WL}$	1,055	1,041	111	124	88	(896, 1,327)
		$\hat{N}_{c_0}$	998	991	79	79	64	(863, 1,155)	$\hat{N}_{c_0}$	987	981	75	76	61	(863, 1,135)	$\hat{N}_{c_0}$	1,000	998	85	85	64	(847, 1,180)
		$\hat{N}_{c_1}$	850	848	50	158	150	(756, 954)	$\hat{N}_{c_1}$	1,082	1,076	72	109	88	(958, 1,212)	$\hat{N}_{c_1}$	1,358	1,356	129	381	358	(1,143, 1,628)
		$\hat{N}_{c_2}$	1,082	1,072	99	129	100	(911, 1,279)	$\hat{N}_{c_2}$	1,561	1,539	163	586	561	(1,295, 1,865)	$\hat{N}_{c_2}$	2,195	2,189	338	1,245	1,195	(1,676, 2,876)
		$\hat{N}_J$	1,183	1,170	112	215	185	(978, 1,397)	$\hat{N}_J$	1,209	1,211	117	240	209	(1,037, 1,498)	$\hat{N}_J$	1,237	1,231	132	272	237	(1,023, 1,534)
.75	1	$\hat{N}_u$	1,042	1,019	115	123	53	(958, 1,298)	$\hat{N}_u$	1,024	1,016	70	74	37	(959, 1,174)	$\hat{N}_u$	1,030	1,017	60	67	40	(959, 1,188)
		$\hat{N}_{WL}$	1,042	1,023	70	82	51	(959, 1,231)	$\hat{N}_{WL}$	1,029	1,018	55	62	40	(961, 1,181)	$\hat{N}_{WL}$	1,041	1,023	63	75	50	(966, 1,203)
		$\hat{N}_{c_0}$	998	995	39	39	30	(930, 1,084)	$\hat{N}_{c_0}$	997	996	34	34	27	(934, 1,068)	$\hat{N}_{c_0}$	1,003	1,000	37	37	30	(934, 1,077)
		$\hat{N}_{c_1}$	1,064	1,058	35	73	65	(1,001, 1,133)	$\hat{N}_{c_1}$	1,130	1,133	41	137	130	(1,053, 1,218)	$\hat{N}_{c_1}$	1,129	1,129	50	139	129	(1,024, 1,229)
		$\hat{N}_{c_2}$	1,179	1,174	55	188	179	(1,084, 1,281)	$\hat{N}_{c_2}$	1,277	1,274	67	286	277	(1,159, 1,419)	$\hat{N}_{c_2}$	1,238	1,234	78	251	238	(1,081, 1,403)
		$\hat{N}_J$	1,173	1,160	76	189	173	(1,086, 1,325)	$\hat{N}_J$	1,179	1,179	42	185	179	(1,100, 1,266)	$\hat{N}_J$	1,200	1,197	43	205	200	(1,117, 1,275)
	1.5	$\hat{N}_u$	1,028	1,026	44	52	37	(966, 1,112)	$\hat{N}_u$	1,026	1,013	57	62	34	(968, 1,158)	$\hat{N}_u$	1,035	1,015	89	96	50	(954, 1,321)
		$\hat{N}_{WL}$	1,033	1,029	46	57	40	(970, 1,145)	$\hat{N}_{WL}$	1,033	1,017	53	62	39	(970, 1,173)	$\hat{N}_{WL}$	1,038	1,018	69	79	50	(956, 1,215)
		$\hat{N}_{c_0}$	1,000	1,000	39	39	32	(934, 1,078)	$\hat{N}_{c_0}$	1,003	1,002	31	31	24	(945, 1,071)	$\hat{N}_{c_0}$	999	998	37	37	30	(926, 1,071)
		$\hat{N}_{c_1}$	1,123	1,124	41	130	123	(1,040, 1,210)	$\hat{N}_{c_1}$	1,149	1,139	63	162	149	(1,060, 1,297)	$\hat{N}_{c_1}$	1,129	1,109	83	154	130	(1,012, 1,340)
		$\hat{N}_{c_2}$	1,301	1,298	72	310	301	(1,163, 1,462)	$\hat{N}_{c_2}$	1,278	1,255	120	304	278	(1,119, 1,573)	$\hat{N}_{c_2}$	1,233	1,183	155	280	233	(1,044, 1,663)
		$\hat{N}_J$	1,168	1,165	51	176	168	(1,086, 1,297)	$\hat{N}_J$	1,185	1,184	38	189	185	(1,115, 1,256)	$\hat{N}_J$	1,195	1,194	47	201	195	(1,102, 1,278)

NOTE: Footnotes are the same as in Table 7.

Table 10. Summary of Relative Performance of Estimators in Simulation I

			MAE[a]						Median-bias[b]						95% coverage[c]					
d	CV	(α, β)	$\underline{\hat{N}_{WL}}$	$\hat{N}_u$	$\hat{N}_{c_0}$	$\hat{N}_{c_1}$	$\hat{N}_{c_2}$	$\hat{N}_J$	$\underline{\hat{N}_{WL}}$	$\hat{N}_u$	$\hat{N}_{c_0}$	$\hat{N}_{c_1}$	$\hat{N}_{c_2}$	$\hat{N}_J$	$\underline{\hat{N}_{WL}}$	$\hat{N}_u$	$\hat{N}_{c_0}$	$\hat{N}_{c_1}$	$\hat{N}_{c_2}$	$\hat{N}_J$
.2	.5	(4, 17.43)	+	−		+		−	−	−			+	+	+	+	+	+	+	−
	.71	(2, 8.47)	+	−			+	−	+	+	−			−	+	+	+	+	+	−
	1	(1, 4)	+		−	−		+	+	+	−	−			−	+	−	−	−	−
	1.41	(.5, 1.78)		+	−	−		+		+	−	−		+	−	+	−	−	−	−
.3	.5	(4, 10.72)		−	−		+	+		+	−	−		+	+	+	+	+	+	+
	.71	(2, 5.12)	+	−	−			+		+	−	−		+	+	+	−	−	+	+
	1	(1, 2.33)		−	−		+	+		+	−	−		+	−	+	−	−	−	+
	1.41	(.5, .96)			−	−	+	+			−	−	+	+	−	+	−	−	−	−
.4	.5	(4, 7.43)		−		+	+	−	+	+	−	−			+	+	+	+	+	+
	.71	(2, 3.44)	+	−	−			+		+	−	−		+	+	+	−	−	+	+
	1	(1, 1.5)	+	−	−			+		+	−	−		+	−	+	+	−	−	+
	1.41	(.5, .56)		−	−	+		+			−	−	+	+	−	+	−	−	−	+
.5	.5	(4, 5.29)		−		+	+	−	+	+	−			−	+	+	+	+	+	+
	.71	(2, 2.41)	+	−	−			+	+		−	−		+	+	+	−	−	+	+
	1	(1, 1)		−	−		+	+			−	−	+	+	+	+	−	−	−	+
	1.41	(.5, .33)	+		−	−		+	+		−	−		+	+	+	−	−	−	+
.75	.5	(4, 2.41)		−		+	+	−	+	+	−			−	+	+	+	+	+	−
	.71	(2, 1)		−	−		+	+	+	+	−	−			+	+	−	−	+	+
	1	(1, .33)	+		−	−		+	+		−	−		+	+	+	−	−	−	+
	1.41	(.5, .067)	+			−	−	+	+			−	−	+	+	+	−	−	−	+
.9	.5	(4, 1.28)	−		+	+	+	−	+	+	−			−	+	+	+	+	+	−
	.71	(2, .46)		+	+	−	+	−	+	+		−	−		+	+	−	−	−	+
	1	(1, .11)	+			−	−	+	+			−	−	+	+	+	−	−	−	+
	1.41	(.5, .01)	+			−	−	+	+			−	−	+	+	+	−	−	−	+

[a]MAE: "+" means that the estimator is the best two with smallest mean absolute error and "−" means the worst two.
[b]Median-Bias: "+" means that the estimator is the best two with smallest median-bias and "−" means the worst two.
[c]95% coverage: "+" means the central 95% percentile interval covers $N = 1{,}000$, and "−" means that it does not.

distribution

$$\hat{Q} = .941\Delta_{\lambda=.37} + .052\Delta_{\lambda=3.51} + .007\Delta_{\lambda=9.99}$$

and $\hat{N}_{WL} = 8{,}919$, the estimates obtained at $\tau = 15$. The results from $\hat{N}_{WL}$, $\hat{N}_u$, $\hat{N}_{c_0}$ $\hat{N}_{c_1}$, and $\hat{N}_{c_2}$ are presented in Table 14. The sampling distribution of these estimators closely confirms the original point estimates from Table 13. For example, the median of $\hat{N}_{WL}$ was 9,158, biased upward by about 239 from the true value 8,919 used in the simulation. The unconditional NPMLE approximant $\hat{N}_u$ had an upward median bias of 435, but was much more variable than $\hat{N}_{WL}$. The other estimators, $\hat{N}_{c_0}$, $\hat{N}_{c_1}$, $\hat{N}_{c_2}$, and $\hat{N}_J$, were all more simulation-biased than $\hat{N}_{WL}$, with their central 95% intervals not covering $\hat{N}_{WL} = 8{,}919$, the true N of the simulation. The estimate $\hat{N}_{c_3}$ failed due to a negative

Table 11. Summary of Relative Performance of Estimators in Simulation II

			MAE						Median-bias						95% coverage					
d	CV	ρ	$\underline{\hat{N}_{WL}}$	$\hat{N}_u$	$\hat{N}_{c_0}$	$\hat{N}_{c_1}$	$\hat{N}_{c_2}$	$\hat{N}_J$	$\underline{\hat{N}_{WL}}$	$\hat{N}_u$	$\hat{N}_{c_0}$	$\hat{N}_{c_1}$	$\hat{N}_{c_2}$	$\hat{N}_J$	$\underline{\hat{N}_{WL}}$	$\hat{N}_u$	$\hat{N}_{c_0}$	$\hat{N}_{c_1}$	$\hat{N}_{c_2}$	$\hat{N}_J$
.3	1	1.5	+	−	−			+		+	−	−	+		+	+	−	−	−	+
		2.5	+	−	−	−		+	+		−	−		+	+	+	+		+	+
		4.7		−		+	−	+	+		−		−	+	+	+	+	+	+	+
	1.5	1.5	+	−		−		+	+	−		−		+	+	+	−	−	−	+
		2.5	+	−		−		+		+	−	−		+	+	+	+	+	−	−
		4.7		−		+	−	+	+		−		−	+	+	+	+	+	−	−
.5	1	1.5	+	−			+	−	+	+	−	−			+	+	−	−	+	+
		2.5			+	+	−	−		+		+	−	−	+	+	+	+	−	−
		4.7	+		+		−	−		+	+		−	−	+	+	+	−	−	−
	1.5	1.5	+		+	−		−		+	+	−		−	+	+	+	−	+	+
		2.5	+		+		−	−		+	+		−	−	+	+	+	+	−	−
		4.7	+		+	−	−			+	+	−	−		+	+	+	−	−	−
.75	1	1.5	+		+		−	−		+	+		−	−	+	+	+	−	−	−
		2.5		+	+		−	−		+	+		−	−	+	+	+	−	−	−
		4.7		+	+		−	−		+	+		−	−	+	+	+	−	−	−
	1.5	1.5		+	+		−	−		+	+		−	−	+	+	+	−	−	−
		2.5		+	+		−	−		+	+		−	−	+	+	+	−	−	−
		4.7	+	+	+		−	−		+	+		−	−	+	+	+	−	−	−

NOTE: Footnotes are the same as in Table 10.

Table 12. Arabidopsis thaliana Root EST Data

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	21	23	24	25^+	D
n_i	2,187	490	133	121	37	51	22	19	7	8	6	7	6	4	5	5	1	4	2	2	2	1	6	3,126

estimate of the duplication factor and so was excluded from this simulation.

6. DISCUSSION

6.1 Comparison of Methods

Accurate estimation of species richness N requires achieving a delicate balance between bias and variability. Each of the methods that we considered had some scheme for tuning; the tuning parameter τ used for $\hat{N}_{c_1}$ and $\hat{N}_{c_2}$, the order used in the jackknife estimator $\hat{N}_J$, and the penalizing functions (and weights) in $\hat{N}_{WL}$ and $\hat{N}_u$ are all devices to this end.

Our simulation results revealed a few lessons about coverage estimators. A good choice of τ is critical for the two coverage estimators because of their sensitivity to this number. For the coverage methods, we note that there is a $i^2 n_i$ term in the CV estimate. Therefore, including large i's can result in great variability of the CV estimate. Nonetheless, a good criterion for choosing τ remains undeveloped. Moreover, the estimator $\hat{N}_{c_2}$ can improve estimation based on $\hat{N}_{c_1}$, but it can also be wildly biased, as shown in Simulation II. A convincing rule for choosing one estimator over the other has not been given.

As to the jackknife estimator, the selection of the order of the estimator based on a sequential testing procedure for $\hat{N}_J$ is an attempt to balance bias and variability. Just the same, we found that even when we use an upper bound on the order of 5, overshooting still occurred frequently.

We have gained considerable understanding of the likelihood-based estimators. The estimator $\hat{N}_u$ was shown to be practically equivalent to $\hat{N}_{UNP}$. We believe that a considerable savings in computation makes it more useful in real data analysis than the approach of Norris and Pollock (1998). However, for either, instability and variability is a big concern, because the penalizing function quickly becomes flat as θ increases. In our analysis, the estimator $\hat{N}_{WL}$ achieved similar freedom from bias, in that its sampling distribution fell near the true values, but with much lower variability due to its greater stability.

The robust accuracy of the sampling distribution is achieved in two senses by the penalized NPMLE estimators, particularly $\hat{N}_{WL}$. First, $\hat{N}_{WL}$ exhibited better robustness to the form of latent distribution Q than the coverage and jackknife approaches. For example, in Simulation II, high skewness and low CV in a discrete Q resulted in large positive bias to $\hat{N}_{c_1}, \hat{N}_{c_2}$, and $\hat{N}_J$, but not so much to $\hat{N}_{WL}$. This feature is highly desirable for an estimator in practice, because typically little prior knowledge about Q is available, and the data themselves carry little information.

A second remarkable feature of the NPMLE-based estimators is their insensitivity to τ. Species data often have a long tail but with majority data points concentrating on n_i for small i's, which usually dominates the likelihood. As a result, the data points at right tail can be informative for $Q(\lambda)$ but are less influential in the estimation of $f(0; Q)$ or $\theta(Q)$. For example, consider fitting NPMLE to a right-skewed data. Large components but with very small weights are usually fit due to these large observations, while contributing little to $f(0; \hat{Q})$. We can take advantage of this insensitivity to fit NPMLE to relatively "rare" species without introducing additional bias. Fitting NPMLE to the "rare" species data often saves substantial computing time, thereby enabling one to obtain point estimates and bootstrap confidence intervals in an acceptable time interval. The average computing time for the penalized NPMLE $\hat{N}_u$ or $\hat{N}_{WL}$ in the simulation study was several seconds or less.

6.2 Other Possibilities

We did not include parametric approaches for comparison in this article. However, a nested family of increasingly rich mixture models could themselves be turned into a nonparametric methodology using the method of sieves. For example, one could consider letting the Kth parametric model for Q consist of discrete distributions with exactly K components, then use some secondary criterion to select the value of K. In doing this, Pledger (2000) recommended selecting the number of components in Q by testing sequentially.

This approach no doubt would have better stability than $\hat{Q}_{UNP}$, because mass points near 0 would tend to be eliminated. But it is not clear to us that this eliminates the problem. There are also practical difficulties with this approach. Even when testing the simplest hypothesis, one component versus two components, the likelihood ratio does not have a simple limiting distribution in the form of chi-squared or a mixture of chi-squareds (Lindsay 1995). Admittedly, one could use a bootstrap approach to find the critical value given a specific null hypothesis. But that may not be practical when using an EM-based computing algorithm due to the long run time. In ad-

Table 13. Results for Root EST Data

τ	$\hat{N}_u$	$\hat{N}_{WL}$	$\hat{N}_{c_1}$	$\hat{N}_{c_2}$
10	9,176	9,155	9,088	13,943
11	9,197	9,179	9,292	14,687
12	9,124	9,111	9,580	15,762
13	9,046	9,036	9,866	16,867
14	8,992	8,984	10,087	17,746
15	8,926	8,919	10,403	19,034
20	9,041	9,028	11,630	24,468
24	9,036	9,023	12,652	29,399

Table 14. Simulation Results for the Arabidopsis thaliana Root EST Data

Estimator	$\bar{\hat{N}}$	$\hat{M}$	s	*RMSE*	*MAE*	*95%*
$\hat{N}_{WL}$	9,249	9,158	695	770	560	(8,113, 10,762)
$\hat{N}_u$	9,811	9,354	1,725	1,944	1,055	(8,272, 13,749)
$\hat{N}_{c_0}$	8,106	8,098	325	878	815	(7,427, 8,725)
$\hat{N}_{c_1}$	10,200	10,211	393	1,343	1,281	(9,474, 10,933)
$\hat{N}_{c_2}$	18,187	18,146	1,138	9,360	9,268	(16,278, 20,541)
$\hat{N}_J$	10,351	10,313	351	1,478	1,432	(9,679, 11,056)

dition, based on our experience, such a testing device tends to fit a most parsimonious model and often yields an overconservative estimate (results not shown).

From the computing perspective, there is an additional advantage to using the NPMLE. When the number of components K is fixed, the likelihood can be multimodal, in which case the EM may converge to a local maximum. The algorithm for NPMLE that we proposed here is guaranteed to converge to a global solution to any penalized likelihood with a linear or convex penalizing function and is unique in the truncated Poisson case.

To allow more flexibility in the parametric approaches, one might want to consider a mixture distribution for Q. For example, in the Poisson–gamma model, one could assume a mixture of gamma. We have considered this model but with the number of components in Q unfixed. The optimal number of components could be found by nonparametric maximum likelihood estimation (Wang 2003). The performance of the resulting estimator remains under investigation.

6.3 Confidence Assessment

We have mentioned two bootstrap approaches to confidence assessment for N. One was Poisson-model based and involved sampling species counts, and the other was multinomial-based, sampling individuals from a population constructed from the model, then aggregating counts.

Another possibility that has a more nonparametric flavor is to generate D non-0 observations of X from the multinomial distribution corresponding to the empirical distribution of the non-0 counts, that is, $f_n(i) = \frac{n_i}{D}$ for $i = 1, \ldots, t$. We call this the *naive nonparametric bootstrap*, because it treats D as a fixed number. In actual sampling, D is a random variable that is critical in the estimation of N. One could fix this problem by using a *hybrid bootstrap*, in which one would first sample D^* from the binomial $(\hat{N}, 1 - f(0; \hat{Q}))$ distribution, then draw D^* times from the empirical multinomial.

We have found that, as expected, the naive nonparametric bootstrap confidence interval is often narrower than that from population method. For example, for the EST data, the former gave a 95% CI of (8,280, 10,549), slightly tighter than the (8,113, 10,762) obtained by our multinomial method. The 95% CI by the Poisson-based bootstrap was (8,260, 10,760), covering the naive one but not as large as the multinomial one. This relationship supports our intuition that the multinomial bootstrap will tend to be more conservative, and hence reliable, than the others.

However, these findings are tentative. Firm comparisons, if such are even possible, would require a very computationally expensive investigation given the structure of the estimators and the likely complex interplay of bias, variability, and coverage.

6.4 Bayesian Methods

We earlier offered a Bayesian interpretation of the likelihood penalty function, but our methods really differ from a standard Bayesian approach in two ways. First, our estimation of tuning parameters in the penalty clearly differs from conventional Bayesian analysis. However, there is a quasi-Bayes interpretation of what we did; we put a hyperprior on the tuning parameters, but instead of integrating them out, we "plugged in" estimators for them. Second, our penalized likelihood has the form $p[\theta(Q)] \prod_i \int f(x_i; \lambda)\, dQ(\lambda)$, where the "prior" $p[\theta(Q)]$, although it affects Q estimation, is unlike Bayes in that it does not correspond to any actual distribution for Q. Just the same, we believe that thinking of it as a "partial prior" does provide some insight, especially into its role in weighting risk over the parameter space. (Although more restrictive in construction, penalties based on fictional data likelihoods offer a simpler interpretation.)

In the conventional Bayesian analysis, prior(s) for Q (or the parameters in Q) and N are completely specified (Blumenthal and Marcus 1975; Blumenthal 1977, 1982; Hill 1979; Lewins and Joanes 1984; Boender and Kan 1987). For example, a popular prior for the relative abundances p_i given N in the multinomial version of model is the Dirichlet distribution (Lewins and Joanes 1984; Boender and Kan 1987). Clearly, such a prior distribution explicitly specifies the form of the distribution of p_i. To make the model more flexible, one can further impose a higher-level prior for the hyper-parameters in the Dirichlet distribution.

For example, for the same butterfly data as used here, Boender and Kan (1987) used a symmetric Dirichlet prior and came up with a posterior distribution for N that had a mode of 940, a median of 1,020 and a mean of 1,054. These are substantially larger than those reported from any of the nonparametric methods in Table 1. Unfortunately, there are very few systematic comparisons or numerical results in the literature evaluating the Bayesian approach. The flat likelihoods that we found in this problem do suggest that the choice of prior can be quite important here.

6.5 Final Thoughts

Whether thought of as Bayes with a partial prior or as a way of improving frequentist risk, penalized likelihood methods have an element of subjectivity similar to Bayes methods. One must select a penalty function and a tuning parameter, and it is nigh impossible to make convincing statements about why one choice should be uniformly superior to another. In fact, in our particular case, because all penalties were functions of θ, our algorithm was based on the fact that all of the estimators of Q could be found by using the linear penalized function ℓ_2 at different values of its tuning parameter. Thus all of the methods were, in a sense, ℓ_2 estimators with different adaptive selections of γ_2. In the end, we were pleased because the methods that we used here turned out to be highly effective over a range of models, but we can hardly make a claim for global optimality.

Finally, the computational methods given here for using penalties in nonparametric likelihood estimation can be readily extended to other applications where a mixture model is used. Currently, application to the capture-mark-recapture problem (Otis, Burnham, White, and Anderson 1978; Pledger 2000) is under investigation.

APPENDIX A: PROOF OF PROPOSITION 1

(a) Suppose that $(\hat{N}, \hat{Q})$ is the unconditional MLE. Because ℓ_c is completely determined by Q, given $\hat{Q}$, $\hat{N}$ must be $\langle D(1 + \hat{\theta})\rangle$; otherwise, one can always increase the marginal likelihood by setting $\hat{N} = \langle D(1 + \hat{\theta})\rangle$. In contrast, given $\theta(Q) = \hat{\theta}$, because ℓ_m is free of Q, $\hat{Q}$ must be the constrained maximizer given $\theta(Q) = \hat{\theta}$; otherwise, L_c can always be increased.

(b) From (I), $\hat{N}$ is completely determined by $\hat{Q}$; therefore, $\sup_{N,Q}\{\ell(N,Q)\} = \sup_Q\{\ell(N = \langle D(1+\hat{\theta})\rangle, Q)\} = \sup_Q\{\ell_m(N, Q)|_{N=\langle D(1+\theta(Q))\rangle} + \ell_c(Q)\}$.

(c) Note that Stirling's approximation for $\hat{N}!$ and $(\hat{N}-D)!$ in L_m gives

$$\begin{aligned}
&\sup_N\{L_m(N|\theta)\} \\
&= \frac{\hat{N}!}{D!(\hat{N}-D)!}\left[\frac{\theta}{1+\theta}\right]^{\hat{N}-D}\left[\frac{1}{1+\theta}\right]^{D} \\
&\approx \frac{\sqrt{2\pi}\hat{N}^{\hat{N}+.5}e^{-\hat{N}}}{D!\sqrt{2\pi}(\hat{N}-D)^{\hat{N}-D+.5}e^{-\hat{N}+D}}\left[\frac{\theta}{1+\theta}\right]^{\hat{N}-D}\left[\frac{1}{1+\theta}\right]^{D}, \quad \text{(A.1)}
\end{aligned}$$

as $D \to \infty$, where the symbol "$\approx$" means that the ratio of the two sides goes to 1. Because $\arg_N \sup(L_m|\theta) = \langle D(1+\theta)\rangle \approx D(1+\theta)$, by plugging $\hat{N} = D(1+\theta)$ into (A.1), we obtain $\sup_N(L_m|\theta) \approx \text{Poisson}(D,D) \times [\frac{1+\theta}{\theta}]^{1/2} = \frac{e^{-D}D^D}{D!}[\frac{1+\theta}{\theta}]^{1/2}$. The condition "$\hat{N}$, $\hat{N}-D \to \infty$" can be unified as "$D \to \infty$," because $\hat{N} = \langle D(1+\theta)\rangle$ and $\hat{N} - D = \langle D\theta\rangle$. It also can be shown that the ratio of the evaluations of (A.1) under $\hat{N} = \langle D(1+\theta)\rangle$ and $\hat{N} = D(1+\theta)$ converges to 1 as $D \to \infty$. Therefore, the difference of the logarithm of both sides must go to 0 as $D \to \infty$.

APPENDIX B: PROOF OF THEOREM 1

Let $\hat{P}^{\gamma_1}$ and $\hat{P}^{\gamma_2}$ be NPMLEs at penalties γ_1 and γ_2 from (6), and let $\gamma_1 \le \gamma_2$. By the definition of NPMLE, we have

$$\ell^{\gamma_1}(\hat{P}^{\gamma_1}) = \ell(\hat{P}^{\gamma_1}) - \gamma_1 h[\theta(\hat{P}^{\gamma_1})] \ge \ell(\hat{P}^{\gamma_2}) - \gamma_1 h[\theta(\hat{P}^{\gamma_2})]$$

and

$$\ell^{\gamma_2}(\hat{P}^{\gamma_2}) = \ell(\hat{P}^{\gamma_2}) - \gamma_2 h[\theta(\hat{P}^{\gamma_2})] \ge \ell(\hat{P}^{\gamma_1}) - \gamma_2 h[\theta(\hat{P}^{\gamma_1})].$$

Adding both sides gives

$$(\gamma_2 - \gamma_1)h[\theta(\hat{P}^{\gamma_1})] \ge (\gamma_2 - \gamma_1)h[\theta(\hat{P}^{\gamma_2})].$$

By monotonicity of the function $h(\theta)$ in θ, $\theta(\hat{P}^{\gamma_1}) \ge \theta(\hat{P}^{\gamma_2})$, and thus $\hat{N}^{\gamma_1} \ge \hat{N}^{\gamma_2}$.

APPENDIX C: PROOF OF COROLLARY 1

(a) Note that $h(\theta) = \log[\frac{1+\theta}{\theta}]$ is a monotone-decreasing function of θ. The conditional estimator $\hat{N}_{CNP}$ corresponds to a penalty $\gamma_1 = 0$ in ℓ_1, where $\hat{N}_u$ to $\gamma_1 = .5$. By Theorem 1, we have $\hat{\theta}_{CNP} \ge \hat{\theta}_u$; thus $\hat{N}_{CNP} \ge \hat{N}_u$.

(b) Let P represent the unknown mixing distribution in the nonparametric case and the unknown parameters in the parametric case (It is straightforward to verify that the monotonicity in Thm. 1 also applies to the parametric case.) In the nonparametric case, we write P or Q interchangeably, that is, $\ell_m(N,Q) \equiv \ell_m(N,P)$ and $\ell_c(Q) \equiv \ell_c(P)$. By Proposition 1, the objective function corresponding to the full log-likelihood can be written in the penalized form as $\ell(P) = \ell_c(P) - \gamma[-\ell_m^*(\theta)]$, where $\gamma = 1$ and $\ell_m^*(\theta) = \ell_m(N,P)|_{N=\langle D(1+\theta)\rangle}$ depending only on θ. Because the un-penalized conditional MLE of θ, $\hat{\theta}_{CNP}$, is the estimate under $\gamma = 0$, it suffices to show that $\ell_m^*(\theta)$ decreases monotonically in θ by Theorem 1 [or that $-\ell_m^*(\theta)$ increases in θ]. Suppose that we have $\theta_1 < \theta_2$ and $N = \langle D(1+\theta_1)\rangle$, $N+k = \langle D(1+\theta_2)\rangle$, for $k > 0$. Let p_1 and p_2 be the success probability in the binomial distribution corresponding to θ_1 and θ_2. Then $\frac{D}{N+1} < p_1 \le \frac{D}{N}$, $\frac{D}{N+k+1} < p_2 \le \frac{D}{N+k}$. Consider the two marginal likelihoods in the binomial form,

$$L_1 = \binom{N}{D} p_1^D (1-p_1)^{N-D},$$

and

$$L_2 = \binom{N+k}{D} p_2^D (1-p_2)^{N+k-D}.$$

It suffices to show that $L_1 > L_2$. Note that because L_1 is a monotone increasing function for $p_1 \in [\frac{D}{N+1}, \frac{D}{N}]$, we have

$$\begin{aligned}
L_1 &= \binom{N}{D} p_1^D (1-p_1)^{N-D} \\
&> \binom{N}{D}\left(\frac{D}{N+1}\right)^D\left(1-\frac{D}{N+1}\right)^{N-D} \quad \text{(C.1)} \\
&\equiv \binom{N+1}{D}\left(\frac{D}{N+1}\right)^D\left(1-\frac{D}{N+1}\right)^{N+1-D} \\
&> \binom{N+1}{D}\left(\frac{D}{N+2}\right)^D\left(1-\frac{D}{N+2}\right)^{N+1-D} \\
&\equiv \binom{N+2}{D}\left(\frac{D}{N+2}\right)^D\left(1-\frac{D}{N+2}\right)^{N+2-D} \\
&\vdots \\
&> \binom{N+k}{D}\left(\frac{D}{N+k}\right)^D\left(1-\frac{D}{N+k}\right)^{N+k-D} \\
&> \binom{N+k}{D} p_2^D (1-p_2)^{N+k-D} \\
&\equiv L_2. \quad \text{(C.2)}
\end{aligned}$$

This proves the monotonicity of $\ell_m^*(\theta)$ in θ, and thus the result follows.

APPENDIX D: PROOF OF THEOREM 2

(a) Given the boundedness and closedness of the image set $(L_1(\lambda), L_2(\lambda), \ldots, \theta(\lambda))$, the existence of a distribution $\hat{P}_\omega$ that maximizes the penalized likelihood $\ell^\omega(P) = \ell(P) - \omega\theta(P)$ is immediate by the convex geometry optimization results of Lindsay (1995).

(b) By definition, $\hat{P}$ is a local maximizer of $\ell(P) - \gamma h[\theta(P)]$ if and only if the gradient inequality is true, that is,

$$\begin{aligned}
D(\hat{P},\lambda) &= \sum_j n_j\left[\frac{g(j;\lambda)}{g(j;\hat{P})} - 1\right] - \gamma h'[\theta(\hat{P})][\theta(\lambda) - \theta(\hat{P})] \\
&\le 0 \qquad \forall \lambda \in \Omega.
\end{aligned}$$

If we let $\omega = \gamma h'[\theta(\hat{P})]$, then the foregoing condition is exactly the sufficient and necessary condition for $\hat{P}$ to be the NPMLE in the linearized version, namely $\ell(P) - \omega\theta(P)$.

(c) If $-\gamma h[\theta(P)]$ is concave in θ, then one can show that the penalized likelihood $\ell(P) - \gamma h[\theta(P)]$ is also concave in any path $(1-\alpha)P + \alpha P^*$, where P and P^* are two arbitrary mixing distributions. First, note that a local solution $\hat{P}$ satisfying the gradient condition from (b) must satisfy $D(\hat{P},\hat{P}^*) \le 0\ \forall \hat{P}^*$ (can be shown easily, see also Lindsay 1995). Thus such a solution $\hat{P}$ must be the global one, because otherwise, if we let $\hat{P}^*$ be the global solution, then the gradient inequality is violated along the path $(1-\alpha)\hat{P} + \alpha\hat{P}^*$ by concavity.

APPENDIX E: PROOF OF COROLLARY 2

(a) The result is almost immediate from Theorem 2. By definition, the global solution $\hat{P}$ must also be local, which implies that $\hat{P}$ must satisfy the gradient criterion as follows:

$$\begin{aligned}
D(\hat{P},\lambda) &= \sum_j n_j\left[\frac{g(j;\lambda)}{g(j;\hat{P})} - 1\right] - \frac{\gamma_1}{\theta(1+\theta)}[\theta(\lambda) - \theta(\hat{P})] \\
&\le 0 \qquad \forall \lambda \in \Omega.
\end{aligned}$$

Note that $\theta(\lambda)$ is not bounded at $\lambda = 0$. However, this is not a concern, because $\theta(\hat{P}) < \infty$ gives $\frac{\gamma_1}{\theta[\hat{P}](1+\theta[\hat{P}])} > 0$. The linearized likelihood is bounded from above. The gradient criterion given earlier is exactly the one for $\hat{P}$ to be the NPMLE (global) for $\ell_2 = \ell(P) - \gamma_2\theta$ at $\gamma_2 = \gamma h'(\theta(\hat{P}))$.

The uniqueness of global solution in the linearized likelihood ℓ_2 is guaranteed because the zero-truncated Poisson distribution is an exponential family distribution and has infinite support points. The mixture model can never fit the data perfectly (see details in Lindsay and Roeder 1993).

(b) If $\theta(\hat{P}_{CNP}) \leq \mu$, then there exists a $\hat{P}_{CNP}$ that maximizes $\ell(P)$ globally without penalty. Therefore, it must maximize ℓ_3 globally. If $\theta(\hat{P}_{CNP}) > \mu$, then we are maximizing the penalized likelihood with a concave penalty $-\gamma_3 h(\theta(P))$. Because the penalized likelihood function is strictly concave, a sufficient and necessary condition for $\hat{P}$ to be a global maximizer is that $\hat{P}$ is a local solution. By the results from part (b) of Theorem 2, the conclusion is immediate. For the same reason as in part (a), unboundedness of θ is not a concern. The uniqueness is obtained for the same reason as in (a).

[Received January 2004. Revised September 2004.]

REFERENCES

Adams, M. D., Kelley, J. M., Gocayne, J. D., Dubnick, M., Polymeropoulos, M. H., Xiao, H., Merril, C. R., Wu, A., Olde, B., Moreno, R. F., Kerlavage, A. R., McCombie, W. R., and Venter, J. C. (1991), "Complementary DNA Sequencing: Expressed Sequence Tags and Human Genome Project," *Science*, 252, 1651–1656.

Asamizu, E., Nakamura, Y., Sato, S., and Tabata, S. (2000), "A Large Scale Analysis of cDNA in *Arabidopsis thaliana*: Generation of 12,028 Non-Redundant Expressed Sequence Tags From Normalized and Size-Selected cDNA Libraries," *DNA Research*, 7, 175–180.

Blumenthal, S. (1977), "Estimating Population Size With Truncated Sampling," *Communications in Statistics, Part A, Theory and Methods*, 6, 297–308.

——— (1982), "Stochastic Expansions for Point Estimation From Truncated Samples," *Sankhyā*, Ser. A, 44, 436–451.

Blumenthal, S., and Marcus, R. (1975), "Estimating Population Size With Exponential Failure," *Journal of the American Statistical Association*, 70, 913–922.

Boender, C., and Kan, A. (1987), "A Multinomial Bayesian Approach to the Estimation of Population and Vocabulary Size," *Biometrika*, 74, 849–856.

Bunge, J., and Fitzpatrick, M. (1993), "Estimating the Number of Species: A Review," *Journal of the American Statistical Association*, 88, 364–373.

Burnham, K. P., and Overton, W. S. (1978), "Estimation of the Size of a Closed Population When Capture Probabilities Vary Among Animals," *Biometrika*, 65, 625–633.

——— (1979), "Robust Estimation of Population Size When Capture Probabilities Vary Among Animals," *Ecology*, 60, 927–936.

Chao, A. (1984), "Nonparametric Estimation of the Number of Classes in a Population," *Scandinavian Journal of Statistics*, 11, 265–270.

Chao, A., and Bunge, J. (2002), "Estimating the Number of Species in a Stochastic Abundance Model," *Biometrics*, 58, 531–539.

Chao, A., Huang, W.-H., Chen, Y.-C., and Kuo, C.-Y. (2000), "Estimating the Number of Shared Species in Two Communities," *Statistica Sinica*, 10, 227–246.

Chao, A., and Lee, S.-M. (1992), "Estimating the Number of Classes via Sample Coverage," *Journal of the American Statistical Association*, 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.

Efron, B. (1981), "Nonparametric Standard Errors and Confidence Intervals," *Canadian Journal of Statistics*, 9, 139–172.

Fisher, R. A., Corbet, A. S., and Williams, C. B. (1943), "The Relation Between the Number of Species and the Number of Individuals in a Random Sample of an Animal Population," *Journal of Animal Ecology*, 12, 42–58.

Haas, P. J., and Stokes, L. (1998), "Estimating the Number of Classes in a Finite Population," *Journal of the American Statistical Association*, 93, 1475–1487.

Hill, B. M. (1979), "Posterior Moments of the Number of Species in a Finite Population and the Posterior Probability of Finding a New Species," *Journal of the American Statistical Association*, 74, 668–673.

Lewins, W. A., and Joanes, D. N. (1984), "Bayesian Estimation of the Number of Species," *Biometrics*, 40, 323–328.

Lindsay, B. G. (1995), *Mixture Models: Theory, Geometry and Applications*, Vol. 5, NSF-CBMS Regional Conference Series in Probability and Statistics, Hayward, CA: Institute of Mathematical Statistics.

Lindsay, B. G., and Roeder, K. (1987), "A Unified Treatment of Integer Parameter Models," *Journal of the American Statistical Association*, 82, 758–764.

——— (1992), "Residual Diagnostics for Mixture Models," *Journal of the American Statistical Association*, 87, 758–764.

——— (1993), "Uniqueness of Estimation and Identifiability in Mixture Models," *Canadian Journal of Statistics*, 87, 139–147.

Mao, C., and Lindsay, B. G. (2002), "Estimating the Number of Classes: Identifiability, Singularity, and One-Sided Inference," unpublished manuscript.

——— (2003), "Tests and Diagnostics for Heterogeneity in the Species Problem," *Computational Statistics and Data Analysis*, 41, 389–398.

Norris, J. L. I., and Pollock, K. H. (1996), "Nonparametric MLE Under Two Closed Capture-Recapture Models With Heterogeneity," *Biometrics*, 52, 639–649.

——— (1998), "Non-Parametric MLE for Poisson Species Abundance Models Allowing for Heterogeneity Between Species," *Environmental and Ecological Statistics*, 5, 391–402.

Otis, D. L., Burnham, K. P., White, G. C., and Anderson, D. R. (1978), "Statistical Inference From Capture Data on Closed Animal Populations," *Wildlife Monographs*, 62, 1–135.

Pledger, S. (2000), "Unified Maximum Likelihood Estimates for Closed Capture-Recapture Models Using Mixtures," *Biometrics*, 56, 434–442.

Sanathanan, L. (1972), "Estimating the Size of a Multinomial Population," *The Annals of Mathematical Statistics*, 43, 142–152.

——— (1977), "Estimating the Size of a Truncated Sample," *Journal of American Statistical Association*, 72, 669–672.

Smith, E. P., and van Belle, G. (1984), "Nonparametric Estimation of Species Richness," *Biometrics*, 40, 119–129.

Wang, J.-P. Z. (2003), "NPMLE in Estimating the Number of Expressed Genes Using EST Data While Accounting for Measurement Error," unpublished doctoral thesis, Pennsylvania State University.

Wang, J.-P. Z., Lindsay, B. G., Leebens-Mack, J., Cui, L., Wall, K., Webb, C. M., and de Pamphilis, C. W. (2004), "EST Clustering Error Evaluation and Correction," *Bioinformatics*, 20, 2973–2984.