Computational Statistics and Data Analysis 56 (2012) 2562–2573

Contents lists available at SciVerse ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Sampling designs via a multivariate hypergeometric-Dirichlet process model for a multi-species assemblage with unknown heterogeneity

Hongmei Zhang [a,*], Kaushik Ghosh [b], Pulak Ghosh [c]

[a] Department of Epidemiology and Biostatistics, University of South Carolina, Columbia, SC 29208, USA

[b] Department of Mathematical Sciences, University of Nevada Las Vegas, Las Vegas, NV 89154, USA

[c] Department of Quantitative Methods & Information System, Indian Institute of Management, Bangalore 560076, India

ARTICLE INFO

Article history:
Received 18 May 2011
Received in revised form 14 February 2012
Accepted 15 February 2012
Available online 24 February 2012

Keywords:
Cluster analysis
Sample size
Dirichlet process
Sequence tags
Multivariate hypergeometric distribution
Multinomial distribution

ABSTRACT

In a sample of mRNA species counts, sequences without duplicates or with small numbers of copies are likely to carry information related to mutations or diseases and can be of great interest. However, in some situations, sequence abundance is unknown and sequencing the whole sample to find the rare sequences is not practically possible. To collect mRNA sequences of interest, or more generally, species of interest, we propose a two-phase Bayesian sampling method that addresses these concerns. The first phase of the design is used to infer sequence (species) abundance levels through a cluster analysis applied to a pilot data set. The clustering method is built upon a multivariate hypergeometric model with a Dirichlet process prior for species relative frequencies. The second phase, through Monte Carlo simulations, infers the sample size necessary to collect a certain number of species of particular interest. Efficient posterior computing schemes are proposed. The developed approach is demonstrated and evaluated via simulations. An mRNA segment data set is used to illustrate and motivate the proposed sampling method.

1. Introduction

The work presented is motivated by an mRNA sequencing project conducted in a cancer therapy company. The company's proprietary technology yielded a maximum of 80,000 sequence tags of fixed length from mRNAs in different cell populations. Different mRNA tags were represented with different frequencies. The primary interest was in the diversity of mRNA species (distinct mRNA tags). By randomly selecting a subset of tags for sequencing, the ultimate goal was to estimate the required sequencing burden to ensure a collection of a certain number of mRNA species of particular interest. In this application, "species of particular interest" refers to rare sequences. However, the abundance levels of these mRNA sequences are unknown apriori. To fulfill the task, a two-phase sampling plan seems necessary: inferring species abundance levels utilizing information in the collected data, followed by designing sequencing plans based on the inferred levels.

In the species diversity literature, inferring species abundance levels is an imperative step. Overlooking the diversity takes the risk of producing biased inferences (Morris et al., 2003; Böhning and Schön, 2005; Wang, 2010). In some cases, proportions of each individual species in a population are inferred to represent species abundance (Boender and Rinnooy Kan, 1987; Brown et al., 1993; Basu and Ebrahimi, 2001; Besbeas et al., 2002; Chao and Shen, 2003; Morris et al., 2003; Ghosh and Norris, 2005). Others move one step further. Since species with different proportions can be at the same abundance level, instead of inferring proportions, a collection of species are assigned to groups according to their abundance levels, for instance, see Norris and Pollock (1996) and Pledger et al. (2003, 2010). This is the direction taken in our work. To

* Corresponding author. Tel.: +1 803 777 3823; fax: +1 803 777 2524.
E-mail addresses: hzhang@sc.edu, hzhang@mailbox.sc.edu (H. Zhang).

0167-9473/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2012.02.013

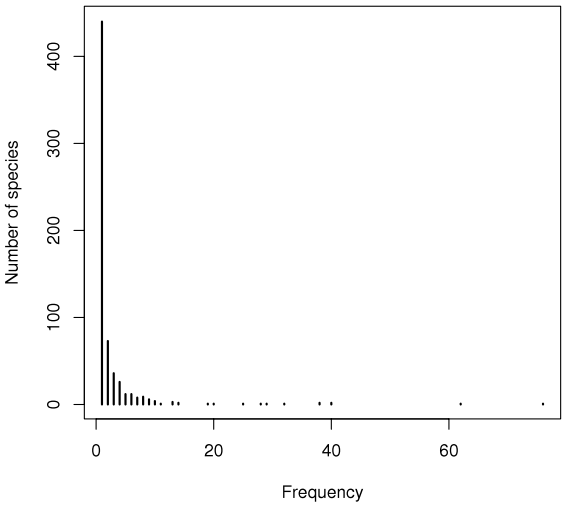

Fig. 1. The distribution of the frequencies for the mRNA sequence data. For instance, the first line indicates 440 mRNA species each with 1 occurrence and the last dot represents 1 mRNA species with 76 occurrences.

determine species abundance levels, some studies rely on information of sample data (Cao et al., 2001; Green and Young, 1993). This can produce misleading inferences, since a species being rare in a sample may not be rare in the population, and *vice versa*. There exists a rich literature on estimating species diversity. Maximum likelihood based nonparametric methods have been proposed; the diversity estimators are rooted in the mixture of binomial distributions (Norris and Pollock, 1996) or built upon mixed Poisson processes (Böhning and Schön, 2005). In these cases, the abundances of species were grouped through a collection of points using point masses. Parametric methods using mixture of distribution functions to describe heterogeneity have been developed as well; for instance, Wang (2010) proposed gamma mixtures with the help of a partial prior approach via exponentials to model diversity, and Pledger et al. (2003, 2010) and Quince et al. (2008) utilized a finite mixture with a distribution function incorporated describing abundance levels.

Most existing methods for inferring species diversity assume infinite population size and, when binomial or multinomial distributions are assumed, each species in the population is assumed to have a sufficiently large number of representatives. However, these assumptions may not be satisfied in many real situations. The mRNA sequence tags data introduced earlier is one of many examples. Fig. 1 shows the distribution pattern of sample frequencies of mRNA species. The sharp decreasing pattern implies that the mRNA population is likely to include a large portion of rare species and very few abundant species. Other examples with a similar nature include data from Serial Analysis of Gene Expression (SAGE) and rare animal species in a geographical region. Furthermore, the existing methods are not able to simultaneously infer abundance levels (clusters) and the number of abundance levels (number of clusters), which to some extent reduced their flexibility and efficiency. To address these limitations, we utilize a multivariate hypergeometric distribution, a direction not sufficiently explored in literature in species heterogeneity modeling. This distribution allows the existence of rare species and admits the fact of finite population size. We will infer the clusters and the number of clusters simultaneously through the Dirichlet Process mixtures (DPM) (Antoniak, 1974; Ferguson, 1973). DPM has been widely applied to cluster populations of interest, for instance, (Dorazio et al., 2008; MacEachern and Müller, 1998; West et al., 1994), or population features under study as in Trippa and Parmigiani (2011), but in many cases DPM was applied to independent means instead of dependent relative frequencies as in our case. In this article, through variable transformation, we apply DPM to dependent relative frequencies. The reversible jump Markov Chain Monte Carlo (RJMCMC) (Green, 1995) is another way to infer the number of clusters, but the computing burden and a possible failure of true convergence limits this method from being widely applied (Neal, 2000). Using DPM, on the other hand, is expected to be more efficient and can result in a more simplified grouping process.

Turning attention to sampling plan designs, various methods to sample species with different abundance levels have been proposed (Christman and Lan, 2001; McArdle, 1990). Commonly used approaches include sequential sampling (which produces a biased sampling approach, see Christman and Lan, 1998), adaptive clustering sampling (Thompson, 1990, 1991; Turk and Borkowski, 2005), and their extensions (Christman and Lan, 2001). However, in these studies, the species to be captured are pre-specified and their abundance is known, a situation different from ours. The sampling scheme utilized in this work is random sampling. In the first step, the sampling effort to collect the pilot data is fixed and the number of individuals observed is random; in the second step, the sampling plan is to collect a certain number of species from one cluster or several clusters inferred in the first step. The sampling effort in this step is random and dependent on the amount of information that needs to be collected from the species of interest. A Monte Carlo simulation-based approach adapted from Zhang and Stern (2009) will be used to design sampling plans for species collection. Methods to collect species regardless of species abundance levels are not the focus of the present work, and interested readers are referred to studies, for instance, by Mao and Colwell (2005), Lijoi et al. (2007a), Quince et al. (2008), and Zhang (2009).

The novelty of our method lies in its ability to classify a species' population without any assumption on the population structure and then collect a group of species of particular interest from the population, e.g. rare species or abundant species. Such inferences can be rather attractive in various areas such as genetics, where gathering sufficient information on mutated genes may be of great interest, microbial taxa recovery (Quince et al., 2008), marine biology with interest in abundance levels of species such as benthic macroinvertebrates (Durell et al., 2005), or ecology where the study interest is an endangered species.

The road map of the remainder of this article is as follows. In Section 2, we develop the hierarchical Bayesian model for species data and describe our choice of prior distributions. Section 2.3 focuses on posterior inferences for the model parameters, and Section 2.4 describes a method of cluster determination. Issues related to the implementation of MCMC are also discussed. In Section 3, we develop a Monte Carlo simulation approach to designing future data collection. Section 4 demonstrates the proposed method through simulations. The sensitivity of results to various population settings is also discussed. We apply our method to the mRNA sequence tags data set in Section 5. Finally we summarize our results in Section 6.

2. A multivariate hypergeometric model with Dirichlet process mixtures

In a population with N_0 individuals representing S species, there exist $N_i,\ i = 1, \ldots, S$, representatives of each species such that $\sum_{i=1}^{S} N_i = N_0$. To focus on inferring abundance levels and designing sampling plans, we assume N_0 and S in the population are known. Inferring N_0 and S is an active research field of its own, for instance, see Chao (1987), Mao (2007a), and Williamson and Slatkin (1999) for population size estimation, and Bunge and Fitzpatrick (1993), Chao and Bunge (2002), Efron and Thisted (1976), Fisher et al. (1943), Mao and Colwell (2005), Mao (2006), Mao (2007b), Quince et al. (2008), and Zhang and Stern (2009) for estimating the number of species.

2.1. Likelihood function

From the population described above, through a random sampling scheme with a fixed sampling effort, M individuals from s_o species are collected. Let $y_i,\ i = 1, \ldots, s_o$, be the number of observed individuals from species i. We use the same index i as that used for the population to label the observed species, since we can always re-label to make the population and the sample consistent. Then $\mathbf{y} = (y_1, \ldots, y_i, \ldots, y_{s_o})$ is one way to represent the observed sample. Note that $\sum_{j=1}^{s_o} y_j = M$.

In many situations, e.g., mRNA sequencing or animal capturing, we are sampling without replacement. In a population with a finite size, the distribution of $\mathbf{y} = (y_1, \ldots, y_{s_o})$ can be reasonably assumed to be multivariate hypergeometric with parameter $\mathbf{N} = (N_1, \ldots, N_S)$, denoted as MH($\mathbf{N}$),

$$\Pr(\mathbf{y}|\mathbf{N}) = \mathrm{MH}(\mathbf{N}) = \frac{\binom{N_1}{y_1} \cdots \binom{N_{s_o}}{y_{s_o}} \binom{N_{s_o+1}}{y_{s_o+1}} \cdots \binom{N_S}{y_S}}{\binom{N_0}{M}}, \tag{1}$$

in which the order of $(N_1, \ldots, N_{s_o})$ corresponds to the order in $\mathbf{y}$. The remaining categories $N_{s_o+1}, \ldots, N_S$ are not present in the sample. Note that mass function (1) can be approximated by a multinomial probability mass function with parameters $(N_0, N_i/N_0 | i = 1, \ldots, S)$ if N_0 is large and N_i is not small. In the following sections, we propose a fully Bayesian approach to draw inference on $\mathbf{N}$ such that each species belongs to a specific abundance group.

2.2. Prior and hyperprior distributions

Prior distribution of $\mathbf{N}$: To select a prior distribution for $\mathbf{N}$, we can treat the population under consideration as a random sample of size N_0 from a hyper-population with S categories. The prior distribution for $\mathbf{N}$ is selected such that $\mathbf{N} - \mathbf{1}$ is multinomially distributed with parameter $\boldsymbol{\theta}$ of dimension S, $\mathbf{N} - \mathbf{1} \sim MN(N_0 - S, \boldsymbol{\theta})$. Thus

$$\Pr(\mathbf{N} - \mathbf{1}|S, \boldsymbol{\theta}) = \frac{(N_0 - S)!}{(N_1 - 1)! \cdots (N_S - 1)!} \theta_1^{(N_1-1)} \cdots \theta_S^{(N_S-1)}, \quad 0 \le \theta_i \le 1,$$

where $\sum_{i=1}^{S} \theta_i = 1$ and each θ_i denotes the proportion of N_0 individuals that are members of the ith species. This prior distribution of $\mathbf{N}$ is non-informative or vague (Section 2.9 on page 61 in Gelman et al., 2003), and selected solely based on the assumption of random sampling scheme. The prior guarantees that each species has at least one and at most $N_0 - S + 1$ (assuming $N_0 > S$) representatives.

Another natural choice of the prior distribution for $\mathbf{N}$ is a "truncated" multinomial distribution,

$$\mathbf{N}|\boldsymbol{\theta} \sim q(\boldsymbol{\theta}) MN(N_0, \boldsymbol{\theta}), \quad 1 \le N_i \le N_0 - S + 1, \tag{2}$$

where the conditional normalizing constant $q(\boldsymbol{\theta})$ is formulated through a summation of all the probabilities such that at least one N_i $(i = 1, \ldots, S)$ in $\mathbf{N}$ is outside the constraints. Specifically, let $A = (N_0, N_0 - 1, \ldots, N_0 - (S-2), 0)$ be the set containing the values of N_i outside the bound $1 \le N_i \le N_0 - S + 1$. Then $q(\boldsymbol{\theta})$ can be written as

$$q(\boldsymbol{\theta}) = \left[1 - \sum_{\substack{\mathbf{N} | N_i \in A, \\ \forall i \in (1,\ldots,S)}} \frac{N_0!}{N_1! \cdots N_S!} \theta_1^{N_1} \cdots \theta_S^{N_S} \right]^{-1}. \tag{3}$$

The number of additions in the evaluation of $q(\boldsymbol{\theta})$ is in the order of $O(S^2 N_0 I_{\{S>2\}} + S)$. Hence, the computational burden of (3) increases quickly with the increase of N_0 and S. This prior distribution thus can only be applied to small populations.

Hyperprior distributions: The hyperprior parameter vector $\boldsymbol{\theta}$ will be inferred non-parametrically utilizing a Dirichlet Process (DP) prior. DP provides a non-parametric prior in the space of distribution functions and gives rise to a more flexible class of models than would be obtained by parametric Bayes approaches. However, since the individual θ_i's, $i = 1, \ldots, S$, are dependent, DP cannot be directly applied. To solve this problem, we consider the method of variable transformation. Let $\mathbf{a} = (a_1, \ldots, a_S)$ be a vector of S independent random variables with $a_i > 0,\ i = 1, \ldots, S$. We assume $(a_1, \ldots, a_S) | G \overset{\text{i.i.d.}}{\sim} G$, and $G|(\alpha, G_0) \sim \text{DP}(\alpha, G_0)$, where α is a precision parameter and G_0 is the base distribution. The quantity G_0 is the "center" of the prior distribution on G and α controls the variability about this center. The precision parameter also controls the degree of heterogeneity of a_i's. G_0 is assumed to be Gamma(τ, ϕ) with uniform priors, $\tau \sim$ Uniform (L_τ, U_τ) and $\phi \sim$ Uniform (L_ϕ, U_ϕ). The lower bounds and upper bounds in the uniform distributions are positive and assumed to be known. In this work, they are selected to achieve large widths of the uniform distributions, hence are less informative, without affecting the efficiency of convergence. Inferences on precision parameter α will be discussed later in this section.

Once $(a_1, \ldots, a_S)$ are obtained, θ_i is determined by

$$\theta_i | (a_1, \ldots, a_S) = \frac{a_i}{a_1 + \cdots + a_S}, \tag{4}$$

since θ_i must satisfy $\sum_{i=1}^{S} \theta_i = 1$. Due to the inherent clustering property of samples drawn from a distribution with DP prior (Escobar and West, 1995; Ferguson, 1973; West et al., 1994), the values of $a_1, \ldots, a_S$ are clustered into different groups. This results in clustering of corresponding θ_i's due to the relationship between θ_i and a_i.

Conditional on α, the hierarchial model proposed above can thus be fully specified as follows:

$$\begin{aligned}
&\mathbf{y}|\mathbf{N} \sim \text{MH}(\mathbf{N}),\\
&\mathbf{N} - \mathbf{1}|\boldsymbol{\theta} \sim MN(N_0 - S, \boldsymbol{\theta}),\\
&\theta_i | (a_1, \ldots, a_S) = \frac{a_i}{a_1 + \cdots + a_S},\\
&a_i | G \overset{\text{i.i.d.}}{\sim} G,\\
&G|G_0,\ \alpha \sim \text{DP}(\alpha,\ G_0),\\
&G_0 = \text{Gamma}(\tau, \phi).
\end{aligned} \tag{5}$$

2.3. Posterior inferences

The posterior distribution of the proposed semiparametric Bayesian model lacks a closed-form analytical expression. Hence, we use Markov chain Monte Carlo (MCMC) methods, specifically the Gibbs sampler with included Metropolis–Hastings steps, to generate observations from full conditional posterior distributions, which are then used to infer the parameters of interest. The conditional posterior distribution of (N_i, N_j) is given by

$$f(N_i, N_j | \cdots) \propto \frac{\binom{N_i}{y_i}\binom{N_j}{y_j}}{(N_i - 1)!(N_j - 1)!} \theta_i^{N_i - 1} \theta_j^{N_j - 1}, \quad i \ne j,\ N_i + N_j = N_0 - \sum_{l \ne i,j} N_l, \tag{6}$$

where "$\cdots$" refers to data and the remaining parameters. To generate proposals of (N_i, N_j), we select a random walk as the proposal distribution. The details on the derivation of the full conditional distribution (6) are given in the Appendix.

The conditional posterior distribution of $(a_1, \ldots, a_S)$ is given by

$$f(a_1, \ldots, a_S | \cdots) \propto \frac{\prod_{i=1}^{S} a_i^{N_i - 1}}{\left(\sum_{i=1}^{S} a_i \right)^{N_0 - S}} \pi(a_1, \ldots, a_S | \alpha, G_0), \tag{7}$$

where $\pi(a_1, \ldots, a_S)$ is the prior of $\mathbf{a}$, given by the Dirichlet process prior as:

$$\pi(a_1, \ldots, a_S|\alpha, G_0) = \prod_{i=1}^{S} \left[\frac{\alpha G_0(da_i) + \sum_{j=1}^{i-1} \delta_{a_j}(da_i)}{\alpha + i - 1} \right].$$

The non-conjugacy in the base prior for the Dirichlet process is handled by using auxiliary parameters, along the line of Algorithm 8 in Neal (2000). The Gibbs sampler proceeds by sequentially repeating the sampling of N_i's according to (6), sampling of a_i's from (7), and updating θ_i using (4).

Finally, we discuss the inference for the precision parameter α. Sensitivity of the posterior inferences of α to its prior choice has been discussed in various applications (Dorazio et al., 2008; Liu, 1996; McAuliffe et al., 2006). As noted earlier, α controls the degree of heterogeneity of the a_i's and plays an important role in the model. Doss (2008) indicates that parameter α is typically the most difficult to estimate or defend as a fixed value. Nevertheless, we follow the suggestion by Dorazio et al. (2008) and adopt their empirical Bayes approach to obtain an estimate of α, $\hat{\alpha}$, which is the maximum likelihood estimate (MLE) of α. The MLE of α is numerically obtained during the Gibbs sampling process by solving $\bar{C} = \sum_{i=1}^{M} \hat{\alpha}/(\hat{\alpha} + i - 1)$, where $\bar{C}$ is the mean of number of clusters after a certain number of MCMC iterations. Details of this empirical Bayes method can be found in Section 3.1 in Dorazio et al. (2008). Other approaches for determining the precision parameter have been discussed elsewhere (Doss, 2008, 2012; Kyung et al., 2010). It is noteworthy that the choice of α based on Dorazio et al. (2008) may be far from the truth due to the possibility of flat likelihood of α (Kyung et al., 2010).

2.4. Determining the number of clusters

At each iteration of the MCMC simulations, the nature of DP automatically clusters species according to abundance levels defined via a_i, $i = 1, \ldots, S$. To make a final selection on the number of clusters together with corresponding estimates of parameters, we consider the following procedure adapted from Dahl (2006), a procedure based on the method of "least-squares clustering":

1. After the MCMC burn-in, continue the MCMC simulations for an additional B iterations. Let A denote an $S \times S$ matrix. The (i, j)th entry of A is the proportion of iterations such that species s_i and s_j $(i, j = 1, \ldots, S)$ are in the same cluster. The matrix A is referred to as an averaged clustering matrix.
2. Continue to run an additional D iterations of the MCMC simulations. For each iteration,
   (a) form an $S \times S$ matrix composed of indicators of clustering for that particular iteration. For instance, if species s_i and s_j are in one cluster, then the (i, j)th entry is 1; otherwise, it is zero.
   (b) calculate the Euclidean distance between the matrix formed above and the averaged clustering matrix A.
3. Sort the Euclidean distances obtained from the D iterations, and the final selection on the number of clusters is in favor of simpler clusters and relatively small Euclidean distances. The parameters will then be inferred accordingly based on the identified clusters. From our extensive simulations, we found that usually the clusters corresponding to the smallest distances are more detailed and the number of clusters is large, which may not be informative enough to differentiate between different species abundance levels.

3. Monte Carlo simulation-based sampling designs

Suppose it is possible to collect additional data beyond the initial M observations. The sampling design discussed below is to infer the size of an additional sample to collect a pre-specified number or a proportion p of species at a specific abundance level.

Through appropriate labeling during the process of clustering, the correspondence between the collected species and the identified clusters can be constructed. After posterior samples are drawn for parameter vectors $\mathbf{N}$ and $\boldsymbol{\theta}$, we are able to identify cluster(s) of interest based on $\boldsymbol{\theta}$ and start collecting species from those clusters.

Assume there are C clusters in $\boldsymbol{\theta}$ such that $\sum_{c=1}^{C} S_c = S$, where S_c is the number of species in cluster c. The clusters are determined using the method discussed in Section 2.4. The probability of collecting p proportion of species from cluster c with additional collection of M_a individuals is given by

$$\begin{aligned}\pi(M_a|\boldsymbol{\theta}, \mathbf{y}) &= P\left(\left[\sum_{i=1}^{s_o} I\{i \in c\} + \sum_{i=s_o+1}^{s_o+S_{\text{new}}} I\{i \in c\} \right] \geq pS_c | M_a, \mathbf{y}, \boldsymbol{\theta} \right) \\ &= \int I\left(\left[\sum_{i=1}^{s_o} I\{i \in c\} + \sum_{i=s_o+1}^{s_o+S_{\text{new}}} I\{i \in c\} \right] \geq pS_c \right) P(\mathbf{y}^*|\mathbf{N}, M_a) P(\mathbf{N}|\mathbf{y}, \boldsymbol{\theta}) d\mathbf{N} d\mathbf{y}^*, \end{aligned} \tag{8}$$

where S_{new} is the number of newly collected species, and $I\{i \in c\}$ is an indicator function with $I\{i \in c\} = 1$ if a collected species is from cluster c.

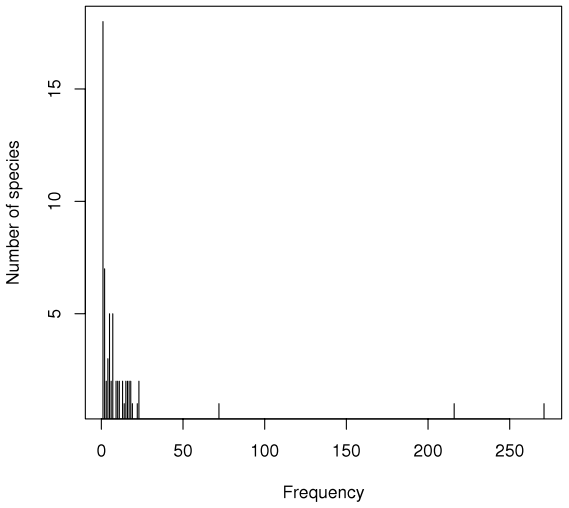

Fig. 2. Distribution of sample frequencies for the simulated data set.

To collect p proportion of species from cluster c with probability 1, we set $\pi(M_a|\boldsymbol{\theta}, \mathbf{y}) = 1$. Because of the redundancy of $\mathbf{N}$ given $\boldsymbol{\theta}$, the posterior distribution of M_a can be estimated based on the posterior distribution of $\boldsymbol{\theta}$ without any inferences on $\mathbf{N}$. A Monte Carlo method can be used for this purpose. Assume there are T posterior samples of $\boldsymbol{\theta}$. For the tth, $t = 1, \ldots, T$, posterior sample,

1. generate a certain number of $\mathbf{N}$'s from $p(\mathbf{N}|\boldsymbol{\theta}, \mathbf{y})$.
2. for each $\mathbf{N}$ generate $\mathbf{y}^*$ from multivariate hypergeometric distribution with parameter $\mathbf{N}$ and M_a.
3. count the number of new categories from cluster c; if the proportion of success is less than 1, we increase M_a by $g \geq 1$, and repeat steps 1–3 until the proportion is 1.

Repeat steps 1–3 for each posterior draw of $\boldsymbol{\theta}$. The choice of g is affected by sampling needs and the feature of a population to be sampled from. In our application, g is chosen to be 10 due to the large population size of mRNA tags. With T posterior samples of $\boldsymbol{\theta}$, we are able to estimate the posterior distribution of M_a, of which the median is used as an estimate of M_a. An empirical 95% posterior interval will be used to evaluate the uncertainty. Alternatively, instead of relying on posterior samples of $\boldsymbol{\theta}$, steps 1 and 2 can be simplified by using posterior samples of $\mathbf{N}$. In this case, we draw multiple samples of $\mathbf{y}^*$ from each posterior sample of $\mathbf{N}$ using multivariate hypergeometric distribution. Based on the discussion given in Section 2.2, these two approaches are equivalent in terms of sample size determinations. The approach based on posterior samples of $\mathbf{N}$, however, avoids the use of the non-standard probability mass function $p(\mathbf{N}|\boldsymbol{\theta}, \mathbf{y})$.

With straightforward modifications, the sampling scheme discussed above can be applied to different situations. For instance, it can be generalized to collect species from multiple abundance levels, or if the sampling interest is on any species, we can simply remove the restriction $I\{i \in c\}$ from (8) to fit the need.

4. Simulations

We demonstrate the proposed method through simulations. We focus on two aspects. One aspect is to examine the inferences on clusters and population properties, and the other is to evaluate the sampling designs to collect additional samples of species of interest. In the following, we present a simulated data set and briefly discuss the results.

4.1. Simulated data

We simulate a population of size $N_0 = 5000$ with $S = 100$ species. The population data are generated from multinomial distributions such that the parameter vector $\boldsymbol{\theta}_0$ has five unique values: $\theta_{0,1}$, $\theta_{0,2} = 0.25$, $\theta_{0,3} - \theta_{0,32} = 0.01$, $\theta_{0,33} - \theta_{0,69} = 0.001$, $\theta_{0,70} = 0.148$, and $\theta_{0,71} - \theta_{0,100} = 0.0005$. Since $\boldsymbol{\theta}_0$ is used to generate population data, conceptually it is different from $\boldsymbol{\theta}$ in the prior distribution of $\mathbf{N}$. With this scenario, $M = 1000$ individuals are sampled and $S_o = 66$ species are observed. Fig. 2 displays the distributions of sample frequencies.

4.2. Results

We apply the method discussed in earlier sections to the simulated data. We run one long Markov Chain, in which 200,000 iterations are used as burn-in, up to 100,000 iterations are used to estimate the precision parameter α, followed by 20,000

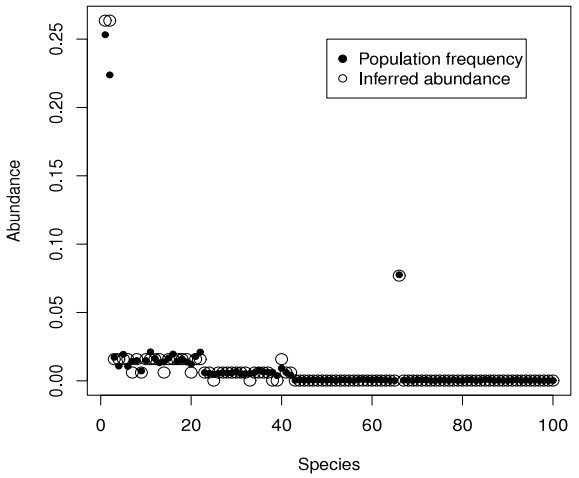

Fig. 3. The true and inferred population abundance.

Table 1
Sizes of additional samples to collect proportion p of rare species ($\hat{\theta}_r = 0.000187$).

p	0.6	0.7	0.8	0.9
$\hat{M}_a$	1120	1735	2480	3320
95% P.I.	(830, 1471)	(1390, 2220)	(2140, 2971)	(2959, 3612)

iterations to calculate the average clustering matrix. Finally, $D = 5000$ iterations are run to determine the number of clusters and make related inferences on parameters.

The estimated precision parameter for this data set is $\hat{\alpha} = 2.63$. Comparing the Euclidean distances (the minimum distance is 14.87) and the complexity of clusters, we decide to choose five clusters corresponding to Euclidean distance of 15.32, the minimum distance among all five clusters. A 95% posterior interval for the number of clusters is (Boender and Rinnooy Kan, 1987; Chao and Bunge, 2002). To visualize the quality of the selected clusters, Fig. 3 displays the pattern of population heterogeneity and that of the inferred (Fig. 3), which indicates that the inferred abundance agrees nicely with the abundance pattern in the population. Table A.1 in the Appendix lists the posterior estimates of $\boldsymbol{\theta}$ together with posterior intervals.

Once clusters are identified, sampling plans are designed based on the posterior samples. As an illustration, we estimate the size of additional samples required to collect proportions $p = 0.60, 0.70, 0.80, 0.90$ of all species belonging to the rarest species category ($\theta_r = 0.000187$) with probability 1, which corresponds to collecting another $m = 10, 15, 22$, and 28 new rare species, respectively. Here, the rare species category is chosen to be the category with the smallest proportion. Some studies pre-specify the cutoff for rare species, for instance, Cao et al. (2001), Green and Young (1993), and Favaro et al. (2011). Zhang (2009) has a discussion on the definition of rare species.

The method discussed in Section 3 is applied to infer the number of additional samples. The value of g is set at $g = 10$. Table 1 lists the estimates of M_a together with 95% posterior intervals. The sample size increases quickly when a larger proportion p of rare species is to be collected.

The sample sizes inferred and listed in Table 1 all correspond to collecting the needed species with probability 1. For a given M_a, the probability of collecting needed species based on M_a additional samples can be calculated using (8). We name this probability as collecting confidence. If the probability is 1, then we have full collecting confidence with M_a additional samples. Following the similar procedure as described in Section 3 for a set of M_a, we are able to evaluate the relationship between the collecting confidence and the sizes of additional samples. Fig. 4 shows the increase of estimated collecting confidence with the increase of additional sample sizes for different values of p. The pattern displayed in Fig. 4 agrees with the findings listed in Table 1 and is as expected. For a given confidence, the sample sizes increase quickly if a large proportion of species are to be collected. Furthermore, for smaller proportions, the confidence reaches 1 more quickly compared to larger proportions.

We performed additional simulation studies to examine the effect of sample size. Different sample sizes were considered and 50 data sets were simulated for each sample size based on the scenario discussed earlier. A clear pattern was identified, that is, the consistency between the truth and the estimated clusters increases dramatically when the sample size increases. For sampling designs, our simulation results indicate that the required future sample sizes decrease along with the increase of the size of the initial sample, which is likely due to the fact that more rare species are collected in the sample (results not shown).

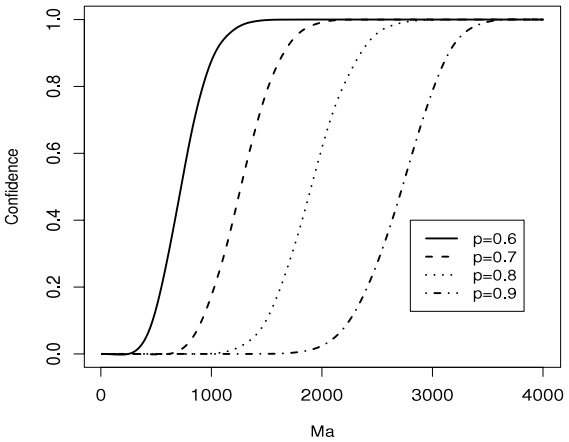

Fig. 4. The relationship between the collecting confidence and additional sample sizes for different values of p.

Table 2
Species clustering and the proportions corresponding to each cluster (mRNA segments data).

Cluster	Abundance level	Number of species (%)
1	0.023	2 (0.078)
2	0.012	7 (0.27)
3	0.0050	20 (0.78)
4	0.0034	106 (4.16)
5	0.0027	45 (1.76)
6	0.0020	56 (2.20)
7	0.0018	24 (0.94)
8	0.0014	32 (1.25)
9	0.0011	52 (2.04)
10	0.00058	2 (0.078)
11	0.00030	1 (0.039)
12	0.000012	2203 (86.39)

We also studied the impact of misspecification of S on the inference of population heterogeneity (results not shown). The inferred species abundance levels are consistent with the truth and the consistency is not affected by the choice of S. Such consistency is also observed in our real data application discussed below.

5. Application

We apply the proposed method to the mRNA sequence tag data introduced in Section 1. A prototype data set was provided with sample size $M = 1677$ and $s_o = 644$. Fig. 1 in Section 1 displays the distribution pattern of sample frequencies, which shows a very sharp decreasing pattern. A few mRNA species were observed with high frequencies, and a very high proportion of the observed species were only observed once. As given earlier, the upper bound of the number of sequences is $N_0 = 80,000$ (the upper bound of population size). We noticed that although using the upper bound is likely not to change the pattern of clusters, in this case, we are inferring the maximum size of additional samples.

The number of mRNA species is unknown. Since estimating the number of species is not the focus of this article, we assume it is known and apply the results from our earlier work (Zhang, 2007). The same data set was discussed there and the number of unique sequences is estimated as $\hat{S} = 2550$ with a 95% posterior interval of (2323, 2784) based on a generalized multinomial distribution. We will first draw inferences with $S = 2550$ and then discuss the change of inferences if S is misspecified.

The estimate of α is $\hat{\alpha} = 2.45$. Based on the consideration of cluster simplicity and the comparison of Euclidean distances, 12 clusters are identified and the ratio between the corresponding distance and the minimum distance is 1.01. A 95% empirical posterior interval is (Chao and Shen, 2003; Ferguson, 1973). Table 2 lists the abundance levels of each cluster and the number of species in each cluster. The percentage of species included in each cluster is given in parentheses. Among these 12 clusters, not surprisingly, most observed species are generally assigned to clusters corresponding to relatively large abundance levels. The first two species corresponding to $y_1 = 62$ and $y_2 = 76$ are clustered together with the highest abundance level of 0.023. Among the 2550 species, 2203 species are in one cluster with the lowest abundance level of 1.19×10^{-5}, of which 1899 such species were not observed. This result implies that the population is likely to have a large number of rare species, which is consistent with the sharp decreasing pattern of the sample frequency distribution (Fig. 1) and supports the findings discussed in Zhang and Stern (2009).

Table 3
Sizes of additional samples to collect proportion p of rare species (mRNA segments data).

Inferences	Proportions (p)			
	0.5	0.7	0.8	0.9
M_a (median)	20 300	36 300	46 700	59 900
95% Post. Interval	(19 095, 21 315)	(34 690, 37 900)	(44 895, 48 300)	(58 100, 61 505)

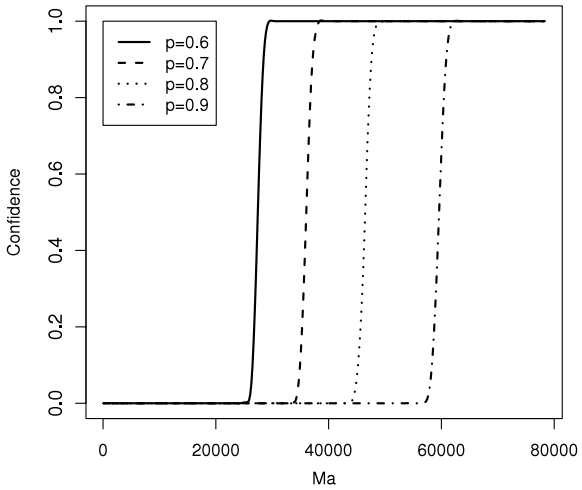

Fig. 5. The relationship between the collecting confidence and additional sample sizes for different values of p (real data).

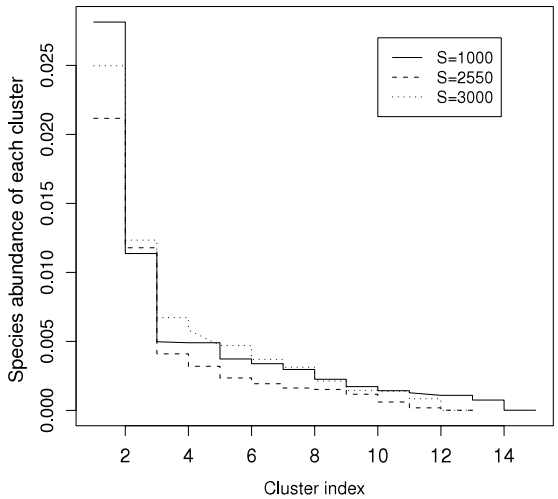

Fig. 6. The abundance pattern for different assumption of S.

The sample size calculation is carried out to sample a certain proportion of mRNA segments from the rarest mRNA species (abundance level of 0.000012). The motivation of sampling this type of mRNA segments is that these sequences may carry mutation information related to the disease under study. Collecting rare species is also of interest to researchers in other fields (Cao et al., 2001; Green and Young, 1993; Cao et al., 2001). Table 3 lists required sample sizes to collect various proportions (p) of the rarest species (with probability 1) from the population. The sample size increases quickly with the increase of p. Clearly, since the species of interest is rare, a larger sample size is needed to collect more such species. Given the large population size of up to 80,000, the required sample size is reasonable and practically useful even to collect 90% of the rare mRNA sequences. In addition, we notice that with p fixed, the increase of sample size seems insensitive to the change of collecting confidence as indicated in Fig. 5, which can be explained as a consequence of the very low abundance level of those species.

The above analysis is based on $S = 2550$. In the following, we investigate the stability of clustering patterns for different choices of S. As given by Zhang (2007), a 95% posterior interval of S is (2323, 2784). We consider two values of S outside the interval, specifically, $S = 1000$ and $S = 3000$. Fig. 6 illustrates the pattern of identified species abundance levels under different assumptions of S. The numbers of clusters vary with S, but overall the patterns of cluster profiles are similar

for different S. This pattern agrees with the findings from our simulation studies. We also observe that regardless of the assumptions on S, the two common species collected in the pilot data are always clustered into one abundance level. In addition, the proportion of rare species increases with S. This is likely due to the severely skewed distribution pattern revealed by the observed data. In summary, the main pattern noted above is consistent with previous findings (Zhang and Stern, 2009) and stays the same regardless of the value of S, i.e., a very small number of species are included in the cluster of common species; and a very large portion of species are clustered as rare species.

6. Discussion

Motivated by an mRNA sequencing project, a multivariate hypergeometric-Dirichlet process model is proposed and applied to count data to collect desired species. The sample size determination is based on a two-phase design. In phase 1, a multivariate hypergeometric distribution is used to model the observed counts and a Dirichlet Process (DP) is applied to group the populations into different clusters based on individual species' relative frequencies. The DP is used to generate the prior distribution for the proportions representing species abundance levels. An empirical Bayes method is used to infer the precision parameter (Dorazio et al., 2008). In the second phase, a Monte Carlo simulation approach is applied to determine the minimum sample size in order to collect additional samples from a set of species of interest.

The developed method has several advantages. First, it is not limited to studies in genetics, but can be applied to other fields as well such as ecology, biology, and literature (Cao et al., 2001; Efron and Thisted, 1976; Fisher et al., 1943; Montagna and Ritter, 2006; Quince et al., 2008). For instance, the inferred clusters can be used to efficiently assess or compare ecological balance in different regions, or the consistency of vocabulary use in different works of the same author. The proposed sampling plans can be applied to infer the possibility of collecting one type of species of interest, which will help researchers to determine the feasibility of their proposed studies, e.g., the feasibility of discovering novel species (Quince et al., 2008; Amaral-Zettler et al., 2009). The sampling plans may also be applied to examine the rareness of species of interest in a specific geographical region. Second, the proposed sampling plan is not limited to rare species collection but can be used to collect species at any abundance levels of interest. Third, the method does not require a large number of representatives of each species in the population, and the transformation of variables makes it possible to apply DP to dependent variables. Finally, the proposed procedure to draw posterior samples of $\mathbf{N}$ dramatically increases the MCMC sampling speed, which overcomes the calculation burden when dealing with multivariate hypergeometric models. Through simulations, it is found that the hypergeometric-Dirichlet process model works well and the sampling plans are feasible over a range of scenarios. An application to an mRNA segments data set further demonstrates the applicability of the proposed method.

The introduced sampling plan is built upon the species abundance levels inferred in the first phase. The correctness of the estimated size of additional samples relies on the credibility of the estimated population heterogeneity. Our simulation results indicate that the estimated population heterogeneity agrees with the truth reasonably well especially when we have large samples. In addition, although it is not the focus of our study, it can be an interesting direction for a future endeavor to infer the number of species in a population with unknown heterogeneity. Lijoi et al. (2007b) brings the idea of Gibbs-type partition, such as DP, into the sampling process to identify new species. This provides a potential to infer the number of unknown species. However, the starting point of Lijoi et al. (2007b) is the randomness of the species proportions and the exchangeability of the population. Under our sampling scheme, there are two noteworthy points. First, with S unknown, the distribution of $\mathbf{y}$ is not multivariate hypergeometric but a generalized multivariate hypergeometric distribution. This distribution is proportional to a summation of a set of multivariate hypergeometric distributions. This may bring in a computational challenge. Second, it is possible to include the reversible jump MCMC (RJMCMC) (Green, 1995; Richardson and Green, 1997) into the clustering process to infer S. However, the split and combine process in RJMCMC potentially disturbs the clustering outcome from DP and consequently may invalidate the subsequent MCMC simulations.

Acknowledgments

The authors thank Professor Edsel Pena at the University of South Carolina for his precious comments and suggestions on the project, and Professor David Dunson at the Duke University for his comments related to the DP sampling. The authors are thankful to the Associate Editor and the referees whose comments and suggestions contributed to many improvements in the manuscript.

Appendix

A.1. Full conditional posterior distributions

The conditional posterior distribution of $\mathbf{N}$ is given by:

$$f(\mathbf{N}|\cdots) \propto \frac{\binom{N_1}{y_1}\cdots\binom{N_{s_o}}{y_{s_o}}\binom{N_{s_o+1}}{y_{s_o+1}}\cdots\binom{N_S}{y_S}}{\binom{N_0}{M}(N_1-1)!\cdots(N_S-1)!}\theta_1^{N_1-1}\cdots\theta_S^{N_S-1}, \tag{9}$$

subject to the restriction that $\sum_{i=1}^{S} N_i = N_0$, $N_i \geq \max(y_i, 1)$, where $y_{s_o+1} = \cdots = y_S \equiv 0$.

Table A.1
Posterior inferences of $\boldsymbol{\theta}$ for the identified five clusters. The numbers in parentheses in the first row are the numbers of species in corresponding clusters. "P.I.": posterior interval.

Inferences	Clusters				
	1 (2)	2 (17)	3 (19)	4 (1)	5 (61)
$\hat{\boldsymbol{\theta}}$	0.264	0.0158	0.00606	0.0770	0.000187
95% P.I.	(0.206, 0.303)	(0.00648, 0.0197)	(0.000320, 0.0163)	(0.0590, 0.0990)	(0.000159, 0.00126)

Because of the non-conjugacy of (9), Metroplis–Hastings (M–H) steps will be implemented in the Gibbs sampler. We can apply the M–H steps directly to the multivariate distribution (9). The proposal (jumping) distribution in the M–H steps to generate candidate posterior samples can be selected as multinomial distributions with constraints applied (i.e., $\sum_{i=1}^{S} N_i = N_0, N_i \geq \max(y_i, 1),\ i = 1, \ldots, S$). However, if the dimension of $\mathbf{N}$ is large, the acceptance rate can be exceptionally low, which will dramatically decrease the efficiency of the Gibbs sampler. It would be much more efficient (in terms of acceptance rate in the M–H steps) if we could update one element of $\mathbf{N}$ at a time. However, due to the restriction among the N_i's, given by $\sum_{i=1}^{S} N_i = N_0$, we have $f(N_i|\cdots) = 1$. To solve this problem, let's first rewrite (9) as follows:

$$f(N_i, N_j|\cdots) \propto \frac{\binom{N_i}{y_i}\binom{N_j}{y_j}}{(N_i-1)!(N_j-1)!}\theta_1^{N_i-1}\theta_S^{N_j-1}, \quad i \neq j, \qquad N_i + N_j = N_0 - \sum_{l\neq i,j} N_l,$$

which is clearly a valid conditional bivariate distribution function with constraint $N_i + N_j = N_0 - \sum_{l\neq i,j} N_l$. It is easy to see from (9) that the full conditional posterior distribution of $\mathbf{N}$ is exchangeable, so the distribution function (6) can be applied to any (N_i, N_j) pairs, $i \neq j$. Of course, in practice, it is easier to do in a sequential order of $N_i,\ i = 1, \ldots, S$.

To generate proposals of (N_i, N_j), we select a random walk as the proposal distribution,

$$N_i^* = N_i + u$$
$$N_j^* = N_j - u,$$

where (N_i^*, N_j^*) denotes the proposal for (N_i, N_j). The step size u of the random walk will be determined based on efficiency of convergence. Note that when N_i is updated, N_j is fixed, and *vice versa*. Therefore, based on the full conditional distribution (6), we update one element of $\mathbf{N}$ at a time and the time spent in updating is also for one element. We expect this sampling scheme will dramatically increase the convergence speed of the MCMC simulations compared to the scheme based on the multivariate distribution (9).

A.2. Posterior inference of $\boldsymbol{\theta}$ in simulated data

See Table A.1.

References

Amaral-Zettler, L.A., McCliment, E.A., Ducklow, H.W., Huse, S.M., 2009. A method for studying protistan diversity using massively parallel sequencing of v9 hypervariable regions of small-subunit ribosomal RNA genes. PLoS One 4, e6372.

Antoniak, C.E., 1974. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. The Annals of Statistics 2, 1152–1174.

Basu, S., Ebrahimi, N., 2001. Bayesian capture-recapture methods for error detection and estimation of population size: heterogeneity and dependence. Biometrika 88, 269–279.

Besbeas, P., Freeman, S.N., Morgan, B.J., Catchpole, E.A., 2002. Integrating mark-recapture-recovery and census data to estimate animal abundance and demographic parameters. Biometrics 58, 540–547.

Boender, C.G.E., Rinnooy Kan, A.H.G., 1987. A multinomial Bayesian approach to the estimation of population and vocabulary size. Biometrika 74, 849–856.

Böhning, D., Schön, D., 2005. Nonparametric maximum likelihood estimation of population size based on the counting distribution. Journal of the Royal Statistical Society, Series C: Applied Statistics 54, 721–737.

Brown, M., Hughey, R., Krogh, A., Mian, I.S., Sjolander, K., Haussler, D., 1993. Using Dirichlet mixture priors to derive hidden Markov models for protein families. In: Proc. Int. Conf. Intell. Syst. Mol. Biol., vol. 1, pp. 47–55.

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

Cao, Y., Larsen, D.P., Thorne, R.S.J., 2001. Rare species in multivariate analysis for bioassessment: some considerations. Journal of the North American Benthological Society 21, 144–153.

Chao, A., 1987. Estimating the population size for capture-recapture data with unequal catchability. Biometrics 43, 783–791.

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

Chao, A., Shen, T., 2003. Nonparametric estimation of Shannon's index of diversity when there are unseen species in sample. Environmental and Ecological Statistics 10, 429–443.

Christman, M., Lan, F., 1998. Sequential adaptive sampling designs to estimate abundance in rare populations. In: Proceedings of the American Statistical Association, Section on Statistics and Environment, pp. 87–96.

Christman, M.C., Lan, F., 2001. Inverse adaptive cluster sampling. Biometrics 57, 1096–1105.

Dahl, D., 2006. Model-based clustering for expression data via a Dirichlet process mixture model. In: Do, K., Müller, P., Vannucci, M. (Eds.), Bayesian Inference for Gene Expression and Proteomics. Cambridge University Press, Cambridge.

Dorazio, R.M., Mukherjee, B., Zhang, L., Ghosh, M., Jelks, H.L., Jordan, F., 2008. Modeling unobserved sources of heterogeneity in animal abundance using a Dirichlet process prior. Biometrics 64, 635–644.

Doss, H., 2008. Estimation of Bayes factors for nonparametric Bayes problems via Radon–Nikodym derivatives. Technical Report, Department of Statistics, University of Florida.

Doss, H., 2012. Hyperparameter and model selection for nonparametric Bayes problems via Radon–Nikodym derivatives. Statistica Sinica 22, 1–26.
Durell, S., McGrorty, S., West, A., Clarke, R., Goss-Custard, J., Stillman, R., 2005. A strategy for baseline monitoring of estuary special protection areas. Biological Conservation 121, 289–301.
Efron, B., Thisted, R., 1976. Estimating the number of unseen species: how many words did Shakespeare know? Biometrika 63, 435–447.
Escobar, M.D., West, M., 1995. Bayesian density estimation and inference using mixtures. Journal of the American Statistical Association 90, 577–588.
Favaro, S., Lijoi, A., Mena, R., Prünster, I., 2011. On some issues related to species sampling problems. Technical Report, University of Torino.
Ferguson, T.S., 1973. A Bayesian analysis of some nonparametric problems. The Annals of Statistics 1, 209–230.
Fisher, R.A., Corbet, A.S., 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.
Gelman, A., Carlin, J.B., Stern, H.S., Rubin, D.B., 2003. Bayesian Data Analysis. Chapman & Hall, CRC.
Ghosh, S.K., Norris, J.L., 2005. Bayesian capture-recapture analysis and model selection allowing for heterogeneity and behavioral effects. Journal of Agricultural, Biological, and Environmental Statistics 10, 35–49.
Green, P.J., 1995. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika 82, 711–732.
Green, R.H., Young, R.C., 1993. Sampling to detect rare species. Ecological Applications 3, 351–356.
Kyung, M., Gill, J., Casella, G., 2010. Estimation in Dirichlet random effects models. Annals of Statistics 38, 979–1009.
Lijoi, A., Mena, R., Prünster, I., 2007a. A Bayesian nonparametric method for prediction in EST analysis. BMC Bioinformatics 8, 10.
Lijoi, A., Mena, R., Prünster, I., 2007b. Bayesian nonparametric estimation of the probability of discovering new species. Biometrika 94, 769–786.
Liu, J.S., 1996. Nonparametric hierarchical Bayes via sequential imputations. The Annals of Statistics 24, 911–930.
MacEachern, S.N., Müller, P., 1998. Estimating mixture of Dirichlet process models. Journal of Computational and Graphical Statistics 7, 223–238.
Mao, C.X., 2006. Inference on the number of species through geometric lower bounds. Journal of the American Statistical Association 101, 1663–1670.
Mao, C.X., 2007a. Estimating population sizes for capture-recapture sampling with binomial mixtures. Computational Statistics and Data Analysis 51, 5211–5219.
Mao, C.X., 2007b. Estimating the number of species with multiple incidence-based subsamples. Statistica Sinica 17, 1591–1600.
Mao, C.X., Colwell, R.K., 2005. Estimation of species richness: mixture models, the role of rare species, and inferential challenges. Ecology 86, 1143–1153.
McArdle, B.H., 1990. When are rare species not there? Oikos 57, 276–277.
McAuliffe, J.D., Blei, D.M., Jordan, M.I., 2006. Nonparametric empirical Bayes for the Dirichlet process mixture model. Statistics and Computing 16, 5–14.
Montagna, P., Ritter, C., 2006. Direct and indirect effects of hypoxia on benthos in Corpus Christi Bay, Texas, USA. Journal of Experimental Marine Biology and Ecology 330, 119–131.
Morris, J.S., Baggerly, K.A., Coombes, K.R., 2003. Bayesian shrinkage estimation of the relative abundance of mRNA transcripts using SAGE. Biometrics 59, 476–486.
Neal, R.M., 2000. Markov chain sampling methods for Dirichlet process mixture models. Journal of Computational and Graphical Statistics 9, 249–265.
Norris, J.I., Pollock, K., 1996. Nonparametric MLE under two closed capture-recapture models with heterogeneity. Biometrics 52, 639–649.
Pledger, S., Pollock, K.H., Norris, J.L., 2003. Open capture-recapture models with heterogeneity: I. Cormack–Jolly–Seber model. Biometrics 59, 786–794.
Pledger, S., Pollock, K.H., Norris, J.L., 2010. Open capture-recapture models with heterogeneity: II. Jolly–Seber model. Biometrics 66, 883–890.
Quince, C., Curtis, T.P., Sloan, W.T., 2008. The rational exploration of microbial diversity. The International Society for Microbial Ecology Journal 2, 997–1006.
Richardson, S., Green, P.J., 1997. On Bayesian analysis of mixtures with an unknown number of components (disc: P758-792) (corr: 1998V60 p661). Journal of the Royal Statistical Society. Series B. Methodological 59, 731–758.
Thompson, S.K., 1990. Adaptive cluster sampling. Journal of the American Statistical Association 85, 1050–1059.
Thompson, S.K., 1991. Adaptive cluster sampling: designs with primary and secondary units. Biometrics 47, 1103–1115.
Trippa, L., Parmigiani, G., 2011. False discovery rates in somatic mutation studies of cancer. The Annals of Applied Statistics 5, 1360–1378.
Turk, P., Borkowski, J.J., 2005. A review of adaptive cluster sampling: 1990–2003. Environmental and Ecological Statistics 12, 55–94.
Wang, J.P., 2010. Estimating species richness by a Poisson-compound gamma model. Biometrika 97, 727–740.
West, M., Müller, P., Escobar, M., 1994. Hierarchical priors and mixture models, with application in regression and density estimation. In: Smith, A.F.M., Freeman, P. (Eds.), Aspects of Uncertainty: A Tribute to D. V. Lindley. John Wiley, New York.
Williamson, E.G., Slatkin, M., 1999. Using maximum likelihood to estimate population size from temporal changes in allele frequencies. Genetics 152, 755–761.
Zhang, H., 2007. Inferences on the number of unseen species and the number of abundant/rare species. Journal of Applied Statistics 34, 725–740.
Zhang, H., 2009. Designing sampling plans to capture rare objects. The Canadian Journal of Statistics / La Revue Canadienne de Statistique 37, 417–434.
Zhang, H., Stern, H., 2009. Sample size calculation for finding unseen species. Bayesian Analysis 4, 763–792.