Bayesian model averaging for estimating the number of classes: applications to the total number of species in metagenomics

Sébastien Li-Thiao-Te, Jean-Jacques Daudin, Stephane Robin

Journal of Applied Statistics
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/cjas20

Bayesian model averaging for estimating the number of classes: applications to the total number of species in metagenomics

Sébastien Li-Thiao-Té [a], Daudin Jean-Jacques [b] & Robin Stéphane [b]

[a] UMR 7539 Institut Galilée/Université Paris 13, Villetaneuse, France

[b] UMR 518 AgroParisTech/INRA, AgroParisTech, Paris, France

Available online: 16 Feb 2012

To cite this article: Sébastien Li-Thiao-Té, Daudin Jean-Jacques & Robin Stéphane (2012): Bayesian model averaging for estimating the number of classes: applications to the total number of species in metagenomics, Journal of Applied Statistics, DOI:10.1080/02664763.2012.658358

To link to this article: http://dx.doi.org/10.1080/02664763.2012.658358

Journal of Applied Statistics
2012, iFirst article

Bayesian model averaging for estimating the number of classes: applications to the total number of species in metagenomics

Sébastien Li-Thiao-Té[a]*, Daudin Jean-Jacques[b] and Robin Stéphane[b]

[a]*UMR 7539 Institut Galilée/Université Paris 13, Villetaneuse, France;* [b]*UMR 518 AgroParisTech/INRA, AgroParisTech, Paris, France*

(*Received 5 May 2011; final version received 13 January 2012*)

The species abundance distribution and the total number of species are fundamental descriptors of the biodiversity of an ecological community. This paper focuses on situations where large numbers of rare species are not observed in the data set due to insufficient sampling of the community, as is the case in metagenomics for the study of microbial diversity. We use a truncated mixture model for the observations to explicitly tackle the missing data and propose methods to estimate the total number of species and, in particular, a Bayesian credibility interval for this number. We focus on computationally efficient procedures with variational methods and importance sampling as opposed to Markov Chain Monte Carlo sampling, and we use Bayesian model averaging as the number of components of the mixture model is unknown.

Keywords: mixture models; Bayesian model averaging; variational methods; truncation; metagenomics

1. Introduction

In this paper, we estimate the number of microbial species that are present in an environmental sample, but not observed in the data set for lack of sufficient sampling. We view the problem in the larger context of estimating the number of classes when some classes are not observed. The statistical origins of the topic date back to the work of Fisher *et al.* [10] on estimating the number of Malaysian butterfly species based on captured specimens. Since then, the statistical methods have been applied to diverse areas, ranging from estimating the number of coins minted in the Middle Ages [9] and the number of words in a language [8] to the number of drug addicts when only a few are reported by their physician [3], see [5,7] for a review.

As in the original work of Fisher, estimating the number of classes has particular applications in ecology. In microbial ecology, for instance, the use of culture-free methods such as Metagenomics and Next Generation Sequencing has revealed large numbers of unknown bacterial species in sea water and soil or in the human gut microbiota [13]. However, most bacterial species in these

*Corresponding author. Email: lithiao@math.univ-paris13.fr

ISSN 0266-4763 print/ISSN 1360-0532 online

http://dx.doi.org/10.1080/02664763.2012.658358
http://www.tandfonline.com

samples are not very abundant and are easily not observed, which motivates the use of statistical methods for estimating the global biodiversity. Surely, the total number of species gives an estimate of how much or how little we understand a given microbial community. In addition, there is ongoing research on whether microbial communities are more efficient or more resilient when more complex [25].

Three methods have been used in the literature for estimating species richness. These are (1) fitting a statistical distribution to either rank-abundance or frequency-count data, (2) extrapolating rarefaction curves and (3) using non-parametric methods. The main statistical distributions used are lognormal and mixtures of exponentials. These have been reviewed and compared in [4]. Techniques based on rarefaction curves are discussed in [11,18]. Non-parametric methods do not restrict the models to a family of distributions. They have led, in particular, to Chao-type estimators [3,6].

When fitting a statistical distribution to frequency-count data, finite mixture models have been shown to perform well for metagenomics data sets [4]. However, choosing the right number of components is not straightforward. In this paper, we tackle two issues in the estimation problem. First, we extend the finite mixture model approach described in [4] to address the uncertainty about the choice of the number of components of the model. Instead of selecting the model order, we use Bayesian model averaging (BMA) to add flexibility to the choice of the model without going all the way to non-parametric methods. Second, we evaluate the level of confidence in the estimated number of species in a Bayesian framework with credibility intervals obtained by importance sampling. These methods are illustrated using a recent study on the microbial diversity of the human gut in relation to eating habits [23].

This paper is organized as follows. In Section 2, we present the probabilistic model that describes the characteristics of the ecological community and how observations for each species are generated given the parameters of the environment. In Section 3, we describe three parameter estimation methods based on the variational Bayesian EM algorithm given in [2] and BMA. In particular, we compute approximate posterior distributions and deduce approximate credibility intervals. In Section 4, we discuss the results of our methods in the context of simulated data and a metagenomics data set.

2. Statistical model

The number of missing species must be extrapolated based on species observed at least once. This cannot be done without a model and has led to questions on the model's representativity of the actual biological phenomenon. In particular, it is not possible to guard against large numbers of species with no representative in the data set ([14] cited in [5]). Therefore, the choice of a model reflects an assumption on the relative number of missing species and observed species. In general, a model is chosen for the species abundance distribution (SAD) which reflects the relative numbers of abundant species with many individuals and rare species.

2.1 *Sampling model*

In the context of metagenomics (large number of species, many rare and unobserved), the following model is often considered (see [5] for a review). Let C denote the total number of species present in the environmental sample and n the number of observed species. Each of these species contributes a Poisson distributed number X_i of individuals to the data set with species-specific rate parameter λ_i. We make the modelling assumption that λ_i are independent identically distributed samples from a random variable Λ. The distribution of Λ is the SAD.

The SAD is not directly observable and is estimated based on the variables X_i. As the species' rates λ_i are not observed, the data set contains samples of a random variable X which are

independent conditionally on λ_i. The relationship between Λ and X is expressed by the following formula (Poisson composition):

$$P_X(x) := \int \frac{\mathrm{e}^{-\lambda}\lambda^x}{x!} f_\Lambda(\lambda)\,\mathrm{d}\lambda.$$

However, species with zero observed individuals are not present in the data set. We explicitly account for these missing data by considering a truncated model, that is, we study the distribution of the variable X_+ which denotes the always positive number of individuals:

$$P_{X_+}(x) := P_X(x|X > 0) = \frac{P_X(x)}{1 - P_X(0)}.$$

Remark As indicated in [19], the truncated model is identifiable, that is, given an infinite number of observations and the associated distribution for X_+, there exists only one corresponding distribution f_Λ.

This model corresponds to the case where individuals and species are sampled and distributed independently. When individuals from the same species are frequently sampled together – for example, in the presence of clusters of individuals – or species are sampled together because of trophic relationships, other models such as the negative binomial can be used [12,20].

2.2 *Truncated mixture model for observed species abundances*

The sampling model described in Section 2.1 indicates how observations are related to the SAD. We now present a parametric model for the SAD, and in Section 2.4, we deduce the total number of species from the model.

Finite mixture models are flexible parametric tools that can adapt to the complexity of each data set. In [4], a systematic comparison to select good statistical models for SADs among the many models available was performed [20]. The truncated mixture of geometric distributions used in this paper was found to be the best choice among the models considered (Poisson, Poisson mixture, inverse Gaussian, etc.).

As in [4], we model the SAD with a finite mixture of exponential distributions $f_\Lambda(\lambda) = \sum_q \beta_q \mathcal{E}(\lambda, \mu_q)$, where μ_q denotes the model parameters and β_q are the mixing proportions. Contrary to that done in Bunge *et al.*, the number of components is not chosen between two and three; we deal with this issue in Section 3.4.

When the SAD follows a mixture model, the observations X_i also follow a mixture model. Due to linearity, composition with the Poisson distribution can be applied to individual components, that is, $P_X(x) = \sum_q \beta_q \int (\mathrm{e}^{-\lambda}\lambda^x/x!) f(\lambda, \mu_q)\,\mathrm{d}\lambda$. When the components are exponential distributions ($f(\lambda, \mu_q) = \mu_q\,\mathrm{e}^{-\lambda\mu_q}$), straightforward computations show that the model further simplifies into a mixture of geometric distributions:

$$P_X(x) = \sum_q \beta_q (1 - \pi_q)(\pi_q)^x,$$

where $\pi_q = 1/(1 + \mu_q)$.

2.3 *The mixture of truncated geometric distributions*

To explicitly take into account missing species in the model, we consider the truncated mixture

$$P_{X_+}(x) = \frac{\sum_q \beta_q (1 - \pi_q)(\pi_q)^x}{1 - P_X(0)} = \frac{\sum_q \beta_q (1 - \pi_q)(\pi_q)^x}{\sum_q \beta_q \pi_q}.$$

Truncated mixtures are cumbersome because the model parameters appear in the denominator. For example, the EM algorithm for truncated geometric mixtures requires root-finding, whereas closed formulae for the updates are available in the un-truncated case.

The results given in [3] allow us to consider a mixture of truncated distributions instead of a truncated mixture model. In the case of the geometric mixture model,

$$P_{X_+}(x) = \frac{\sum_q \beta_q(1-\pi_q)(\pi_q)^x}{\sum_q \beta_q \pi_q} = \sum_q \alpha_q(1-\pi_q)(\pi_q)^{x-1}, \tag{1}$$

where

$$\alpha_q = \frac{\beta_q/\pi_q}{\sum_q \beta_q/\pi_q}. \tag{2}$$

In particular, we can perform maximum-likelihood inference on the mixture of truncated geometric distributions and deduce the maximum-likelihood solution for the truncated mixture model by a simple application of the change of parameters given above.

2.4 *Computation of the total number of species*

A simple formula for the total number of species computes the expected number of observed species as $n = C \times P_X(x > 0)$ from which we deduce the estimator $\hat{C} = n/(1 - \widehat{P_X}(0))$, where n is the number of observed species and $\hat{P_X}(0)$ is estimated from the SAD. $\hat{C}$ is a maximum-likelihood estimator obtained conditionally on the estimates of the model parameters (see [4] for a complete derivation). It is asymptotically equivalent to complete the maximum-likelihood estimation where the total number of species C and the model parameters are optimized at the same time [21].

Based on the previous derivation, we compute $\hat{C}$ by estimating the SAD with a mixture of truncated geometric distributions. We then change the parameters to obtain a truncated mixture and compute $\hat{C} = n/(1 - \sum_q \widehat{\beta_q}(1 - \widehat{\pi_q}))$. A simpler formula directly computes the number of species from the parameters of the mixture of truncated geometric distributions as $\hat{C} = n \sum_q (\widehat{\alpha_q}/\widehat{\pi_q})$ (simple derivation, also in [3]).

The proposed estimation formula for $\hat{C}$ relies on the estimation of the mixture parameters $\widehat{\alpha_q}$ and $\widehat{\pi_q}$. Consequently, large numbers of observed species are more important than large numbers of individuals for each species. In cases where the sampling method puts a limit on the species counts, several data sets can be pooled to alleviate the problem provided that Poisson sampling is still in effect.

3. Parameter inference

Most of the approaches to date have been performed in the frequentist framework [4]. As a consequence, variance estimates and confidence intervals are asymptotic. In this paper, we operate in the Bayesian framework to obtain non-asymptotic approximate credibility intervals.

In the context of metagenomics, the data sets are too large for straightforward Markov Chain Monte Carlo (MCMC) procedures such as using a Gibbs sampler. We propose to approximate the posterior distribution with variational methods. As discussed in Section 3.6, variational methods are not precise enough, so importance sampling is introduced to improve the estimate of the posterior distribution.

3.1 *Bayesian framework*

In the proposed model, the observations X_i are samples from a mixture of truncated geometric distributions and we introduce the hidden variable Z_i that represents the class membership of species i among the M components.[1] We propose the following Bayesian priors for the parameters (α_q, π_q) of the mixture of truncated geometric distributions:

$$\begin{aligned} \alpha &\sim \text{Dirichlet}(a) \\ \pi_q &\sim \text{Beta}(b_q, c_q) \\ Z_i &\sim \mathcal{M}(1; \alpha) \\ X_i | Z_{iq} = 1 &\sim g_{q+}(x, \pi_q), \end{aligned}$$

where $g_{q+}(x, \pi_q) = (1 - \pi_q)(\pi_q)^{x-1}$ is the truncated geometric distribution. Note that a is a vector in $\mathbb{R}^M$. A Dirichlet prior on α is a standard choice because it is the conjugate prior for the parameters of multinomial distributions. We use Beta priors for the parameters π_q for the same reason.

Bayesian parameter inference provides distributions over (α_q, π_q) rather than on the parameters of the truncated mixture. When studying the SAD, the parameters of the truncated mixture (and hence their distribution) can be obtained from Equation (2). We are rather interested in the posterior distribution of the total number of species, which we obtain directly from the parameters of the mixture of truncated geometrics using the formula $C = n \sum_q (\alpha_q / \pi_q)$ (see Section 2.4). In the remainder of this section, we deal with the approximation of this posterior distribution.

3.2 *Variational Bayes-EM approach*

We provide a brief description of the VB-EM algorithm for approximate Bayesian inference. For more information, the reader can refer to [2]. Variational methods decompose the complete likelihood into

$$\log P(X) = \int Q(Z, \vartheta) \log \frac{P(X, Z, \vartheta)}{Q(Z, \vartheta)} \, \mathrm{d}Z \, \mathrm{d}\vartheta + KL(Q(\vartheta, Z), P(\vartheta, Z|X)),$$

where Q is an arbitrary distribution, Z are hidden variables and ϑ are model parameters. Variational methods maximize the first term, which is equivalent to minimizing the Kullback–Leibler divergence $KL(Q(\vartheta, Z), P(\vartheta, Z|X))$.

To perform the maximization, the distribution $Q(\vartheta, Z)$ is approximated by $Q(\vartheta, Z) \sim Q_\vartheta(\vartheta) Q_Z(Z)$. The maximization in $Q_\vartheta(\vartheta)$ and $Q_Z(Z)$ successively leads to the following update formulae that are iterated until convergence:

$$\begin{aligned} Q_Z(Z) &\propto \exp\left[\int P(X|Z, \vartheta) P(Z|\vartheta) Q_\vartheta(\vartheta) \, \mathrm{d}\vartheta \right], \\ Q_\vartheta(\vartheta) &\propto P(\vartheta) \exp\left[\int P(X|Z, \vartheta) P(Z|\vartheta) Q_Z(Z) \, \mathrm{d}Z \right]. \end{aligned}$$

For the sake of readability, we write $Q(\vartheta)$ and $Q(Z)$ where applicable.

3.3 *Application of the VB-EM method to a mixture of truncated geometric distributions*

We apply the VB-EM method to compute an approximate posterior distribution $Q(\alpha, \pi) \sim P(\alpha, \pi | X)$. In the case of conjugate exponential models, the updates can be performed by simply updating the hyperparameters. The mixture of truncated geometric models falls into this category

because truncated geometric distributions are still geometric distributions. Therefore, it suffices to update the parameters a of the Dirichlet prior distribution on the mixing proportions and the parameters (b_q, c_q) of the Beta prior distributions on the parameters for each geometric component.

For the mixture of truncated geometric distributions, the VB-EM algorithm corresponds to the following update formulae:

$$a_q^{(n+1)} = a_q^0 + \sum_i \tau_{iq}^{(n)},$$

$$b_q^{(n+1)} = b_q^0 + \sum_i \tau_{iq}^{(n)}(X_i - 1),$$

$$c_q^{(n+1)} = c_q^0 + \sum_i \tau_{iq}^{(n)},$$

where (a^0, b^0, c^0) are the prior parameters and $\tau_{iq}^{(n)} = Q_{Z_i}(q)$. We denote $\hat{a}_q, \hat{b}_q$ and $\hat{c}_q$ as the hyperparameters of the approximate posterior distribution $Q(\alpha, \pi)$.

As we have used conjugate priors, the approximate posterior distributions belong to the same family to which the priors belong:

$$\alpha|X \sim \text{Dirichlet } (\hat{a}),$$

$$\pi_q|X \sim \text{Beta } (\hat{b}_q, \hat{c}_q).$$

3.4 *Bayesian model averaging*

The selection of the number of components of the mixture model can be done based on a penalized likelihood with penalization terms such as AIC [1] and BIC [22]. However, we have decided to not choose a single model and rather combine the estimates of several models with BMA. Data collected in real-world applications are unlikely to belong to the family of truncated mixtures of geometric distributions, so there is no true value for the number of components. Moreover, averaging over several models is always potentially better than choosing a single model, because the single model solution is contained in the set of possible averages.

Let us briefly recall the basics of the BMA approach. More details can be found in [15]. We consider the set $\mathcal{S}$ of all mixture models and partition it according to the number of components $\mathcal{S} = \cup_m \{\text{models with } m \text{ components}\}$. The BMA model is a linear combination $d_{\text{BMA}} = \sum_m w_m d_m$ such that d_m is the posterior density of the observations given a model with m components. To compute the weights w_m, let us define the hidden random variable M that indicates the number of components. Then, conditionally on M, the model for the data set is d_M with weight $w_M = P(M|X)$.

The weights are computed by $w_M \propto P(X|M)P(M)$ according to the Bayes formula. $P(M)$ is the *a priori* distribution on M. In our context, we assume that M is bounded by M_0 and that it has a uniform distribution over the interval $(1, M_0)$. The integrated likelihood $P(X|M) = \iint P(X|\vartheta, Z, M)P(Z|\vartheta, M)P(\vartheta|M)\,\mathrm{d}Z\,\mathrm{d}\vartheta$ is the likelihood of the observations X in the mixture model with M components and can be approximated using the variational approach by neglecting the Kullback–Leibler term:

$$\log P(X|M) \simeq \iint Q(Z, \vartheta) \log \frac{P(X, Z, \vartheta|M)}{Q(Z, \vartheta)}\,\mathrm{d}Z\,\mathrm{d}\vartheta.$$

In the case of conjugate exponential models, we provide a closed-form formula to compute $\log P(X|M)$ in Theorem 3.1 and the practical computation in Corollary 3.2 for mixture models.

THEOREM 3.1 *Suppose that $P(Z,\vartheta,X)$ belongs to the family of conjugate exponential models*:

$$\log P(\vartheta) = \log h(\eta,\nu) + \eta \log g(\vartheta) + \Phi(\vartheta)'\nu$$
$$\log P(X,Z|\vartheta) = \log f(X,Z) + \log g(\vartheta) + \Phi(\vartheta)' u(X,Z).$$

We note $Q(Z)$ and $Q(\vartheta)$ as the fixed-point solutions of the variational problem $\bar{\Phi} = \mathbb{E}_Q[\Phi(\vartheta)]$ and $\bar{u}(X) = \mathbb{E}_Q[u(X,Z)]$. Then,

$$\log P(X|M) \simeq -\bar{\Phi}'\bar{u}(X) - \log \frac{h(\eta+1,\nu+\bar{u}(X))}{h(\eta,\nu)} + \log \int f(X,Z) \exp^{\bar{\Phi}' u(X,Z)} \, \mathrm{d}Z.$$

Proof We compute $\log P(X|M) \simeq \iint Q(Z,\vartheta) \log(P(X,Z,\vartheta|M)/Q(Z,\vartheta)) \, \mathrm{d}Z \, \mathrm{d}\vartheta$ when $P(Z,\vartheta,X)$ belongs to the family of conjugate exponential models and Q is obtained by the VB-EM algorithm.

According to Section 3.2, after convergence, we obtain $Q(Z,\vartheta|M) = Q(\vartheta)Q(Z)$ such that

$$\log Q(\vartheta) = \log h(\eta+1,\nu+\bar{u}(X)) + (\eta+1)\log g(\vartheta) + \Phi(\vartheta)'(\nu+\bar{u}(X))$$
$$\log Q(Z) = \log f(X,Z) + \bar{\Phi}' u(X,Z) + \log g_{\bar{\Phi}},$$

where $\bar{u}(X) = \int u(X,Z)Q(Z)\,\mathrm{d}Z$ and $\bar{\Phi} = \int \Phi(\vartheta)Q(\vartheta)\,\mathrm{d}\vartheta$ and $g_{\bar{\Phi}} = 1/\int f(X,Z)\,\mathrm{e}^{\bar{\Phi}' u(X,Z)}\,\mathrm{d}Z$. Note that η, f and g are scalar, but ν, $u(X)$, Φ, $\bar{u}(X)$ and $\bar{\Phi}$ are vectors.

We now compute the KL distance with M fixed:

$$\begin{aligned}
\log P(X|M) &\simeq \iint Q(Z,\vartheta) \log \frac{Q(Z,\vartheta)}{P(Z,\vartheta,X)} \,\mathrm{d}Z\,\mathrm{d}\vartheta \\
&= \iint Q(\vartheta,Z)[\log Q(\vartheta,Z) - \log P(X,Z|\vartheta) - \log P(\vartheta)]\,\mathrm{d}Z\,\mathrm{d}\vartheta \\
&= \iint Q(\vartheta,Z)\log Q(\vartheta,Z)\,\mathrm{d}Z\,\mathrm{d}\vartheta \\
&\quad + \iint Q(\vartheta,Z)[-\log h(\eta,\nu) - (\eta+1)\log g(\vartheta) \\
&\quad - \log f(X,Z) - \Phi(\vartheta)'\nu - \Phi(\vartheta)' u(X,Z)]\,\mathrm{d}Z\,\mathrm{d}\vartheta \\
&= \int Q(\vartheta)\log Q(\vartheta)\,\mathrm{d}\vartheta + \int Q(Z)\log Q(Z)\,\mathrm{d}Z - \log h(\eta,\nu) \\
&\quad - (\eta+1)\int Q(\vartheta)\log g(\vartheta)\,\mathrm{d}\vartheta - \int \log f(X,Z)Q(Z)\,\mathrm{d}Z - \bar{\Phi}'\nu - \bar{\Phi}'\bar{u}(X) \\
&= \log \frac{h(\eta+1,\nu+\bar{u}(X))}{h(\eta,\nu)} + (\eta+1)\int Q(\vartheta)\log g(\vartheta)\,\mathrm{d}\vartheta + \bar{\Phi}'\nu + \bar{\Phi}'\bar{u}(X) \\
&\quad + \int \log f(X,Z)Q(Z)\,\mathrm{d}Z + \bar{\Phi}'\bar{u}(X) + \log g_{\bar{\Phi}} \\
&\quad - \log h(\eta,\nu) - (\eta+1)\int Q(\vartheta)\log g(\vartheta)\,\mathrm{d}\vartheta \\
&\quad - \int \log f(X,Z)Q(Z)\,\mathrm{d}Z - \bar{\Phi}'\nu - \bar{\Phi}'\bar{u}(X) \\
&= \bar{\Phi}'\bar{u}(X) + \log \frac{h(\eta+1,\nu+\bar{u}(X))}{h(\eta,\nu)} - \log \int f(X,Z)\exp^{\bar{\Phi}' u(X,Z)}\,\mathrm{d}Z. \qquad \blacksquare
\end{aligned}$$

COROLLARY 3.2 *The evidence in Theorem* 3.1 *can easily be computed in the case of mixture models*:

- $\bar{\Phi}$ *is given directly by the VB-EM algorithm.*
- $\bar{u}(X)$ *is given by the equations of the VB-EM algorithm* (*e.g. replacing* Z_{iq} *by* τ_{iq} *in the expression of* $u(X, Z)$). *Easier still, the VB-EM algorithm provides* $\nu + \bar{u}(X)$ *where* ν *is the given prior.*
- *As* $\log P(\vartheta|\eta, \nu) = \log h(\eta, \nu) + \eta \log g(\vartheta) + \Phi(\vartheta)'\nu$, *for any value of* ϑ,

$$\log \frac{h(\eta + 1, \nu + \bar{u}(X))}{h(\eta, \nu)} = \log P(\vartheta|\eta + 1, \nu + \bar{u}(X)) - \log P(\vartheta|\eta, \nu) - \log g(\vartheta) - \Phi(\vartheta)'\bar{u}(X).$$

- *The last item is just a sum because* Z *is a multinomial.*

For a mixture of truncated geometric distributions,

$$f(X, Z) = 1$$
$$g(X, Z) = 1$$
$$h(\eta, \nu) = h(a, b, c) = \frac{\Gamma(\sum_q a_q)}{\prod_q \Gamma(a_q)} \frac{1}{\prod_q B(b_q, c_q)},$$

where B is the Beta function.

3.5 *Posterior distribution averaging*

The VB-EM method described in Section 3.3 computes the posterior distribution of the parameters of the mixture model when the number of components is provided. In this study, we are rather interested in the posterior distribution of the total number of species $C|X$. To our knowledge, this does not correspond to a well-studied distribution. As a consequence, we generate samples from the posterior distribution of the parameters and obtain samples from the posterior distribution of the number of species with the formula $C = \sum_q \alpha_q/\pi_q$.

When the number of components M is unknown, BMA uses a mixture of posterior distributions for several values of M. The weights $(\widehat{w_M})$ of this mixture are computed according to Section 3.4. We call the *BMA posterior* the approximate posterior distribution obtained by aggregating n samples generated from the posterior distribution $C|X, M$ for each value of M. More precisely, we first draw the multinomial sample (n_M) according to the weights $(\widehat{w_M})$ and successively generate n_M samples of $C|X, M$. From the total n samples, we compute summary statistics such as the posterior mean estimate of the total number of species and a credibility interval based on the quantiles of the distribution.

3.6 *Posterior distribution from importance sampling*

As reported previously by Keribin [16] and Titterington and Wang [24] the posterior distribution obtained from the VB-EM algorithm usually underestimates the true variance of the parameters and hence the variance of the total number of species $C|X$. Consequently, credibility intervals computed from the BMA posterior are too narrow, and their real level is lower than anticipated (see the simulations described in Section 4.1).

To improve the approximation of the posterior distribution and credibility intervals, we propose to replace the samples obtained from the VB-EM hyperparameters with samples drawn from importance sampling. We then aggregate the importance samples to obtain a *BMA IS posterior*.

For a given number of components, we generate the importance samples $(\alpha_j, \pi_j)_{j=1}^n$ from any proposal distribution Q, (absolutely continuous with respect to P). According to the Law of Large Numbers, for any function h,

$$\frac{1}{n}\sum_j h(\alpha_j, \pi_j)\frac{P(\alpha_j, \pi_j|X)}{Q(\alpha_j, \pi_j)} \longrightarrow \int h(\alpha, \pi)\frac{P(\alpha, \pi|X)}{Q(\alpha, \pi)}\,\mathrm{d}Q(\alpha, \pi) = \mathbb{E}[h(\alpha, \pi)|X].$$

This means that the sum of Dirac masses $(1/n)\sum_j (P(\alpha_j, \pi_j|X)/Q(\alpha_j, \pi_j))\delta_{(\alpha_j, \pi_j)}$ converges in distribution towards $P(\alpha, \pi|X)$. Hence, samples $C(\alpha_j, \pi_j)$ with weights $P(\alpha_j, \pi_j|X)/Q(\alpha_j, \pi_j)$ are samples from an approximation of the posterior distribution of the total number of species.

Importance sampling provides good results when the importance weights $P(\alpha_j, \pi_j|X)/Q(\alpha_j, \pi_j)$ are close to one, that is, when the proposal distribution Q is a good approximation of the posterior P. Consequently, we use the approximate posterior obtained from the VB-EM algorithm described in Section 3.3 as a proposal distribution.

We compute the weights, and in particular the quantity $P(\vartheta|X)$, with the Bayes formula:

$$P(\vartheta|X) = \frac{P(\vartheta, X)}{P(X)} = \frac{P(X|\vartheta)P(\vartheta)}{P(X)}.$$

$P(\vartheta)$ is the prior distribution. $P(X)$ is a multiplicative constant. $P(X|\vartheta)$ can be obtained by integrating in the missing variable Z :

$$P(X|\vartheta) = \int P(X, Z|\vartheta)\,\mathrm{d}Z.$$

In the case of mixture models, the missing variables Z_i are independent, and the computation of the integral is efficient.

3.7 *Improvements to the VB-EM posterior for importance sampling*

The fact that the VB-EM posterior does not have enough variance is reflected in the BMA IS posterior: the tails of the distribution of the parameters are not adequately sampled, and this is the same for the tails of the distribution of the number of species. We propose a method to increase the variance of the proposal distribution that consists in multiplying the hyperparameters by a factor v.

Recall that the proposal distribution Q obtained from the VB-EM algorithm is the product of a Dirichlet distribution over the mixture parameters α and Beta distributions over the parameters π_q. It is characterized by the hyperparameters $\hat{a}$ for the Dirichlet distribution and the hyperparameters $(\hat{b}_q, \hat{c}_q)$ for the Beta distributions.

For any positive number v, we obtain a proposal distribution $Q_{/v}$ by considering the hyperparameters $\hat{a}/v$ for the Dirichlet distribution, and the hyperparameters $(\hat{b}_q/v, \hat{c}_q/v)$ for the Beta distributions. More precisely,

- Dirichlet distributions with hyperparameters $\hat{a}$ and $\hat{a}/v$ have the same mean value.
- Beta distributions with hyperparameters $(\hat{b}_q, \hat{c}_q)$ and $(\hat{b}_q/v, \hat{c}_q/v)$ have the same mean value.
- A Beta distribution with hyperparameters $(\hat{b}_q, \hat{c}_q)$ has variance $V = \hat{b}_q\hat{c}_q/(\hat{b}_q + \hat{c}_q)^2(\hat{b}_q + \hat{c}_q + 1)$. After dividing by v, the variance becomes $v((\hat{b}_q + \hat{c}_q + 1)/(\hat{b}_q + \hat{c}_q + v))V$.
- The marginals of a Dirichlet distribution are Beta distributed, so the same kind of expansion factor is applied to the marginals of the Dirichlet distribution.

In particular, we consider the factor v to be greater than 1 so that the variances are increased. According to the update formulae given in Section 3.3, the hyperparameters $\hat{a}$ and $(\hat{b}_q, \hat{c}_q)$ are

proportional to the number of samples and much larger than the dividing factor v. Consequently, the term $((\hat{b}_q + \hat{c}_q + 1)/(\hat{b}_q + \hat{c}_q + v))$ is approximately 1 and dividing by v roughly corresponds to multiplying the variance by v.

In the following, we denote the *BMA IS/v posterior* as the approximate posterior distribution obtained in the same way as the BMA IS posterior with the proposal distribution $Q_{/v}$.

4. Results and simulations

In this section, we consider the following approximate posterior distributions:

- MCMC samples,
- BMA posterior,
- BMA IS posterior, and
- BMA IS/v posterior.

We evaluate these in terms of their estimated value of the total number of species and a credibility interval for that number. MCMC provides an asymptotically correct reference.

We illustrate the performance on simulated data and a real metagenomics data set. We implemented the variational algorithms using GNU R. For MCMC computations, we used the Gibbs sampler implemented in the `jags` package in R.

4.1 *Simulations*

We first present an example to illustrate how BMA operates and mixes several potential models. We generated a data set with 2000 species according to the truncated mixture of geometric distribution models with three components (parameters $(0.4, 0.8, 0.95)$ for the geometric distributions and mixing weights $(0.6, 0.3, 0.1)$). In Figure 1, we plot the density of the observed number of species and the true model on a log scale. To this, we add the log-densities of the mixtures of geometric distributions obtained with the VB-EM method and the BMA model.

When drawn on a log scale, the density of a mixture of k geometric distributions behaves like a piecewise linear function with k pieces. For instance, a single component appears as a line, whereas higher complexity models add linear sections. The model with three components provides a better fit to the true model (known) than the models of orders 1 and 2. The order 4

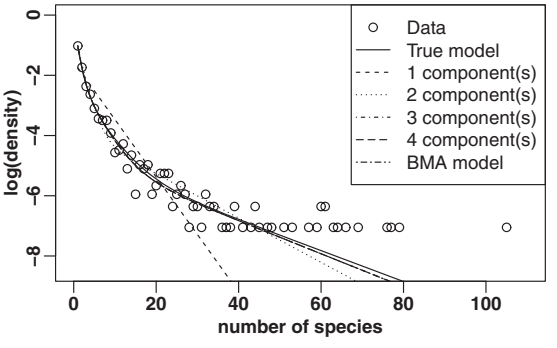

Figure 1. A simulated data set with 2000 species generated under a truncated three-component mixture model. Models of increasing complexity are compared with the true model and the empirical distribution. We use the posterior mean as model parameters to draw the model densities.

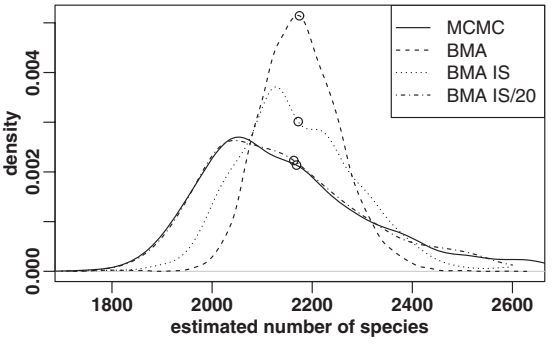

Figure 2. The MCMC, BMA, BMA IS and BMA IS/20 posterior distributions.

model is superimposed on the order 3 model; this is not surprising, as the models are nested. In this case, $\widehat{w_3} \simeq 0.997$, so the BMA model is also superimposed on the order 3 model.

In Figure 2, we show the posterior distributions and the corresponding estimated total number of species (mean *a posteriori*). This shows the level of precision that can be expected from BMA, with respect to the inherent uncertainty of the data set, as represented by MCMC. The estimated number of species coincides in general; this suggests that most of the variance in the posterior mean estimate from BMA is related to the random sample. The importance sampling posterior distributions shown on this example coincide with the MCMC posterior as expected, but the BMA IS/v posterior is a better approximation.

As shown in Figure 2, the posterior distribution obtained from the VB-EM method and hence the BMA method does not have enough variance. As a consequence, the credibility intervals obtained from BMA will be too small. However, the scaled importance sampling method BMA IS/v allows us to retrieve the correct posterior distribution.

To evaluate the bias, precision and the actual level of the credibility intervals obtained from the posterior distributions, we generated 1000 data sets with the same parameters and computed the posterior distribution for each method. Due to timing constraints, the MCMC posterior was only computed up to 2000 total species, as this already required 5 min per data set. In Table 1, we present the median of the estimated number of species. This shows that the estimated total number of species is unbiased. In Table 2, we present the median average deviation relative to the true value to evaluate the precision of the estimated total number of species. The results show that the precision is comparable for all posterior distributions, which suggests that most of the variation is in fact due to the random sample and not due to the estimation method. However, precision decreases with the BMA IS/v posterior and large values of v because the proposal distribution has too much variance and many samples have negligible importance weight.

Table 1. Median of the estimated total number of species in 1000 data sets.

C	BMA	BMA IS	BMA IS/10	BMA IS/20	BMA IS/30	MCMC
200	187.8	188.2	191.8	195.6	201.8	213.5
2000	2011.3	2010.5	2011.8	2037.4	2053	2059.7
20,000	20,030.6	20,032.7	20,028.9	20,041	20,050.2	0
200,000	200,106.5	200,111.5	200,059.9	200,077.5	200,016.2	0

Notes: For all methods, the values are close to the theoretical number of species, which shows that the bias is low. The bias decreases with the total number of species because the data set is larger.

Table 2. Median average deviation of the total number of species relative to the true number.

C	BMA	BMA IS	BMA IS/10	BMA IS/20	BMA IS/30	MCMC
200	0.133	0.126	0.132	0.164	0.22	0.171
2000	0.084	0.083	0.083	0.092	0.109	0.071
20,000	0.02	0.02	0.021	0.021	0.02	0
200,000	0.007	0.007	0.007	0.007	0.007	0

Notes: The precision of MCMC, BMA and BMA IS is the same because the estimate is nearly identical. BMA IS/v can be less precise when too many samples have negligible importance weight.

Table 3. Empirical level of the computed credibility intervals for a target level of 95%.

C	BMA	BMA IS	BMA IS/10	BMA IS/20	BMA IS/30	MCMC
200	0.656	0.731	0.82	0.816	0.809	0.971
2000	0.637	0.73	0.86	0.874	0.875	0.955
20,000	0.617	0.75	0.876	0.902	0.916	0
200,000	0.606	0.734	0.888	0.902	0.915	0

Notes: The MCMC posterior is exact. The BMA posterior distribution computes credibility intervals that are too narrow. The BMA IS methods retrieve most of the true variance.

We evaluated the actual level of the confidence interval by counting how many times the true value belongs to the computed interval. Table 3 shows the results when computing a 95% credibility interval. The reference method MCMC provides accurate confidence intervals, as the empirical level matches the desired level (0.95). However, this method is run with the correct number of components in the mixture model and is not applicable for large data sets.

As shown in Figure 2, the BMA posterior does not have enough variance, and this is reflected in the empirical level of the confidence (around 0.63). The empirical level appears to be constant across varying numbers of species, which suggests that the problem is related to the variational approximation described in Section 3.2 rather than to the insufficient number of samples.

The use of the VB-EM posterior as a proposal distribution for importance sampling improves the quality of the credibility intervals to around 0.74 (expected 0.95). This is a definite improvement, but coverage can be further improved by increasing the variance of the proposal distribution. The obtained BMA IS/v posterior reaches more than 90% coverage. With too much increased variance, however, the estimated total number of species is less precise because many importance samples have negligible weights (see Table 2). In our experience, the value $v = 20$ provides good results in practice.

4.2 *Influence of the prior choice*

The posterior distribution is dependent on the choice of the prior distribution, and we checked that the results are mostly the same with Figures 3 and 4. For the sake of simplicity, the prior is described in our implementation by a single value t such that $a_q = b_q = c_q = t$ for all components q and the same value is used for all the models considered by the BMA algorithm. When t is close to 1, the Dirichlet and Beta priors are approximately uniform on $[0, 1]$. Higher values of t put more weight a priori on 0.5. As shown by the example in Figure 3, different choices of t have little influence on the total number of species in the simulations.

In Figure 4, we plot the lower and upper bounds of the credibility intervals as well as the posterior mean estimate for the BMA IS/20 method and several values of t. The box plots are computed from 1000 random data sets. The plots show no significant difference for the confidence

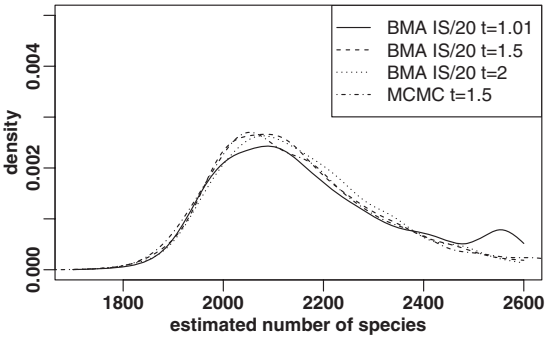

Figure 3. BMA IS/20 approximate posterior distribution for several choices of the prior distribution parameter t. The MCMC posterior distribution is provided for reference. The posterior distribution is not affected by prior choice.

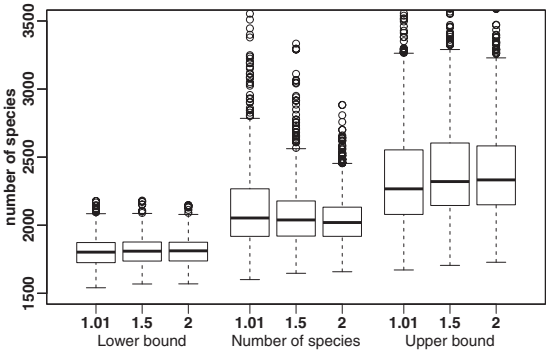

Figure 4. Influence of prior choice on the lower bound of the confidence interval, the estimated number of species and the upper bound of the confidence interval. The interval is robust to prior choice, but high numbers of species are more likely with prior choice $t = 1.01$.

intervals; most of the variation is in fact due to the random data set. When the prior parameter t is close to 1, the parameters π_q close to 0 are more probable and so are high numbers of species.

4.3 *Real data set*

In this section, we apply the BMA method to estimate the total number of microbial strains in the human gut microbiome from the data set published in [23]. Faecal samples from 17 healthy subjects with a vegetarian diet or an omnivorous diet were studied. Using bacterial primers, 16S rDNA genes were extracted and then amplified before sequencing. In total, 10,456 unambiguous sequences were pooled into 3180 OTUs with RapidOTU [17]. In the following, we pool the 17 individuals and use the number of sequences as a measure of the abundance of an OTU.

Figure 5 shows that the mixture model is adequate for this data set. In [4], it was indicated that models with two and three components are usually enough in practice. In this case, the best fit appears to be the four-component model with weight $\widehat{w_4} \simeq 0.971$, but the three-component model also has non-negligible weight $\widehat{w_3} \simeq 0.026$.

Figure 6 shows the BMA posterior distribution and the BMA IS/20 posterior distribution. The MCMC posterior distribution with four components is provided for reference. The posterior mean estimates coincide at around 25,700 species. The 95% confidence interval is [19421, 36355]. In [23], the Chao1 estimate which provides a lower bound of 9940 OTUs was used.

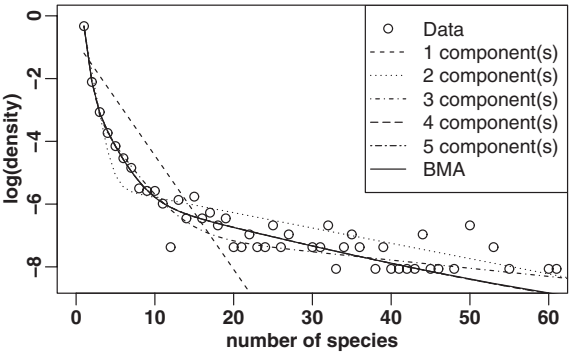

Figure 5. Mixture models with up to five components fit to the metagenomics data set. The BMA method indicates that four components are appropriate for this data set.

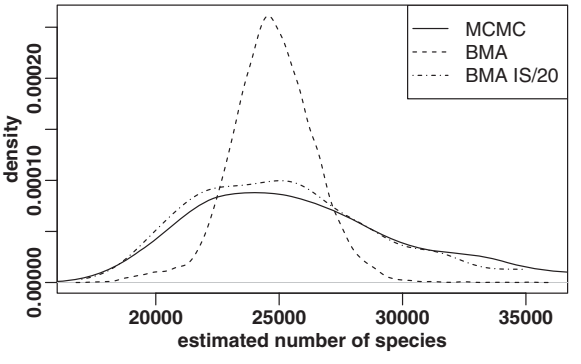

Figure 6. The MCMC, BMA and BMA IS/20 approximate posterior distributions of the number of species. The MCMC posterior is computed with four components.

In Figure 7, we check that the BMA IS/20 posterior distribution is not sensitive to the prior choice. In this data set, BMA averages between the three-component and the four-component model, so the BMA IS/20 posterior distribution does not exactly correspond to the MCMC posterior. As the three-component model infers less species, the BMA IS/20 posterior also estimates a lower number of species than MCMC.

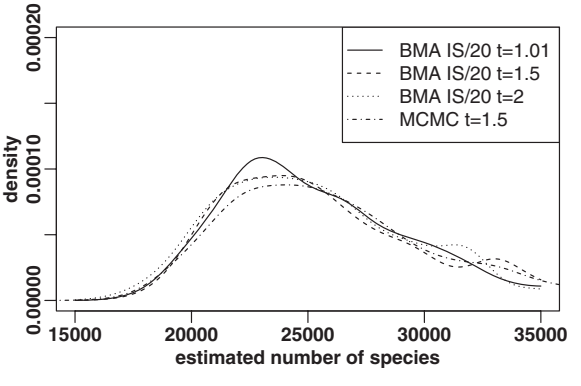

Figure 7. Sensitivity of the BMA IS/20 posterior distribution to prior choice. The MCMC posterior with four components is provided for comparison.

Table 4. Computational time in seconds required for generating posterior samples for a data set with C species.

C	BMA	BMA IS	BMA IS/10	BMA IS/20	BMA IS/30	MCMC
200	2.42	5.12	5.13	5.14	5.13	31.57
2000	5.47	5.72	5.72	5.73	5.74	307.44
20,000	16.58	5.91	5.91	5.91	5.93	0
200,000	76.18	7.08	7.08	7.08	7.08	0

4.4 *Computational time*

The computations were run on a workstation with a Core 2 Duo E7300 processor with 3 GB of RAM. Table 4 presents the cpu time used for generating the benchmark given in Section 4.1. The computation of the MCMC posterior requires roughly linear time and takes 5 min per data set at $C = 2000$. Samples for larger values of C were not generated due to the large amounts of computational time required.

The reported cpu time for the BMA method includes running the VB-EM algorithm for every number of components, as well as generating the samples and computing the confidence intervals. Most of the time is consumed in the VB-EM loop. For the BMA IS methods, the number corresponds to the time required for generating the samples and computing the confidence interval, but the estimation of the BMA model is left out. Importance sampling is implemented efficiently and only depends on the number of generated samples. Most of the time is in fact spent on sampling Dirichlet and Beta variables. There is a small additional amount of time required for pre-processing large data sets.

For the metagenomics data set, the BMA method is about 60 times faster than the MCMC sampling. Fitting all the five models in the BMA algorithm took around 7 s and an additional 16 s for generating the BMA IS/20 samples. In comparison, the MCMC simulation for the four-component model took around 1357 s (23 min). Speed improvement is variable between data sets, as it depends on the convergence of the VB-EM algorithm, which is slower for complex models. In any case, the MCMC posterior requires a significant number of iterations for burn-in and large steps between samples to obtain uncorrelated samples.

5. Conclusion

The truncated mixture model is an appropriate tool for estimating the SAD in ecological studies. In this paper, we have shown how to perform Bayesian estimation of the parameters and obtain non-asymptotic credibility intervals when estimating the number of species that are missing from the data set due to limited sampling. The model as well as the variational methods can be applied in larger contexts.

BMA is a useful tool that bridges parametric and non-parametric statistical approaches. In particular, it removes potential biases in model selection and is more flexible. A future perspective would be to average over several classes of models of different forms. For instance, we would like to compare mixtures of geometric distributions with lognormal models.

Acknowledgements

This work was supported by the French Research National Agency (ANR) in the CBME project.

Note

1. The number of components M of the mixture model is fixed in this section. It becomes a random variable, associated with index m given in Section 3.4 when considering model averaging.

References

[1] H. Akaike, *A new look at the statistical model identification*, IEEE Trans. Autom. Control 19(6) (1974), pp. 716–723.

[2] M. Beal and Z. Ghahramani, *The variational Bayesian EM algorithm for incomplete data: With application to scoring graphical model structures*, Bayesian Statist. 7 (2003), pp. 453–464.

[3] D. Bohning and R. Kuhnert, *Equivalence of truncated count mixture distributions and mixtures of truncated count distributions*, Biometrics 62(4) (2006), pp. 1207–1215.

[4] J. Bunge and K. Barger, *Parametric models for estimating the number of classes*, Biom. J. 50(6) (2008), pp. 971–982.

[5] J. Bunge and M. Fitzpatrick, *Estimating the number of species: A review*, J. Amer. Statist. Assoc. 88(421) (1993), pp. 364–373.

[6] A. Chao, *Nonparametric estimation of the number of classes in a population*, Scand. J. Statist. 11(4) (1984), pp. 265–270.

[7] A. Chao, *Species richness estimation*, Encyclopedia Statist. Sci. 12 (2005), pp. 7907–7916.

[8] B. Efron and R. Thisted, *Estimating the number of unseen species: How many words did Shakespeare know?* Biometrika 63(3) (1976), pp. 435–447.

[9] W. Esty, *Estimation of the size of a coinage: A survey and comparison of methods*, Numismatic Chronicle 146 (1986), pp. 185–215.

[10] R. Fisher, A. Corbet, and C. Williams, *The relation between the number of species and the number of individuals in a random sample of an animal population*, J. Anim. Ecol. 12(1) (1943), pp. 42–58.

[11] N. Gotelli, M. Anderson, H. Arita, A. Chao, R. Colwell, S. Connolly, D. Currie, R. Dunn, G. Graves, J. Green, J.A. Grytnes, Y.H. Jiang, W. Jetz, S. Kathleen Lyons, C.M. McCain, A.E. Magurran, C. Rahbek, T.F. Rangel, J. Soberón, C.O. Webb, and M.R. Willig, *Patterns and causes of species richness: A general simulation model for macroecology*, Ecol. Lett. 12(9) (2009), pp. 873–886.

[12] J. Green and J. Plotkin, *A statistical theory for sampling species abundances*, Ecol. Lett. 10(11) (2007), pp. 1037–1045.

[13] J. Handelsman, J. Tiedje, L. Alvarez-Cohen, M. Ashburner, I. Cann, E. DeLong, W. Ford Doolittle, C.M. Fraser-Liggett, A. Godzik, J.I. Gordon, M. Riley, and M.B. Schmid, *The New Science of metagenomics: Revealing the secrets of our microbial planet* (2007). http://www.nap.edu/openbook.php?record_id=11902&page=R5

[14] B. Harris, *Determining bounds on integrals with applications to cataloging problems*, Ann. Math. Statist. 30(2) (1959), pp. 521–548.

[15] J. Hoeting, *Methodology for Bayesian model averaging: An update*, International Biometrics Conference Proceedings, Citeseer, 2002.

[16] C. Keribin, *Méthodes bayésiennes variationnelles: Concepts et applications en neuroimagerie*, Journal de la Société Française de Statistique 151(2) (2010), pp. 107–131.

[17] L. Legrand, J. Tap, C. Gauthey, J. Doré, C. Caron, and M. Leclerc, *RapidOTU: A fast pipeline to analyse 16S rDNA sequences by alignment or tetranucleotide frequency*, Proc. Gut Microbiome Symp., 6th congress INRA Rowett Research Institute, Clermont-Ferrand, 2008, p. 35.

[18] J. Longino, J. Coddington, and R. Colwell, *The ant fauna of a tropical rain forest: Estimating species richness three different ways*, Ecology 83(3) (2002), pp. 689–702.

[19] C. Mao and B. Lindsay, *Estimating the number of classes*, Ann. Statist. 35(2) (2007), pp. 917–930.

[20] B. McGill, R. Etienne, J. Gray, D. Alonso, M. Anderson, H. Benecha, M. Dornelas, B. Enquist, J. Green, F. He, H.H. Hurlbert, A.E. Magurran, P.A. Marquet, B.A. Maurer, A. Ostling, C.U. Soykan, K.I. Ugland, and E.P. White, *Species abundance distributions: Moving beyond single prediction theories to integration within an ecological framework*, Ecol. Lett. 10(10) (2007), pp. 995–1015.

[21] L. Sanathanan, *Estimating the size of a truncated sample*, J. Amer. Statist. Assoc. 72(359) (1977), pp. 669–672.

[22] G. Schwarz, *Estimating the dimension of a model*, Ann. Statist. 6(2) (1978), pp. 461–464.

[23] J. Tap, S. Mondot, F. Levenez, E. Pelletier, C. Caron, J. Furet, E. Ugarte, R. Muñoz-Tamayo, D. Paslier, R. Nalin, J. Dore, and M. Leclerc, *Towards the human intestinal microbiota phylogenetic core*, Environ. Microbiol. 11(10) (2009), pp. 2574–2584.

[24] D.M. Titterington and B. Wang, *Inadequacy of interval estimates corresponding to variational Bayesian approximations*, in *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics*, R.G. Cowell and Z. Ghahramani, eds., Society for AISTATS, The Savannah Hotel, Barbados, 2005, pp. 373–380.

[25] V. Torsvik and L. Øvreås, *Microbial diversity and function in soil: From genes to ecosystems*, Curr. Opin. Microbiol. 5(3) (2002), pp. 240–245.