Mixture models of truncated data for estimating the number of species.

Li-Thiao-Té Sébastien, Jean-Jacques Daudin, Stéphane Robin

Equipe Statistique et Génome, UMR 518 AgroParisTech / INRA MIA

19th COMPSTAT symposium, 23rd August 2010

Context

Situation

- individuals are sampled from a population then classified into species
- Goal 1 : estimate the species abundance distribution
- Goal 2 : estimate the number of species with no sampled individual

Applications

- ecological surveys : number of species of butterflies [Fisher et al., 1943]
- metagenomics (our interest, large number of unobserved species, large datasets)
- other : number of words in a language, number of unreported drug addicts

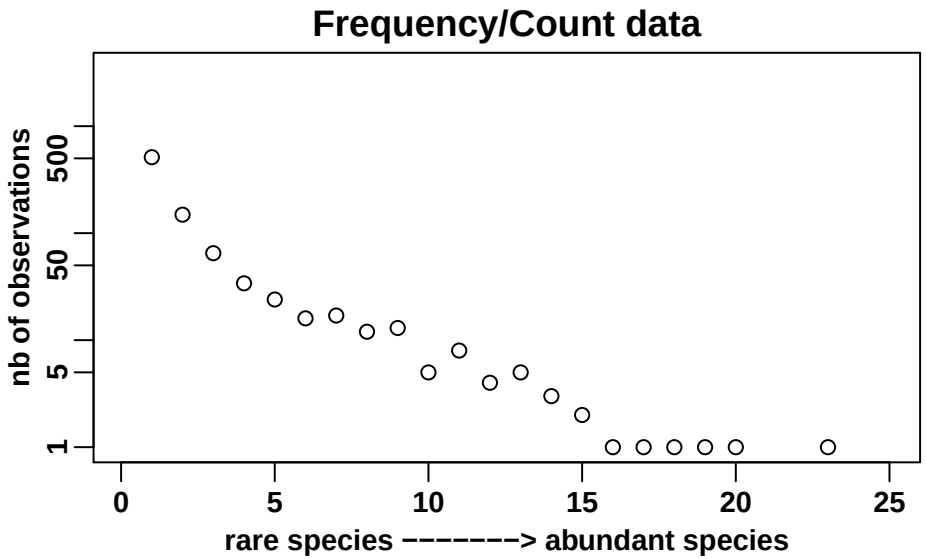
Example

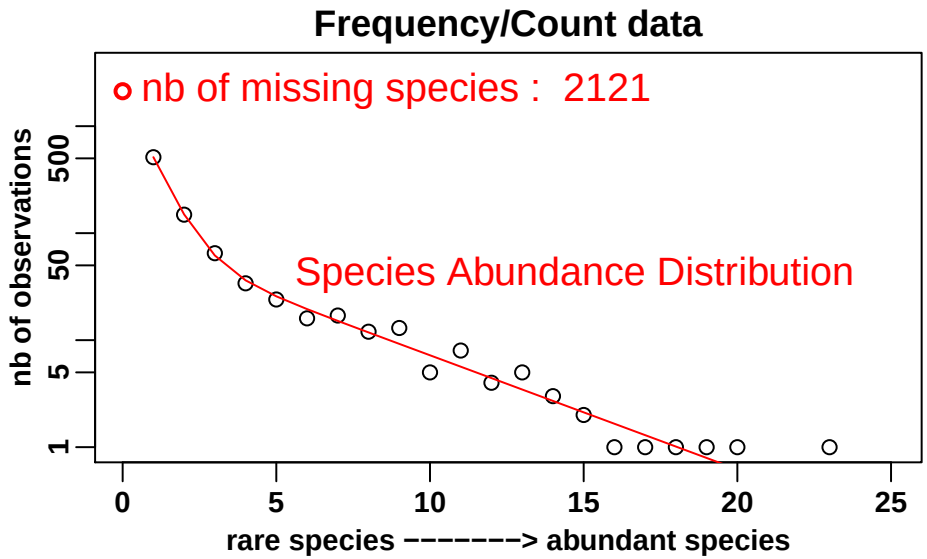
Observations

Species	A	B	C	D	E	...
Number of individuals	10	430	10	289	3	...

Species abundance distribution

Number of individuals	1	2	3	4	5	...
Number of species	513	149	65	34	24	...





Sampling model

- species abundance distribution :

$$f(\lambda) = \sum_q \alpha_q f_q(\lambda)$$

- X_i individuals are observed for species i conditionally on its abundance λ_i (Poisson distributed number)

$$f(X_i|\lambda_i) = \frac{\exp^{-\lambda_i} \lambda_i^{X_i}}{X_i!}$$

- only positive numbers of individuals are recorded in the data set :

$$f(X_+|\lambda_i) = f(X_i|\lambda_i, X_i > 0)$$

Truncated model ($\vartheta = \{\alpha_q, \pi_q\}$) :

$$X_+ \sim f(x, \vartheta) = \frac{\sum_q^Q \alpha_q f_q(x, \pi_q)}{1 - \sum_q^Q \alpha_q f_q(0, \pi_q)}$$

Bayesian model

$$X_+ \sim f(x, \pi_q) = \frac{\sum_q^Q \alpha_q f_q(x, \pi_q)}{1 - \sum_q^Q \alpha_q f_q(0, \pi_q)}$$

A priori :

$$\begin{aligned} \alpha &\sim \text{Dirichlet}(\vec{a}) \\ \pi_q &\sim \text{Beta}(b_q, c_q) \\ Z &\sim \text{Multinom}(\vec{a}) \\ X|Z &\sim \text{Geom}(\pi_q) \end{aligned}$$

Approximate a posteriori distribution :

$$\begin{aligned} \alpha|X &\sim \text{Dirichlet}(\tilde{a}) \\ \pi_q|X &\sim \text{Beta}(\tilde{b}_q, \tilde{c}_q) \end{aligned}$$

The hyper parameters $\tilde{a}$, $\tilde{b}$ and $\tilde{c}$ provide an approximation of the a posteriori distribution and hence confidence intervals.

Variational framework

Theorem

The log-likelihood can be decomposed into :

$$\log P(X) = \log \iint P(X, Z, \vartheta)\, \mathrm{d}Z\, \mathrm{d}\vartheta$$
$$= \mathcal{F}(X, Q) + KL(Q, P(.|X))$$

where $\mathcal{F}(X, Q) := \iint \log \frac{P(X, Z, \vartheta)}{Q(Z, \vartheta)} Q(Z, \vartheta) \mathrm{d}Z \mathrm{d}\vartheta.$

Consequently :

- $\log P(X) \geq \mathcal{F}(X, Q)$
- *if* $Q = \text{argmax } \mathcal{F}(X, Q)$ *then* $Q = \text{argmin } KL(Q, P(.|X))$.
- $\text{argmax } \mathcal{F}(X, Q) = P(Z, \vartheta|X)$

VB-EM algorithm

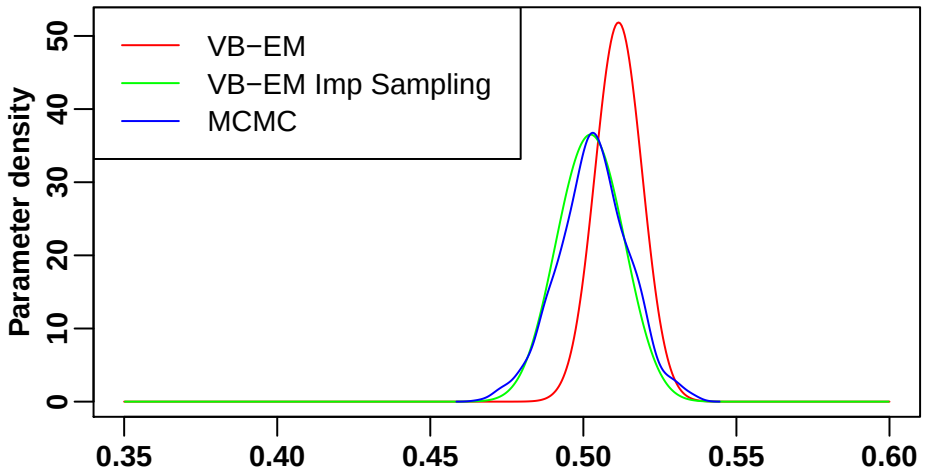
Application of [Beal and Ghahramani, 2003] leads to the following update formulae :

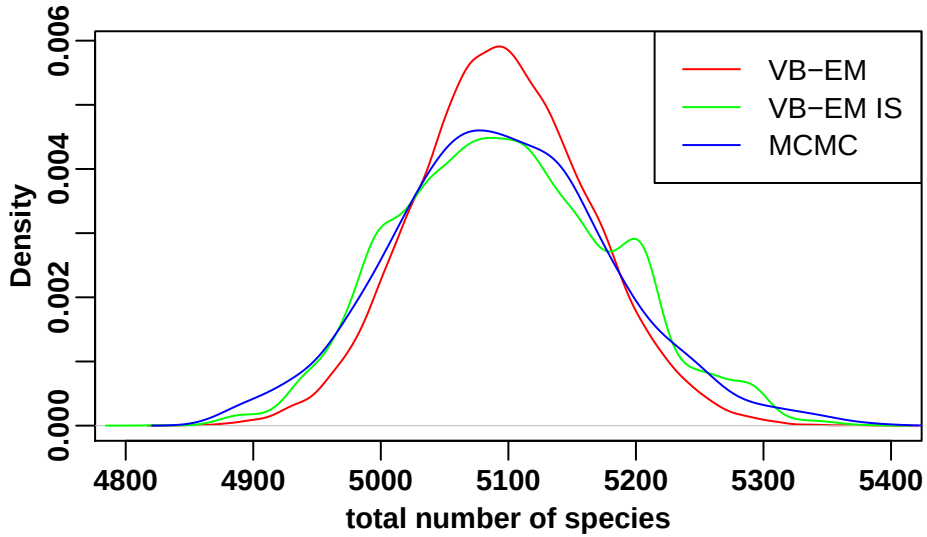
$$\begin{cases} a_q^{(n+1)} & = a_q^0 + \sum_i \tau_{iq}^{(n)} \\ b_q^{(n+1)} & = b_q^0 + \sum_i \tau_{iq}^{(n)} (X_i - 1) \\ c_q^{(n+1)} & = c_q^0 + \sum_i \tau_{iq}^{(n)} \end{cases}$$

where $\tau_{iq}^{(n)} = \mathbb{P}_{Q_{Z_i}}(Z_i = q)$.

Consequences :

- Approximate posterior distribution
- approximate non asymptotic credibility intervals
- proposal distribution for importance sampling





Bayesian Model Averaging

Let M denote the (random) number of components in the mixture model. Then the BMA model is

$$f_{\text{BMA}} = \sum_m \mathbb{P}(M = m|X) f_m$$

where f_m is the posterior density of the observations given a model with m components.

The weights $\mathbb{P}(M = m|X)$ can be computed based on the Bayes formula :

$$\mathbb{P}(M = m|X) \propto \mathbb{P}(X|M)\mathbb{P}(M)$$

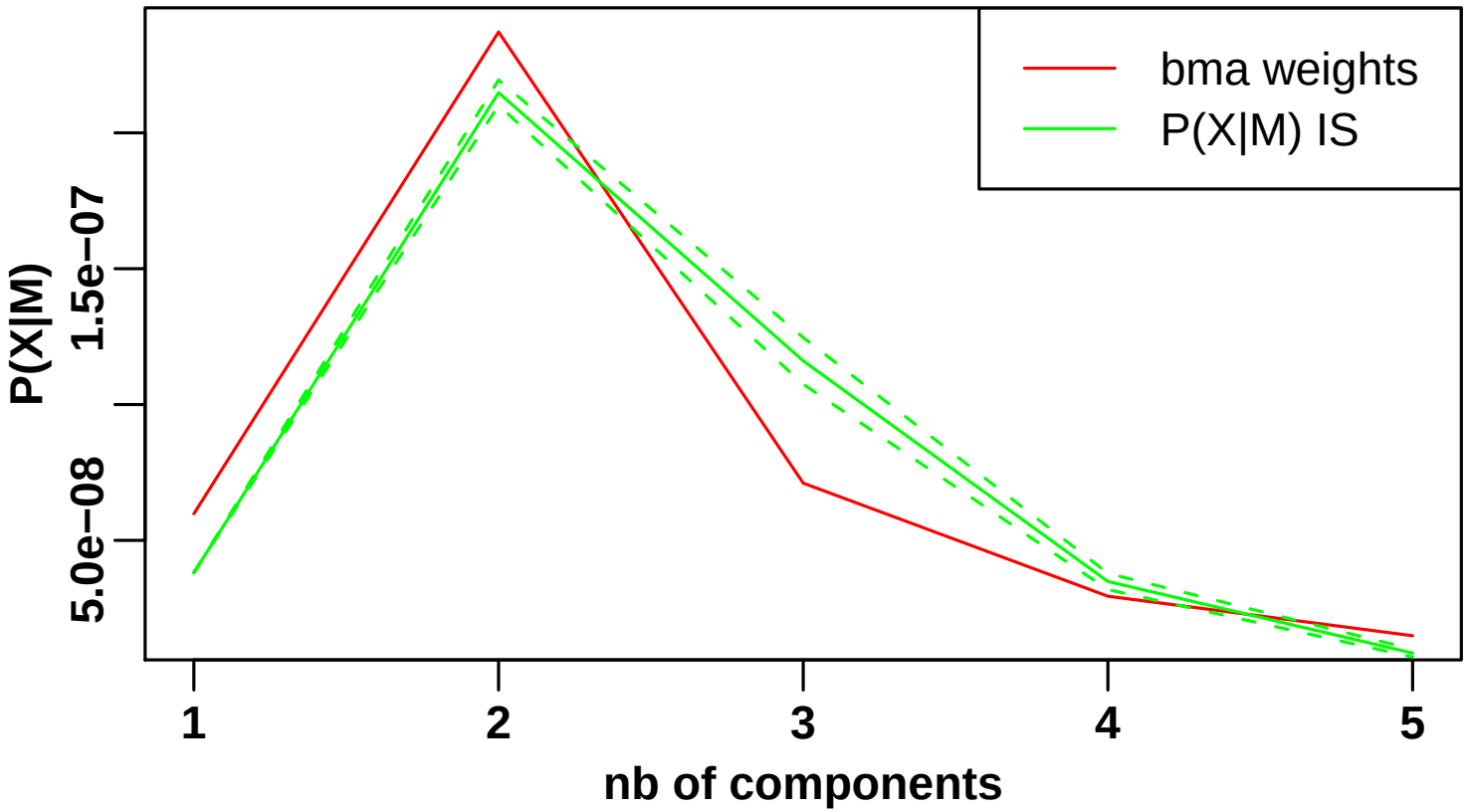
where $\mathbb{P}(M)$ is the a priori distribution on M.

The evidence $\mathbb{P}(X|M)$ is hard to compute in general ; the VB-EM algorithm provides the approximation

$$\log P(X) \sim \mathcal{F}(X, Q) = \iint \log \frac{P(X, Z, \vartheta)}{Q(Z, \vartheta)} Q(Z, \vartheta) \mathrm{d}Z \mathrm{d}\vartheta$$

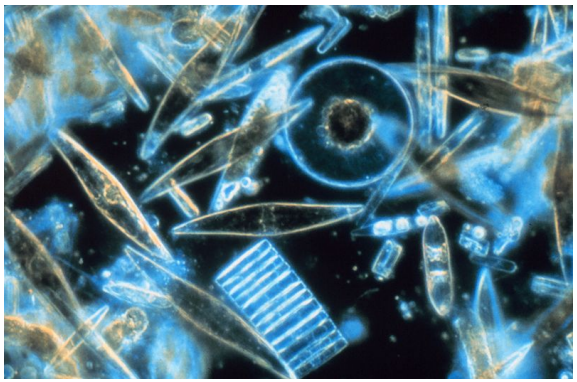
where the error term $KL(Q, P(.|X))$ has been neglected.

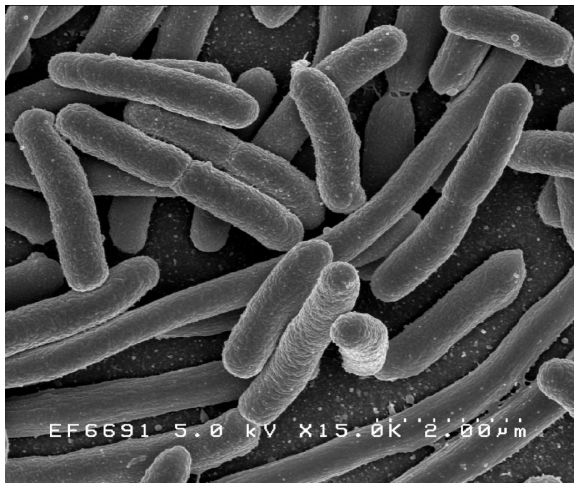
Bayesian Model Averaging (example)



Metagenomics

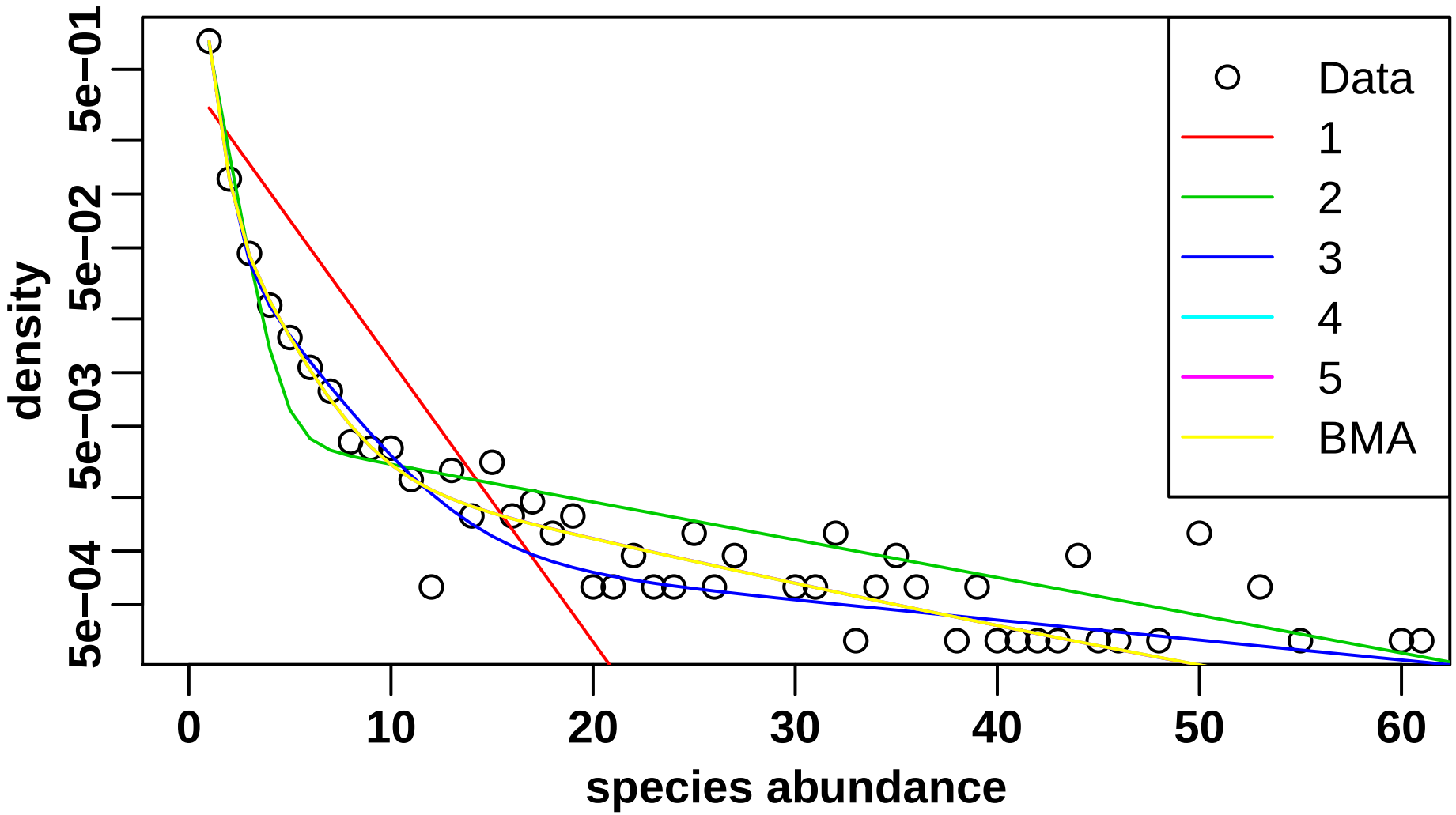
- High throughput DNA sequencing
- Complex environmental samples : soil, seawater, intestine microflora





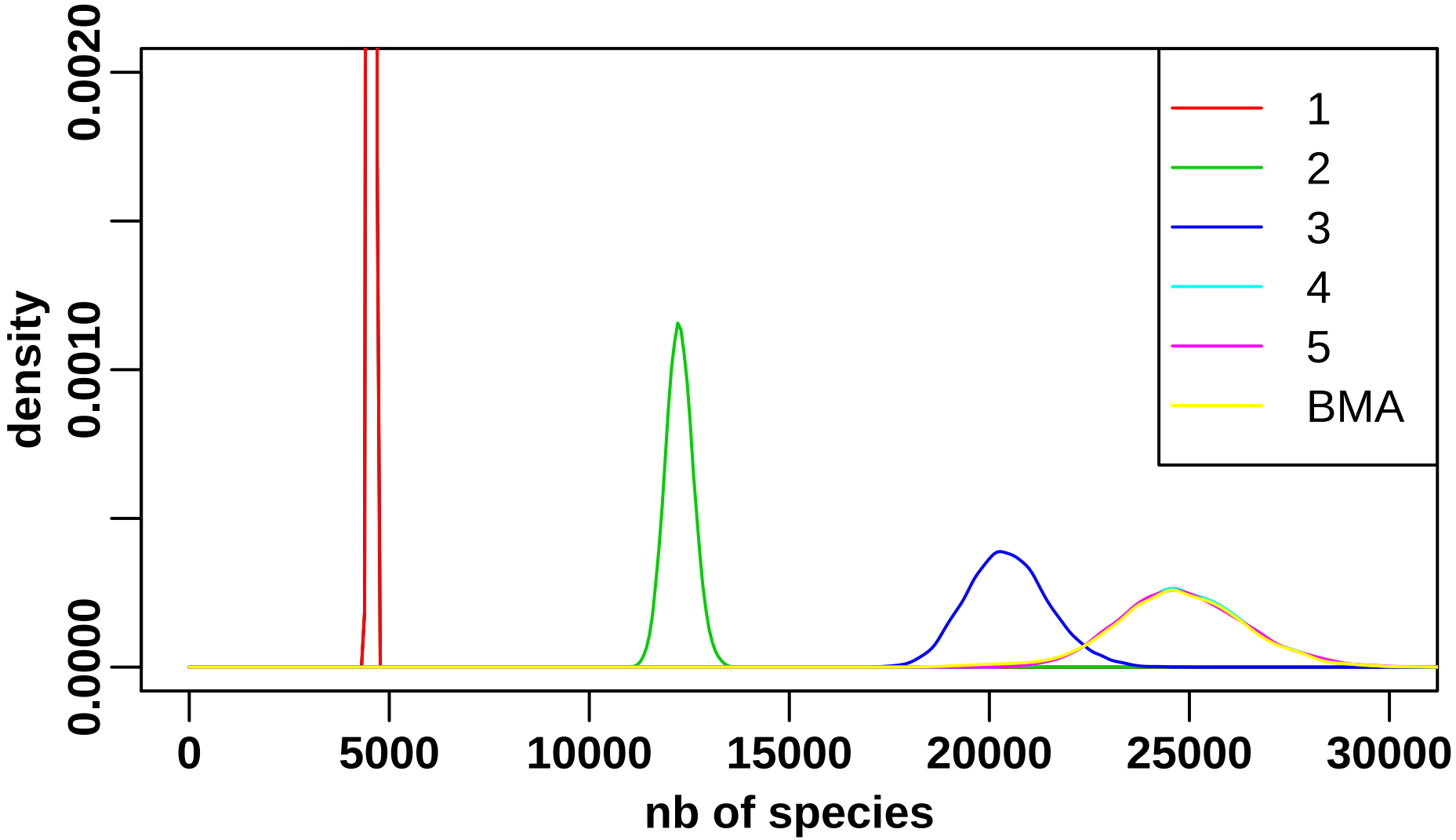
Real dataset example

Model fit to the data (human gut microbiota [Tap et al., 2009]) :



Real dataset example

Estimated number of species and approximate posterior distributions :



 Beal, M. and Ghahramani, Z. (2003). The variational Bayesian EM algorithm for incomplete data : with application to scoring graphical model structures. Bayesian Statistics 7 (pp. 453–464).

 Fisher, R., Corbet, A., and Williams, C. (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(1) :42–58.

 Tap, J., Mondot, S., Levenez, F., Pelletier, E., Caron, C., Furet, J., Ugarte, E., Muñoz-Tamayo, R., Paslier, D., Nalin, R., et al. (2009). Towards the human intestinal microbiota phylogenetic core. *Environmental Microbiology*, 11(10) :2574–2584.