Estimating Species Abundance. Application to Metagenomics

S. Robin

AgroParisTech / INRA

X, Rencontres MMB, February 2013

Species abundance

How many species are there?

An old ecological problem when exploring a given environment: how many species are not observed?

- X_i = number of observed individuals from species i,
- C_x = number of species with x observed individuals,
- C = total number of species $= \sum_{x \geq 0} C_x$.

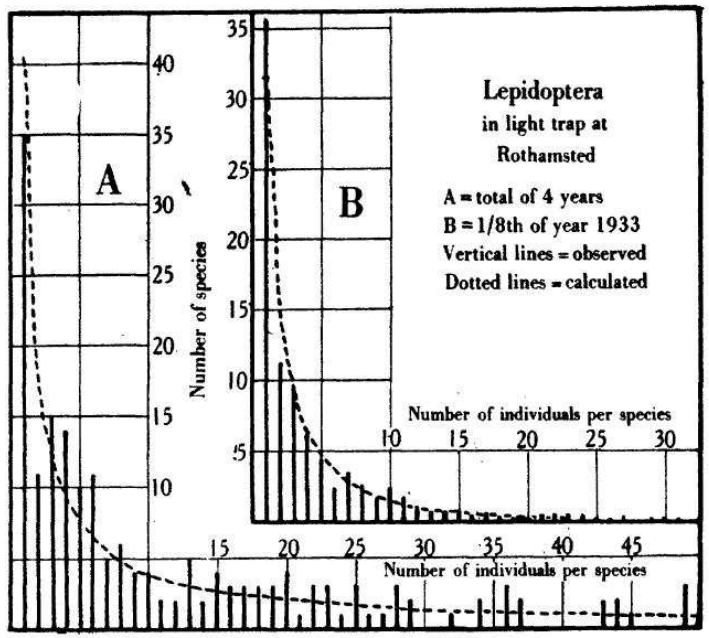
Problem: $\widehat{C}_0 = ?$, $\widehat{C} = ?$

How many species are there?

An old ecological problem when exploring a given environment: how many species are not observed?

- X_i = number of observed individuals from species i,
- C_x = number of species with x observed individuals,
- C = total number of species $= \sum_{x \geq 0} C_x$.

Problem: $\widehat{C}_0 = ?$, $\widehat{C} = ?$



Fisher et al. *(1943)*

Bacterial communities

Biological context:

- Many bacterial species can not be grown artificially out of their natural environment.
- Sets species can only be studied all together, within their environment, e.g. ocean, human gut, soil, cheese surface, etc.

Bacterial communities

Biological context:

- Many bacterial species can not be grown artificially out of their natural environment.
- Sets species can only be studied all together, within their environment, e.g. ocean, human gut, soil, cheese surface, etc.

Their diversity and functions can be studied via NGS by sampling and sequencing DNA (or RNA) from all species (*McHardy and Rigoutsos (2007)*).

Bacterial communities

Biological context:

- Many bacterial species can not be grown artificially out of their natural environment.
- Sets species can only be studied all together, within their environment, e.g. ocean, human gut, soil, cheese surface, etc.

Their diversity and functions can be studied via NGS by sampling and sequencing DNA (or RNA) from all species (*McHardy and Rigoutsos (2007)*).

Data:

- X_i = number of reads from species i (if the genome is available)
- X_i = number of reads from gene i (whatever the species)

Species abundance distribution

General strategy: The observed counts $\{X_i\}$ are truncated, meaning that 0's are not observed.

1. Suppose that the 'complete' counts are iid, with distribution g:

$$g = \text{species abundance distribution (SAD)};$$

2. The observed counts $\{X_i\}$ are iid with truncated SAD g^+

$$g^+(x) = \frac{g(x)}{1 - g(0)}, \qquad \text{for } x > 0;$$

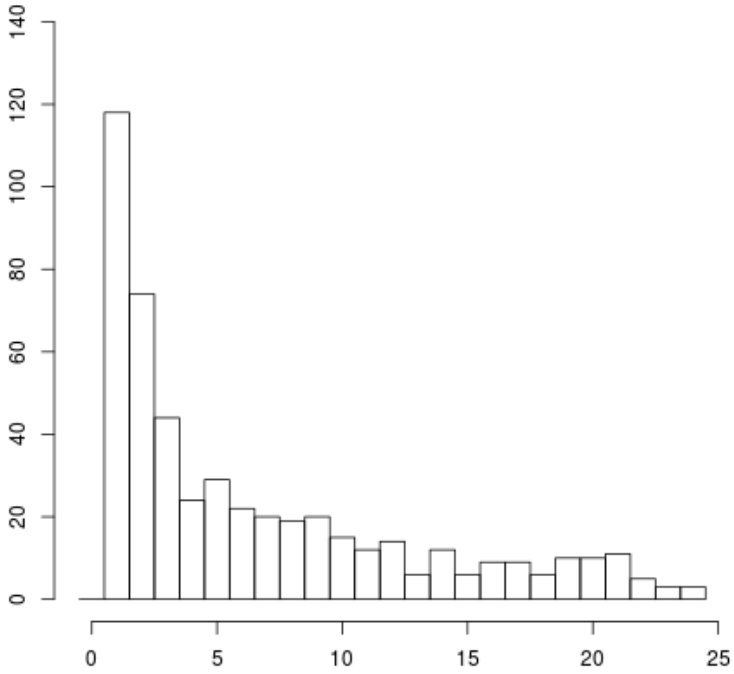
3. Fit some (parametric?) distribution to the $\{X_i\} \to \widehat{g}^+(\cdot) = g^+(\cdot; \widehat{\gamma})$;
4. Estimate $g(0)$ with the Horwitz-Thomson estimate

$$\widehat{C} = c \,/[1 - \widehat{g}(0)] \,.$$

Estimation of abundance

Standard strategy.

Data: $X = \{X_i\}$, X_i = number of individuals from species i.



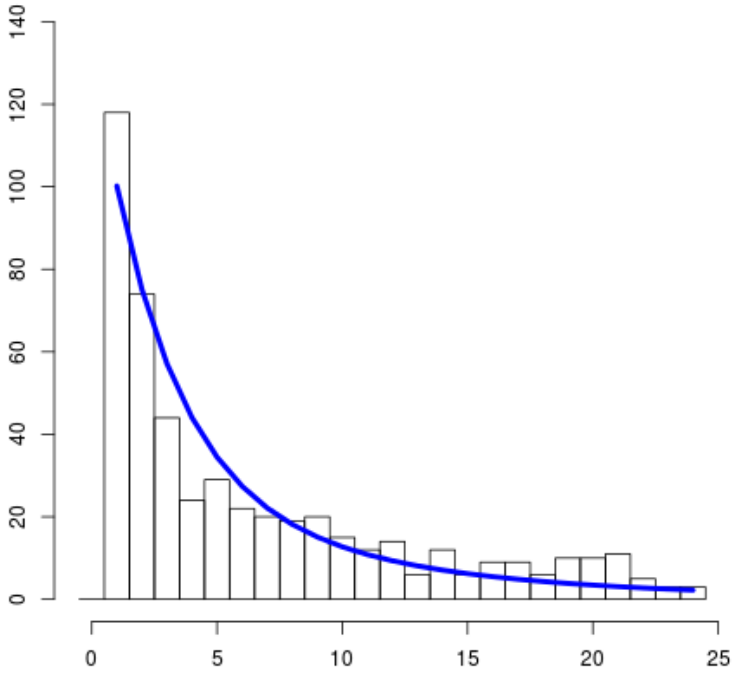
Estimation of abundance

Standard strategy.

Data: $X = \{X_i\}$, X_i = number of individuals from species i.

Fit some (truncated) distribution g^+ to X:

$$g^+(x) = g(x)\,/[1 - g(0)]\,, \quad x > 0.$$



Estimation of abundance

Standard strategy.

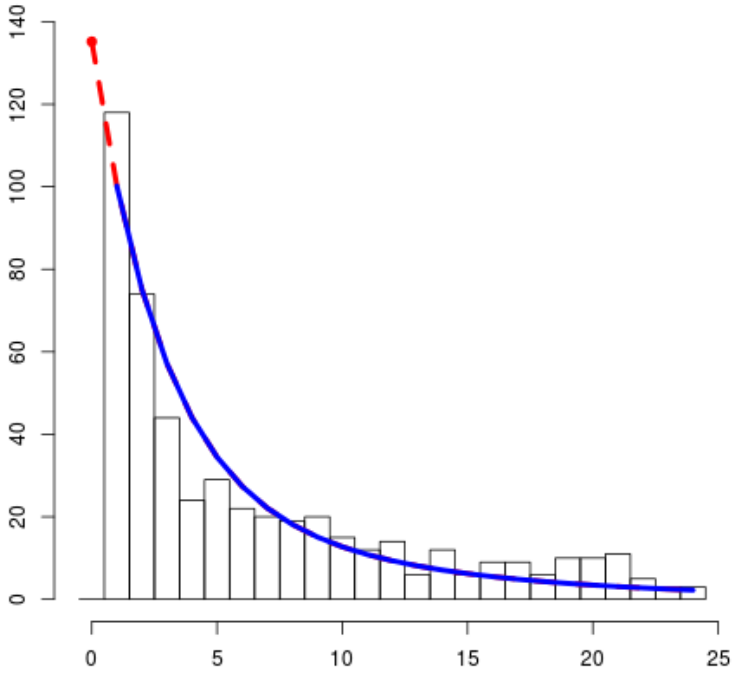
Data: $X = \{X_i\}$, X_i = number of individuals from species i.

Fit some (truncated) distribution g^+ to X:

$$g^+(x) = g(x) / [1 - g(0)] , \quad x > 0.$$

Estimate C with the Horwitz-Thomson estimate

$$\widehat{C} = c / [1 - \widehat{g}(0)] .$$



Estimation of abundance

Standard strategy.

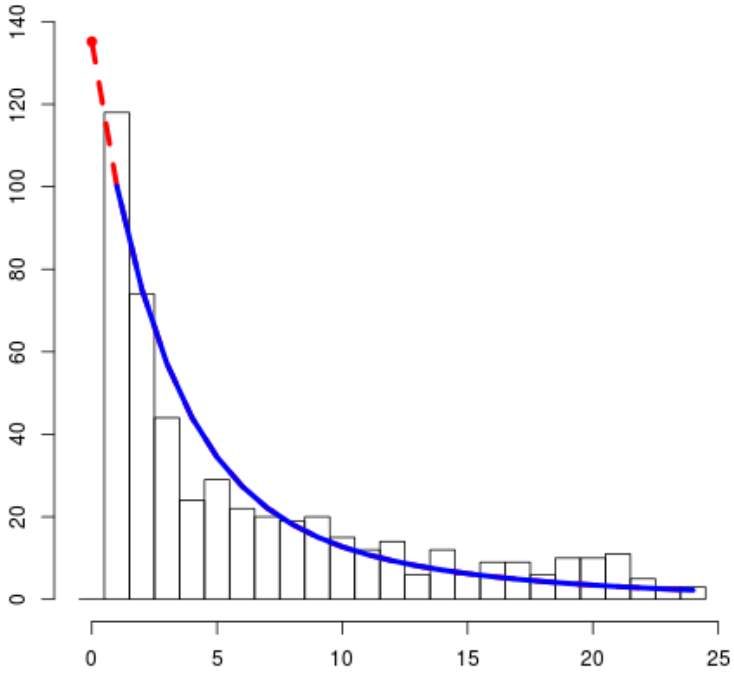
Data: $X = \{X_i\}$, X_i = number of individuals from species i.

Fit some (truncated) distribution g^+ to X:

$$g^+(x) = g(x) / [1 - g(0)] , \quad x > 0.$$

Estimate C with the Horwitz-Thomson estimate

$$\widehat{C} = c / [1 - \widehat{g}(0)] .$$



Estimation of abundance

Standard strategy.

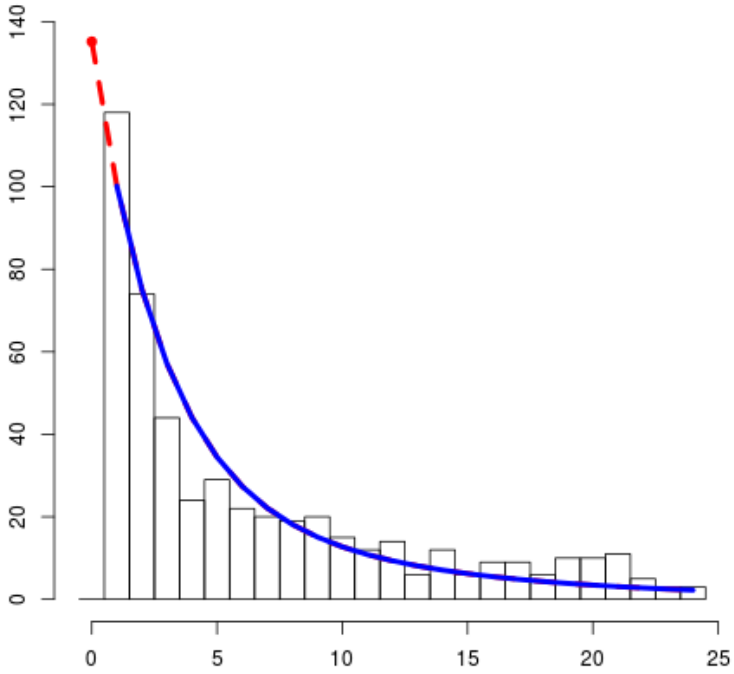
Data: $X = \{X_i\}$, X_i = number of individuals from species i.

Fit some (truncated) distribution g^+ to X:

$$g^+(x) = g(x) / [1 - g(0)] , \quad x > 0.$$

Estimate C with the Horwitz-Thomson estimate

$$\widehat{C} = c / [1 - \widehat{g}(0)] .$$



Species abundance distribution (SAD)

Some classical distributions:

- Poisson;
- Log-normal (*Doroghazi and Buckley (2008)*);
- Poisson-Gamma (*Fisher* et al. *(1943)*, *Hooper* et al. *(2010)*) = Poisson counts with Gamma intensities;
- Mixture of discrete distributions $f(\cdot;\gamma)$:

$$g(x) = \int f(x;\gamma)\pi(\gamma)\mathrm{d}\gamma$$

Species abundance distribution (SAD)

Some classical distributions:

- Poisson;
- Log-normal (*Doroghazi and Buckley (2008)*);
- Poisson-Gamma (*Fisher* et al. *(1943)*, *Hooper* et al. *(2010)*) = Poisson counts with Gamma intensities;
- Mixture of discrete distributions $f(\cdot;\gamma)$:

$$g(x) = \int f(x;\gamma)\pi(\gamma)\mathrm{d}\gamma$$

Interest of the SAD:

- Modeling the SAD allows to guaranty identifiability.
- SAD provides the saturation curve

$$\Pr\{X_i > 0\}$$

which is useful to design experiments.

In this talk

Goal:

- Provide an estimate of $g(0)$
- With confidence bounds.

1. Bayesian averaging of mixture models

2. A 'true' non-parametric estimate

In this talk

Goal:

- Provide an estimate of $g(0)$
- With confidence bounds.

1. Bayesian averaging of mixture models

2. A 'true' non-parametric estimate

Bayesian averaging of mixture models

Joint work with

- S. Li-Thiao-Té,
- J.-J. Daudin

Mixture models

'Non-parametric' = mixture model: *Norris and Pollock (1998)*

$$\pi(\gamma) = \sum_k \pi_k \delta_{\gamma_k}(\gamma) \quad \Rightarrow \quad g(x) = \sum_k \pi_k f(x; \gamma_k).$$

Mixture models

'Non-parametric' = mixture model: *Norris and Pollock (1998)*

$$\pi(\gamma) = \sum_k \pi_k \delta_{\gamma_k}(\gamma) \quad \Rightarrow \quad g(x) = \sum_k \pi_k f(x;\gamma_k).$$

Truncated mixture vs Mixture of truncated. The distribution of the observed counts can be expressed in two equivalent ways:

$$g^+(x) = \sum_k \pi_k f(x;\gamma_k) \Big/ \left[1 - \sum_k \pi_k f(0;\gamma_k)\right] \tag{1}$$

$$\text{or} \qquad g(x) = \sum_k \pi_k^+ f^+(x;\gamma_k). \tag{2}$$

Incomplete data model

A mixture model can be rewritten as:

$$\begin{aligned} (Z_i)_i \text{ iid:} \quad Z_i &\sim \mathcal{M}(1;\pi), \\ (X_i)_i \text{ indep. } |(Z_i)_i : \quad X_i|Z_i = k &\sim f^+(\cdot;\gamma_k) \end{aligned}$$

where Z_i is the unknown group to which species i belongs.

Notations:

$X = (X_i)_i$ observed counts,

$Z = (Z_i)_i$ unobserved groups,

$\theta = (\pi, \gamma)$ parameter $(\gamma_k)_k$.

Incomplete data model

A mixture model can be rewritten as:

$$\begin{aligned} (Z_i)_i \text{ iid:} \quad Z_i &\sim \mathcal{M}(1;\pi), \\ (X_i)_i \text{ indep. } |(Z_i)_i : \quad X_i|Z_i = k &\sim f^+(\cdot;\gamma_k) \end{aligned}$$

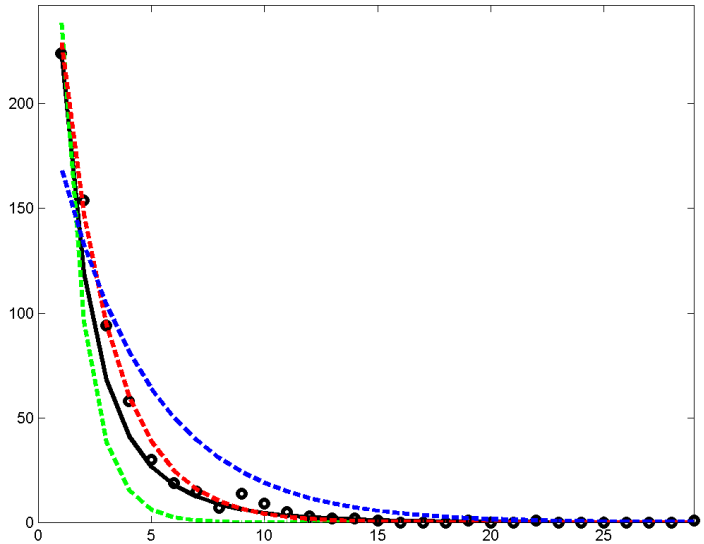
where Z_i is the unknown group to which species i belongs.

Notations:

$X = (X_i)_i$ observed counts,

$Z = (Z_i)_i$ unobserved groups,

$\theta = (\pi, \gamma)$ parameter $(\gamma_k)_k$.



- We need get an estimate $\widehat{\theta}$
- or to calculate the posterior $P(\theta|X)$.

Inference

Inference on truncated data:

- Inference of mixture of truncated (1) is often easier than this of truncated mixture (2).
- MLE estimates for (1) and (2) are equivalent (*Bohning and Kuhnert (2006)*) in the Poisson case.

Inference

Inference on truncated data:

- Inference of mixture of truncated (1) is often easier than this of truncated mixture (2).
- MLE estimates for (1) and (2) are equivalent (*Bohning and Kuhnert (2006)*) in the Poisson case.

Bayesian inference

- Bayesian inference provides credibility interval through the posterior $P(\theta|X)$.
- Exact Bayesian inference with incomplete data requires computationally intensive MCMC.
- Variational Bayes provides an (optimal) approximation of the joint posterior $P(\theta, Z|X)$.

Exponential family / Conjugate prior

Exponential family:

$$P(X, Z|\theta) \propto \exp[\psi(\theta)' u(X, Z)]$$

includes distributions like geometric, Poisson, truncated geometric ... but not truncated Poisson (while they can still be handled...)

Exponential family / Conjugate prior

Exponential family:

$$P(X, Z|\theta) \propto \exp[\psi(\theta)' u(X, Z)]$$

includes distributions like geometric, Poisson, truncated geometric ... but not truncated Poisson (while they can still be handled...)

Conjugate prior.

$$P(\theta) \propto \exp[\psi(\theta)' \boldsymbol{\nu}]$$

that is

- Dirichlet for the multinomial distribution (Z),
- Gamma for Poisson or Beta for the geometric $(X|Z)$,

$$\Rightarrow \qquad P(\theta|X, Z) \propto \exp\{\psi(\theta)'[u(X, Z) + \boldsymbol{\nu}]\}.$$

Variational Bayes E-M

Best approximation. As $P(\theta, Z|X)$ is intractable, we look for the best 'manageable' approximation:

$$\begin{aligned} Q^*(\theta, Z) &= \arg\min_{Q \in \mathcal{Q}} KL[Q(Z,\theta); P(Z,\theta|X)] \\ &= \arg\min_{Q \in \mathcal{Q}} \mathcal{H}(Q) - \mathbb{E}_Q[\log P(X,Z,\theta)] + \text{cst} \end{aligned}$$

Variational Bayes E-M

Best approximation. As $P(\theta, Z|X)$ is intractable, we look for the best 'manageable' approximation:

$$\begin{aligned} Q^*(\theta, Z) &= \arg\min_{Q \in \mathcal{Q}} KL[Q(Z,\theta); P(Z,\theta|X)] \\ &= \arg\min_{Q \in \mathcal{Q}} \mathcal{H}(Q) - \mathbb{E}_Q[\log P(X,Z,\theta)] + \text{cst} \end{aligned}$$

Factorisable distributions. When considering the class

$$\mathcal{Q} = \{Q(\theta, Z) = Q_\theta(\theta) Q_Z(Z)\},$$

the optimal $Q^* \in \mathcal{Q}$ can be recovered via (*Beal and Ghahramani (2003)*)

$$\text{'M'-step:} \quad Q_\theta(\theta) \propto \exp\left(\psi(\theta)' \left[\mathbb{E}_{Q_Z} u(X,Z) + \boldsymbol{\nu}\right]\right)$$

$$\text{'E'-step:} \quad Q_Z(Z) \propto \exp(\mathbb{E}_{Q_\theta} \psi(\theta)' u(X,Z)]$$

Bayesian model averaging

Number of components.

- The number of components K is unknown
- ... but the existence of a 'true' number of component is questionable.

Bayesian model averaging

Number of components.

- The number of components K is unknown
- ... but the existence of a 'true' number of component is questionable.

Bayesian model averaging (BMA). Consider a parameter of interest $\Delta = \Delta(\theta)$ that can be defined for a series of models $1, \ldots, K \ldots$. Denoting

$$\mathbb{E}(\Delta|X, K) = \int \Delta(\theta) P(\theta|X, K) \mathrm{d}\theta$$

we have

$$\mathbb{E}(\Delta|X) = \sum w_k \mathbb{E}(\Delta|X, K)$$

where

$$w_K = P(K|X),$$

the calculation of which is an issue.

Evaluating the weights

Optimal variational approximation. Optimal weights can be obtained by direct minimisation of

$$KL[Q(K, Z, \theta), P(K, Z, \theta|X)]$$

to get (*Volant* et al. *(2012)*)

$$\tilde{w}_K \quad \propto \quad P(K|X) \exp\left\{-KL[Q^*(Z, \theta|K); P(Z, \theta|X, K)]\right\}.$$

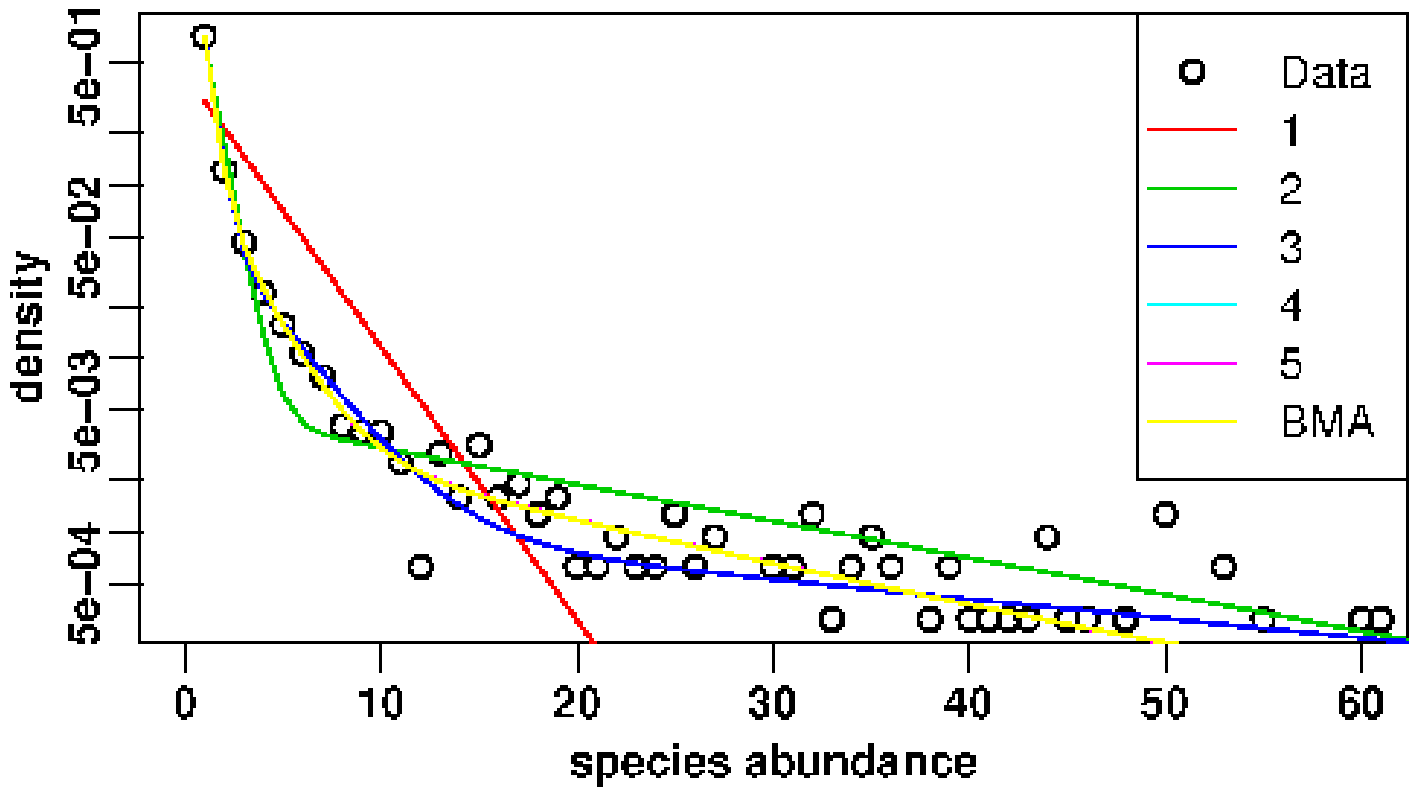
which combines

- the posterior probability of the model $P(K|X)$
- with the quality of the variational inference within the model

(although none of the two can be computed).

Microbial diversity in human gut (*Tap* et al. *(2009)*)

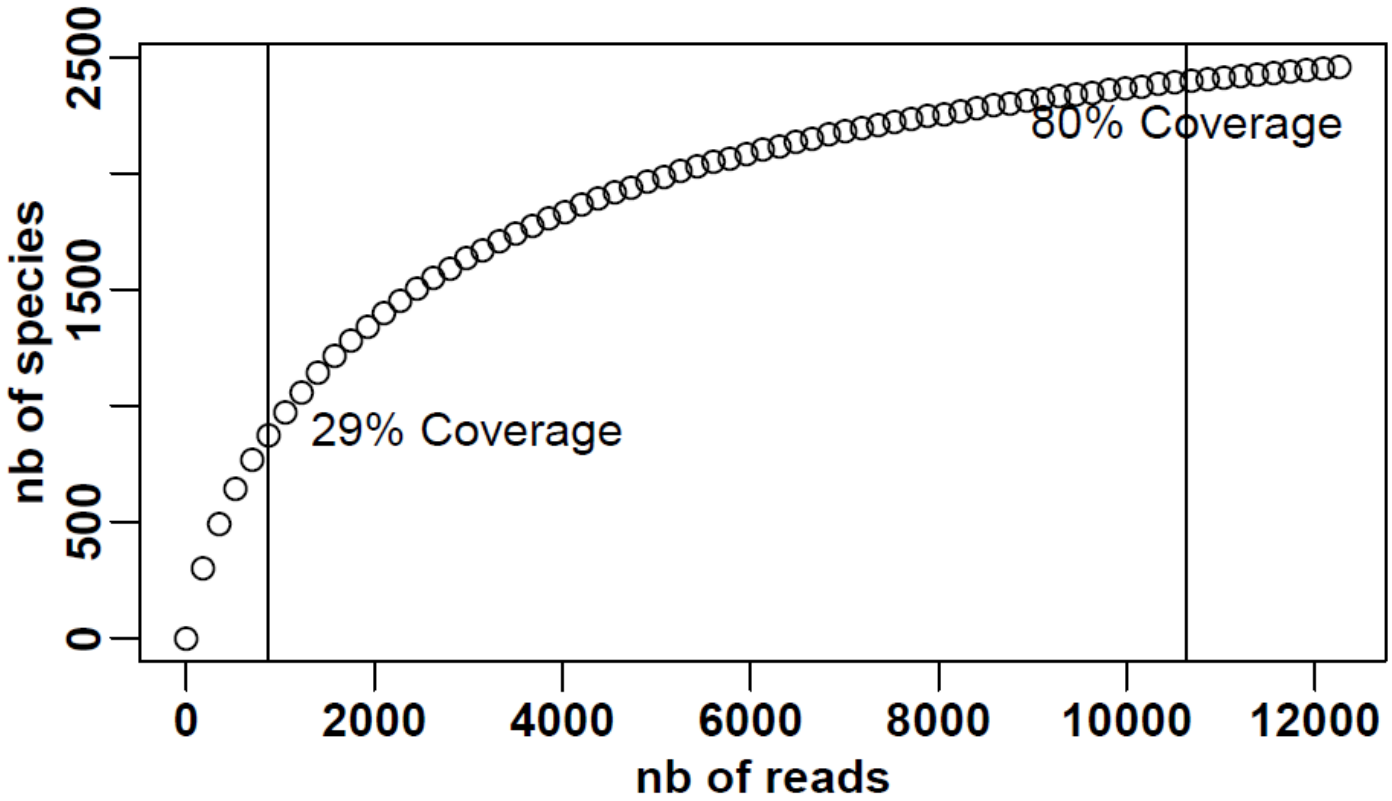
Fit of different geometric mixtures $K = 1, \ldots 5$: $\widehat{\theta}_K$ = mode of $Q_\theta(\theta)$.



Mixture: $\widehat{g}^{+K}(x) = \sum_k \widehat{\pi}_K f^+(x; \widehat{\gamma}_K)$, BMA: $\widetilde{g}^{+}(x) = \sum_K w_K \widehat{f}^{+K}(x)$.

Saturation curve

Reverse use of $\widetilde{f}^{+}(x)$: Design of NGS metagenomics experiment



Li-Thiao-Té et al. *(2012)*

Confidence interval for the number of species

Geometric distribution. The proportion of absent species under the geometric distribution is

$$\widehat{g}(0) = \widehat{\gamma}$$

for which the approximate posterior $Q_K^*(\gamma)$ is a Beta distribution.

Confidence interval for the number of species

Geometric distribution. The proportion of absent species under the geometric distribution is

$$\widehat{g}(0) = \widehat{\gamma}$$

for which the approximate posterior $Q_K^*(\gamma)$ is a Beta distribution.

Mixture of geometric, we get

$$\widehat{g}_K(0) = \sum_{k=1}^{K} \widehat{\pi}_k \widehat{\gamma}_k.$$

Confidence interval for the number of species

Geometric distribution. The proportion of absent species under the geometric distribution is

$$\widehat{g}(0) = \widehat{\gamma}$$

for which the approximate posterior $Q_K^*(\gamma)$ is a Beta distribution.

Mixture of geometric, we get

$$\widehat{g}_K(0) = \sum_{k=1}^{K} \widehat{\pi}_k \widehat{\gamma}_k.$$

Number of absent species. The Horwitz-Thomson is

$$\widehat{C}_K = c/[1 - \widehat{g}_K(0)].$$

BMA can also be applied:

$$\widetilde{C} = \sum^{K_{\max}} w_K \widehat{C}_K.$$

Importance sampling

Approximate posterior.

- Variational Bayes only provides an approximate posterior $Q_\theta(\theta)$.
- which is known to often under-estimate the posterior variances.

Importance sampling

Approximate posterior.

- Variational Bayes only provides an approximate posterior $Q_\theta(\theta)$.
- which is known to often under-estimate the posterior variances.

Importance sampling (IS). For any distribution Q, taking $\{\theta^b\}$ iid $\sim Q$,

$$\begin{aligned}\int_{\mathcal{I}} P(X|\theta)P(\theta)\mathrm{d}\theta &= \int_{\mathcal{I}} P(X|\theta)\frac{P(\theta)}{Q(\theta)}Q(\theta)\mathrm{d}\theta \\ &\simeq \frac{1}{B}\sum_{\theta^b\in\mathcal{I}}\frac{P(\theta^b)}{Q(\theta^b)}P(X|\theta^b) \quad =: \quad \widehat{P}(\theta\in\mathcal{I}|X).\end{aligned}$$

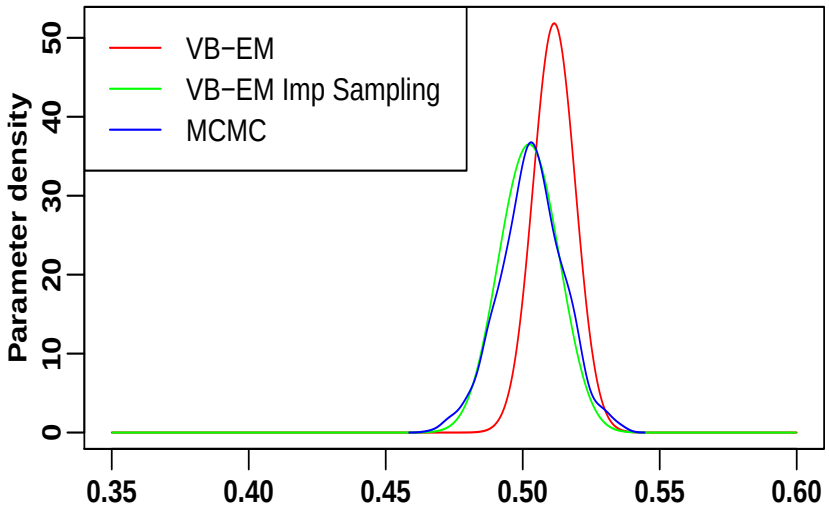
The variance gets smaller when Q gets closer to $P(\theta|X)$.

$\rightarrow$ The variational approximation $Q^*(\theta)$ can be used as a proxy.

Approximate posterior distribution

A Gibbs sampler is used as a gold standard for $\widehat{P}(\cdot|X)$.

Simulated data: $\widehat{g}(0)$

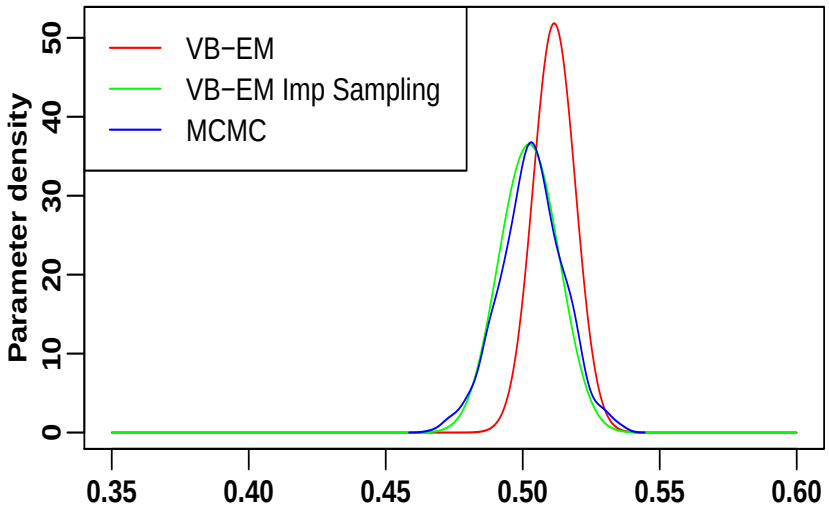


Li-Thiao-Té et al. *(2012)*

Approximate posterior distribution

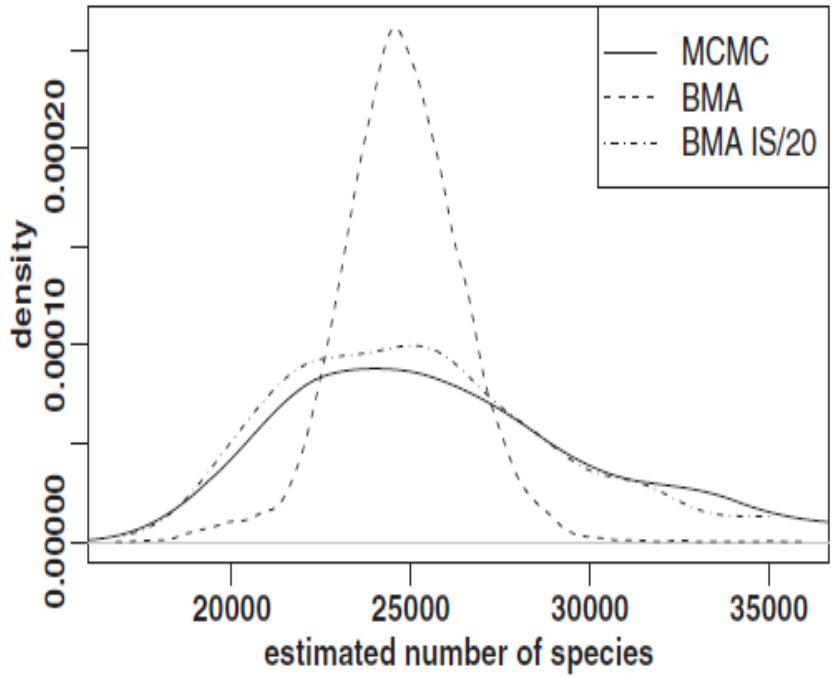
A Gibbs sampler is used as a gold standard for $\widehat{P}(\cdot|X)$.

Simulated data: $\widehat{g}(0)$



Li-Thiao-Té et al. *(2012)*

Human gut: $\widehat{C}_0 = 25,700$



$\text{CI}_{95\%} = [19,421; 36,355]$.

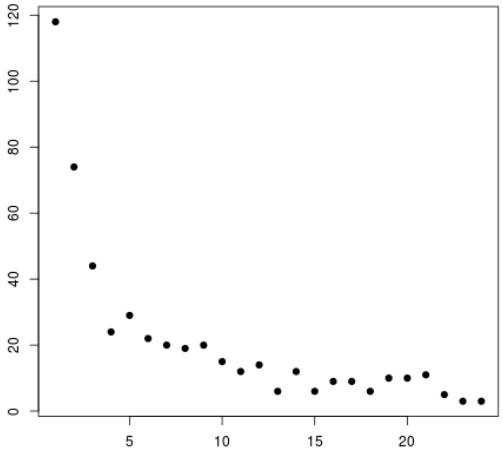
A 'true' non-parametric estimate

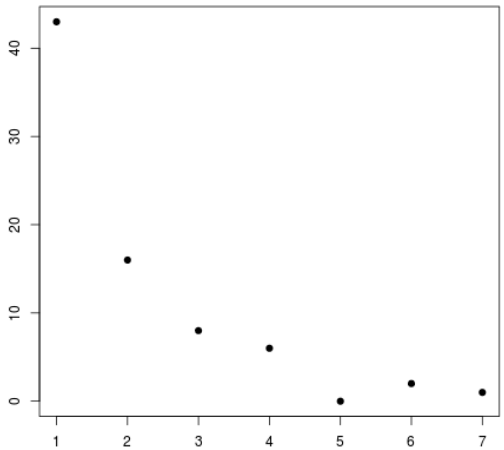
Joint work with

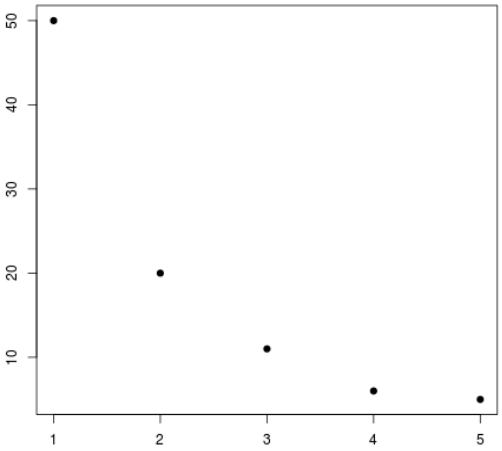
- C. Durot,
- F. Koladjo,
- S. Huet

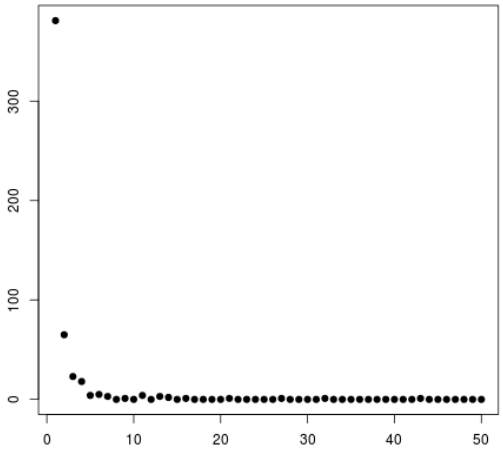
Convexity assumption

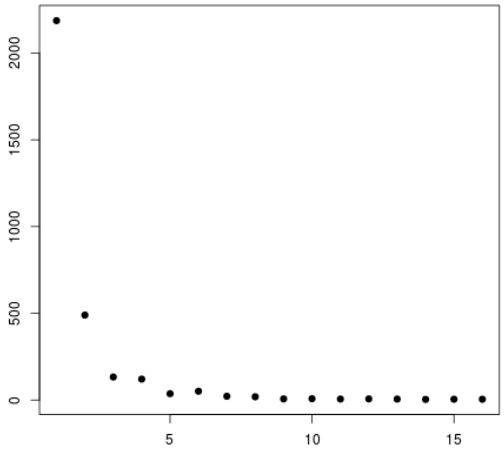
Most real-life SAD seem to be convex.









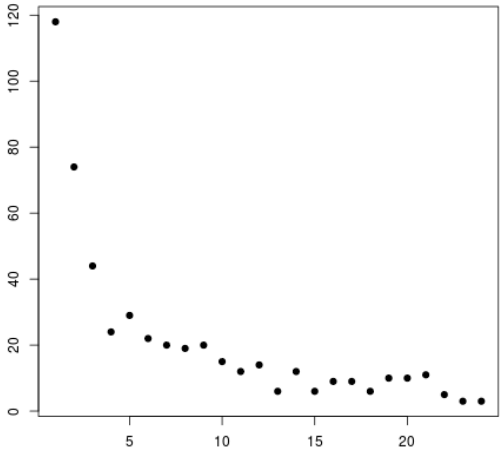


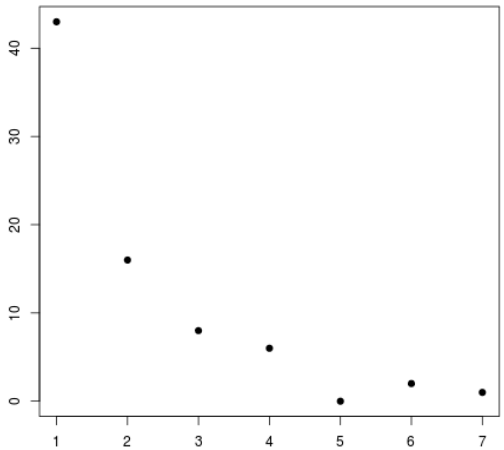
Convexity assumption

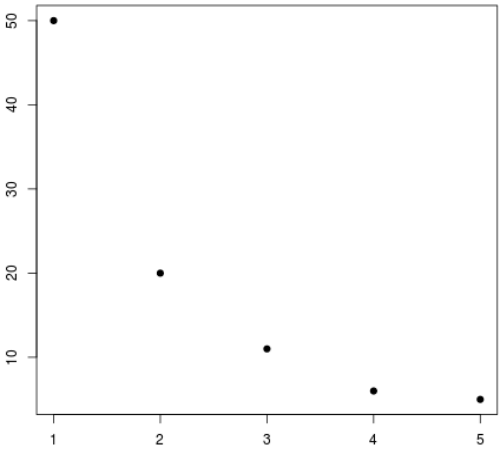
Most real-life SAD seem to be convex.

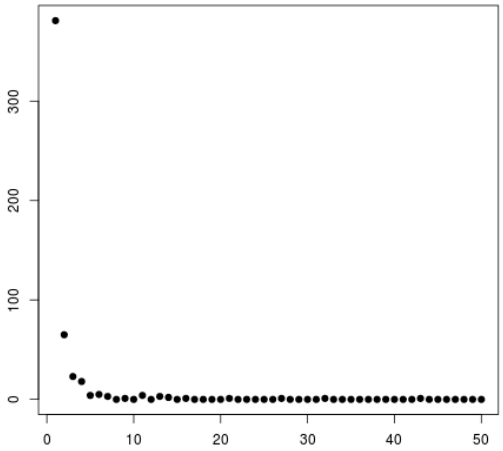
$\rightarrow$ Assumption:

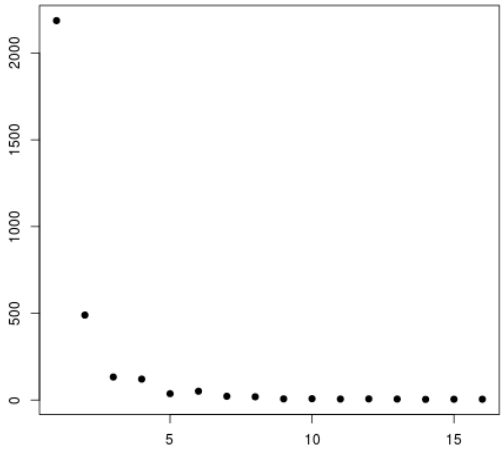
$g(\cdot)$ is convex.











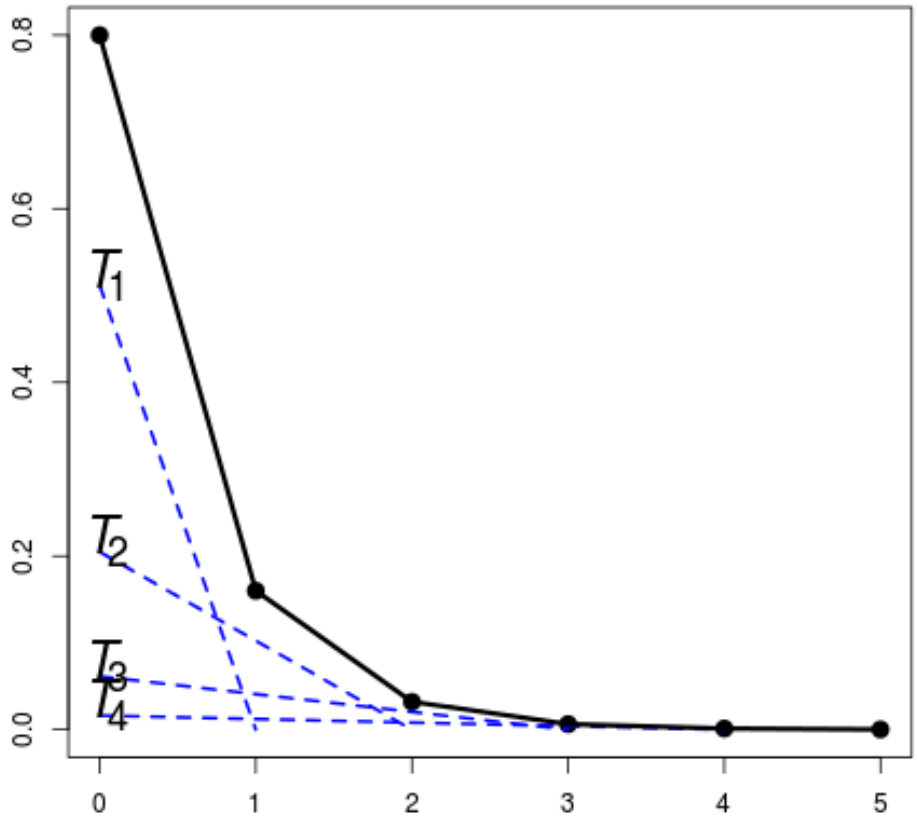
Decomposition of convex distributions

Any convex distribution g can be decomposed as a mixture

$$g(x) = \sum_j \pi_j T_j(x)$$

where the T_j are triangular distributions[1]

$$T_j(x) = \frac{2(j-x)}{j(j+1)}.$$



[1] this also holds for continuous convex distributions.

A definition of convex SAD

Mixture interpretation. Species are spread into groups

$$\begin{aligned} (Z_i) \text{ iid} &\sim \mathcal{M}(1;\pi) \\ (X_i) \text{ indep}|(Z_i): \quad X_i|Z_i = j &\sim T_j \end{aligned}$$

A definition of convex SAD

Mixture interpretation. Species are spread into groups

$$\begin{aligned} (Z_i) \text{ iid} &\sim \mathcal{M}(1;\pi) \\ (X_i) \text{ indep}|(Z_i): \quad X_i|Z_i = j &\sim T_j \end{aligned}$$

Interpretation of group 1. T_1 is Dirac mass on 0

$\rightarrow$ Species from group 1 can only display $X_i = 0$

$\rightarrow$ Such species can be thought of as ... absent species.

A definition of convex SAD

Mixture interpretation. Species are spread into groups

$$\begin{aligned} (Z_i) \text{ iid} &\sim \mathcal{M}(1;\pi) \\ (X_i) \text{ indep}|(Z_i): \quad X_i|Z_i = j &\sim T_j \end{aligned}$$

Interpretation of group 1. T_1 is Dirac mass on 0

$\rightarrow$ Species from group 1 can only display $X_i = 0$

$\rightarrow$ Such species can be thought of as ... absent species.

Definition. (*Durot* et al. *(2012)*) g is a convex SAD if

(*i*) g is convex discrete distribution.

(*ii*) The proportion of T_1 is null: $\pi_1 = 0$.

Non-parametric (convex) estimate of g

Empirical truncated distribution.

$$\widetilde{g}_n^+(x) = n^{-1} \sum_i \mathbb{I}\{X_i = x\}, \qquad x > 0$$

Least-square truncated convex SAD estimate.

$$\widehat{g}_n^+ = \arg\min_{g \in \mathcal{C}} \|g - \widetilde{g}_n^+\|^2$$

where $\mathcal{C}$ denotes the set of truncated convex SAD.

Non-parametric (convex) estimate of g

Empirical truncated distribution.

$$\widetilde{g}_n^+(x) = n^{-1} \sum_i \mathbb{I}\{X_i = x\}, \qquad x > 0$$

Least-square truncated convex SAD estimate.

$$\widehat{g}_n^+ = \arg\min_{g \in \mathcal{C}} \|g - \widetilde{g}_n^+\|^2$$

where $\mathcal{C}$ denotes the set of truncated convex SAD.

Inference. $\widehat{g}_n^+$ can be obtained via an extension of the support reduction algorithm (*Groeneboom* et al. *(2001)*) to an unknown support for g^+.

Some properties of $\widehat{g}_n^+$

1. The support $\widehat{s}_n$ of $\widehat{g}^+$ is finite.

Some properties of $\widehat{g}_n^+$

1. The support $\widehat{s}_n$ of $\widehat{g}^+$ is finite.
2. If g^+ is convex, $\widehat{g}^+$ is consistent at rate $\sqrt{n}$:

$$\sqrt{n}\|\widehat{g}^+ - g^+\|_r = O_P(1), \qquad \text{for } r \geq 2.$$

Some properties of $\widehat{g}_n^+$

1. The support $\widehat{s}_n$ of $\widehat{g}^+$ is finite.
2. If g^+ is convex, $\widehat{g}^+$ is consistent at rate $\sqrt{n}$:

$$\sqrt{n}\|\widehat{g}^+ - g^+\|_r = O_P(1), \qquad \text{for } r \geq 2.$$

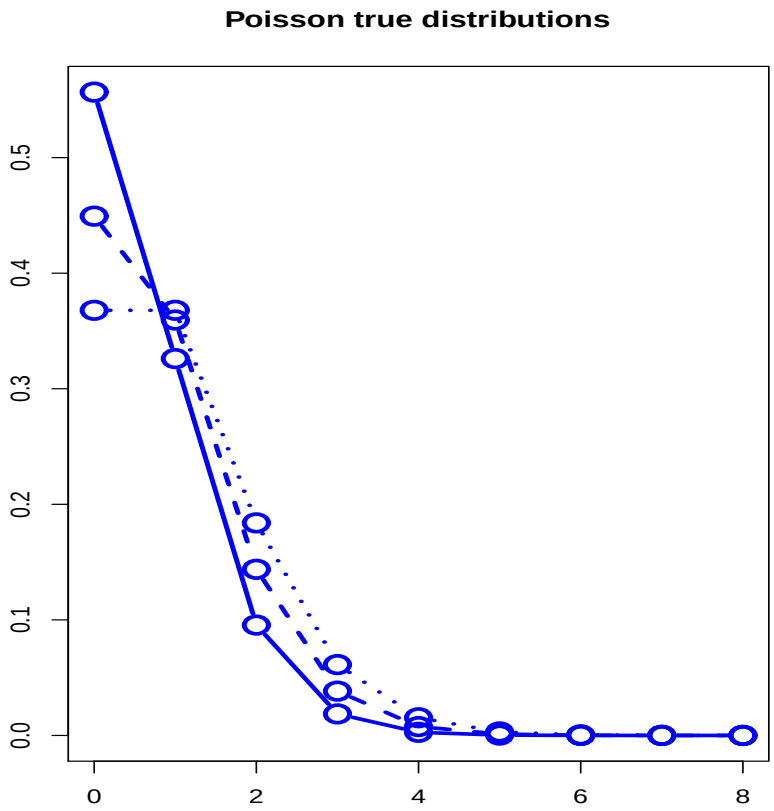
3. If g^+ is not convex, $\widehat{g}^+$ converge towards the projection of g^+ onto $\mathcal{C}$:

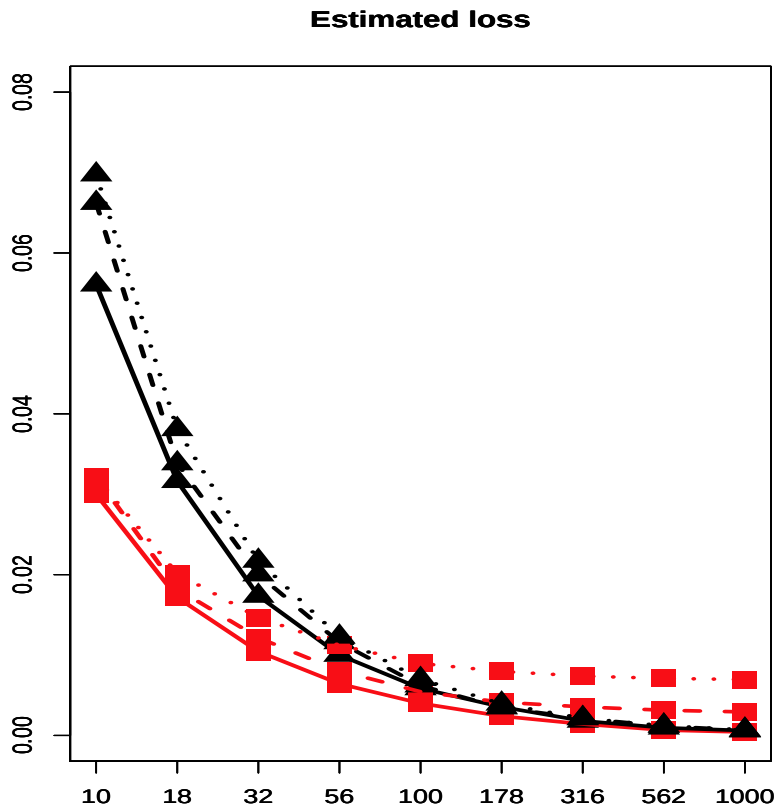
$$\sqrt{n}\|\widehat{g}^+ - \Pi_{\mathcal{C}} g^+\|_r = O_P(1), \qquad \text{for } r \geq 2.$$

Some properties of $\widehat{g}_n^+$

1. The support $\widehat{s}_n$ of $\widehat{g}^+$ is finite.
2. If g^+ is convex, $\widehat{g}^+$ is consistent at rate $\sqrt{n}$:
$$\sqrt{n}\|\widehat{g}^+ - g^+\|_r = O_P(1), \qquad \text{for } r \geq 2.$$
3. If g^+ is not convex, $\widehat{g}^+$ converge towards the projection of g^+ onto $\mathcal{C}$:
$$\sqrt{n}\|\widehat{g}^+ - \Pi_{\mathcal{C}} g^+\|_r = O_P(1), \qquad \text{for } r \geq 2.$$
4. Absolute moments are larger for $\widehat{g}^+$ than for $\widetilde{g}^+$.

Sensitivity to non-convexity





Estimated ℓ_2 loss for the empirical pdf $\widehat{g}$ and the convex estimate $\widehat{g}$ as a function of n for set of non-convex Poisson distribution ($\lambda \leq 2 - \sqrt{2}$).

Proportion of unobserved species

Estimate of $g(0)$. Using the definition of convex SAD (i.e. $\pi_1 = 0$):

$$\widehat{g}(0) = \frac{\widehat{\theta}}{1+\widehat{\theta}} \qquad \text{where} \quad \widehat{\theta} = 2\widehat{g}^{+}(1) - \widehat{g}^{+}(2).$$

[2] The asymptotic distribution of $\widetilde{\theta}$ is standard

Proportion of unobserved species

Estimate of $g(0)$. Using the definition of convex SAD (i.e. $\pi_1 = 0$):

$$\widehat{g}(0) = \frac{\widehat{\theta}}{1+\widehat{\theta}} \qquad \text{where} \quad \widehat{\theta} = 2\widehat{g}^{+}(1) - \widehat{g}^{+}(2).$$

Ongoing work.

- Asymptotic variance of $\widehat{\theta}$: no closed form.
- $\sqrt{n}(\widehat{\theta} - \theta)$ converges in distribution towards a non-standard distribution[2].
  $\rightarrow$ Bootstrap procedure.

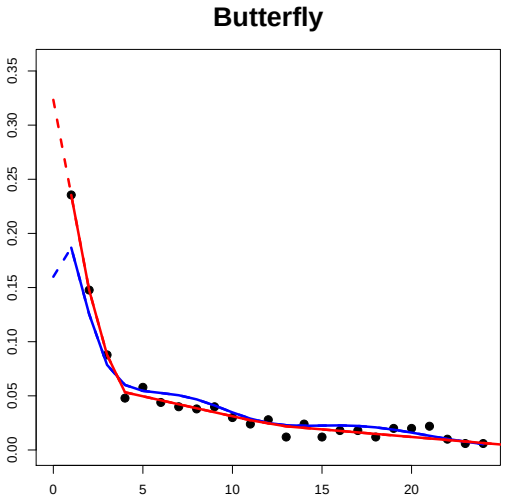
[2] The asymptotic distribution of $\widetilde{\theta}$ is standard

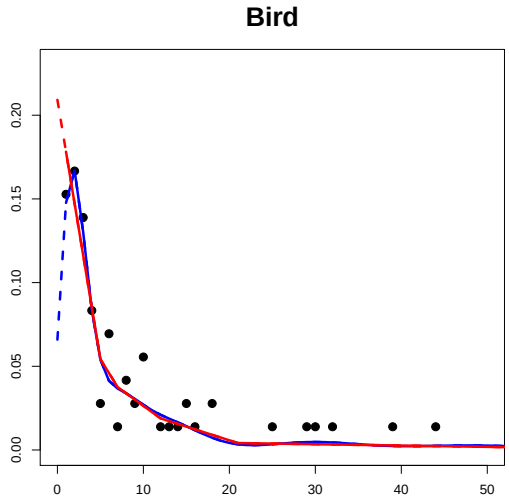
Some examples

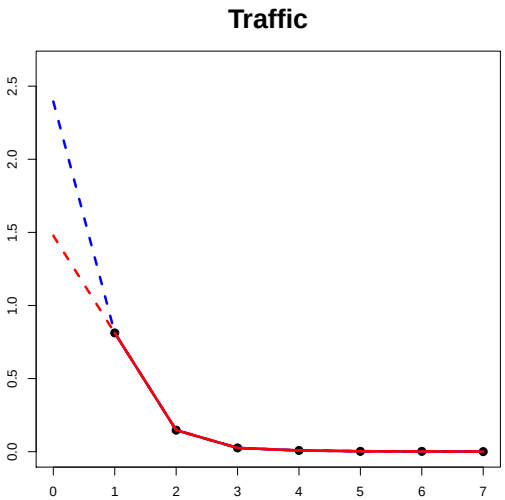
Poisson mixture

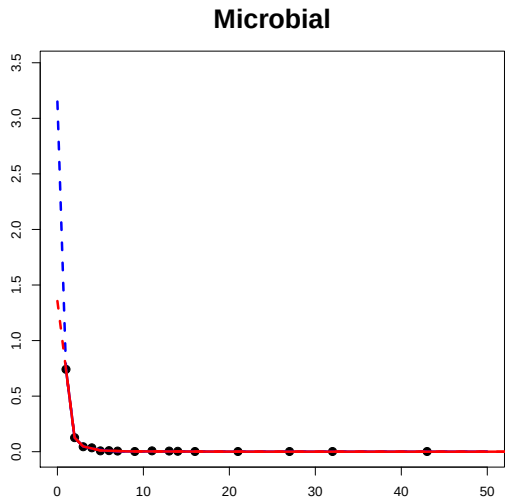
and

convex estimate









Sensitivity to truncation

As SAD are often long-tailed, *Chao and Shen (2004)* suggest truncation at some τ to infer $g(0)$.

τ	$\widehat{C}_{mCNP}$	$\widehat{C}_{u}$	$\widehat{C}_{UNP}$	$\widehat{C}_{WL}$	$\widehat{C}_{CONV}$
10	716	715	715	716	782
11	711	715	715	739	782
12	729	723	722	730	782
13	731	724	724	728	782
14	726	723	723	724	782
15	724	722	722	724	782
20	721	718	718	725	782
24	721	719	719	722	782

Estimates of N on Fisher's butterfly data.
$\widehat{N}_{mCNP}$, $\widehat{N}_{u}$, $\widehat{N}_{UNP}$ and $\widehat{N}_{WL}$ reported from *Wang and Lindsay (2005)*.

Conclusion & Future works

Species abundance is an old statistical problem revisited by metagenomics.

Conclusion & Future works

Species abundance is an old statistical problem revisited by metagenomics.

First estimate: Parametric with Bayesian inference

- Mixture models $\rightarrow$ flexible modeling of the SAD;
- Variational Bayes Model Averaging $\rightarrow$ approximate posterior distribution;
- Importance sampling $\rightarrow$ exact posterior, less computationally demanding than MCMC.

Conclusion & Future works

Species abundance is an old statistical problem revisited by metagenomics.

First estimate: Parametric with Bayesian inference

- Mixture models $\rightarrow$ flexible modeling of the SAD;
- Variational Bayes Model Averaging $\rightarrow$ approximate posterior distribution;
- Importance sampling $\rightarrow$ exact posterior, less computationally demanding than MCMC.

Second estimate: Non-parametric with frequentist inference

- Convexity $\rightarrow$ natural assumption for SAD;
- Triangular decomposition $\rightarrow$ definition of convex SAD;
- Asymptotic distribution of $\widehat{g}(0) \rightarrow$ under study.


Beal, J., M. and Ghahramani, Z. (2003). The variational Bayesian EM algorithm for incomplete data: with application to scoring graphical model structures. *Bayes. Statist.* **7** 543–52.


Bohning, D. and Kuhnert, R. (Dec, 2006). Equivalence of truncated count mixture distributions and mixtures of truncated count distributions. *Biometrics*. **62** 1207–1215.


Chao, A. and Shen, T.-J. (2004). Nonparametric prediction in species sampling. *Journal of Agricultural, Biological, and Environmental Statistics*. **9** 253–269.


Doroghazi, J. R. and Buckley, D. H. (2008). Evidence from GC-TRFLP that bacterial communities in soil are lognormally distributed. *PLoS ONE*. **3** e2910.


Durot, C., Koladjo, F., Huet, S. and Robin, S. (Feb., 2012). Estimation of a convex discrete distribution.


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. *J. Animal Ecol.* **12 (1)** 42–58.


Groeneboom, P., Jongbloed, G. and Wellner, J. A. (2001). Estimation of a convex function: characterizations and asymptotic theory. *Ann. Statist.* **29** 1653–1698.


Hooper, S. D., Dalevi, D., Pati, A., Mavromatis, K., Ivanova, N. N. and Kyrpides, N. C. (Feb, 2010). Estimating DNA coverage and abundance in metagenomes using a Gamma approximation. *Bioinformatics*. **26** 295–301.


Li-Thiao-Té, S., Jean-Jacques, D. and Stéphane, R. (2012). Bayesian model averaging for estimating the number of classes: applications to the total number of species in metagenomics. *Journal of Applied Statistics*. **39 (7)** 1489–1504.


McHardy, A. C. and Rigoutsos, I. (Oct, 2007). What's in the mix: phylogenetic classification of metagenome sequence samples. *Curr. Opin. Microbiol.* **10** 499–503.


Norris, J. L. I. and Pollock, K. H. (1998). Non-parametric MLE for poisson species abundance models allowing for heterogeneity between species. *Envir. Ecol. Statist.* **5** 391–402.

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.

Volant, S., Magniette, M.-L. M. and Robin, S. (2012). Variational bayes approach for model aggregation in unsupervised classification with markovian dependency. *Comput. Statis. & Data Analysis*. **56 (8)** 2375 – 2387.

Wang, J.-P. Z. and Lindsay, B. G. (2005). A penalized nonparametric maximum likelihood approach to species richness estimation. *Journal of the American Statistical Association*. **100 (471)** 942–959.