Bayesian nonparametric dependent model for the study of diversity for species data

Journées jeunes probabilistes et statisticiens

J. Arbel, K. Mengersen, J. Rousseau, C. King, B. Raymond
julyan.arbel@carloalberto.org

Moncalieri, Italy

April 10, 2014

Collegio Carlo Alberto

Project's bio

Authors

- Judith Rousseau (ENSAE, Université Paris-Dauphine, CREST, Paris)
- Kerrie L. Mengersen (Mathematical Sciences, Queensland University of Technology, Brisbane)
- Cath King (Australian Antarctic Division, Kingston, Tasmania 7050)
- Ben Raymond (Australian Antarctic Division, Kingston, Tasmania 7050)

Status

Chapter of my PhD thesis. Two submitted manuscripts Arbel et al. (2013b) and Arbel et al. (2013a).

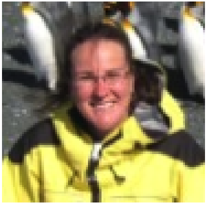
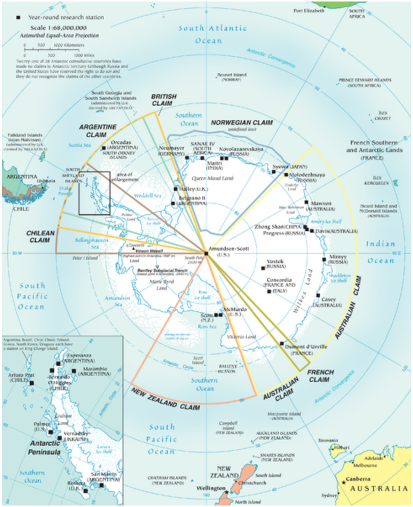
   

Table of Contents

Table of Contents

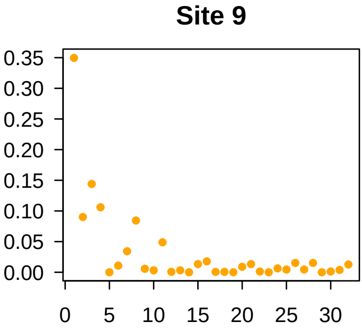
Context in ecology

- Series of measurements at different places around Casey Station, permanent base in Antarctica



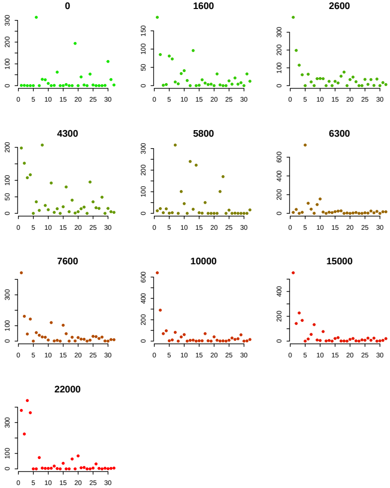
Context in ecology

- Series of measurements at different places around Casey Station, permanent base in Antarctica
- At each site: pollution level (total petroleum hydrocarbon (TPH) in mg/kg of soil), and abundance of microbes.



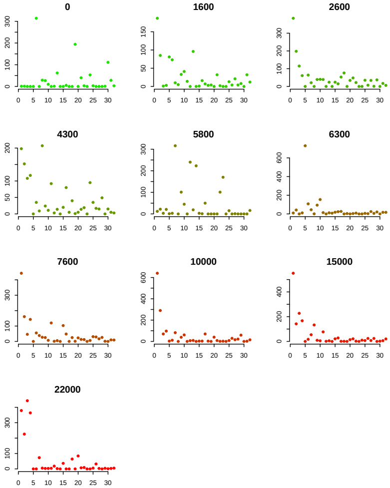
Context in ecology

- Series of measurements at different places around Casey Station, permanent base in Antarctica
- At each site: pollution level (total petroleum hydrocarbon (TPH) in mg/kg of soil), and abundance of microbes.



Context in ecology

- Series of measurements at different places around Casey Station, permanent base in Antarctica
- At each site: pollution level (total petroleum hydrocarbon (TPH) in mg/kg of soil), and abundance of microbes.
- Goals: *Assess the impact of a pollutant on the soil composition / biodiversity*, e.g. compute *effective concentration* values at level $x\%$, EC_x



Data collection

Soil samples were collected from a range of sites across a fuel contamination gradient at Australias Casey Station in East Antarctica (110° 32′ E, 66° 17′ S). The data comprise counts of a large number (of the order of 1 800) of microbial taxa, referred to as OTUs (operational taxonomic units; see Schloss et al., 2009), collected at 60 sites, across a range of hydrocarbon contamination (Siciliano et al., 2014). Genomic DNA extracted from samples was sequenced on a 454 Titanium FLX+ instrument (Roche, Brandford, CT, USA) at the Research and Testing facility (Lubbock, TX, USA) using the universal bacterial primers 28F and 519R (Dowd et al., 2008). Pyrosequencing data were processed using the mothur software package (Schloss et al., 2009). This involved removal of short reads (<150bp), excessive homoploymeric reads (>8bp repeats) and denoising with AmpliconNoise (min/max flows 360/720) (Quince et al., 2011). Preclustering at 1% was performed to negate the per base error rate of the 454 platforms. Seed sequences were then aligned to the SILVA 16S rRNA gene database alignment using a NAST alignment algorithm (Pruesse et al., 2007; Caporaso et al., 2010). Reads were then chimaera-checked (Edgar et al., 2011) and clustered into OTUs at 96% sequence similarity to achieve approximately species-level units as derived by Kim et al. (2011). Seed sequences from each OTU were then classified using a Naïve Bayesian classifier in mothur against the Greengenes 16S reference database (October 2012 version, see McDonald et al., 2012).

Diversity indices

Shannon index

$$H_{\text{Shan}}(\mathbf{p}) = -\sum p_j \log p_j$$

Simpson index

$$H_{\text{Simp}}(\mathbf{p}) = 1 - \sum p_j^2$$

Good index

$$H_{\text{Good},\alpha,\beta}(\mathbf{p}) = -\sum p_j^\alpha \log^\beta p_j$$

Diversity indices of microbial data

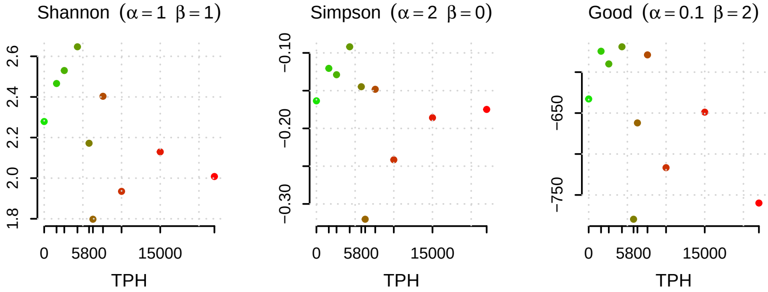


Table of Contents

Model

Model the distribution of microbes in the soil:

The nth observation at site i is species j with probability $p_j(X_i)$

Model

Model the distribution of microbes in the soil:

The nth observation at site i is species j with probability $p_j(X_i)$

Model

For every site i

$$Y_{n,i} \mid \mathbf{p}(X_i), X_i \overset{\text{ind}}{\sim} \sum_{j=1}^{\infty} p_j(X_i)\delta_j$$

Model

Model the distribution of microbes in the soil:

The nth observation at site i is species j with probability $p_j(X_i)$

Model

For every site i

$$Y_{n,i} \mid \mathbf{p}(X_i), X_i \overset{\text{ind}}{\sim} \sum_{j=1}^{\infty} p_j(X_i)\delta_j$$

Parameters: $\mathbf{p} = \big(\mathbf{p}(X_1), \ldots, \mathbf{p}(X_I)\big) = \big(p_j(X_i)\big)_{i,j}$

Model

Model the distribution of microbes in the soil:

The nth observation at site i is species j with probability $p_j(X_i)$

Model

For every site i

$$Y_{n,i} \,|\, \mathbf{p}(X_i), X_i \overset{\text{ind}}{\sim} \sum_{j=1}^{\infty} p_j(X_i)\delta_j$$

Parameters: $\mathbf{p} = \big(\mathbf{p}(X_1), \ldots, \mathbf{p}(X_I)\big) = \big(p_j(X_i)\big)_{i,j}$

- Holmes et al. (2012): Dirichlet Multinomial Mixtures: Generative Models for Microbial Metagenomics (PloS one)

Model

Model the distribution of microbes in the soil:

The nth observation at site i is species j with probability $p_j(X_i)$

Model

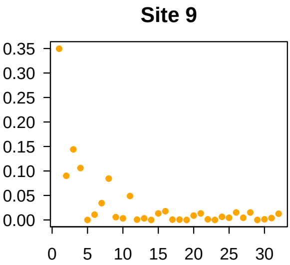
For every site i

$$Y_{n,i} \mid \mathbf{p}(X_i), X_i \overset{\text{ind}}{\sim} \sum_{j=1}^{\infty} p_j(X_i)\delta_j$$

Parameters: $\mathbf{p} = \big(\mathbf{p}(X_1), \ldots, \mathbf{p}(X_I)\big) = \big(p_j(X_i)\big)_{i,j}$

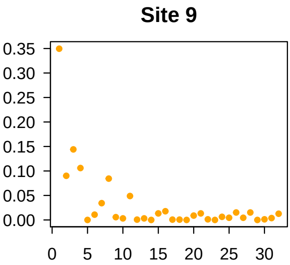
- Holmes et al. (2012): Dirichlet Multinomial Mixtures: Generative Models for Microbial Metagenomics (PloS one)
- Lijoi et al. (2007): Bayesian nonparametric estimation of the probability of discovering new species (Biometrika)

Randomizing the weights $p_j(X_i)$



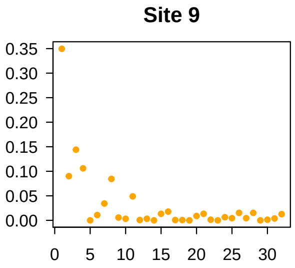
Randomizing the weights $p_j(X_i)$

- Use the distribution of the weights in a Dirichlet process, obtained by a stick-breaking construction



Randomizing the weights $p_j(X_i)$

- Use the distribution of the weights in a Dirichlet process, obtained by a stick-breaking construction



Stick-breaking construction

$$p_1 = V_1, \quad p_j = V_j \prod_{l<j} (1 - V_l),$$

with $V_j \overset{\text{iid}}{\sim} \text{Beta}(1, M)$.

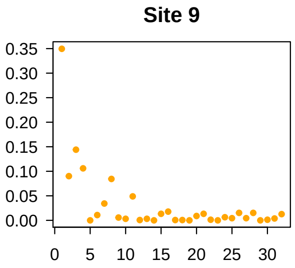
Randomizing the weights $p_j(X_i)$

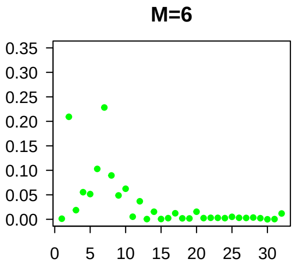
- Use the distribution of the weights in a Dirichlet process, obtained by a stick-breaking construction

Stick-breaking construction

$$p_1 = V_1, \quad p_j = V_j \prod_{l<j} (1 - V_l),$$

with $V_j \overset{\text{iid}}{\sim} \text{Beta}(1, M)$.





Randomizing the weights $p_j(X_i)$

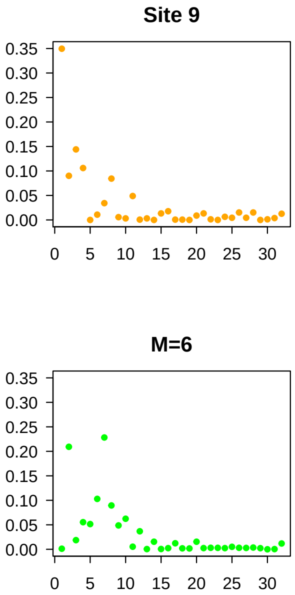
- Use the distribution of the weights in a Dirichlet process, obtained by a stick-breaking construction

Stick-breaking construction

$$p_1 = V_1, \quad p_j = V_j \prod_{l<j} (1 - V_l),$$

with $V_j \overset{\text{iid}}{\sim} \text{Beta}(1, M)$.

It is denoted $\mathbf{p} \sim \text{GEM}(M)$.



On convergence rates

$\Big($

On convergence rates

Need control in probability tail sums

$$\sum_{j=N+1}^{\infty} p_j$$

On convergence rates

$\Big($Need control in probability tail sums

$$\sum_{j=N+1}^{\infty} p_j$$

or partial product or partial sum

$$\prod_{j=1}^{N}(1 - V_j)$$

$$\sum_{j=1}^{N} -\log(1 - V_j)$$

On convergence rates

(Need control in probability tail sums

$$\sum_{j=N+1}^{\infty} p_j$$

or partial product or partial sum

$$\prod_{j=1}^{N}(1 - V_j)$$

$$\sum_{j=1}^{N} -\log(1 - V_j)$$

which is shown to be $\text{Ga}(N, M)$ when $V_j \overset{\text{iid}}{\sim} \text{Beta}(1, M)$

On convergence rates

$\Big($Need control in probability tail sums

$$\sum_{j=N+1}^{\infty} p_j$$

or partial product or partial sum

$$\prod_{j=1}^{N}(1-V_j)$$

$$\sum_{j=1}^{N} -\log(1-V_j)$$

which is shown to be $\mathrm{Ga}(N, M)$ when $V_j \overset{\text{iid}}{\sim} \mathrm{Beta}(1, M)$

Trickier when

$V_j \overset{\text{ind}}{\sim} \mathrm{Beta}(a, b+cj),\ a \neq 1$

On convergence rates

(Need control in probability tail sums

$$\sum_{j=N+1}^{\infty} p_j$$

or partial product or partial sum

$$\prod_{j=1}^{N}(1 - V_j)$$

$$\sum_{j=1}^{N} -\log(1 - V_j)$$

which is shown to be $\mathrm{Ga}(N, M)$ when $V_j \overset{\text{iid}}{\sim} \mathrm{Beta}(1, M)$

Trickier when

$$V_j \overset{\text{ind}}{\sim} \mathrm{Beta}(a, b + cj), \ a \neq 1$$

Need a central limit theorem for

$$\sum_{j=1}^{N} -\log(1 - V_j)$$

On convergence rates

(Need control in probability tail sums

$$\sum_{j=N+1}^{\infty} p_j$$

or partial product or partial sum

$$\prod_{j=1}^{N}(1 - V_j)$$

$$\sum_{j=1}^{N} -\log(1 - V_j)$$

which is shown to be $\text{Ga}(N, M)$ when $V_j \overset{\text{iid}}{\sim} \text{Beta}(1, M)$

Trickier when

$$V_j \overset{\text{ind}}{\sim} \text{Beta}(a, b + cj),\ a \neq 1$$

Need a central limit theorem for

$$\sum_{j=1}^{N} -\log(1 - V_j)$$

Limit distribution seems to be Gumbel...

On convergence rates

(Need control in probability tail sums

$$\sum_{j=N+1}^{\infty} p_j$$

or partial product or partial sum

$$\prod_{j=1}^{N} (1 - V_j)$$

$$\sum_{j=1}^{N} - \log(1 - V_j)$$

which is shown to be $\mathrm{Ga}(N, M)$ when $V_j \overset{\text{iid}}{\sim} \mathrm{Beta}(1, M)$

Trickier when

$$V_j \overset{\text{ind}}{\sim} \mathrm{Beta}(a, b + cj), \; a \neq 1$$

Need a central limit theorem for

$$\sum_{j=1}^{N} - \log(1 - V_j)$$

Limit distribution seems to be Gumbel...)

Construction of the prior

- With the strick-breaking relation, a Dep − GEM prior is obtained from a Beta process.

Construction of the prior

- With the strick-breaking relation, a $\mathrm{Dep} - \mathrm{GEM}$ prior is obtained from a Beta process.
- Such a dependent Beta process is obtained by a transformed Gaussian process (Rasmussen and Williams, 2006)
  $\rightarrow$ Denote by $Z \sim \mathrm{N}(0, \sigma^2)$ a Gaussian random variable, by Φ_{σ_Z} its CDF and by F_M a $\mathrm{Beta}(1, M)$ CDF. Then:

$$\Phi_{\sigma_Z}(Z) \sim \mathrm{Unif}(0,1) \text{ and } V = F_M^{-1} \circ \Phi_{\sigma_Z}(Z) \sim \mathrm{Beta}(1, M),$$

Construction of the prior

- With the strick-breaking relation, a $\text{Dep} - \text{GEM}$ prior is obtained from a Beta process.
- Such a dependent Beta process is obtained by a transformed Gaussian process (Rasmussen and Williams, 2006)
$\rightarrow$ Denote by $Z \sim \mathrm{N}(0, \sigma^2)$ a Gaussian random variable, by Φ_{σ_Z} its CDF and by F_M a $\text{Beta}(1, M)$ CDF. Then:

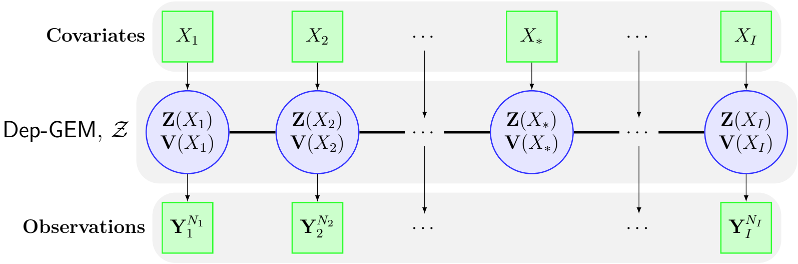
$$\Phi_{\sigma_Z}(Z) \sim \text{Unif}(0,1) \text{ and } V = F_M^{-1} \circ \Phi_{\sigma_Z}(Z) \sim \text{Beta}(1, M),$$

- Dependence specified by covariance function

$$K(X_i, X_j) = \text{Cov}\big(\mathcal{Z}(X_i), \mathcal{Z}(X_j)\big).$$

Covariance function	$\tilde{K}_\lambda(X_1, X_2)$
Squared Exponential (SE)	$\exp\big(-(X_1 - X_2)^2/(2\lambda^2)\big)$
Ornstein-Uhlenbeck (OU)	$\exp\big(-\lvert X_1 - X_2\rvert/\lambda\big)$
Rational Quadratic (RQ)	$\big(1 + (X_1 - X_2)^2/(2\lambda^2)\big)^{-1}$

Graphical model representation for the Dep − GEM model



Algorithm: Metropolis within Gibbs

Algorithm 1 Dep – GEM algorithm (Gibbs)

1: Update $\mathbf{Z}$ given $(\sigma_{\mathbf{Z}}, \lambda, M)$
2: Update $\sigma_{\mathbf{Z}}$ given $(\mathbf{Z}, \lambda, M)$
3: Update λ given $(\mathbf{Z}, \sigma_{\mathbf{Z}}, M)$
4: Update M given $(\mathbf{Z}, \sigma_{\mathbf{Z}}, \lambda)$

Algorithm: Metropolis within Gibbs

Algorithm 3 Dep – GEM algorithm (Gibbs)

1: Update $\mathbf{Z}$ given $(\sigma_{\mathbf{Z}}, \lambda, M)$
2: Update $\sigma_{\mathbf{Z}}$ given $(\mathbf{Z}, \lambda, M)$
3: Update λ given $(\mathbf{Z}, \sigma_{\mathbf{Z}}, M)$
4: Update M given $(\mathbf{Z}, \sigma_{\mathbf{Z}}, \lambda)$

Algorithm 4 MH algorithm

1: Given $\boldsymbol{\theta}$, propose $\boldsymbol{\theta}' \sim Q_{\boldsymbol{\theta}}(\cdot \mid \boldsymbol{\theta})$
2: Compute $\rho_{\boldsymbol{\theta}} = \frac{P_{\boldsymbol{\theta}}(\boldsymbol{\theta}')}{P_{\boldsymbol{\theta}}(\boldsymbol{\theta})} \frac{Q_{\boldsymbol{\theta}}(\boldsymbol{\theta}|\boldsymbol{\theta}')}{Q_{\boldsymbol{\theta}}(\boldsymbol{\theta}' \mid \boldsymbol{\theta})}$
3: Accept $\boldsymbol{\theta}'$ *wp* $\min(\rho_{\boldsymbol{\theta}}, 1)$, otherwise keep $\boldsymbol{\theta}$

Predictive distribution

- Predictive distribution of $\mathbf{Z}_*$ obtained by integrating out $\mathbf{Z}$ in the conditional distribution according to the posterior distribution $\pi(\mathbf{Z}|Y,X)$:

$$\pi(\mathbf{Z}_* \,|\, X_*, Y) = \int \pi(\mathbf{Z}_* \,|\, X_*, X, \mathbf{Z})\pi(\mathbf{Z}|Y,X)\mathrm{d}\mathbf{Z}.$$

- No particular computational burden:

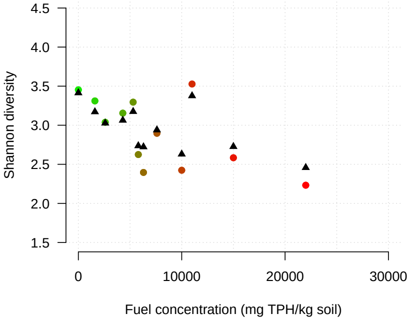
Algorithm 5 Predictive distribution simulation

1: Sample $\mathbf{Z}$ from the posterior distribution $\pi(\mathbf{Z}|Y,X)$

2: Given $\mathbf{Z}$, sample $\mathbf{Z}_*$ from the conditional distribution $\pi(\mathbf{Z}_* \,|\, X_*, X, \mathbf{Z})$

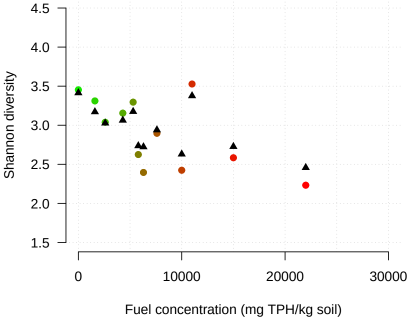
Table of Contents

Comparison of the Dep – GEM and indep. GEM models

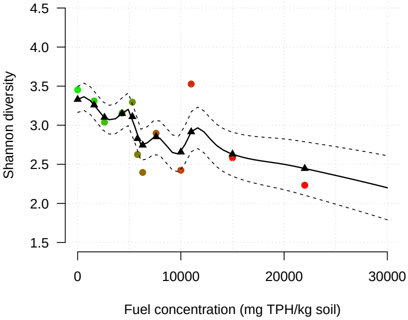


(a) GEM

Comparison of the Dep − GEM and indep. GEM models

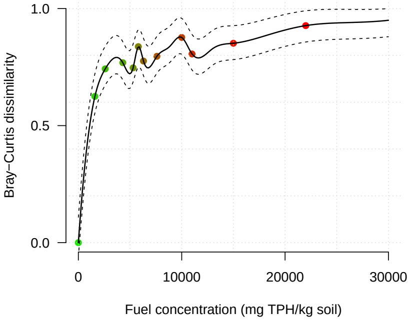


(c) GEM



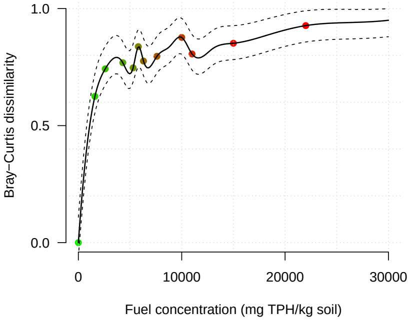
(d) Dep − GEM

Effective concentration estimation EC_x

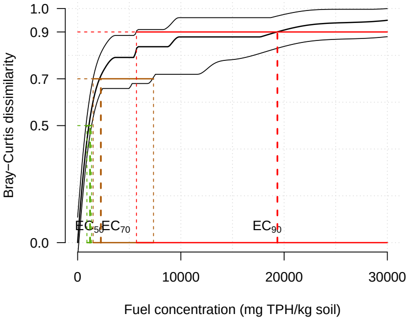


(e) Bray-Curtis dissimilarity

Effective concentration estimation EC_x



(g) Bray-Curtis dissimilarity



(h) Illustration of EC_x estimation

Future work

- Extension to multiple covariates

  $\longrightarrow$ by using Gaussian random fields instead of Gaussian processes

  $\longrightarrow$ model choice

Future work

- Extension to multiple covariates

  $\longrightarrow$ by using Gaussian random fields instead of Gaussian processes

  $\longrightarrow$ model choice

- Use of finer stick-breaking distributions

  $\longrightarrow$ e.g. Beta(a, b) or Gibbs-type priors instead of Beta$(1, M)$

Thank you for your attention!

Arbel, J., Mengersen, K., Raymond, B., and King, C. (2013a). Ecotoxicological data study of diversity using a dependent Bayesian nonparametric model. *Manuscript under preparation*.

Arbel, J., Mengersen, K., and Rousseau, J. (2013b). Bayesian nonparametric dependent models for the study of diversity in species data. *Manuscript under preparation*.

Holmes, I., Harris, K., and Quince, C. (2012). Dirichlet Multinomial Mixtures: Generative Models for Microbial Metagenomics. *PloS one*, 7(2):e30126.

Lijoi, A., Mena, R. H., and Prünster, I. (2007). Bayesian nonparametric estimation of the probability of discovering new species. *Biometrika*, 94(4):769–786.

Lijoi, A., Prünster, I., and Walker, S. G. (2008). Bayesian nonparametric estimators derived from conditional Gibbs structures. *The Annals of Applied Probability*, 18(4):1519–1547.

Rasmussen, C. E. and Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning*. MIT Press.