On diversity under a Bayesian nonparametric dependent model

Julyan Arbel, Kerrie L. Mengersen, Judith Rousseau

On diversity under a Bayesian nonparametric dependent model

Sulla diversità per un modello bayesiano nonparametrico dipendente

Julyan Arbel, Kerrie Mengersen and Judith Rousseau

Abstract We present a dependent Bayesian nonparametric model for the probabilistic modelling of species-by-site data, *i.e.* population data where observations at different sites are classified into distinct species. We use a dependent version of the Griffiths-Engen-McCloskey distribution, the distribution of the weights of the Dirichlet process, in the same lines as the Dependent Dirichlet process is defined. The prior is thus defined via the stick-breaking construction. Some distributional properties of this model are presented.

Abstract Presentiamo un modello parametrico bayesiano dipendente per la modellazione probabilistica dei dati specie per siti, cioé i dati sulla popolazione in cui le osservazioni in diversi siti sono classificati in specie distinte. Usiamo una versione dipendente della distribuzione Griffiths-Engen-McCloskey, la distribuzione dei pesi del processo di Dirichlet, in analogia con la definizione del processo di Dirichlet dipendente. La distribuzione a priori é definita tramite la costruzione stick-breaking. Alcune proprietà distributive di questo modello sono presentate.

Key words: Bayesian nonparametrics, Covariate-dependent model, Griffiths-Engen-McCloskey distribution, Stick-breaking representation.

Julyan Arbel
Collegio Carlo Alberto, Moncalieri, Italy e-mail: julyan.arbel@carloalberto.org

Kerrie Mengersen
Queensland University of Technology, Brisbane, Australia e-mail: k.mengersen@qut.edu.au

Judith Rousseau
ENSAE, Université Paris-Dauphine, France e-mail: rousseau@ceremade.dauphine.fr

1 Introduction

In this paper we announce results which will be extensively presented and proved in Arbel et al. (2013b) about distributional properties of a Bayesian nonparametric dependent model. The construction of (covariate) dependent random probability measures for Bayesian inference has been a very active line of research in the last 15 years. This was settled by the pioneering works by MacEachern (1999, 2000) who introduced a general class of dependent Dirichlet processes. The literature on dependent processes was developed in numerous models, such as nonparametric regression, time series data, meta-analysis, to cite but a few, and applied to a wealth of fields such as, e.g., epidemiology, bioassay problems, genomics, finance. Most contributions to this line of research rely on random probability measures defined by means of a stick-breaking procedure, a popular method set forth in its generality by Ishwaran and James (2001). This is also the approach retained here, where the dependence among different stick-breaking priors is obtained by transforming Gaussian processes. An advantage of such a construction is in the variety of dependence structures that can be handled across the covariate thanks to the covariance function or the Gaussian processes. Some distributional properties of the introduced dependent processes are studied. An application of the model in ecotoxicology to a microbial dataset species collected in Antarctica is conducted in Arbel et al. (2013a).

The modelling of species data under a Bayesian nonparametric framework has already been introduced by Lijoi et al. (2007, 2008). We follow the same approach, and add a dependence structure to covariates. In studies oriented towards species sampling and abundance measures, diversity is also often a notion of interest. The question of measuring diversity arises in many fields, *e.g.* ecology as in the present study, but also biology, engineering or probability theory. There are numerous ways to study the diversity of a population divided into groups or species. We retain here the Simpson index H_{Simp} defined by $H_{\mathsf{Simp}}(\mathbf{p}) = 1 - \sum_j p_j^2$ for a discrete probability distribution $\mathbf{p} = (p_1, p_2, \ldots)$ where p_j is the probability of observing species j. For a discussion of different diversity indices, see for example Cerquetti (2012).

The paper is organized as follows: the typical framework of the sampling model is described in Section 2, along with the dependent prior. A review of some of the distributional properties of this model is then presented in Section 3.

2 Data and model

We describe here the notation and sampling process of covariate dependent species-by-site count data. To each site $i = 1, \ldots, I$ corresponds a covariate value $X_i \in \mathscr{X}$, where the space $\mathscr{X}$ is a subset of $\mathbb{R}^d$. We focus here on univariate factors, *i.e.* $d = 1$. Individual observations at site i are species, indexed by natural numbers $j \in \mathbb{N}^*$. No hypothesis is made on the unknown total number of species in the population of interest, which might be infinite. We observe $(\mathbf{X}, \mathbf{Y}) = (X_i, \mathbf{Y}_i^{N_i})_{i=1,\ldots,I}$ where $\mathbf{Y}_i^{N_i} = (Y_{n,i})_{n=1,\ldots,N_i}$ are observations at site i with total abundance (number of observations)

N_i and factor X_i. We model the probabilities $\mathbf{p} = (\mathbf{p}(X_i))_{i=1\ldots I} = (p_j(X_i)_{j\in\mathbb{N}^*})_{i=1\ldots I}$ by the following. For $i = 1\ldots I$ and $n = 1\ldots N_i$:

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

We now proceed to a brief description of the dependent prior. For a complete account comprising details about posterior sampling, the reader is referred to Arbel et al. (2013b). We use the same construction as MacEachern (2000) to extend the GEM distribution in order to define the following dependent version, abbreviated Dep-GEM. Starting with independent stochastic processes $(V_j(X), X \in \mathscr{X})$, $j \in \mathbb{N}^*$, satisfying $V_j(X) \overset{\text{iid}}{\sim} \mathsf{Beta}(1, M)$ marginally in X, the Dep-GEM distribution on the weights is defined by:

$$p_j(X) = V_j(X)\prod_{l<j}(1 - V_l(X)), \tag{2}$$

where $(V_j(X), X \in \mathscr{X})$, $j \in \mathbb{N}^*$ are independent. Note that (2) can be easily extended to the two-parameter Poisson-Dirichlet process, denoted by $\mathsf{PD}(\alpha, M)$. It follows the same stick-breaking construction as in Equation (2), with V_j's defined with independent $\mathsf{Beta}(1-\alpha, M+j\alpha)$ distributions (*i.e.* the V_j's are not identically distributed), where $\alpha > 0$ and $M > -\alpha$. The advantage of the $\mathsf{PD}(\alpha, M)$ process compared to the $\mathsf{DP}(M)$ is a more flexible predictive structure.

3 Distributional properties

The purpose of this section is to present some elementary distributional properties of the Dep-GEM prior. It has continuous sample-paths, as stated in the following proposition.

Proposition 1 *Let* $\mathbf{p} \sim$Dep-GEM (M). *Then* $\mathbf{p}$ *is stationary and marginally,* $\mathbf{p}(X) \sim \mathsf{GEM}(M)$. *Also,* $\mathbf{p}$ *has continuous paths (*i.e. $X \to (p_1(X), p_2(X), \ldots)$ *is continuous for the sup norm), and its marginal moments are*

$$\begin{aligned}
\mathsf{E}(p_j(X)) &= \frac{M^{j-1}}{(M+1)^j}, \quad \mathsf{E}(p_j^n(X)) = \frac{n!}{M_{(n)}}\left(\frac{M}{M+n}\right)^j, \\
\mathsf{Var}(p_j(X)) &= \frac{2M^{j-1}}{(M+1)(M+2)^j} - \frac{M^{2(j-1)}}{(M+1)^{2j}}, \\
\mathsf{Cov}(p_j(X), p_k(X)) &= \frac{M^{(j\vee k)-1}}{(M+1)^{|j-k|+1}(M+2)^{j\wedge k}} - \frac{M^{j+k-2}}{(M+1)^{j+k}}, k \neq j,
\end{aligned}$$

for any $j,k \geq 1$, $n \geq 0$, *and where* $M_{(n)} = M(M+1)\ldots(M+n-1)$ *denotes the ascending factorial.*

We now review some results on size-biased permutations which are useful for deriving distributional properties of the process. Let $\mathbf{p} = (p_1, p_2, \ldots)$ be a probability. A size-biased permutation of $\mathbf{p}$, is a sequence $\tilde{\mathbf{p}} = (\tilde{p}_1, \tilde{p}_2, \ldots)$ obtained by reordering $\mathbf{p}$ by a permutation σ with particular probabilities. Namely, the first index appears with a probability equal to its weight, $\mathsf{P}(\sigma_1 = j) = p_j$; the subsequent indices appear with a probability proportional to their weight in the remaining indices, *i.e.* for k distinct integers $j_1, \ldots, j_k$,

$$\mathsf{P}(\sigma_k = j_k | \sigma_1 = j_1, \ldots, \sigma_{k-1} = j_{k-1}) = \frac{p_{j_k}}{1 - p_{j_1} - \cdots - p_{j_{k-1}}}. \tag{3}$$

The following lemma extends Pitman's result (for example Equation (2.23) of Pitman, 2006):

$$\mathsf{E}\Big(\sum f(p_j)\Big) = \mathsf{E}\Big(\sum f(\tilde{p}_j)\Big) = \mathsf{E}\left(\frac{f(\tilde{p}_1)}{\tilde{p}_1}\right), \tag{4}$$

for any measurable function f.

Lemma 2 *Let $\tilde{\mathbf{p}}$ is a size-biased permutation of $\mathbf{p}$. For any measurable function f and any integer $k \geq 1$, we have*

$$\mathsf{E}\left(\sum_{(*)} f(p_{i_1}, \ldots, p_{i_k})\right) = \mathsf{E}\left(f(\tilde{p}_1, \ldots, \tilde{p}_k) \prod_{i=1}^{k} (1 - \tilde{p}_1 - \cdots - \tilde{p}_{i-1})/\tilde{p}_i\right), \tag{5}$$

where the sum $(*)$ runs over all distinct $i_1, \ldots, i_k$, and with the convention that the product in the right-hand side of Equation (5) *equals $1/\tilde{p}_1$ when $i = 1$.*

When it comes to averaging sums of transforms of k weights $p_{i_1}, \ldots, p_{i_k}$ over all distinct $i_1, \ldots, i_k$, the lemma shows that all required information is encoded by the first k picks $\tilde{p}_1, \ldots, \tilde{p}_k$. The case $k = 2$ was proved by Archer et al. (2013).

One can gain further insight into the Dep-GEM process by studying the exchangeable partition probability function (EPPF) for the random variables $\mathbf{Y}_1^n = (Y_{1,1}, \ldots, Y_{n,1})$ and $\mathbf{Y}_2^m = (Y_{1,2}, \ldots, Y_{m,2})$ observed at covariates X_1 and X_2. See for instance Pitman (1995, 2006) for a summary of the importance of partition probability functions. The observations partition $[n] = \{1, 2, \ldots, n\}$ and $[m] = \{1, 2, \ldots, m\}$ into $k + k_1 + k_2$ clusters of distinct values where

- k clusters are commonly observed, with respective frequencies $\mathbf{n} = (n_1, \ldots, n_k)$ and $\mathbf{m} = (m_1, \ldots, m_k)$,
- k_1 (resp. k_2) clusters are observed only at the site of covariate X_1 (resp. X_2), with frequencies $\tilde{\mathbf{n}} = (\tilde{n}_1, \ldots, \tilde{n}_{k_1})$ (resp. $\tilde{\mathbf{m}} = (\tilde{m}_1, \ldots, \tilde{m}_{k_2})$).

The EPPF can be expressed as follows

$$p(\mathbf{n},\tilde{\mathbf{n}},\mathbf{m},\tilde{\mathbf{m}}) = \mathsf{E}\Bigg(\sum_{(*)} p_{i_1}^{n_1}(X_1)p_{i_1}^{m_1}(X_2)\dots p_{i_k}^{n_k}(X_1)p_{i_k}^{m_k}(X_2) \times p_{j_1}^{\tilde{n}_1}(X_1)\dots p_{j_{k_1}}^{\tilde{n}_{k_1}}(X_1)\times p_{l_1}^{\tilde{m}_1}(X_2)\dots p_{l_{k_2}}^{\tilde{m}_{k_2}}(X_2)\Bigg) \tag{6}$$

where the sum $(*)$ runs over all $(k+k_1+k_2)$-uples $(i_1,\dots,i_k,j_1,\dots,j_{k_1},l_1,\dots,l_{k_2})$ with pairwise distinct elements.

In *non covariate-dependent models*, the EPPF can be typically derived as follows. First, express the definition of the EPPF in terms of the first few elements of a size-biased permutation $\tilde{\mathbf{p}}$ given $\mathbf{p}$ by application of Lemma 2 where $f(p_1,\dots,p_k) = p_1^{n_1}\dots p_k^{n_k}$. Second, use the invariance under size-biased permutation (ISBP) property that characterizes the GEM distribution (*cf.* Pitman, 1996), in order to replace the first few elements of the size-biased permutation $\tilde{\mathbf{p}}$ by the first few elements of $\mathbf{p}$. And finally use the stick-breaking representation of $\mathbf{p}$ with independent Beta random variables $\mathbf{V}$ and compute the moments of Beta random variables in order to obtain a compact expression of the EPPF.

In the case of *covariate-dependent models* as in (6), the hindrance to further computation of a closed-form expression for $p(\mathbf{n},\tilde{\mathbf{n}},\mathbf{m},\tilde{\mathbf{m}})$ is, to the best of our knowledge, twofold: (i) the sum in Equation (6) does not reduce to any conditional expectation of the first few elements of a size-biased permutation of $\mathbf{p}$, and (ii) the invariance under size-biased permutation property is not straightforward to generalize to covariate-dependent distributions, hence equality in distribution between $(\tilde{p}_1(X_1),\tilde{p}_1(X_2))$ and $(p_1(X_1),p_1(X_2))$ is not a known property (whereas it is marginally true).

Notwithstanding this, EPPF have been obtained in the covariate-dependent literature, though not for stick-breaking constructions, but when the dependent process is defined by normalizing random probability measures such as completely random measures. See for instance Lijoi et al. (2013); Kolossiatis et al. (2013); Griffin et al. (2013), and Müller et al. (2011) for an approach based on product partition models.

The following proposition now gives the joint law for the first picks at two sites under the Dep-GEM prior.

Proposition 3 *Let the samples $\mathbf{Y}_1^n = (Y_{1,1},\dots,Y_{n,1})$ and $\mathbf{Y}_2^m = (Y_{1,2},\dots,Y_{m,2})$ at two sites X_1 and X_2, given the process $\mathbf{p}\sim$Dep-GEM (M). The joint law of $Y_{1,1}$ and $Y_{1,2}$ is:*

$$\mathsf{P}(Y_{1,1}=j,Y_{1,2}=k) = (M+1-\mu_M)M^{|j-k|-1}(M^2-1+\mu_M)^{(j\wedge k)-1}/(M+1)^{j+k}, \tag{7}$$

for $k\neq j$ and

$$\mathsf{P}(Y_{1,1}=j,Y_{1,2}=j) = \mu_M(M^2-1+\mu_M)^{j-1}/(M+1)^{2j}, \tag{8}$$

where $\mu_M(X_1,X_2) = (M+1)^2\mathsf{E}\big(V(X_1)V(X_2)\big)$.

We proceed now to the investigation of the dependence at the diversity level. Under a GEM(M) prior on $\mathbf{p}$ without dependence, the prior expectation of the Simp-

son diversity follows from Lemma 2 and equals $\mathsf{E}(H_{\mathsf{Simp}}) = M/(M+1)$ (Cerquetti, 2012). The following proposition characterizes the dependence induced in the Simpson diversity index in terms of the covariance between the indices at two sites.

Proposition 4 *The covariance between the Simpson diversity indices at two sites, $H_{\mathsf{Simp}}(X_1)$ and $H_{\mathsf{Simp}}(X_2)$, and the variance of the Simpson diversity index, induced by the* Dep-GEM *distribution, are as follows*

$$\mathsf{Cov}(H_{\mathsf{Simp}}(X_1), H_{\mathsf{Simp}}(X_2)) = \frac{\nu_{2,2}(1-\omega_{2,0}) + 2\nu_{2,0}\gamma_{2,2}}{(1-\omega_{2,0})(1-\omega_{2,2})} - \nu_{1,0}^2, \tag{9}$$

$$\mathsf{Var}(H_{\mathsf{Simp}}(X)) = \frac{M+6}{(M+1)_{(3)}} - \frac{1}{(M+1)^2} = \frac{2M}{(M+1)(M+1)_{(3)}}, \tag{10}$$

where $\nu_{i,j} = \mathsf{E}[V^i(X_1)V^j(X_2)]$, $\omega_{i,j} = \mathsf{E}[(1-V(X_1))^i(1-V(X_2))^j]$, *and* $\gamma_{i,j} = \mathsf{E}[V^i(X_1)(1-V(X_2))^j]$.

The values of $\nu_{i,j}, \omega_{i,j}, \gamma_{i,j}$ cannot be computed in a closed-form expression when $i \times j \neq 0$, but can be approximated numerically. The asymptotics of $\mathsf{Cov}(H_{\mathsf{Simp}}(X_1), H_{\mathsf{Simp}}(X_2))$ w.r.t. $|X_1 - X_2|$ are as follows

- $|X_1 - X_2| \to 0$: the covariance in Proposition 4 converges to $\mathsf{Var}(H_{\mathsf{Simp}}(X_1))$, a property that is also inherited from the continuity of the sample paths of $\mathbf{p}$.
- $|X_1 - X_2| \to \infty$: it can be checked that in the *independent* case, the covariance vanishes to 0.

The variations of the Simpson diversity w.r.t. the precision parameter M are as follows

- $M \to 0$: the prior degenerates to a single species with probability 1, hence $H_{\mathsf{Simp}} \to 0$.
- $M \to \infty$: the prior tends to favour infinitely many species, and $H_{\mathsf{Simp}} \to 1$. In both cases, $\mathsf{Var}(H_{\mathsf{Simp}}) \to 0$. We see that the covariance vanishes also in these two cases by using the inequality

$$\begin{aligned} |\mathsf{Cov}(H_{\mathsf{Simp}}(X_1), H_{\mathsf{Simp}}(X_2))| &\leq \left[\mathsf{Var}(H_{\mathsf{Simp}}(X_1))\mathsf{Var}(H_{\mathsf{Simp}}(X_2))\right]^{1/2} \\ &= \mathsf{Var}(H_{\mathsf{Simp}}(X_1)). \end{aligned}$$

- $M = 1/2$: the variance (10) of the Simpson index is maximum for the precision parameter value $M = 1/2$.

Despite the fact that the first moments of the diversity indices under a GEM prior can be easily derived, a full description of the distribution seems hard to achieve. For instance, the distribution of the Simpson index involves the small-ball like probabilities $\mathsf{P}(\sum_j p_j^2 < a)$ for which, to the best of our knowledge, no result is known under the GEM distribution.

Acknowledgment

This work was supported by the European Research Council (ERC) through StG "N-BNP" 306406.

References

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

Arbel, J., Mengersen, K., and Rousseau, J. (2013b). Bayesian nonparametric dependent model for the study of diversity in species data. `http://arxiv.org/abs/1402.3093`.

Archer, E., Park, I. M., and Pillow, J. (2013). Bayesian Entropy Estimation for Countable Discrete Distributions. *Preprint*.

Cerquetti, A. (2012). Bayesian nonparametric estimation of Simpson's evenness index under $\alpha-$Gibbs priors. *arXiv preprint arXiv:1203.1666*.

Griffin, J. E., Kolossiatis, M., and Steel, M. F. (2013). Comparing distributions by using dependent normalized random-measure mixtures. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*.

Ishwaran, H. and James, L. (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, 96(453):161–173.

Kolossiatis, M., Griffin, J. E., and Steel, M. F. (2013). On Bayesian nonparametric modelling of two correlated distributions. *Statistics and Computing*, 23(1):1–15.

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., Nipoti, B., and Prünster, I. (2013). Bayesian inference with dependent normalized completely random measures. *Bernoulli*, to appear.

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.

MacEachern, S. (1999). Dependent nonparametric processes. *ASA Proceedings of the Section on Bayesian Statistical Science*, pages 50–55.

MacEachern, S. (2000). Dependent Dirichlet processes. Technical report, Ohio State Uni- versity. Association.

Müller, P., Quintana, F. A., and Rosner, G. L. (2011). A product partition model with regression on covariates. *Journal of Computational and Graphical Statistics*, 20(1).

Pitman, J. (1995). Exchangeable and partially exchangeable random partitions. *Probability Theory and Related Fields*, 102(2):145–158.

Pitman, J. (1996). Random discrete distributions invariant under size-biased permutation. *Advances in Applied Probability*, pages 525–539.

Pitman, J. (2006). *Combinatorial stochastic processes*, volume 1875. Springer-Verlag.