Contributions to Bayesian nonparametric statistic

Julyan Arbel

Université Paris-Dauphine

Thèse doctorale

Contributions à la statistique bayésienne non-paramétrique

Thèse soutenue par Julyan Arbel
pour le titre de Docteur en Mathématiques Appliquées

Université Paris-Dauphine
Ecole Doctorale de Dauphine
CEREMADE
CEntre de REcherche en MAthématiques de la DEcision

le 23 septembre 2013

Jury

Judith Rousseau	Université Paris-Dauphine	*Directrices*
Ghislaine Gayraud	Université de Technologie de Compiègne	*de thèse*
David Dunson	Duke University	*Rapporteurs*
Antonio Lijoi	University of Pavia	
Ismaël Castillo	LPMA CNRS	*Examinateurs*
François Caron	INRIA Bordeaux	
Vincent Rivoirard	Université Paris-Dauphine	

Résumé

Un résumé des différents chapitres est proposé ci-dessous, et se trouve galement en dbut de chaque chapitre.

Chapitre 1 : La thèse est divisée en deux parties portant sur deux aspects relativement différents des approches bayésiennes non-paramétriques. Dans la première partie, nous nous intéressons aux propriétés fréquentistes de lois a posteriori pour des paramètres appartenant à l'ensemble des suites réelles de carré sommable. Dans la deuxième partie, nous nous intéressons à des approches non-paramétriques modélisant des données d'espèces et leur diversité en fonction de certaines variables explicatives. La présente introduction reflète cette dichotomie. Dans un premier temps, nous rappelons les principaux résultats concernant les propriétés fréquentistes asymptotiques des lois a posteriori en dimension infinie. Puis nous décrivons les outils autour des modèles de mesures de probiblité aléatoires qui nous servirons dans la modélisation de la diversité d'espèces.

Chapitre 2 (co-écrit avec Ghislaine Gayraud et Judith Rousseau) : On propose une forme générique de distributions a priori pour obtenir des résultats de vitesse de contraction de la loi a posteriori dans plusieurs modèles. Ces lois a priori sont appelées *sieve priors*. Elles permettent de plus d'obtenir des vitesses qui s'adaptent à la régularité du paramètre, sans que cette régularité soit utilisée dans la méthode d'estimation. Les résultats sont illustrés sur les modèles de densité, de régression, d'autorégression d'ordre 1 et de bruit blanc Gaussien. On montre en outre qu'une approche adaptative pour une fonction de perte donnée (par exemple globale) peut s'avérer sous-optimale pour une autre fonction de perte (par exemple locale) dans le cas du modèle de bruit blanc Gaussien.

Chapitre 3 (co-écrit avec Judith Rousseau et Kerrie Mengersen) : On introduit dans ce chapitre un modèle bayésien non-paramétrique pour étudier de manière probabiliste des données d'espèces par site, c'est à dire des données de population pour lesquelles les individus observés par site par site appartiennent à différentes espèces. Ces données peuvent être représentées par une matrice constituée du nombre d'occurrences de chaque espèce sur chaque site. Notre but est d'étudier l'impact de facteurs, ou variables explicatives,

additionnels, tels que des variables environnementales, sur la structure des données, et en particulier sur la diversité. A cet effet, on introduit de la dépendance a priori selon les variables explicatives, et on montre que cela améliore l'inférence a posteriori. On utilise une version dépendante de la distribution GEM, qui représente la distribution des poids du processus de Dirichlet, de la même manière que sont dfinis les processus de Dirichlet dépendants. La loi a priori est définie à partir de la construction *stick-breaking*, dans laquelle on obtient les poids en transformant un processus gaussien, et la dépendance découle de la fonction de variance-covariance de ce dernier. On explicite des propriétés de distribution du modèle, telle que sa fonction de probabilité de partition échangeable jointe. On décrit un algorithme de Monte-Carlo par chaîne de Markov pour l'échantillonnage a posteriori, ainsi que l'échantillonnage de la loi prédictive pour des facteurs inobservés. Les deux algorithmes sont illustrés sur les données simulées et sur des données d'expériences réalisées par des prélèvements dans le sol en Antarctique.

Chapitre 4 (co-écrit avec Judith Rousseau, Kerrie Mengersen, Ben Raymond et Catherine King) : On étudie dans ce chapitre le modèle bayésien non-paramétrique du Chapitre 3 dans une perspective plus appliquée, avec comme domaine d'application l'écotoxicologie. Ici, les espèces sont des microbes, et le facteur est une variable de contamination environnementale importante appelée Hydrocarbure de Pétrole Total. On étudie son impact sur les données sous des angles différents: en terme de diversité de Shannon, de clustering (en des groupes de microbes qui réagissent de manière similaire au contaminant), et en terme de décroissance de la proportion relative des espèces (l'estimation de quantités appelées IC_{50}, qui correspondent au niveau de contamination pour lequel la proportion relative est divisée par deux par rapport à la proportion relative à une valeur de contamination de référence). Ce modèle, étudié sur des données microbiennes mesurées dans le sol en Antarctique, est applicable plus généralement à de nombreux autres problèmes dans lesquels la structure des données et les questions inférentielles sont similaires.

Chapitre 5 : Ce chapitre présente des travaux en lien avec la statistique bayésienne non-paramétrique qui ont été commencés pendant la thèse et seront poursuivis par la suite. Le premier projet concerne l'estimation adaptative spatiale. On se place dans le cas où la régularité du paramètre varie dans l'espace de définition de ce dernier. On recherche des lois a priori qui sont adaptatives optimales dans toutes les régions de l'espace, c'est à dire dont la loi a posteriori converge avec une vitesse associée à la régularité de la région de l'espace considérée. Le second projet concerne l'estimation de densité dans un cadre multivarié dans lequel les variables ou composantes des observations ne sont pas directement comparables. Le cas de grandeurs physiques avec des unités de mesure différentes, tel que l'espace des phases constitué des variables de

position et de moment, représentent un exemple typique. On montre par des simulations que l'estimation de la densité à partir de mélanges de processus de Dirichlet ont des propriétés d'invariance qui font de ces mélanges une solution adaptée à ce type de problèmes.

Abstract

A summary of each chapter, which can also be found in the abstract of some of them, is proposed below.

Chapter 1: This thesis is divided in two parts on rather different aspects of Bayesian statistics. In the first part, we deal with frequentist properties of posterior distributions for parameters which belong to the space of real square sommable sequences. In the second part, we deal with nonparametric approaches modelling species data and the diversity of these data with respect to covariates. This introduction reflects this dichotomy. First, we recall the most important results about asymptotic frequentist properties of posterior distributions in infinite dimension. Second, we describe the tools based on random probability measures that will be utilised in modelling species diversity.

Chapter 2 (joint work with Ghislaine Gayraud and Judith Rousseau): A generic shape of prior distributions which result in optimal posterior contraction rate in various models is proposed. These prior distributions are called *sieve priors*. These allow derivation of rates which are adaptive to the parameter smoothness, without using this smoothness knowledge in the estimation. The results are illustrated on the density, regression, order 1 autoregressive, and Gaussian white noise models. We show in addition that an adaptive approach for a given loss function (for instance global) can prove to be severely sub-optimal for another loss function (for instance local) in the case of the Gaussian white noise model.

Chapter 3 (joint work with Judith Rousseau and Kerrie Mengersen): We introduce 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 in distinct species. These data can be represented as a frequency matrix giving the number of times each species is observed in each site. Our aim is to study the impact of additional factors (covariates), for instance environmental factors, on the data structure, and in particular on the diversity. To that purpose, we introduce dependence a priori across the covariates, and show that it improves posterior inference. We use a dependent version of the GEM distribution, which is 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, where the weights are obtained by transforming a Gaussian process, and the dependence stems from the covariance function of the latter. Some distributional properties of the model are derived, such as its joint exchangeable partition probability function. A Markov chain Monte Carlo algorithm for posterior sampling is described, along with the sampling scheme of the predictive distribution for unobserved factors. Both samplers are illustrated on simulated data and on a real data set obtained in experiments conducted in Antarctica soil.

Chapter 4 (joint work with Judith Rousseau, Kerrie Mengersen, Ben Raymond and Catherine King): We study the Bayesian nonparametric model of Chapter 3 from a more applied perspective, with a focus in the field of ecotoxicology. Here, the species are microbes, and the factor is an important environmental contaminant called Total Petroleum Hydrocarbon. Its impact on the data is studied from different points of view: in term of Shannon diversity, of clustering (into groups with similar behaviour with respect to the contaminant), and of species relative proportion decrease (estimation of a quantity called IC_{50}, the covariate level for which the relative proportion is divided by 2 compared to the relative proportion at a given covariate value). The model, which is studied on soil microbial data collected in Antarctica, is broadly applicable to a range of other problems with the same data structure and inferential requirements.

Chapter 5: We present some work in Bayesian nonparametric statistic which were started during the PhD and will be continued afterwards. The first project deals with spatially adaptive estimation. The typical framework is the estimation of parameters with a smoothness that varies over its domain. We are looking for priors which are optimal adaptive in any region of the domain, that is to say whose posterior distribution contracts with a rate which depends on the parameter smoothness specific to any region of the domain. The second project deals with density estimation in a multivariate space where the variables are not directly comparable. The typical case relates to physical quantities with different units of measurement, such as the phase space which consists of position and momentum variables. We show by a simulation study that density estimation based on Dirichlet process mixtures have appropriate invariance properties which make these mixtures good candidates for this kind of problems.

Remerciements

Ghislaine et Judith, votre support pendant mes années de thèse m'a été précieux. Vous avez été source d'inspiration et de motivation, disponibles pour m'aider dans les moments de doute. Soyez en chaleureusement remerciées ! Merci également aux personnes de l'autre côté de la planète, Kerrie Mengersen, Ben Raymond et Cath King, avec qui nous collaborons et qui ont répondu à mes questions. Jean-Bernard, merci pour ta relecture top chrono ! Votre aide a été décisive lors de la période de rédaction.

Merci à David Dunson et à Antonio Lijoi d'avoir accepté d'être rapporteurs de cette thèse, ainsi qu'à François Caron, Ismaël Castillo et Vincent Rivoirard d'avoir accepté de faire partie du jury.

De nombreuses personnes, amis, collègues, professeurs, ont joué un rôle important pendant ma thèse. Amis de tous horizons, danseurs, sportifs, et autres, collègues de l'ENSAE, du CREST ou de l'INSEE, ou rencontrés en séjour à Rochebrune, Valencià, Milan, Pavie, ou en conférence à l'autre bout du monde. Des rencontres formidables dans des lieux parfois extraordinaires, des plages du Yucatán au milieu de la nuit, un cocktail à la main, aux remontées mécaniques de Megève, skis aux pieds. Vous saurez vous reconnatre j'espère, vous êtes tous un peu pour quelque chose dans ce manuscript (mais ne sauriez être tenus responsables des erreurs qui s'y sont glissées). Je vous remercie pour ces très bons moments ! Merci à vous, les enseignants qui ont jalonné ma scolarité et qui ont su me donner le goût de la recherche.

Merci Bernardo, bien malgrè toi tu as joué un rôle déterminant dans cette thèse, et merci Igor, Antonio et l'équipe du *Collegio Carlo Alberto* pour votre confiance.

Enfin, à ma famille et ma belle-famille : merci pour votre soutien inconditionnel depuis le début. Christelle, merci d'avoir accepté tous ces sacrifices. Pour tout ce temps passé à me remplacer à la maison, tu mériterais d'être co-auteure de cette thèse. Jeanne, merci à toi, tu m'as donné beaucoup d'énergie ces derniers mois. Papa va bientôt laisser un peu son ordinateur et venir jouer avec toi !

Contents

Chapter 1

Introduction

La thèse est divisée en deux parties portant sur deux aspects relativement différents des approches bayésiennes non-paramétriques. Dans la première partie, nous nous intéressons aux propriétés fréquentistes de lois a posteriori pour des paramètres appartenant à l'ensemble des suites réelles de carré sommable. Dans la deuxième partie, nous nous intéressons à des approches non-paramétriques modélisant des données d'espèces et leur diversité en fonction de certaines variables explicatives. La présente introduction reflète cette dichotomie. Dans un premier temps, nous rappelons les principaux résultats concernant les propriétés fréquentistes asymptotiques des lois a posteriori en dimension infinie. Puis nous décrivons les outils autour des modèles de mesures de probiblité aléatoires qui nous servirons dans la modélisation de la diversité d'espèces.

1.1 General introduction

Bayesian nonparametric modelling has recently attracted a lot of attention. Bernardo and Smith (2009) define a "Bayesian nonparametric model" as a probability model with *infinitely many parameters*, also referred to as a model with *massively many parameters* in Müller and Mitra (2013). Bayesian nonparametric methods are motivated by the aim of considering models that are not limited to finite parametrizations. Data $\mathbf{Y}^{(n)} = (Y_1, \ldots, Y_n) \sim P_{\boldsymbol{\theta}}^{(n)}$ is modelled by some infinite dimensional parameter $\boldsymbol{\theta}$ in a space Θ. We want to make inference typically on a curve, *eg* a density, a spectral density, a regression function, a cumulative distribution function, a hazard function, a link function, etc. The basic steps of the Bayesian machinery consist in (i) constructing a prior distribution Π on the space Θ of infinite dimension ; (ii) computing the posterior distribution $\Pi\big(\boldsymbol{\theta}|\mathbf{Y}^{(n)}\big)$. Practically, step (ii) involves Monte Carlo methods such as Markov chain Monte Carlo algorithms (MCMC) or sequential Monte Carlo algorithms. Theoretically, the performance of estimation can be established via asymptotic properties such as consistency, rates of convergence, Bernstein-von Mises property, etc. Loosely speaking, consistency means that if the data are generated according to a model with a true parameter $\boldsymbol{\theta}_0$, then the posterior $\Pi\big(\boldsymbol{\theta}|\mathbf{Y}^{(n)}\big)$ should concentrate around $\boldsymbol{\theta}_0$ as the number of observations n goes to infinity. The rate of contraction of the posterior distribution characterizes how fast the convergence operates.

This introductory chapter reviews general aspects of Bayesian nonparametric methods. Their theoretical large sample properties are presented in Section 1.2, while Section 1.3 is devoted to methodological aspects of random probability measures. The monograph Ghosh and Ramamoorthi (2003) covers many aspect of Bayesian asymptotic statistics. Surveys of Bayesian nonparametric models include Walker et al. (1999), Müller and Quintana (2004), and a broad cover on random probability measures models can be found in the monograph Hjort et al. (2010).

1.2 Large sample properties

We present in this section general theoretical results on the asymptotic behaviour of posterior distributions in nonparametric models. Looking at the asymptotic behaviour of the posterior distribution helps understanding how the prior acts on the likelihood. In particular it sheds light on the most influential aspects of the prior, which are those which do not disappear asymptotically. Typically in finite dimensional models, to first order, the posterior is asymptotically independent of the prior and any prior leads to asymptotically equivalent inference. In infinite dimensional models, this is not the case

any more and the prior does not looses its influence, even to first order. Understanding its influence is thus of crucial importance, given the complexity of the models and studying the frequentist properties of the posterior distribution is one way to do so. Besides it allows for a comparison with frequentist estimators.

1.2.1 Posterior consistency and posterior concentration rates

We now present briefly the various notions of consistency that are encountered in this thesis. Let $\mathbb{Y}$ and Θ be complete and separable metric spaces endowed with their Borel σ–algebras. We assume that observations $\mathbf{Y}^{(n)}$ are random variables on $\mathbb{Y}$. Let $d(\,\cdot\,,\,\cdot\,)$ be a loss function over Θ, and U any neighbourhood of a given point $\theta_0 \in \Theta$ associated to this loss function, in other words U has the form $U = \{\boldsymbol{\theta}; d(\boldsymbol{\theta}_0, \boldsymbol{\theta}) < \epsilon\}$, for some $\epsilon > 0$. We then write $U = U_\epsilon$ to make this dependence explicit. The loss $d(.,.)$ function can be a metric, such as the Hellinger metric, see Equation (1.3), in the case of independent and identically distributed (i.i.d.) observations when the parameter is the density of the observations or a semi-parametric loss where only some aspects of the model are of interest, such as the square error loss of a linear functional of the parameter. Let $p_{\boldsymbol{\theta}}^{(n)}$ denote the density of the probability measure $P_{\boldsymbol{\theta}}^{(n)}$ with respect to some measure μ (independent of $\boldsymbol{\theta}$), the posterior probability of any measurable subset B of Θ and associated to a prior distribution Π on Θ can be written as

$$\Pi(B|\mathbf{Y}^{(n)}) = \frac{\int_B p_{\boldsymbol{\theta}}^{(n)}(\mathbf{Y}^{(n)})\mathrm{d}\Pi(\boldsymbol{\theta})}{\int_\Theta p_{\boldsymbol{\theta}}^{(n)}(\mathbf{Y}^{(n)})\mathrm{d}\Pi(\boldsymbol{\theta})}. \tag{1.1}$$

In the case of independent and identically distributions, for instance, $\boldsymbol{\theta} = \boldsymbol{f}$ the density with respect to a given measure say $\mathrm{d}x$ and $p_{\boldsymbol{\theta}}^{(n)}(\mathbf{Y}^{(n)}) = \prod_{i=1}^n f(Y_i)$. We can then define posterior consistency with respect to $d(.,.)$ in the following way (see for instance Ghosh and Ramamoorthi, 2003).

Definition 1.1. *The posterior distribution is said to be consistent at $\boldsymbol{\theta}_0 \in \Theta$ if for any $\epsilon > 0$, the posterior probability of an ϵ neighbourhood of $\boldsymbol{\theta}_0$, $U_\epsilon = \{\boldsymbol{\theta} \in \Theta; d(\boldsymbol{\theta}_0, \boldsymbol{\theta}) < \epsilon\}$, converges to 1 under $P_{\boldsymbol{\theta}_0}^{(n)}$:*

$$\Pi\left(U_\epsilon|\mathbf{Y}^{(n)}\right) \to 1,$$

either in $P_{\boldsymbol{\theta}_0}^{(n)}$-probability, or $P_{\boldsymbol{\theta}_0}^{(\infty)}$ almost surely, as n tends to infinity.

Note that speaking of a true parameter means we adopt what is called a *frequentist-Bayesian* point of view. An argument for consistency is that it allows a correct identification of the mechanism that generated the data when unlimited data is available. In particular when posterior consistency is not verified, interpretation of the posterior

distribution is problematic. Moreover, in the case of exchangeable data Diaconis and Freedman (1986) have shown that posterior consistency is equivalent to weak merging of the posterior distributions associated to any proper prior. An early result of Doob (1949) implies that when $d(.,.)$ is a metric and (Θ, d) is separable and complete and under ergodicity conditions, any posterior distribution is consistent at $\boldsymbol{\theta}_0$, Π almost surely. This is a weak result since it does not provide the identification of the set of $\boldsymbol{\theta}_0 \in \Theta$ such that posterior consistency holds at $\boldsymbol{\theta}_0$.

General approaches to prove consistency are due to Schwartz (1965) in the case of independent and identically distributed observations, and by Barron (1988) in a generalized case. These are based on two types of conditions, (i) conditions on the size of the model, and (ii) prior support conditions for the Kullback–Leibler divergence (abbreviated KL). Loosely speaking, Schwartz's theorem states that the posterior is consistent if the model is controlled by some tests, and the support of the prior is large in the sense of KL. Recall that the Kullback–Leibler divergence between the distributions $P_{\boldsymbol{\theta}_1}^{(n)}$ and $P_{\boldsymbol{\theta}_2}^{(n)}$ is defined by

$$\mathsf{KL}(P_{\boldsymbol{\theta}_1}^{(n)}, P_{\boldsymbol{\theta}_2}^{(n)}) = \int_{\mathcal{Y}_n} p_{\boldsymbol{\theta}_1}^{(n)}(y) \log \left(\frac{p_{\boldsymbol{\theta}_1}^{(n)}}{p_{\boldsymbol{\theta}_2}^{(n)}} \right)(y) \mathrm{d}\mu(y). \tag{1.2}$$

We also denote by $\mathsf{KL}(\boldsymbol{f}_1, \boldsymbol{f}_2)$ the Kullback - Leibler divergence between the distributions with density $\boldsymbol{f}_1$ and $\boldsymbol{f}_2$ with respect to a given measure μ. Schwartz's Theorem then states:

Theorem 1.2 (Schwartz (1965))**.** *Assume that the Y_i's are independent and identically distributed with density $\boldsymbol{f} \in \mathcal{F}$ on $\mathcal{Y}$. Let Π be a prior on $\mathcal{F}$ such that any KL neighbourhood of the true density $\boldsymbol{f}_0$ has positive prior probability:*

$$\Pi\left(\mathsf{KL}(\boldsymbol{f}_0, \boldsymbol{f}) < \epsilon\right) > 0 \quad \forall \epsilon > 0.$$

If there exists a sequence of tests $\phi_n(\mathbf{Y}^{(n)})$ (said to be exponentially consistent) such that for any neighbourhood U of $\boldsymbol{f}_0$, there exist $C, \beta > 0$ such that

$$\mathsf{E}_{\boldsymbol{f}_0} \phi(\mathbf{Y}^{(n)}) \leq C e^{-n\beta}, \quad \sup_{\boldsymbol{f} \in U^c} \mathsf{E}_{\boldsymbol{f}} \left(1 - \phi(\mathbf{Y}^{(n)})\right) \leq C e^{-n\beta},$$

then

$$\Pi(U | \mathbf{Y}^{(n)}) \longrightarrow 1 \quad P_{\boldsymbol{f}_0}^{(n)} \text{ a.s..}$$

In the density problem, an example of metric d is the Hellinger metric. If $P_{\boldsymbol{\theta}_1}^{(n)}$ and $P_{\boldsymbol{\theta}_2}^{(n)}$ are two probability measures, of respective densities $p_{\boldsymbol{\theta}_1}^{(n)}$ and $p_{\boldsymbol{\theta}_2}^{(n)}$ with respect to

a common measure μ, the Hellinger metric between $P_{\boldsymbol{\theta}_1}^{(n)}$ and $P_{\boldsymbol{\theta}_2}^{(n)}$ is defined by

$$h^2(P_{\boldsymbol{\theta}_1}^{(n)}, P_{\boldsymbol{\theta}_2}^{(n)}) = \frac{1}{2}\int \left(\sqrt{p_{\boldsymbol{\theta}_1}^{(n)}(y)} - \sqrt{p_{\boldsymbol{\theta}_2}^{(n)}(y)}\right)^2 \mathrm{d}\mu(y). \tag{1.3}$$

There are alternative approaches to Schwartz's which do not require the use of uniformly exponentially consistent tests. For instance Walker (2004), Walker et al. (2005) uses a martingale method. Another technique makes use of so-called power-posterior distributions: the Bayes formula is modified by raising the likelihood to a power $\alpha \in (0, 1)$. It is thus an altered version of the Bayes paradigm, which has the advantage to be consistent under the KL condition only, without having to resort to tests. See Walker and Hjort (2001) for the original paper and the chapter **?** of Hjort et al. (2010) for a concise account on the principle of the method.

A more refined asymptotic property than consistency is the so-called posterior concentration (or contraction) rates, measuring how fast the posterior distribution shrinks around the true parameter. In studying posterior consistency, we get a better understanding on some aspects of the prior or at least on how it operates with respect to the likelihood. In particular, in large or infinite dimensional models the prior does not completely vanish asymptotically even to first order, as opposed to what happens in the parametric case. The aspects, in the prior which are influential asymptotically are bound to have a strong influence for finite samples and should be treated with particular care, in applications. We now define the posterior concentration or contraction rate.

Definition 1.3. *Given the prior* Π*, a rate of contraction of the posterior distribution in* $P_{\boldsymbol{\theta}_0}^{(n)}$ *probability with respect to a semimetric* d_n *on* Θ *is defined as a sequence* $(\epsilon_n)_{n\geq 1}$ *such that*

$$\Pi\left(\boldsymbol{\theta} : d_n^2(\boldsymbol{\theta}, \boldsymbol{\theta}_0) \geq M_n \epsilon_n^2 | Y^n\right) \underset{n\to\infty}{\longrightarrow} 0, \tag{1.4}$$

in $P_{\boldsymbol{\theta}_0}^{(n)}$ *probability, for some* $\boldsymbol{\theta}_0 \in \Theta$ *and every sequence* $M_n \to \infty$ *as* $n \to \infty$*.*

The best possible (i.e. the smallest) sequence $(\epsilon_n)_{n\geq 1}$ *satisfying Equation (1.4) is called the* optimal rate of contraction*.*

In their seminal paper Ghosal et al. (2000) and in its extension to non i.i.d. settings, Ghosal and van der Vaart (2007), the authors have developed a generic methodology to obtain posterior concentration rates, following the ideas of Barron (1988) and Schwartz (1965) (see also Shen and Wasserman, 2001). We give and discuss below a version of the general result from Ghosal and van der Vaart (2007). First we define the following kth centred moment of KL between the distributions $P_{\boldsymbol{\theta}_1}^{(n)}$ and $P_{\boldsymbol{\theta}_2}^{(n)}$ or equivalently their

densities $p_{\boldsymbol{\theta}_i}^{(n)}$, $i = 1, 2$, for $k \geq 2$,

$$\mathsf{V}_k\big(P_{\boldsymbol{\theta}_1}^{(n)}, P_{\boldsymbol{\theta}_2}^{(n)}\big) = \int p_{\boldsymbol{\theta}_1}^{(n)}(y) \left| \log\left(\frac{p_{\boldsymbol{\theta}_1}^{(n)}}{p_{\boldsymbol{\theta}_2}^{(n)}}\right)(y) - \mathsf{KL}\big(P_{\boldsymbol{\theta}_1}^{(n)}, P_{\boldsymbol{\theta}_2}^{(n)}\big) \right|^k \mathrm{d}\mu(y),$$

and let the following KL neighbourhood

$$\mathcal{B}_n(\boldsymbol{\theta}_0, \epsilon_n, k) = \left\{ \boldsymbol{\theta} : \mathsf{KL}\big(P_{\boldsymbol{\theta}_0}^{(n)}, p_{\boldsymbol{\theta}}^{(n)}\big) \leq n\epsilon_n^2, \mathsf{V}_k\big(P_{\boldsymbol{\theta}_0}^{(n)}, p_{\boldsymbol{\theta}}^{(n)}\big) \leq (n\epsilon_n^2)^{k/2} \right\}.$$

Theorem 1.4 (Theorem 3 of Ghosal and van der Vaart, 2007)**.** *Let $d(.,.)$ be a semimetric (possibly depending on n) on Θ, $\epsilon_n > 0$, $\epsilon_n \to 0$, $n\epsilon_n^2 \to \infty$, $k > 1$, $K > 0$, and $\Theta_n \subset \Theta$. If there exists a (measurable) sequence of test functions $\phi_n : \mathbb{Y} \to [0, 1]$ such that for every sufficiently large integer j*

$$\mathsf{E}_{\boldsymbol{\theta}_0}^{(n)} \phi_n \to 0, \qquad \sup_{\boldsymbol{\theta} \in \Theta_n : j\epsilon_n < d(\boldsymbol{\theta}, \boldsymbol{\theta}_0) \leq 2j\epsilon_n} \mathsf{E}_{\boldsymbol{\theta}}^{(n)}(1 - \phi_n) \leq e^{-Kj^2 n\epsilon_n^2}, \tag{1.5}$$

$$\frac{\Pi\big(\boldsymbol{\theta} \in \Theta_n : j\epsilon_n < d(\boldsymbol{\theta}, \boldsymbol{\theta}_0) \leq 2j\epsilon_n\big)}{\Pi\big(\mathcal{B}_n(\boldsymbol{\theta}_0, \epsilon_n, k)\big)} \leq e^{Kn\epsilon_n^2 j^2/2}, \tag{1.6}$$

then for every $M_n \to 0$, we have that

$$\Pi\big(\boldsymbol{\theta} \in \Theta_n : d(\boldsymbol{\theta}, \boldsymbol{\theta}_0) \geq M_n\epsilon_n | Y^{(n)}\big) \to 0,$$

as $n \to \infty$ in $P_{\boldsymbol{\theta}_0}^{(n)}$-probability.

We give a brief intuition of the conditions required in Theorem 1.4. A more general discussion can be found in Ghosal et al. (2000), Ghosal and van der Vaart (2007). The Θ_n, called *sieve* spaces, allow focusing on well-behaved parameters (in terms of complexity, of size, *etc*), and to avoid the problematic part of the space, $\Theta \backslash \Theta_n$, whose size is controlled. The tests in condition (1.5) can be thought of as separating $\boldsymbol{\theta}_0$ with any θ supported by the prior and away from $\boldsymbol{\theta}_0$. Often the construction of ϕ_n involves a covering of Θ_n into small balls $V_{n,l}$ with individual tests $\phi_{n,l}$ satisfying

$$E_{\boldsymbol{\theta}_0}^{(n)}[\phi_{n,l}] \leq e^{-cnd(\boldsymbol{\theta}_0, \boldsymbol{\theta}_l)^2}, \qquad \sup_{\boldsymbol{\theta} \in V_{n,l}} E_{\boldsymbol{\theta}}\left(1 - \phi_{n,l}\right) \leq e^{-cnd(\boldsymbol{\theta}_0, \boldsymbol{\theta}_l)^2},$$

$\boldsymbol{\theta}_l$ some point in $V_{n,l}$. Then $\phi_n = \max_l \phi_{n,l}$ satisfies condition (1.5) as soon as the covering number $N_{n,l}$ of $\{\boldsymbol{\theta} \in \Theta_n; j\epsilon_n < d(\boldsymbol{\theta}, \boldsymbol{\theta}_0) \leq 2j\epsilon_n\}$ by these balls is bounded by $e^{cnj^2\epsilon_n^2/2}$. Furthermore, a sufficient condition which ensures that condition (1.6) holds is that the prior Π puts some minimal mass on the Kullback-Leibler neighbourhoods $\mathcal{B}_n(\boldsymbol{\theta}_0, \epsilon_n, k)$ of $\boldsymbol{\theta}_0$.

From Theorem 1.4, we see that posterior concentration rates depend on the ability to approximate the true model in the prior model. This is influenced typically by either some shape constraints on the parameter or by smoothness assumptions on the curve when the parameter is a curve. Hence, it is common practice to determine posterior concentration rates uniformly over functional (or parameter) classes and the posterior is said to concentration at a minimax rate if the posterior concentration rate over this class corresponds to the minimax convergence rate over the same class and under the same loss function $d(.,.)$. In the last decade posterior concentration rates have been derived for various types of families priors and models and minimax posterior concentration rates have been achieved in many cases (up to $\log n$ terms). These notions are described in the following section.

Adaptation and minimax posterior concentration rates

In some cases, we assume some restrictions on the parameter space Θ. Usually, these restrictions are intended to focus on particular well-behaved classes. The subspace $\Theta_\beta \subset \Theta$ can be for example a class of a given smoothness β (*eg*, if the parameters are curves, these include Hölder, Sobolev and Besov classes as leading examples).

The theory of contraction rates can be related to the classical (or frequentist) theory of optimal rates of convergence. A celebrated criterion is the *minimax criterion*: given a functional class Θ_0, the *minimax risk* is defined by the maximal risk of an estimator with minimal risk among all estimators. More precisely the minimax rate for estimating θ under the loss function $d(.,.)$ over the class $\Theta_0 \subset \Theta$ is defined by any sequence v_n satisfying (see *eg* Tsybakov, 2009)

$$\liminf_n \inf_{\hat{\boldsymbol{\theta}}_n} \sup_{\boldsymbol{\theta}_0 \in \Theta_0} v_n^{-1} E_{\boldsymbol{\theta}_0}^{(n)} \left[d(\hat{\boldsymbol{\theta}}_n, \boldsymbol{\theta}_0) \right] > 0$$

and there exists an estimator $\tilde{\boldsymbol{\theta}}_n$ such that

$$\limsup_n \sup_{\boldsymbol{\theta}_0 \in \Theta_0} v_n^{-1} E_{\boldsymbol{\theta}_0}^{(n)} \left[d(\tilde{\boldsymbol{\theta}}_n, \boldsymbol{\theta}_0) \right] < +\infty.$$

A posterior distribution is said to concentrate at the minimax rate over the class Θ_0 in terms of the loss $d(,.,.)$ if its posterior concentration is the corresponding minimax estimation rate. It is shown (*eg* in Ghosal et al., 2000) that the posterior yields a point estimate that converges at the same rate as the posterior contraction rate, at least for convex and bounded loss functions. Hence, the optimal rate of contraction cannot be better than the minimax rate, uniformly over the class. For instance various types of priors have been studied in the context of density estimation for independent

and identically distributed data. The nonparametric mixture models have proved in particular to lead to minimax (up to a $\log n$ term) posterior concentration rates for Hölder functional classes, see for instance Ghosal and van der Vaart (2001, 2007) or Kruijer et al. (2010), Shen et al. (2013) in the case of Gaussian mixtures, Ghosal (2001), Kruijer and Van der Vaart (2008), Rousseau (2010) and McVinish et al. (2009) for mixtures of Beta distributions or triangular densities. Posterior concentrations for other types of curves such as regression, autoregressions, spectral densities have also been studied. For the former, priors are often based on Gaussian processes, see for instance van der Vaart and van Zanten (2008), while histograms have been considered for the second type in Ghosal and van der Vaart (2007) and series expansions for the later in Rousseau et al. (2012). If one wants to estimate a univariate β–smooth function, say β-Hölder, for instance a density or a regression, then the minimax rate under the squared error loss is typically of order $n^{-\beta/(2\beta+1)}$. When the smoothness parameter β is unknown, it may be hard to construct a good estimator, in the sense that its rate of contraction is the minimax rate. It is desirable to build estimators which do not depend explicitly on β. In other words, from a Bayesian point of view, it means that the prior should be constructed without the knowledge of β, and one then speaks of adaptive posterior distribution. When the posterior concentration rate is the same as the adaptive minimax estimation rate over the given collection of smoothness classes then we say that it concentrates at the minimax adaptive rate.

This problem of adaptive optimality is studied in different settings. The first result on posterior adaptation has been obtained when the unknown parameter is assumed to belong to a discrete set, in Belitser and Ghosal (2003). Later more general aspects of adaptations have been considered. In the context of density and regression function estimation, some criteria have been obtained by Huang (2004) and Ghosal et al. (2008) and specific families of priors have been studied by Scricciolo (2006), van der Vaart and van Zanten (2009), Rivoirard and Rousseau (2012), Rousseau (2010) and Kruijer et al. (2010), de Jonge and van Zanten (2010), in the context of spectral density estimation minimax adaptive posterior contraction rates have been derived by Rousseau and Kruijer (2011) and Rousseau et al. (2012). Recent results also concern adaptation with respect to the dimension of the problem, see for instance Bhattacharya et al. (2013), nonparametric testing problems as in Salomond (2013) or some empirical Bayes approaches as in Szabó et al. (2013). For a general discussion on the impact of the loss function and on posterior adaptation see also the recent work of Hoffmann et al. (2013).

1.2.2 Contributions to asymptotic aspects of Bayesian nonparametric approaches in Chapter 2

The motivation of Chapter 2 stems from the observation that in the Bayesian nonparametric asymptotic literature, the computations involved to derive the posterior concentration rates under priors on curves based on series expansions used the same types of ingredients, for various types of models. Hence, we have proposed in Chapter 2, which is also published in Arbel et al. (2013), a generic study of such families of priors, called *sieve prior*. In this chapter we propose a general theorem, in a similar spirit of the works of van der Vaart and van Zanten (2008, 2009) and we apply this theorem to a series of models. The sieve priors can be defined as follows: let $\Theta = \ell_2$, the set of real sequences $\boldsymbol{\theta} = (\theta_n)_n$ satisfying $\sum_n \theta_n^2 < +\infty$ and consider the following hierarchical prior on Θ

$$\boldsymbol{\theta} \sim \Pi(.) = \sum_{k=1}^{\infty} \pi(k)\Pi_k(.),$$

where $\sum_{k=1}^{\infty} \pi(k) = 1$ and for each k, $\Pi_k(.)$ is a probability on $\Theta_k = \mathbb{R}^k$. In Chapter 2 we restrict our attention to the case of conditionally independent prior, in other words for each k

$$\forall \boldsymbol{\theta}_k = (\theta_1, \cdots, \theta_k) \in \Theta_k,\ \Pi_k(\boldsymbol{\theta}_k) = \prod_{j=1}^{k} \frac{1}{\tau_j} g(\theta_j/\tau_j),$$

with possibly the restriction over some ball A across Θ. In the above definition $\pi(.)$, τ_j and g may depend on n. Such prior can be used for instance in a regression or autoregression setting where the regression function is assumed to be square integral, as is common practice. It can also be used combined with some nonlinear transformation, as in the case of density estimation or for modelling a spectral density. In all these cases $\boldsymbol{\theta}$ represents the vector of the coefficients of the expansion of the function (or some transformation of it) on a basis. Under some weak conditions on g, $\pi(.)$ and τ_j, adaptive posterior concentration rates of order $n^{-\beta/(2\beta+1)}(\log n)^{\kappa}$, for some positive κ, are obtained over collections of Sobolev balls in the form $\Theta_\beta(L) = \{\boldsymbol{\theta} \in \ell_2; \sum_{j=1}^{\infty}(1+j)^{2\beta}\theta_j^2 \leq L\}$, when both the loss function and the Kullback - Leibler divergence can be compared in some weak sense with the l_2 norm on $\boldsymbol{\theta}$, see Theorem 2.2 and Corollary 2.3 in Chapter 2. This theorem is applied in various contexts, where this rate is minimax up to a $\log n$ term and some novel results on regression models and nonlinear auto-regressive models are provided. Moreover a lower and an upper bound is obtained in the case of the white noise model with the local loss function

$$d(\boldsymbol{\theta}, \boldsymbol{\theta}') = \left(\sum_{j=1}^{\infty} \theta_j - \theta_j'\right)^2,$$

showing that for such local loss functions these priors lead to suboptimal posterior concentration rates and estimates. This is the motivating example of the work by Hoffmann et al. (2013), which has lead these authors to analyse the impact of the loss function in the determination of posterior concentration rates.

The family of priors considered in Chapter 2 are useful for modelling curves but rather limited for density estimation, in particular when they are used after exponentiation :

$$\boldsymbol{f}_{\boldsymbol{\theta}}(x) = \exp\left(\sum_{j=1}^{\infty} \theta_j \phi_j(x) - c(\boldsymbol{\theta})\right),$$

since it forces (i) to have compact support and (ii) strong conditions on the behaviour of the densities at the boundary of their support. For density estimation or distribution modelling, models based on nonparametric mixtures for the former or completely random measures for the latter are typically favoured. In the former case, the densities are then written as

$$\boldsymbol{f}_P = \int_{\Theta} K_{\boldsymbol{\theta}}(x) \mathrm{d}P(\boldsymbol{\theta}),$$

where for each θ, K_θ is a density kernel and P is a probability measure which is assumed to be a realisation, under the prior, of a normalized completely random measure or some other random probability measure (abbreviated RPM). Models based on random probability measures will be examined in detail in Section 1.3.

Rates of convergence in the density model under RPM mixtures have been widely studied in the last decade, as presented in Section 1.2.1, the special case of DPM have been considered in particular by Ghosal and van der Vaart (2007) in the Gaussian univariate case, Shen et al. (2013) in the Gaussian multivariate using earlier results of Kruijer et al. (2010), the case of Pitman-Yor processes and of normalized inverse-Gaussian processes (introduced by Lijoi et al., 2005) by Scricciolo (2012). Consistency under the large class of Gibbs-type priors is studied by De Blasi et al. (2012).

RPM can also be used to model directly the observations, species sampling models is a typical example of this. The asymptotic properties of these models are studied by Jang et al. (2010).

Section 1.3.2 introduces dependent RPMs, which also are central to Chapter 3. Since the seminal paper by MacEachern (1999) and the introduction of Dependent Dirichlet process (DDP), many extensions were proposed in the Bayesian nonparametric literature of covariate dependent models (see references in Section 1.3.2). It is shown by Barrientos et al. (2012b) that the DDP has full weak support. A mean of comparison between the models is to study their asymptotic properties, a recent line of research initiated by Norets and Pelenis (2011). They obtain posterior consistency under kernel

stick-breaking processes (KSBP) for modelling the mixing probabilities in conditional distribution estimation. In the same problem, Pati et al. (2013) use dependent mixtures of Gaussian linear regressions.

We now turn to review some properties about the Dirichlet process and extensions. We will mainly focus on dependent extensions, which are the subject of Chapter 3 and Chapter 4. As an application, diversity measures arising in Bayesian nonparametric dependent models will be presented at the end of the section.

1.3 Random probability measures

We now turn to review some properties about the Dirichlet process and extensions. We will mainly focus on dependent extensions, which are the subject of Chapter 3 and Chapter 4. Diversity measures arising in species sampling models will be presented at the end of the section.

1.3.1 Discrete random probability measures

A step of the Bayesian nonparametric approach consists in constructing prior distributions on infinite dimensional spaces, like functions, or probability distributions. Random probability measures play a key role because their probability distribution precisely acts as priors for Bayesian inference as stated by the celebrated *de Finetti's representation theorem* (de Finetti, 1937). Suppose that data $(Y_1, \ldots, Y_n)$, sampled in the measurable space $(\mathbb{Y}, \mathcal{Y})$, are *exchangeable*, that is, for any $n \geq 1$, for any permutation σ of $\{1, 2, \ldots, n\}$, $(Y_1, \ldots, Y_n)$ and $(Y_{\sigma(1)}, \ldots, Y_{\sigma(n)})$ are equal in distribution. Then *de Finetti's representation theorem* states that there exists a probability measure Π on the space of probability distributions $\left(\mathbb{P}_{\mathbb{Y}}, \mathcal{P}_{\mathbb{Y}}\right)$ on $(\mathbb{Y}, \mathcal{Y})$ such that

$$\mathsf{P}\big(Y_1 \in A_1, \ldots, Y_n \in A_n\big) = \int_{\mathbb{P}_{\mathbb{Y}}} \prod_{i=1}^{n} p(A_i) \Pi(\mathrm{d}p). \tag{1.7}$$

The probability measure Π is called de Finetti's measure. Conditionally on p, it is clear from Equation (1.7) that the Y_i's are independent and identically distributed (i.i.d.) with common distribution p. Hence, the exchangeability assumption enables to write the following model

$$\begin{aligned} &Y_i \,|\, p \overset{\text{iid}}{\sim} p \quad \text{for } i \geq 1, \\ &p \sim \Pi, \end{aligned} \tag{1.8}$$

where Π acts as a prior distribution for Bayesian inference as it is the law of a random probability measure p.

Dirichlet process

The most celebrated example of prior Π is the Dirichlet process prior (DP), introduced by Ferguson (1973). A DP is a distribution on probability measures that can be defined as follows. Let $M > 0$ and G_0 be a probability measure on a space Θ. The law of the process can be written in a *constructive* way (see Sethuraman, 1994) as the law of the following random probability measure G, known as the stick-breaking representation:

$$G = \sum_{j=1}^{\infty} p_j \delta_{\theta_j}, \tag{1.9}$$

$$p_j = V_j \prod_{l<j} (1 - V_l), \text{ with } V_j \overset{\text{iid}}{\sim} Beta(1, M) \text{ and } \theta_j \overset{\text{iid}}{\sim} G_0, \tag{1.10}$$

mutually independently, where δ_{θ_j} stands for the Dirac point mass at θ_j. We write $G \sim \mathsf{DP}(M, G_0)$.

One of the most general use of the DP is as a mixing prior for the density problem, what is called Dirichlet process mixtures (DPM). For instance, in the context of Section 5.2, let some bivariate data as represented on the left part of Figure 1.1. A DPM prior can be used to estimate the density of the data, and is illustrated on Figure 1.1.

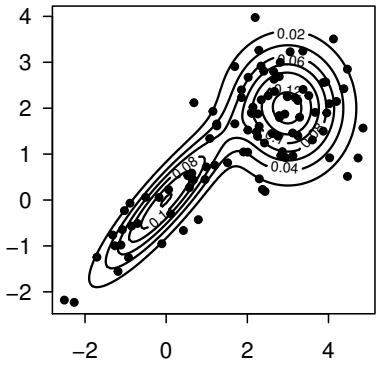

FIGURE 1.1: Data (100 observations) sampled in the distribution (5.2) (see Chapter 5) and contour of a Dirichlet process mixtures estimate.

PD and GEM distributions

Chapter 3 and Chapter 4, we are mainly dealing with different types of models, corresponding to probability measures on $\mathbb{N}$, the set of positive integers. Let $\mathcal{P}$ be the space

of probability measures on the positive integers:

$$\mathcal{P} = \{\mathbf{p} = (p_j)_j \, : \, p_j \geq 0, \sum_{j=1}^{\infty} p_j = 1\}.$$

Define the two following RPMs on $\mathcal{P}$:

Definition 1.5 (PD(M), Kingman, 1975)**.** *Let $\Gamma_{(1)} > \Gamma_{(2)} > \ldots$ be the points of a Poisson random measure on $(0, \infty)$ with mean measure $Mx^{-1}e^{-x}dx$. The distribution of the sequence $\mathbf{p} = (p_j)$ defined by*

$$p_j = \Gamma_{(j)}/\Sigma \quad \textit{where} \quad \Sigma = \sum_j \Gamma_{(j)},$$

is called the Poisson Dirichlet distribution with parameter M*, abbreviated* PD(M)*. It satisfies $p_1 > p_2 > \ldots$ and $\sum_j p_j = 1$ almost surely.*

Definition 1.6 (GEM(M), Ewens, 1990)**.** *The distribution of a sequence $\mathbf{p} = (p_j)$ which can be written as*

$$p_1 = V_1, \; p_j = V_j \prod_{l<j} (1 - V_l), \quad j \geq 2,$$

where the V_j are i.i.d. variables with $Be(1, M)$ distribution $\mathsf{P}(V_j \in dx) = M(1-x)^{M-1}\,dx$, is called the GEM($M$) *distribution, after Griffiths-Engen-McCloskey.*

The weights of a Dirichlet process DP(M, G_0) have the GEM(M) distribution, which does not depend on G_0. The history and background of the GEM distribution is detailed in Chapter 41 of Johnson et al. (1997).

Remark 1.1. Pitman (2006) generalized both distributions PD(M) and GEM(M) to two-parameter distributions PD(α, M) and GEM(α, M). The associated RPM on a generic ambient space with base measure G_0 is then called the *Pitman-Yor process*, abbreviated PY(α, M, G_0).

Size-biased permutations

Let $\mathbf{p} = (p_1, p_2, \ldots)$ be a probability in $\mathcal{P}$. A size-biased permutation (SBP) 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 size, $\mathsf{P}(\sigma_1 = j) = p_j$; the subsequent indices appear with a probability proportional to their size in the remaining indices, *i.e.* for k distinct integers $j_1, \ldots, j_k$,

$$\mathsf{P}(\sigma_k = j_k | \sigma_1 = j_1, \ldots, \sigma_{k-1} = j_{k-1}) = \frac{p_{j_k}}{1 - p_{j_1} - \ldots - p_{j_{k-1}}}. \tag{1.11}$$

The following theorem sheds light on the close link between GEM and PD distributions:

Theorem 1.7 (McCloskey, 1965)**.** *Let* $\tilde{\mathbf{p}}$ *be a size-biased permutation of a sequence of random variables* $\mathbf{p}$ *such that* $p_1 > p_2 > \ldots > 0$ *with* $\sum_{j=1}^{\infty} p_j = 1$*. Then*

$$\tilde{p}_j = V_j \prod_{l<j} (1 - V_l),$$

for a sequence of i.i.d. random variables (V_j) *if and only if* $\mathbf{p}$ *has* $\mathsf{PD}(M)$ *distribution for some* $M > 0$*. Then the common distribution of the* V_j *is* $Be(1, M)$*, and* $\tilde{\mathbf{p}}$ *has the* $\mathsf{GEM}(M)$ *distribution.*

The correspondence described in the latter theorem can be illustrated by the following diagram between both distributions:

$$\mathsf{GEM}(M) \underset{\mathsf{SBP}}{\overset{\mathsf{Rank}}{\rightleftharpoons}} \mathsf{PD}(M).$$

As a consequence, the GEM distribution is invariant under size-biased permutations: spelled out, this means that if $\mathbf{p}$ has the GEM distribution, then a size-biased permutation of $\mathbf{p}$ also has the GEM distribution.

Species population data

Ewens (1990) used the GEM distribution in the fields of genetics and ecology. An early occurrence of species sampling model was proposed by Fisher et al. (1943). This example, along with Engen (1978), shows that the discreteness of the GEM and PD distributions is appreciated in models in ecology. It allows the handling of the problem of species sampling. Pitman (1996) describes the following sampling. Suppose that a sample $Y_1, Y_2, \ldots$ is drawn from a large population where individuals are divided into species. The Y_k represent the species of the kth individual sampled. The species are actually labelled by tags, in an arbitrary space. This tagging allows transforming the *random partition* of n individuals into species, into the *random sequence* $Y_1, Y_2, \ldots$. Such a sampling process will be studied in Chapter 3 and Chapter 4. From now on we will describe some distributional properties of this sampling process which will be used in these chapters.

Ewens formula and EPPF

The partition structure of a sample $\mathbf{Y}^{(n)} = (Y_1, Y_2, \ldots, Y_n)$ obtained as described above can be characterized in two ways when the labels are unimportant. The sample induces

a partition $\mathbf{N}_n = (n_1, n_2, \ldots, n_k)$ of the n first integers $[n] = \{1, 2, \ldots, n\}$ into $K_n = k$ species, where the first one that appeared in the sampling process appeared n_1 times, the second one appeared n_2 times, etc. Another representation of the partition structure of the sample with no mention of the labels is $\mathbf{A}_n = (a_1, a_2, \ldots, a_n)$, where a_k counts the number of species which appeared k times in the sample of size n. So $\mathbf{N}_n$ and $\mathbf{A}_n$ satisfy

$$\sum_{j=1}^{k} = n, \quad \sum_{j=1}^{n} j a_j = n.$$

The distributions of $\mathbf{N}_n$ and $\mathbf{A}_n$ under the $\mathsf{GEM}(M)$ distribution for the relative frequencies of species are known as the exchangeable partition probability function (EPPF) and the Ewens formula (see Ewens, 1972, Antoniak, 1974). Exchangeable partition probability functions are a tool for studying clustering in Bayesian analysis. For an account about random partitions, see for instance Pitman (2006).

Definition 1.8. *When* $\mathbf{p} \sim \mathsf{GEM}(M)$*, the exchangeable partition probability function of* $\mathbf{Y}^{(n)}$ *is:*

$$\mathsf{P}(\mathbf{N}_n = (n_1, n_2, \ldots, n_k)) = p(n_1, n_2, \ldots, n_k) = \frac{M^k}{M_{(n)}} \prod_{j=1}^{k} (n_j - 1)!, \tag{1.12}$$

where $M_{(n)} = M(M+1)\ldots(M+n-1)$*. Ewens sampling formula is*

$$\mathsf{P}(\mathbf{A}_n = (a_1, a_2, \ldots, a_n)) = \frac{n!}{M_{(n)}} \prod_{j=1}^{n} \left(\frac{M}{j}\right)^{a_j} \frac{1}{a_j!} \tag{1.13}$$

The EPPF (1.12) can be recovered by Ewens formula (1.13) by enumeration and expectation of all possible configurations of labels:

$$p(n_1, n_2, \ldots, n_k) = \mathsf{E}\left(\sum_{(*)} p_{i_1}^{n_1} \ldots p_{i_k}^{n_k} \right),$$

where the sum $(*)$ runs over all distinct $i_1, \ldots, i_k$. Corollary 7 in Pitman (1995) provides a more straightforward representation of $p(n_1, n_2, \ldots, n_k)$ in terms of $\mathbf{p}$ as follows:

$$p(n_1, n_2, \ldots, n_k) = \mathsf{E}\left[\left(\prod_{i=1}^{k} p_i^{n_i - 1} \right) \prod_{i=1}^{k-1} \left(1 - \sum_{j=1}^{i} p_j\right) \right].$$

By using the stick-breaking representation of Definition 1.6 and the relation $\sum_{l<j} p_l + \prod_{l<j}(1 - V_l) = 1$, one obtains:

$$p(n_1, n_2, \ldots, n_k) = \mathsf{E}\left[\prod_{i=1}^{k} V_i^{n_i - 1} (1 - V_i)^{n_{i+1} + \cdots + n_k} \right], \tag{1.14}$$

and by independence of the V_i and computation of moments derived from the beta distributions in Equation (6.1) in Appendix:

$$p(n_1, n_2, \ldots, n_k) = \prod_{i=1}^{k} \frac{1^{(n_i-1)} M^{n_{i+1}+\cdots+n_k}}{(M+1)^{n_i+\cdots+n_k-1}},$$

which coincides with the EPPF formula (1.12).

Section 1.3.2 introduces dependent random probability measures. In this setting also the EPPF is of interest, since it can be a tool for defining posterior sampling schemes We define such a joint EPPF for a dependent GEM distribution in Chapter 3.

Note that $p(n_1, n_2, \ldots, n_k)$ is a symmetric function, *i.e.* $p(n_{\sigma(1)}, n_{\sigma(2)}, \ldots, n_{\sigma(k)}) = p(n_1, n_2, \ldots, n_k)$ for any permutation σ of $\{1, 2, \ldots, k\}$. This illustrates the exchangeability of the sequence $(Y_1, Y_2, \ldots, Y_n)$. The formulae of Definition 1.8 are given for the two-parameter Poisson-Dirichlet distribution $\mathsf{PD}(\alpha, M)$ by Pitman (1992).

A recent extension of the EPPF is provided by Broderick et al. (2013). Instead of corresponding to a single cluster or species, each data point is allowed to belong to an arbitrary number of groups. The group is called a feature, or a topic. The paper provides an extension of the EPPF for this model called the exchangeable feature probability functions (EFPF).

Pólya urn and Chinese Restaurant process

Now we turn to define the Pólya urn (PU) and the Chinese Restaurant process (CRP) which can be seen as predictive rules. They are of interest here since some of the random probability measures which will be reviewed later are defined by their predictive rule.

In the generic exchangeable model (1.8), define the *predictive rule* of the random probability measure Π as the distribution of the first element Y_1 in the sample, and then, for $n \geq 1$, the distribution of Y_{n+1} conditional to the observed sample $(Y_1, Y_2, \ldots, Y_n)$.

Definition 1.9. *The predictive rule of the Dirichlet process given in Equation (1.9) with a sample $(Y_1, Y_2, \ldots, Y_{n+1})$ is called the* Blackwell–MacQueen Urn scheme*, and has the following form*

$$\begin{aligned} &\mathsf{P}(Y_1 \in \cdot\,) = \nu(\cdot\,), \quad \textit{and for } n \geq 1, \\ &\mathsf{P}(Y_{n+1} \in \cdot\,|Y_1, Y_2, \ldots, Y_n) = \frac{M}{M+n} G_0(\cdot\,) + \frac{1}{M+n} \sum_{j=1}^{n} \delta_{Y_j}(\cdot\,). \end{aligned} \tag{1.15}$$

This predictive rule is also called the Pólya Urn *(abbreviated* PU*), or* Hoppe's Urn*, from the point of view of a process on colors. It is obtained as follows, which is equivalent to (1.15). Consider an urn which initially contains a black ball of mass M. Ball will be successively drawn with probabilities proportional to their masses: if a black ball is drawn, it is return with a ball of a new color of mass 1; if a coloured ball is drawn, it is returned with a ball of the same color of mass 1.*

Note that there exist equivalent ways to present the same sampling process:

Definition 1.10. *The* Chinese Restaurant process *(abbreviated* CRP*) is a discrete-time partition valued process on partitions of the first n integers $[n] = \{1, 2, \ldots, n\}$ at time n. It has a positive parameter M. Its probability distribution is popularly defined with the restaurant analogy. At time 1, the first customer sits at table 1 with probability 1,* i.e. $\mathsf{P}(\{1\}) = 1$. *At time $n+1$, the $(n+1)$th customer sits*

- *at an occupied table j with probability proportional to the number of sitting customers,*
- *at a new table $k+1$ with probability proportional to M.*

In the Pólya urn, if instead of picking a new color one picks a random value in a diffuse base distribution G_0, the resulting distribution over labels is the same as the distribution over draws from a Dirichlet process.

The Chinese Restaurant process and the Pólya urn define the same process with a slightly different point of view: the first one focuses on partitions while the second one focuses on the predictive of a new element of the process. Consider the marginal distribution of the Y_i's in the following exchangeable model

$$\begin{aligned} Y_i \,|\, G &\overset{\text{iid}}{\sim} G \quad \text{for } i \geq 1, \\ G &\sim \mathsf{DP}, \end{aligned}$$

which boils down to marginalizing out thee DP distribution. Then the Pólya urn is the conditional distribution of any of the Y_i's given all others.

Note also that the link between EPPF and the predictive probability function in species sampling models is studied by Lee et al. (2013).

1.3.2 Dependent random probability measures

Up to now, we have reviewed problems where a single distribution is assigned a non-parametric prior distribution. In many applications, it is desirable to model a collection

of distributions, $\mathcal{G} = \{G_X, X \in \mathcal{X}\}$, where X is a covariate of a space $\mathcal{X}$, such as a treatment, time, a spatial coordinate, etc. Two extreme solutions consist in

1. assume a single distribution everywhere, $G_X = G$ for all X: this is restrictive as it does not allow for variations for varying predictors.
2. assume independence across X: this prevents from sharing common components between X's.

An extension proposed by MacEachern (1999) as the Dependent Dirichlet process, DDP, allows the weights p_j and/or the clusters θ_j to vary with a predictor X, according to stochastic processes $p_j(X)$ and $\theta_j(X)$:

$$G_{X_1} = \sum_{j=1}^{\infty} p_j(X)\delta_{\theta_j(X)},$$

where $p_j(X) = V_j(X)\prod_{l<j}(1 - V_l(X))$ with $V_j(X) \overset{\text{iid}}{\sim} Beta(1, M)$ and $\theta_j(X) \overset{\text{iid}}{\sim} G_0$ mutually independently. For any fixed X, one recovers Equations(1.10), so this yields a DP distribution for G_X.

The DDP is an extension of the DP, in the case of more structured data, when the usual exchangeability assumption does not hold. Previous reference to predictor-dependent DP models include Cifarelli and Regazzini (1978) and Muliere and Petrone (1993), where the centring measure of an independent collection of DP is based on a regression. Incorporating the dependence in the base measure of Dirichlet processes however limits the flexibility of the model to capturing the structure of dependence of the regression. In the HDP extension, Teh et al. (2006) propose a hierarchical model in which the base measure is itself a draw from a DP. Since draws from a DP are almost surely discrete, it forces a sharing of clusters between the different probability measures.

There has been increasing interest since MacEachern (1999) in the construction of predictor dependent probability measures. See the chapter Dunson (2010) for a general review of the methods, with a focus on biostatistics applications. The case with fixed weights $p_j(X) = p_j$ (called single-p) was implemented in a wide range of applications. To name but a few, De Iorio et al. (2004) use a DDP in the case of categorical predictors, which allows defining an ANOVA model for unknown densities. It is used in spatial applications by Gelfand et al. (2005), Duan et al. (2007) in dynamic models by Caron et al. (2006), in variable selection by Chung and Dunson (2009b) and in testing settings by Dunson and Peddada (2008). Extensions with varying weights include order-based DDP, or πDDP (Griffin and Steel, 2006), local DP, or lDP (Chung and Dunson, 2009a),

weighted mixtures of DP (Dunson and Park, 2008), and kernel stick-breaking processes, or KSBP (Dunson et al., 2007).

In a recent work, Wade et al. (2013) define a DDP with a particular focus on the random partitions it creates. The focus is on the regression problem in dimension 1, tackled with Dirichlet process mixtures. They impose an order constraint that if two subjects with covariates X and X' are clustered together then all subjects whose covariates are between X and X' are in the same cluster. This constraint is effective only in dimension 1, hence limits this approach to univariate covariates X.

The types of dependent priors used in the literature are diverse. For instance, dependent Bernstein polynomials are used by Barrientos et al. (2012a) in the regression problem. Completely random measures are used by Lijoi et al. (2013) which focus on clustering properties, and by Chen et al. (2013), Barrientos et al. (2012a). Dependence is introduced via Gaussian processes by Williamson et al. (2010), Palla et al. (2013). This is also the tool which we use in Chapter 3 for defining dependent GEM distributions, although from a different angle.

Measuring dependence under the DDP

Now that we have introduced dependent random probability measures, we turn to describe simple measures of dependence that arise from these processes, and allow a better understanding of these. The results are given in the case of the DDP, and constitute an introduction for the results that are provided in Chapter 3 for the dependent GEM process.

First, denote by $c_M = c_M(|X_1 - X_2|)$ the dependence factor between the process at two covariate points X_1 and X_2 defined by:

$$c_M(|X_1 - X_2|) = (M+1)^2 \mathsf{E}(V(X_1)V(X_2)),$$

We identify two extreme cases denoted by:

- (I): *independence*, $V(X_1) \perp\!\!\!\perp V(X_2)$ (*eg* $|X_1 - X_2| \to \infty$), then $c_M = 1$.
- (E): *equality*, $X_1 = X_2$, *i.e.* $V(X_1) = V(X_2)$ in distribution, then $c_M = 2(M+1)/(M+2) = 1 + M/(M+2)$,

Suppose that $G_X \sim \mathsf{DDP}(MG_0)$, and let A be a measurable subset of $\mathbb{Y}$. A definition of the DP, equivalent to (1.9), given by Ferguson (1973), entails that $G_X(A) \sim$

$\mathsf{Beta}(MG_0(A), MG_0(A^c))$, and thus

$$\begin{aligned}\mathsf{E}(G_{X_1}(A)) &= \frac{MG_0(A)}{MG_0(A) + MG_0(A^c)} = G_0(A), \\ \mathsf{Var}(G_{X_1}(A)) &= \frac{1}{M+1} G_0(A)(1 - G_0(A)).\end{aligned} \tag{1.16}$$

The covariance is given by

$$\begin{aligned}&\mathsf{Cov}(G_{X_1}(A), G_{X_2}(A)) = C_V(X_1, X_2)\big(\mathsf{P}(\theta(X_1) \in A \cap \theta(X_2) \in A) - G_0(A)^2\big), \\ &\text{where } C_V(X_1, X_2) = \frac{\mathsf{E}(V(X_1)V(X_2))}{\mathsf{E}(V(X_1)) + \mathsf{E}(V(X_2)) - \mathsf{E}(V(X_1)V(X_2))}.\end{aligned}$$

By independence between the weights and the clusters in a DP, the structure of the covariance between $G_{X_1}(A)$ and $G_{X_2}(A)$ appears to be separated in two parts, with the weights in the one hand with the term $C_V(X_1, X_2)$, and with the clusters in the other hand. Equation (1.16) shows that the process on the clusters, $\theta(X)$, plays a major role in the dependence structure since $\mathsf{Cov}(G_{X_1}(A), G_{X_2}(A))$ can vanish only if there is independence between $\theta(X_1)$ and $\theta(X_2)$. The process on the weights intervene as a multiplicative constant: $C_V(X_1, X_2)$ is maximum when $X_1 = X_2$, equal to $1/(M+1)$, and is minimum when there is independence between $V(X_1)$ and $V(X_2)$ (which can occur for instance when the distance between X_1 and X_2 goes to ∞), and then its value tends to $1/(2M+1)$.

In the case of a single–θ DDP, the covariance between $G_{X_1}(A)$ and $G_{X_2}(A)$ can be written by

$$\begin{aligned}\mathsf{Cov}(G_{X_1}(A), G_{X_2}(A)) &= C_V(X_1, X_2)\big(G_0(A) - G_0(A)^2\big), \\ &= (M+1)C_V(X_1, X_2)\big(\mathsf{Var}(G_{X_1}(A))\mathsf{Var}(G_{X_2}(A))\big)^{1/2},\end{aligned}$$

hence the correlation between $G_{X_1}(A)$ and $G_{X_2}(A)$ will not depend on A, and we can define this quantity as the correlation between G_{X_1} and G_{X_2}, which is

$$\mathsf{Corr}(G_{X_1}, G_{X_2}) = (M+1)C_V(X_1, X_2).$$

Also, denote by $Y_1|G_{X_1} \sim G_{X_1}$ and $Y_2|G_{X_2} \sim G_{X_2}$, $X_1 \neq X_2$ two conditionally independent draws in the Dependent Dirichlet process $G_{\mathcal{X}}$. Then

$$\begin{aligned}\mathsf{E}(Y_1) &= \mathsf{E}(\theta(X_1)) \\ \mathsf{Var}(Y_1) &= \frac{1}{M+1}\mathsf{Var}(\theta(X_1)) \\ \mathsf{Cov}(Y_1, Y_2) &= \frac{1}{M+1}\mathsf{Cov}(\theta(X_1), \theta(X_2)).\end{aligned}$$

Note that one can link the dependence at the level of G_X and the dependence at the level of draws $Y_X \,|\, G_X \sim G_X$ by the notion of α–dependence defined as follows:

Definition 1.11 (Bradley et al., 1986)**.** *Let two random variables Y_1 and Y_2 on the measurable space $\mathbb{Y}$, and let A be a measurable subset of $\mathbb{Y}$. Then α_A–dependence between Y_1 and Y_2 is defined by*

$$\alpha_A(Y_1, Y_2) = \mathsf{P}(Y_1 \in A, Y_2 \in A) - \mathsf{P}(Y_1 \in A)\mathsf{P}(Y_2 \in A).$$

Then we have the following corollary

Corollary 1.12. *We have*

$$\alpha_A(Y_1, Y_2) = \mathsf{Cov}(G_{X_1}(A), G_{X_2}(A)) = C_V(X_1, X_2)\ \alpha_A(\theta_{X_1}, \theta_{X_2}). \tag{1.17}$$

Proof. $P(Y_1 \in A) = E(P(Y_1 \in A|G_{X_1})) = E(G_{X_1}(A)) = G_0(A)$, and $P(Y_1 \in A, Y_2 \in A) = E(P(Y_1 \in A, Y_2 \in A|G_{X_1}, G_{X_2})) = E(G_{X_1}(A)G_{X_2}(A))$ by conditional independence of Y_1 and Y_2. □

This means that the covariance can be interpreted at the level of draws from the priors.

The Dirichlet process has the simple Pólya urn predictive rule, as described in Definition 1.9. The additional layer due to the predictor-dependent feature in dependent models makes the prediction more involved. In several single-θ processes, predictive rules can be obtained by marginalising out the process $G_{\mathcal{X}}$. The following theorem by Dunson and Park (2008) holds for the particular case of the kernel stick-breaking process, when the clusters $\theta_j(X)$ are fixed through X (*i.e.* single–θ case), when the base measure is diffuse and when the weights are stationary (*i.e.* their distribution is fixed over X):

Theorem 1.13 (Dunson and Park, 2008)**.** *Let $G_{\mathcal{X}}$ be a kernel stick-breaking process in the above special case, and $Y_i \,|\, X_i \overset{ind}{\sim} G_{X_i}$. The predictive rule of the* KSBP *is given by*

$$\mathsf{P}(Y_i \in .\,|Y_1, \ldots, Y_{i-1}, X_1, \ldots, X_i) = \pi_0 G_0(.) + \sum_{j=1}^{i-1} \pi_j \delta_{Y_j}(.), \tag{1.18}$$

where the probability weights π_j are defined as follows: let $\mathcal{N}_i^{(r,s)}$ (resp. $\mathcal{N}_{i,j}^{(r,s)}$) be the set of all r-dimensional subsets of $\{1, \ldots, s\}$ including i (resp. including i and j). Define $\mu_{\mathcal{I}} = \mathsf{E}\left(\prod_{i \in \mathcal{I}} V_{X_i}\right)$ and $\omega_{\mathcal{I}} = \frac{\mu_{\mathcal{I}}}{\sum_{t=1}^{|\mathcal{I}|}(-1)^{t-1} \sum_{\mathcal{J} \in \mathcal{I}_t} \mu_{\mathcal{J}}}$, where $\mathcal{I}_t$ is the set of all t-element

subsets of $\mathcal{I}$. *Then*

$$\pi_0 = 1 - \sum_{r=2}^{i}(-1)^r \sum_{\mathcal{I}\in\mathcal{N}_i^{(r,s)}} \omega_{\mathcal{I}}, \quad \pi_j = \sum_{r=2}^{i}(-1)^r \sum_{\mathcal{I}\in\mathcal{N}_{i,j}^{(r,s)}} \frac{\omega_{\mathcal{I}}}{r-1} \text{ for } j \geq 1.$$

In the case when there is a unique covariate value $X_i = X$, then the prediction rule (1.18) reduces to the standard Pólya urn of Equation (1.15) with $\pi_j = \frac{1}{M+i-1}$ for $j \geq 1$.

Note that the results hold for stationary Dependent Dirichlet process $G_X \sim \mathsf{DDP}(MG_0)$ and can be generalized to the non stationary case $G_X \sim \mathsf{DDP}(M_X G_{0,X})$. Such an extension can better accommodate for changes in the data structure.

Measuring dependence under the GEM

Chapter 3 defines a dependent version of the GEM distribution called the Dep – GEM. Specific properties of this prior, in terms of dependence and predictive rule, are presented in Section 3.5 therein.

1.3.3 Diversity indices

We present here a Bayesian nonparametric approach to the study of species diversity. It is based on the choice of a random probability measure as a prior distribution for the unknown relative abundance frequencies of species, and is developed in Chapter 3 by using a dependent GEM distribution. Applications to the field of ecotoxicology are given in Chapter 4.

Diversity in populations that are classified into groups were studied in the seminal papers Fisher et al. (1943), Simpson (1949). See Pielou (1975) for an account in ecology. Diversity was first modelled in a Bayesian framework by Gill and Joanes (1979), and in a Bayesian nonparametric framework in Lijoi et al. (2007).

Data classified into groups are common in many fields, for instance ecological data (species are microbes), biological data, population genetics (species are alleles). Several indices can account for a measure of diversity between species. One of such is the Shannon index defined by

$$H_{\mathsf{Shan}}(\mathbf{p}) = -\sum_j p_j \log p_j,$$

where the observations are species indexed by positive integers j with respective probabilities p_j. Many other diversity indices can be thought of, such as the Simpson index

$$H_{\mathsf{Simp}}(\mathbf{p}) = 1 - \sum_j p_j^2, \tag{1.19}$$

and the generalized diversity index proposed by Good (1953) in the form of

$$H_{\mathsf{Good},\alpha,\beta}(\mathbf{p}) = -\sum_j p_j^\alpha \log^\beta p_j, \tag{1.20}$$

for non-negative integer values of α and β. It includes both the Shannon index $H_{\mathsf{Good},1,1}$ and the Simpson index $H_{\mathsf{Good},2,0} + 1$. It was further extended to values for (α, β) in the real plane, where only several regions lead to sensible indices (see *eg* Baczkowski et al., 1998) that satisfy the following basic properties (stated *eg* in Pielou, 1975):

- for fixed J, the index increases as the relative abundances become more equal,
- for equal relative abundances, the index is an increasing function of J.

Figure 3.1 illustrates three diversity indices on the real data set studied in Chapter 3.

We present now the prior diversity induced by a dependent GEM prior. In Chapter 3, we also give a joint distribution of the prior diversity under this prior. To this purpose, some of the material introduced before is needed, such as size-biased permutations of a sample defined in Equation (1.11). We denote a size-biased permutation of an infinite vector of probabilities $\mathbf{p} = (p_1, p_2, \ldots)$ by $\tilde{\mathbf{p}} = (\tilde{p}_1, \tilde{p}_2, \ldots)$, whose first element $\tilde{p}_1$ is called the *size-biased pick*. The following result is proved in Equation (2.23) of Pitman (2006)

$$\mathsf{E}\Big(\sum f(p_j)\Big) = \mathsf{E}\Big(\sum f(\tilde{p}_j)\Big) = \mathsf{E}\left(\frac{f(\tilde{p}_1)}{\tilde{p}_1}\right). \tag{1.21}$$

Hence the distribution of the size-biased pick $\tilde{p}_1$ encodes much information about $\mathbf{p}$. It is sufficient in order to compute the expectation of any additive transform of the form $\sum f(p_j)$, for example the generalized diversity index given in Equation (1.20). In the case of a $\mathsf{GEM}(M)$ prior on $\mathbf{p}$, the prior expectation of Simpson diversity is found by Cerquetti (2012)

$$\mathsf{E}(H_{\mathsf{Simp}}) = \frac{M}{1+M}. \tag{1.22}$$

The result for the Shannon diversity index is given in an unpublished work by Cerquetti

$$\mathsf{E}(H_{\mathsf{Shan}}) = \psi(M+1) - \psi(1), \tag{1.23}$$

where ψ is the digamma function, *i.e.* the derivative of the log of the gamma function. The prior expectation of the diversity in both cases is an increasing function of the

precision parameters M: it vanishes when M goes to 0, and is maximum when M goes to ∞, as illustrated in Figure 1.2.

Extended properties of the diversity under the dependent prior $\mathsf{Dep}-\mathsf{GEM}$ are presented in Section 3.5.

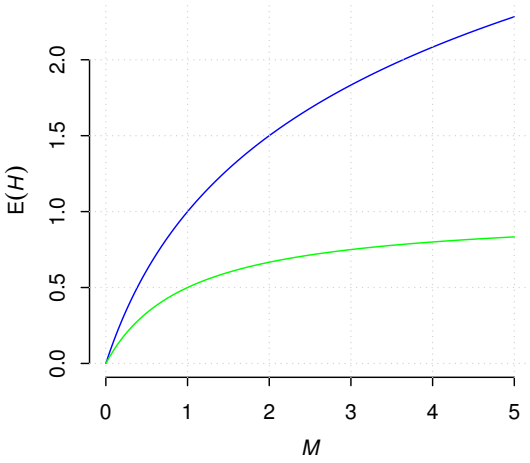

FIGURE 1.2: Prior expectation of the Simpson index (1.22) (*in green*) and Shannon index (1.23) (*in blue*) under GEM distribution.

Bibliography

Antoniak, C. E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The annals of statistics*, pages 1152–1174. 27

Arbel, J., Gayraud, G., and Rousseau, J. (2013). Bayesian optimal adaptive estimation using a sieve prior. *Scandinavian Journal of Statistics*. 21

Baczkowski, A., Joanes, D., and Shamia, G. (1998). Range of validity of α and β for a generalized diversity index $H(\alpha, \beta)$ due to Good. *Mathematical biosciences*, 148(2):115–128. 35

Barrientos, A., Jara, A., and Quintana, F. A. (2012a). Fully nonparametric regression for bounded data using dependent Bernstein polynomials. *Manuscript under preparation*. 31

Barrientos, A. F., Jara, A., and Quintana, F. A. (2012b). On the support of MacEachern's dependent Dirichlet processes and extensions. *Bayesian Analysis*, 7(2):277–310. 22

Barron, A. R. (1988). The exponential convergence of posterior probabilities with implications for Bayes estimators of density functions. Technical report, University of Illinois at Urbana-Campaign. 16, 17

Belitser, E. and Ghosal, S. (2003). Adaptive Bayesian inference on the mean of an infinite-dimensional normal distribution. *Ann. Statist.*, 31(2):536–559. 20

Bernardo, J. M. and Smith, A. F. (2009). *Bayesian theory*, volume 405. Wiley. 14

Bhattacharya, A., Pati, D., and Dunson, D. B. (2013). Anisotropic function estimation using multi-bandwidth Gaussian processes. Technical report, Duke University. 20

Bradley, R., Dehling, H., Doukhan, P., and Neumann, M. H. (1986). Dependence in probability and statistics. *Eberlin Tassu*, pages 165–192. 33

Broderick, T., Pitman, J., and Jordan, M. I. (2013). Feature allocations, probability functions, and paintboxes. *arXiv preprint arXiv:1301.6647*. 28

Caron, F., Davy, M., Doucet, A., Duflos, E., and Vanheeghe, P. (2006). Bayesian inference for dynamic models with Dirichlet process mixtures. In *Proc. International Conference on Information Fusion*. Citeseer. 30

Cerquetti, A. (2012). Bayesian nonparametric estimation of Simpson's evenness index under $\alpha-$Gibbs priors. *arXiv preprint arXiv:1203.1666*. 35

Chen, C., Rao, V., and Buntine, W. (2013). Dependent Normalized Random Measures. *Manuscript under preparation*. 31

Chung, Y. and Dunson, D. (2009a). The local Dirichlet process. *Annals of the Institute of Statistical Mathematics*, pages 1–22. 30

Chung, Y. and Dunson, D. B. (2009b). Nonparametric Bayes conditional distribution modeling with variable selection. *Journal of the American Statistical Association*, 104(488). 30

Cifarelli, D. and Regazzini, E. (1978). Problemi statistici non parametrici in condizioni di scambialbilita parziale e impiego di medie associative. *Tech. rep.* 30

De Blasi, P., Lijoi, A., and Prünster, I. (2012). An asymptotic analysis of a class of discrete nonparametric priors. *Statistica Sinica*. 22

de Finetti, B. (1937). La prévision: ses lois logiques, ses sources subjectives. In *Annales de l'institut Henri Poincaré*, volume 7, pages 1–68. Presses universitaires de France. 23

De Iorio, M., Mueller, P., Rosner, G., and MacEachern, S. (2004). An ANOVA model for dependent random measures. *Journal of the American Statistical Association*, 99(465):205–215. 30

de Jonge, R. and van Zanten, J. (2010). Adaptive nonparametric Bayesian inference using location-scale mixture priors. *Ann. Statist.*, 38(6):3300–3320. 20

Diaconis, P. and Freedman, D. (1986). On the consistency of Bayes estimates. *The Annals of Statistics*, pages 1–26. 16

Doob, J. L. (1949). Application of the theory of martingales. *Le calcul des probabilites et ses applications*, pages 23–27. 16

Duan, J., Guindani, M., and Gelfand, A. (2007). Generalized spatial Dirichlet process models. *Biometrika*, 94(4):809. 30

Dunson, D. and Park, J. (2008). Kernel stick-breaking processes. *Biometrika*, 95(2):307. 31, 33

Dunson, D., Pillai, N., and Park, J. (2007). Bayesian density regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(2):163–183. 31

Dunson, D. B. (2010). Nonparametric Bayes applications to biostatistics. *In Hjort et al. (2010)*, 28:223–273. 30

Dunson, D. B. and Peddada, S. D. (2008). Bayesian nonparametric inference on stochastic ordering. *Biometrika*, 95(4):859–874. 30

Engen, S. (1978). *Stochastic abundance models, with emphasis on biological communities and species diversity*. Chapman and Hall Ltd. 26

Ewens, W. J. (1972). The sampling theory of selectively neutral alleles. *Theoretical population biology*, 3(1):87–112. 27

Ewens, W. J. (1990). Population genetics theory-the past and the future. In *Mathematical and statistical developments of evolutionary theory*, pages 177–227. Springer. 25, 26

Ferguson, T. (1973). A Bayesian analysis of some nonparametric problems. *The annals of statistics*, 1(2):209–230. 24, 31

Fisher, R. A., Corbet, A. S., and Williams, C. B. (1943). The relation between the number of species and the number of individuals in a random sample of an animal population. *The Journal of Animal Ecology*, pages 42–58. 26, 34

Gelfand, A., Kottas, A., and MacEachern, S. (2005). Bayesian nonparametric spatial modeling with Dirichlet process mixing. *Journal of the American Statistical Association*, 100(471):1021–1035. 30

Ghosal, S. (2001). Convergence rates for density estimation with Bernstein polynomials . *Annals of Statistics*, 29:1264–1280. 20

Ghosal, S., Ghosh, J. K., and van der Vaart, A. W. (2000). Convergence rates of posterior distributions. *Ann. Statist.*, 28(2):500–531. 17, 18, 19

Ghosal, S., Lember, J., and van der Vaart, A. W. (2008). Nonparametric Bayesian model selection and averaging. *Electron. J. Stat.*, 2:63–89. 20

Ghosal, S. and van der Vaart, A. W. (2001). Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities. *The Annals of Statistics*, 29(5):1233–1263. 20

Ghosal, S. and van der Vaart, A. W. (2007). Convergence rates of posterior distributions for noniid observations. *Ann. Statist.*, 35(1):697–723. 17, 18, 20

Ghosal, S. and van der Vaart, A. W. (2007). Posterior convergence rates of Dirichlet mixtures at smooth densities. *The Annals of Statistics*, 35(2):697–723. 20, 22

Ghosh, J. K. and Ramamoorthi, R. (2003). *Bayesian nonparametrics*. Springer. 14, 15

Gill, C. A. and Joanes, D. N. (1979). Bayesian estimation of Shannon's index of diversity. *Biometrika*, 66(1):81–85. 34

Good, I. J. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika*, 40(3-4):237–264. 35

Griffin, J. and Steel, M. (2006). Order-based dependent Dirichlet processes. *Journal of the American statistical Association*, 101(473):179–194. 30

Hjort, N. L., Holmes, C., Müller, P., and Walker, S. G. (2010). *Bayesian nonparametrics*, volume 28. Cambridge University Press. 14, 17

Hoffmann, M., Rousseau, J., and Schmidt-Hieber, J. (2013). On adaptive posterior concentration rates. *arXiv preprint arXiv:1305.5270*. 20, 22

Huang, T. (2004). Convergence rates for posterior distributions and adaptive estimation. *Ann. Statist.*, 32(4):1556–1593. 20

Jang, G. H., Lee, J., and Lee, S. (2010). Posterior consistency of species sampling priors. *Statistica Sinica*, 20(2):581. 22

Johnson, N. L., Kotz, S., and Balakrishnan, N. (1997). *Discrete multivariate distributions*, volume 165. Wiley New York. 25

Kingman, J. F. (1975). Random discrete distributions. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–22. 25

Kruijer, W., Rousseau, J., and van der Vaart, A. W. (2010). Adaptive Bayesian density estimation with location-scale mixtures. *Electron. J. Stat.*, 4:1225–1257. 20, 22

Kruijer, W. and Van der Vaart, A. (2008). Posterior convergence rates for Dirichlet mixtures of beta densities. *Journal of Statistical Planning and Inference*, 138(7):1981–1992. 20

Lee, J., Quintana, F. A., Müller, P., and Trippa, L. (2013). Defining predictive probability functions for species sampling models. *Statistical science*, 28(2):209–222. 29

Lijoi, A., Mena, R. H., and Prünster, I. (2005). Hierarchical mixture modeling with normalized inverse-Gaussian priors. *Journal of the American Statistical Association*, 100(472):1278–1291. 22

Lijoi, A., Mena, R. H., and Prünster, I. (2007). Bayesian nonparametric estimation of the probability of discovering new species. *Biometrika*, 94(4):769–786. 34

Lijoi, A., Nipoti, B., and Prünster, I. (2013). Bayesian inference with dependent normalized completely random measures. *Bernoulli*, to appear. 31

MacEachern, S. (1999). Dependent nonparametric processes. *ASA Proceedings of the Section on Bayesian Statistical Science*, pages 50–55. 22, 30

McCloskey, J. W. (1965). *A model for the distribution of individuals by species in an environment*. PhD thesis, Michigan State University. Department of Statistics. 26

McVinish, R., Rousseau, J., and Mengersen, K. (2009). Bayesian goodness of fit testing with mixtures of triangular distributions. *Scand. J. Stat.*, 36(2):337–354. 20

Muliere, P. and Petrone, S. (1993). A Bayesian predictive approach to sequential search for an optimal dose: Parametric and nonparametric models. *Statistical Methods and Applications*, 2(3):349–364. 30

Müller, P. and Mitra, R. (2013). Bayesian Nonparametric Inference–Why and How. *Bayesian Analysis*, 8(2):323–356. 14

Müller, P. and Quintana, F. A. (2004). Nonparametric Bayesian data analysis. *Statistical science*, pages 95–110. 14

Norets, A. and Pelenis, J. (2011). Posterior consistency in conditional density estimation by covariate dependent mixtures. Technical report, Economics Series, Institute for Advanced Studies. 22

Palla, K., Knowles, D. A., and Ghahramani, Z. (2013). A dependent partition-valued process for multitask clustering and time evolving network modelling. *arXiv preprint arXiv:1303.3265*. 31

Pati, D., Dunson, D. B., and Tokdar, S. T. (2013). Posterior consistency in conditional distribution estimation. *Journal of Multivariate Analysis*. 23

Pielou, E. C. (1975). *Ecological diversity*. Wiley New York. 34, 35

Pitman, J. (1992). The two-parameter generalization of Ewens random partition structure. Technical report, Technical Report 345, Dept. Statistics, UC Berkeley. 28

Pitman, J. (1995). Exchangeable and partially exchangeable random partitions. *Probability Theory and Related Fields*, 102(2):145–158. 27

Pitman, J. (1996). Some developments of the Blackwell-MacQueen urn scheme. *Lecture Notes-Monograph Series*, pages 245–267. 26

Pitman, J. (2006). *Combinatorial stochastic processes*, volume 1875. Springer-Verlag. 25, 27, 35

Rivoirard, V. and Rousseau, J. (2012). Bernstein-von Mises theorem for linear functionals of the density. *Ann. Statist.*, 40(3):1489–1523. 20

Rousseau, J. (2010). Rates of convergence for the posterior distributions of mixtures of betas and adaptive nonparametric estimation of the density. *Ann. Statist.*, 38(1):146–180. 20

Rousseau, J., Chopin, N., and Liseo, B. (2012). Bayesian nonparametric estimation of the spectral density of a long or intermediate memory Gaussian process. *Ann. Statist.*, 40(2):964–995. 20

Rousseau, J. and Kruijer, W. (2011). Adaptive Bayesian Estimation of a spectral density. *Preprint*. 20

Salomond, J.-B. (2013). Adaptive Bayes test for monotonicity. *arXiv preprint arXiv:1303.6466*. 20

Schwartz, L. (1965). On bayes procedures. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 4(1):10–26. 16, 17

Scricciolo, C. (2006). Convergence rates for Bayesian density estimation of infinite-dimensional exponential families. *Ann. Statist.*, 34(6):2897–2920. 20

Scricciolo, C. (2012). Adaptive Bayesian density estimation using Pitman-Yor or normalized inverse-Gaussian process kernel mixtures. *arXiv preprint arXiv:1210.8094*. 22

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica*, 4:639–650. 24

Shen, W., Tokdar, S., and Ghosal, S. (2013). Adaptive Bayesian multivariate density estimation with Dirichlet mixtures. *arXiv preprint arXiv:1109.6406*. 20, 22

Shen, X. and Wasserman, L. (2001). Rates of convergence of posterior distributions. *Ann. Statist.*, 29(3):687–714. 17

Simpson, E. H. (1949). Measurement of diversity. *Nature*, 163(4148):688. 34

Szabó, B., van der Vaart, A. W., and van Zanten, J. (2013). Empirical Bayes scaling of Gaussian priors in the white noise model. *Electronic Journal of Statistics*, 7:991–1018. 20

Teh, Y., Jordan, M., Beal, M., and Blei, D. (2006). Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581. 30

Tsybakov, A. (2009). *Introduction to nonparametric estimation*. Springer Verlag. 19

van der Vaart, A. W. and van Zanten, J. (2008). Rates of contraction of posterior distributions based on Gaussian process priors. *Ann. Statist.*, 36(3):1435–1463. 20, 21

van der Vaart, A. W. and van Zanten, J. (2009). Adaptive Bayesian estimation using a Gaussian random field with inverse Gamma bandwidth. *The Annals of Statistics*, 37(5B):2655–2675. 20, 21

Wade, S., Walker, S. G., and Petrone, S. (2013). A predictive study of Dirichlet process mixture models for curve fitting. *Manuscript in preparation*. 31

Walker, S. G. (2004). New approaches to Bayesian consistency. *The Annals of Statistics*, 32(5):2028–2043. 17

Walker, S. G., Damien, P., Laud, P. W., and Smith, A. F. (1999). Bayesian nonparametric inference for random distributions and related functions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):485–527. 14

Walker, S. G. and Hjort, N. L. (2001). On bayesian consistency. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(4):811–821. 17

Walker, S. G., Lijoi, A., and Prünster, I. (2005). Data tracking and the understanding of Bayesian consistency. *Biometrika*, 92(4):765–778. 17

Williamson, S., Orbanz, P., and Ghahramani, Z. (2010). Dependent Indian buffet processes. In *Proc. International Conference on Artificial Intelligence and Statistics*. Citeseer. 31

Chapter 2

Bayesian optimal adaptive estimation using a sieve prior

On propose une forme générique de distributions a priori pour obtenir des résultats de vitesse de contraction de la loi a posteriori dans plusieurs modèles. Ces lois a priori sont appelées *sieve priors*. Elles permettent de plus d'obtenir des vitesses qui s'adaptent à la régularité du paramètre, sans que cette régularité soit utilisée dans la méthode d'estimation. Les résultats sont illustrés sur les modèles de densité, de régression, d'autorégression d'ordre 1 et de bruit blanc Gaussien. On montre en outre qu'une approche adaptative pour une fonction de perte donnée (par exemple globale) peut s'avérer sous-optimale pour une autre fonction de perte (par exemple locale) dans le cas du modèle de bruit blanc Gaussien.

Authors

- Julyan Arbel (Université Paris-Dauphine, CREST, Paris)
- Ghislaine Gayraud (Université de Technologie de Compiègne, CREST, Paris)
- Judith Rousseau (ENSAE, Université Paris-Dauphine, CREST, Paris)

Status

Article Arbel et al. (2013) to appear in the *Scandinavian Journal of Statistics*.

Abstract

We derive rates of contraction of posterior distributions on nonparametric models resulting from sieve priors. The aim of the paper is to provide general conditions to get posterior rates when the parameter space has a general structure, and rate adaptation when the parameter space is, *e.g.*, a Sobolev class. The conditions employed, although standard in the literature, are combined in a different way. The results are applied to density, regression, nonlinear autoregression and Gaussian white noise models. In the latter we have also considered a loss function which is different from the usual l^2 norm, namely the pointwise loss. In this case it is possible to prove that the adaptive Bayesian approach for the l^2 loss is strongly suboptimal and we provide a lower bound on the rate.

Keywords: adaptation, minimax criteria, nonparametric models, rate of contraction, sieve prior, white noise model.

2.1 Introduction

The asymptotic behaviour of posterior distributions in nonparametric models has received growing consideration in the literature over the last ten years. Many different models have been considered, ranging from the problem of density estimation in i.i.d. models (Barron et al., 1999, Ghosal et al., 2000), to sophisticated dependent models (Rousseau et al., 2012). For these models, different families of priors have also been considered, where the most common are Dirichlet process mixtures (or related priors), Gaussian processes (van der Vaart and van Zanten, 2008), or series expansions on a basis (such as wavelets, see Abramovich et al., 1998).

In this paper we focus on a family of priors called *sieve priors*, introduced as *compound priors* and discussed by Zhao (1993, 2000), and further studied by Shen and Wasserman (2001). It is defined for models $(\mathcal{X}^{(n)}, A^{(n)}, P_{\boldsymbol{\theta}}^{(n)} : \boldsymbol{\theta} \in \Theta)$, $n \in \mathbb{N}\backslash\{0\}$, where $\Theta \subseteq \mathbb{R}^{\mathbb{N}}$, the set of sequences. Let A be a σ-field associated to Θ. The observations are denoted X^n, where the asymptotics are driven by n. The probability measures $P_{\boldsymbol{\theta}}^{(n)}$ are dominated by some reference measure μ, with density $p_{\boldsymbol{\theta}}^{(n)}$. Remark that such an infinite-dimensional parameter $\boldsymbol{\theta}$ can often characterize a functional parameter, or a curve, $\boldsymbol{f} = \boldsymbol{f}_{\boldsymbol{\theta}}$. For instance, in regression, density or spectral density models, $\boldsymbol{f}$ represents a regression function, a log density or a log spectral density respectively, and $\boldsymbol{\theta}$ represents its coordinates in an appropriate basis $\boldsymbol{\psi} = (\psi_j)_{j\geq 1}$ (*e.g.*, a Fourier, a wavelet, a log spline, or an orthonormal basis in general). In this paper we study frequentist properties of the posterior distributions as n tends to infinity, assuming that data X^n are generated by a measure $P_{\boldsymbol{\theta}_0}^{(n)}$, $\boldsymbol{\theta}_0 \in \Theta$. We study in particular rates of contraction of the posterior distribution and rates of convergence of the risk.

A sieve prior Π is expressed as

$$\boldsymbol{\theta} \sim \Pi(\,\cdot\,) = \sum_{k=1}^{\infty} \pi(k) \Pi_k(\,\cdot\,), \tag{2.1}$$

where $\sum_k \pi(k) = 1$, $\pi(k) \geq 0$, and the Π_k's are prior distributions on so-called sieve spaces $\Theta_k = \mathbb{R}^k$. Set $\boldsymbol{\theta}_k = (\theta_1, \ldots, \theta_k)$ the finite-dimensional vector of the first k entries of $\boldsymbol{\theta}$. Essentially, the whole prior Π is seen as a hierarchical prior, see Figure 2.1. The hierarchical parameter k, called threshold parameter, has prior π. Conditionally on k, the prior on $\boldsymbol{\theta}$ is Π_k which is supposed to have mass only on Θ_k (this amounts to say that the priors on the remaining entries θ_j, $j > k$, are point masses at 0). We assume that Π_k is an independent prior on the coordinates θ_j, $j = 1, \ldots, k$, of $\boldsymbol{\theta}_k$ with a unique probability density g once rescaled by positive $\boldsymbol{\tau} = (\tau_j)_{j\geq 1}$. Using the same notation Π_k for probability and density with Lebesgue measure or $\mathbb{R}^k$, we have

$$\forall \boldsymbol{\theta}_k \in \Theta_k, \quad \Pi_k\left(\boldsymbol{\theta}_k\right) = \prod_{j=1}^{k} \frac{1}{\tau_j} g\left(\frac{\theta_j}{\tau_j}\right). \tag{2.2}$$

Note that the quantities Π, Π_k, π, $\boldsymbol{\tau}$ and g could depend on n. Although not purely Bayesian, data dependent priors are quite common in the literature. For instance, Ghosal and van der Vaart (2007) use a similar prior with a deterministic cutoff $k = \lfloor n^{1/(2\alpha+1)} \rfloor$ in application 7.6.

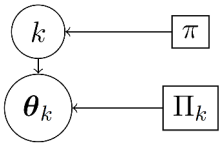

FIGURE 2.1: Graphical representation of the hierarchical structure of the *sieve prior* given by Equation (2.1)

We will also consider the case where the prior is truncated to an l^1 ball of radius $r_1 > 0$ (see the nonlinear AR(1) model application in Section 2.2.3)

$$\forall \boldsymbol{\theta}_k \in \Theta_k, \quad \Pi_k(\boldsymbol{\theta}_k) \propto \prod_{j=1}^{k} \frac{1}{\tau_j} g\left(\frac{\theta_j}{\tau_j}\right) \mathbb{I}\left(\sum_{j=1}^{k} |\theta_j| \leq r_1\right). \tag{2.3}$$

The posterior distribution $\Pi(\cdot|X^n)$ is defined by, for all measurable sets B of Θ,

$$\Pi(B|X^n) = \frac{\int_B p_{\boldsymbol{\theta}}^{(n)}(X^n) d\Pi(\boldsymbol{\theta})}{\int_\Theta p_{\boldsymbol{\theta}}^{(n)}(X^n) d\Pi(\boldsymbol{\theta})}. \tag{2.4}$$

Given the sieve prior Π, we study the rate of contraction of the posterior distribution in $P_{\boldsymbol{\theta}_0}^{(n)}-$probability with respect to a semimetric d_n on Θ. This rate is defined as the best possible (*i.e.* the smallest) sequence $(\epsilon_n)_{n\geq 1}$ such that

$$\Pi\left(\boldsymbol{\theta} : d_n^2(\boldsymbol{\theta}, \boldsymbol{\theta}_0) \geq M\epsilon_n^2 | X^n\right) \underset{n\to\infty}{\longrightarrow} 0,$$

in $P_{\boldsymbol{\theta}_0}^{(n)}$ probability, for some $\boldsymbol{\theta}_0 \in \Theta$ and a positive constant M, which can be chosen as large as needed. We also derive convergence rates for the posterior loss $\Pi(d_n^2(\boldsymbol{\theta}, \boldsymbol{\theta}_0)|X^n)$ in $P_{\boldsymbol{\theta}_0}^{(n)}-$probability.

The posterior concentration rate is optimal when it coincides with the minimax rates of convergence, when $\boldsymbol{\theta}_0$ belongs to a given functional class, associated to the same semimetric d_n. Typically these minimax rates of convergence are defined for functional classes indexed by a smoothness parameter Sobolev, Hölder, or more generally Besov spaces.

The objective of this paper is to find mild generic assumptions on the sieve prior Π of the form (2.1), on models $P_{\boldsymbol{\theta}}^{(n)}$ and on d_n, such that the procedure adapts to the optimal rate in the minimax sense, both for the posterior distribution and for the risk. Results in Bayesian nonparametrics literature about contraction rates are usually of two kinds. Firstly, general assumptions on priors and models allow to derive rates, see for example Shen and Wasserman (2001), Ghosal et al. (2000), Ghosal and van der Vaart

(2007). Secondly, other papers focus on a particular prior and obtain contraction rates in a particular model, see for instance Belitser and Ghosal (2003) in the white noise model, De Jonge and van Zanten (2010) in regression, and Scricciolo (2006) in density. The novelty of this paper is that our results hold for a family of priors (sieve priors) without a specific underlying model, and can be applied to different models.

An additional interesting property that is sought at the same time as convergence rates is adaptation. This means that, once specified a loss function (a semimetric d_n on Θ), and a collection of classes of different smoothnesses for the parameter, one constructs a procedure which is independent of the smoothness, but which is rate optimal (under the given loss d_n), within each class. Indeed, the optimal rate naturally depends on the smoothness of the parameter, and standard straightforward estimation techniques usually use it as an input. This is all the more an important issue that relatively few instances in the Bayesian literature are available in this area. That property is often obtained when the unknown parameter is assumed to belong to a discrete set, see for example Belitser and Ghosal (2003). There exist some results in the context of density estimation by Huang (2004), Scricciolo (2006), Ghosal et al. (2008), van der Vaart and van Zanten (2009), Rivoirard and Rousseau (2012a), Rousseau (2010) and Kruijer et al. (2010), in regression by De Jonge and van Zanten (2010), and in spectral density estimation by Rousseau and Kruijer (2011). What enables adaptation in our results is the thresholding induced by the prior on k: the posterior distribution of parameter k concentrates around values that are the typical efficient size of models of the true smoothness.

As seen from our assumptions in Section 2.2.1 and from the general results (Theorem 2.2 and Corollary 2.3), adaptation is relatively straightforward under sieve priors defined by (2.1) when the semimetric is a global loss function which acts like the Kullback-Leibler divergence, the l^2 norm on $\boldsymbol{\theta}$ in the regression problem, or the Hellinger distance in the density problem. If the loss function (or the semimetric) d_n acts differently, then the posterior distribution (or the risk) can be quite different (suboptimal). This is illustrated in Section 2.3.2 for the white noise model (2.16) when the loss is a local loss function as in the case of the estimation of $\boldsymbol{f}(t)$, for a given t, where $d_n(\boldsymbol{f}, \boldsymbol{f}_0) = (\boldsymbol{f}(t) - \boldsymbol{f}_0(t))^2$. This phenomenon has been encountered also by Rousseau and Kruijer (2011). It is not merely a Bayesian issue: Cai et al. (2007) show that an optimal estimator under global loss cannot be locally optimal at each point $\boldsymbol{f}(t)$ in the white noise model. The penalty between global and local rates is at least a $\log n$ term. Abramovich et al. (2004) and Abramovich et al. (2007a) obtain similar results with Bayesian wavelet estimators in the same model.

The paper is organized as follows. Section 2.2 first provides a general result on rates of contraction for the posterior distribution in the setting of sieve priors. We also derive a result in terms of posterior loss, and show that the rates are adaptive optimal for Sobolev smoothness classes. The section ends up with applications to the density, the regression function and the nonlinear autoregression function estimation. In Section 2.3, we study more precisely the case of the white noise model, which is a benchmark model. We study in detail the difference between global or pointwise losses in this model, and provide a lower bound for the latter loss, showing that sieve priors lead to suboptimal contraction rates. Proofs are deferred to the Appendix.

Notations

We use the following notations. Vectors are written in bold letters, for example $\boldsymbol{\theta}$ or $\boldsymbol{\theta}_0$, while light-face is used for their entries, like θ_j or θ_{0j}. We denote by $\boldsymbol{\theta}_{0k}$ the projection of $\boldsymbol{\theta}_0$ on its first k coordinates, and by $p_{0k}^{(n)}$ and $p_0^{(n)}$ the densities of the observations in the corresponding models. We denote by d_n a semimetric, by $\|\cdot\|_2$ the l^2 norm (on vectors) in Θ or the L^2 norm (on curves $\boldsymbol{f}$), and by $\|\cdot\|_{2,k}$ the l^2 norm restricted to the first k coordinates of a parameter. Expectations $\mathsf{E}_0^{(n)}$ and $\mathsf{E}_{\boldsymbol{\theta}}^{(n)}$ are defined with respect to $P_{\boldsymbol{\theta}_0}^{(n)}$ and $P_{\boldsymbol{\theta}}^{(n)}$ respectively. The same notation $\Pi(\cdot|X^n)$ is used for posterior probability or posterior expectation. The expected posterior risk and the frequentist risk relative to d_n are defined and denoted by $\mathcal{R}_n^{d_n}(\boldsymbol{\theta}_0) = \mathsf{E}_0^{(n)}\Pi(d_n^2(\boldsymbol{\theta},\boldsymbol{\theta}_0)|X^n)$ and $R_n^{d_n}(\boldsymbol{\theta}_0) = \mathsf{E}_0^{(n)}(d_n^2(\widehat{\boldsymbol{\theta}},\boldsymbol{\theta}_0))$ respectively (for an estimator $\widehat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}_0$), where the mention of $\boldsymbol{\theta}_0$ might be omitted (*cf.* Robert, 2007, Section 2.3). We denote by φ the standard Gaussian probability density.

Let K denote the Kullback-Leibler divergence $K(f,g) = \int f\log(f/g)d\mu$, and $V_{m,0}$ denote the m^{th} centered moment $V_{m,0}(f,g) = \int f|\log(f/g) - K(f,g)|^m d\mu$, with $m \geq 2$.

Define two additional divergences $\widetilde{K}$ and $\widetilde{V}_{m,0}$, which are expectations with respect to $p_0^{(n)}$, $\widetilde{K}(f,g) = \int p_0^{(n)}|\log(f/g)|d\mu$ and $\widetilde{V}_{m,0}(f,g) = \int p_0^{(n)}|\log(f/g) - K(f,g)|^m d\mu$.

We denote by C a generic constant whose value is of no importance and we use $\lesssim$ for inequalities up to a multiple constant.

2.2 General case

In this section we give a general theorem which provides an upper bound on posterior contraction rates ϵ_n. Throughout the section, we assume that the sequence of positive

numbers $(\epsilon_n)_{n\geq 1}$, or $(\epsilon_n(\beta))_{n\geq 1}$ when we point to a specific value of smoothness β, is such that $\epsilon_n \underset{n\to\infty}{\longrightarrow} 0$ and $n\epsilon_n^2/\log n \underset{n\to\infty}{\longrightarrow} \infty$.

We introduce the following numbers

$$j_n = \lfloor j_0 n\epsilon_n^2/\log(n)\rfloor,\ k_n = \lfloor M_0 j_n \log(n)/L(n)\rfloor, \tag{2.5}$$

for $j_0 > 0, M_0 > 1$, where L is a slow varying function such that $L \leq \log$, hence $j_n \leq k_n$. We use k_n to define the following approximation subsets of Θ

$$\Theta_{k_n}(Q) = \left\{\boldsymbol{\theta} \in \Theta_{k_n} : \|\boldsymbol{\theta}\|_{2,k_n} \leq n^Q\right\},$$

for $Q > 0$. Note that the prior actually charges a union of spaces of dimension k, $k \geq 1$, so that $\Theta_{k_n}(Q)$ can be seen as a union of spaces of dimension $k \leq k_n$. Lemma 2.13 provides an upper bound on the prior mass of $\Theta_{k_n}(Q)$.

It has been shown (Ghosal et al., 2000, Ghosal and van der Vaart, 2007, Shen and Wasserman, 2001) that an efficient way to derive rates of contraction of posterior distributions is to bound from above the numerator of (2.4) using tests (and k_n for the increasing sequence $\Theta_{k_n}(Q)$), and to bound from below its denominator using an approximation of $p_0^{(n)}$ based on a value $\boldsymbol{\theta} \in \Theta_{j_n}$ close to $\boldsymbol{\theta}$. The latter is done in Lemma 2.14 where we use j_n to define the finite component approximation $\boldsymbol{\theta}_{0j_n}$ of $\boldsymbol{\theta}_0$, and we show that the prior mass of the following Kullback-Leibler neighbourhoods of $\boldsymbol{\theta}_0$, $\mathcal{B}_n(m)$, $n \in \mathbb{N}^*$, are lower bounded by an exponential term:

$$\mathcal{B}_n(m) = \left\{\boldsymbol{\theta} :\ K\left(p_0^{(n)}, p_{\boldsymbol{\theta}}^{(n)}\right) \leq 2n\epsilon_n^2, V_{m,0}\left(p_0^{(n)}, p_{\boldsymbol{\theta}}^{(n)}\right) \leq 2^{m+1}(n\epsilon_n^2)^{m/2}\right\}.$$

Define two neighbourhoods of $\boldsymbol{\theta}_0$ in the sieve space Θ_{j_n}, $\widetilde{\mathcal{B}}_n(m)$, similar to $\mathcal{B}_n(m)$ but using $\widetilde{K}$ and $\widetilde{V}_{m,0}$, and $\mathcal{A}_n(H_1)$, an l^2 ball of radius n^{-H_1}, $H_1 > 0$:

$$\widetilde{\mathcal{B}}_n(m) = \left\{\boldsymbol{\theta} \in \Theta_{j_n} :\ \widetilde{K}\left(p_{0j_n}^{(n)}, p_{\boldsymbol{\theta}}^{(n)}\right) \leq n\epsilon_n^2, \widetilde{V}_{m,0}\left(p_{0j_n}^{(n)}, p_{\boldsymbol{\theta}}^{(n)}\right) \leq \left(n\epsilon_n^2\right)^{m/2}\right\},$$
$$\mathcal{A}_n(H_1) = \left\{\boldsymbol{\theta} \in \Theta_{j_n} :\ \|\boldsymbol{\theta}_{0j_n} - \boldsymbol{\theta}\|_{2,j_n} \leq n^{-H_1}\right\}.$$

2.2.1 Assumptions

The following technical assumptions are involved in the subsequent analysis, and are discussed at the end of this section. Recall that the true parameter is $\boldsymbol{\theta}_0$, under which the observations have density $p_0^{(n)}$.

A_1 Condition on $p_0^{(n)}$ and ϵ_n. For n large enough and for some $m > 0$,

$$K\left(p_0^{(n)}, p_{0j_n}^{(n)}\right) \leq n\epsilon_n^2 \quad \text{and} \quad V_{m,0}\left(p_0^{(n)}, p_{0j_n}^{(n)}\right) \leq \left(n\epsilon_n^2\right)^{m/2}.$$

A_2 Comparison between norms. The following inclusion holds in Θ_{j_n}

$$\exists H_1 > 0, \text{ s.t. } \mathcal{A}_n(H_1) \subset \widetilde{\mathcal{B}}_n(m).$$

A_3 Comparison between d_n and l^2. There exist three non negative constants D_0, D_1, D_2 such that, for any two $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \Theta_{k_n}(Q)$,

$$d_n(\boldsymbol{\theta}, \boldsymbol{\theta}') \leq D_0 k_n^{D_1} \left\|\boldsymbol{\theta} - \boldsymbol{\theta}'\right\|_{2,k_n}^{D_2}.$$

A_4 Test Condition. There exist two positive constants c_1 and $\zeta < 1$ such that, for every $\boldsymbol{\theta}_1 \in \Theta_{k_n}(Q)$, there exists a test $\phi_n(\boldsymbol{\theta}_1) \in [0,1]$ which satisfies

$$\begin{aligned} &\mathsf{E}_0^{(n)}(\phi_n(\boldsymbol{\theta}_1)) \leq e^{-c_1 n d_n^2(\boldsymbol{\theta}_0, \boldsymbol{\theta}_1)} \quad \text{and} \\ &\sup_{d_n(\boldsymbol{\theta}, \boldsymbol{\theta}_1) < \zeta d_n(\boldsymbol{\theta}_0, \boldsymbol{\theta}_1)} \mathsf{E}_{\boldsymbol{\theta}}^{(n)}\left(1 - \phi_n(\boldsymbol{\theta}_1)\right) \leq e^{-c_1 n d_n^2(\boldsymbol{\theta}_0, \boldsymbol{\theta}_1)}. \end{aligned}$$

A_5 On the prior Π. There exist positive constants $a, b, G_1, G_2, G_3, G_4, H_2, \alpha$ and τ_0 such that π satisfy

$$\forall k = 1, 2, \ldots, \quad e^{-akL(k)} \leq \pi(k) \leq e^{-bkL(k)}, \tag{2.6}$$

where the function L is a slow varying function introduced in Equation (2.5); g satisfy

$$\forall \theta \in \mathbb{R}, \quad G_1 e^{-G_2|\theta|^\alpha} \leq g(\theta) \leq G_3 e^{-G_4|\theta|^\alpha}. \tag{2.7}$$

The scales $\boldsymbol{\tau}$ defined in Equation (2.2) satisfy the following conditions

$$\max_{j \geq 1} \tau_j \leq \tau_0, \tag{2.8}$$

$$\min_{j \leq k_n} \tau_j \geq n^{-H_2}, \tag{2.9}$$

$$\sum_{j=1}^{j_n} |\theta_{0j}|^\alpha / \tau_j^\alpha \leq C j_n \log n. \tag{2.10}$$

Remark 2.1.

- *Conditions* $\boldsymbol{A_1}$ *and* $\boldsymbol{A_2}$ *are local in that they need to be checked at the true parameter* $\boldsymbol{\theta}_0$ *only. They are useful to prove that the prior puts sufficient mass around Kullback-Leibler neighbourhoods of the true probability. Condition* $\boldsymbol{A_1}$ *is a limiting factor to the rate: it characterizes* ϵ_n *through the capacity of approximation of* $p_0^{(n)}$ *by* $p_{0j_n}^{(n)}$*: the smoother* $p_0^{(n)}$*, the closer* $p_0^{(n)}$ *and* $p_{0j_n}^{(n)}$*, and the faster* ϵ_n*. In many models, they are ensured because* $K(p_0^{(n)}, p_{\boldsymbol{\theta}_{j_n}}^{(n)})$ *and* $V_{m,0}(p_0^{(n)}, p_{\boldsymbol{\theta}_{j_n}}^{(n)})$ *can be written locally (meaning around* $\boldsymbol{\theta}_0$*) in terms of the* l^2 *norm* $\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}_{j_n}\|_2$ *directly. Smoothness assumptions are then typically required to control* $\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}_{j_n}\|_2$*.*

  It is the case for instance for Sobolev and Besov smoothnesses (cf. Equation (2.12)). The control is expressed with a power of j_n*, whose comparison to* ϵ_n^2 *provides in turn a tight way to tune the rate (cf. the proof of Proposition 2.6).*

  Note that the constant H_1 *in Condition* $\boldsymbol{A_2}$ *can be chosen as large as needed: if* $\boldsymbol{A_2}$ *holds for a specified positive constant* H_0*, then it does for any* $H_1 > H_0$*. This makes the condition quite loose. A more stringent version of* $\boldsymbol{A_2}$*, if simpler, is the following.*

  $\boldsymbol{A_2'}$ ***Comparison between norms.*** *For any* $\boldsymbol{\theta} \in \Theta_{j_n}$

$$\begin{aligned} \widetilde{K}\left(p_{0j_n}^{(n)}, p_{\boldsymbol{\theta}}^{(n)}\right) &\leq Cn\,\|\boldsymbol{\theta}_{0j_n} - \boldsymbol{\theta}\|_{2,j_n}^2 \quad \text{and} \\ \widetilde{V}_{m,0}\left(p_{0j_n}^{(n)}, p_{\boldsymbol{\theta}}^{(n)}\right) &\leq Cn^{m/2}\,\|\boldsymbol{\theta}_{0j_n} - \boldsymbol{\theta}\|_{2,j_n}^m . \end{aligned}$$

  This is satisfied in the Gaussian white noise model (see Section 2.3).

- *Condition* $\boldsymbol{A_3}$ *is generally mild. The reverse is more stringent since* d_n *may be bounded, as is the case with the Hellinger distance.* $\boldsymbol{A_3}$ *is satisfied in many common situations, see for example the applications later on. Technically, this condition allows to switch from a covering number (or entropy) in terms of the* l^2 *norm to a covering number in terms of the semimetric* d_n*.*

- *Condition* $\boldsymbol{A_4}$ *is common in the Bayesian nonparametric literature. A review of different models and their corresponding tests is given in Ghosal and van der Vaart (2007) for example. The tests strongly depend on the semimetric* d_n*.*

- *Condition* $\boldsymbol{A_5}$ *concerns the prior. Equations (2.6) and (2.7) state that the tails of* π *and* g *have to be at least exponential or of exponential type. For instance, if* π *is the geometric distribution,* $L = 1$*, and if it is the Poisson distribution,* $L(k) = \log(k)$ *(both are slow varying functions). Laplace and Gaussian distributions are covered by* g*, with* $\alpha = 1$ *and* $\alpha = 2$ *respectively. These equations aim at controlling the prior mass of* $\Theta_{k_n}^c(Q)$*, the complement of* $\Theta_{k_n}(Q)$ *in* Θ *(see Lemma 2.13). The case where the scale* $\boldsymbol{\tau}$ *depends on* n *is considered in Babenko and Belitser (2009, 2010) in the white noise model. Here the constraints on* $\boldsymbol{\tau}$ *are rather mild since*

they are allowed to go to zero polynomially as a function of n, and must be upper bounded. Rivoirard and Rousseau (2012a) study a family of scales $\boldsymbol{\tau} = (\tau_j)_{j\geq 1}$ that are decreasing polynomially with j. Here the prior is more general and encompasses both frameworks. Equations (2.6) - (2.10) are needed in Lemmas 2.13 and 2.14 for bounding respectively $\Pi(\mathcal{B}_n(m))$ from below and $\Pi(\Theta_{k_n}^c(Q))$ from above. A smoothness assumption on $\boldsymbol{\theta}_0$ is usually required for Equation (2.10).

2.2.2 Results

Concentration and posterior loss

The following theorem provides an upper bound for the rate of contraction of the posterior distribution.

Theorem 2.2. *If Conditions $\boldsymbol{A_1}$ - $\boldsymbol{A_5}$ hold, then for M large enough and for L introduced in Equation (2.5),*

$$\mathsf{E}_0^{(n)}\Pi\left(\boldsymbol{\theta}:\ d_n^2(\boldsymbol{\theta},\boldsymbol{\theta}_0) \geq M\frac{\log n}{L(n)}\epsilon_n^2 | X^n\right) = \mathcal{O}\left((n\epsilon_n^2)^{-m/2}\right) \underset{n\to\infty}{\longrightarrow} 0.$$

Proof. See the Appendix. □

The convergence of the posterior distribution at the rate ϵ_n implies that the expected posterior risk converges (at least) at the same rate ϵ_n, when d_n is bounded.

Corollary 2.3. *Under the assumptions of Theorem 2.2, with a value of m in Conditions $\boldsymbol{A_1}$ and $\boldsymbol{A_2}$ such that $(n\epsilon_n^2)^{-m/2} = \mathcal{O}(\epsilon_n^2)$, and if d_n is bounded on Θ, then the expected posterior risk given $\boldsymbol{\theta}_0$ and Π converges at least at the same rate ϵ_n*

$$\mathcal{R}_n^{d_n} = \mathsf{E}_0^{(n)}\Pi(d_n^2(\boldsymbol{\theta},\boldsymbol{\theta}_0)|X^n) = \mathcal{O}\left(\frac{\log n}{L(n)}\epsilon_n^2\right).$$

Proof. Denote D the bound of d_n, *i.e.* for all $\boldsymbol{\theta},\ \boldsymbol{\theta}' \in \Theta$, $d_n(\boldsymbol{\theta},\boldsymbol{\theta}') \leq D$. We have

$$\begin{aligned}\mathcal{R}_n^{d_n} \leq & \quad M\frac{\log n}{L(n)}\epsilon_n^2 + \mathsf{E}_0^{(n)}\Pi\left(\mathsf{I}\left(d_n^2(\boldsymbol{\theta},\boldsymbol{\theta}_0) \geq M\frac{\log n}{L(n)}\epsilon_n^2\right) d_n^2(\boldsymbol{\theta},\boldsymbol{\theta}_0)|X^n\right)\\ \leq & \quad M\frac{\log n}{L(n)}\epsilon_n^2 + D\mathsf{E}_0^{(n)}\Pi\left(\boldsymbol{\theta}:\ d_n^2(\boldsymbol{\theta},\boldsymbol{\theta}_0) \geq M\frac{\log n}{L(n)}\epsilon_n^2 | X^n\right)\end{aligned}$$

so $\mathcal{R}_n^{d_n} = \mathcal{O}(\log n/L(n)\epsilon_n^2)$ by Theorem 2.2 and the assumption on m. □

Remark 2.4. *The condition on m in Corollary 2.3 requires $n\epsilon_n^2$ to grow as a power of n. When $\boldsymbol{\theta}_0$ has Sobolev smoothness β, this is the case since ϵ_n^2 is typically of order*

$(n/\log n)^{-2\beta/(2\beta+1)}$. *The condition on* m *boils down to* $m \geq 4\beta$. *When* $\boldsymbol{\theta}_0$ *is smoother, e.g. in a Sobolev space with exponential weights, the rate is typically of order* $\log n/\sqrt{n}$. *Then a common way to proceed is to resort to an exponential inequality for controlling the denominator of the posterior distribution of Equation (2.4) (see e.g. Rivoirard and Rousseau, 2012b).*

Remark 2.5. *We can note that this result is meaningful from a non Bayesian point of view as well. Indeed, let* $\widehat{\boldsymbol{\theta}}$ *be the posterior mean estimate of* $\boldsymbol{\theta}$ *with respect to* Π. *Then, if* $\boldsymbol{\theta} \to d_n^2(\boldsymbol{\theta}, \boldsymbol{\theta}_0)$ *is convex, we have by Jensen's inequality*

$$d_n^2(\widehat{\boldsymbol{\theta}}, \boldsymbol{\theta}_0) \leq \Pi\left(d_n^2(\boldsymbol{\theta}, \boldsymbol{\theta}_0)|X^n\right),$$

so the frequentist risk converges at the same rate ϵ_n

$$R_n^{d_n} = \mathsf{E}_0^{(n)}(d_n^2(\widehat{\boldsymbol{\theta}}, \boldsymbol{\theta}_0)) \leq \mathsf{E}_0^{(n)}\Pi\left(d_n^2(\boldsymbol{\theta}, \boldsymbol{\theta}_0)|X^n\right) = \mathcal{R}_n^{d_n} = \mathcal{O}\left(\frac{\log n}{L(n)}\epsilon_n^2\right).$$

Note that we have no result for general pointwise estimates $\widehat{\boldsymbol{\theta}}$, *for instance for the MAP. This latter was studied in Abramovich et al. (2007b, 2010).*

Adaptation

When considering a given class of smoothness for the parameter $\boldsymbol{\theta}_0$, the minimax criterion implies an optimal rate of convergence. Posterior (resp. risk) adaptation means that the posterior distribution (resp. the risk) concentrates at the optimal rate for a class of possible smoothness values.

We consider here Sobolev classes $\Theta_\beta(L_0)$ for univariate problems defined by

$$\Theta_\beta(L_0) = \left\{\boldsymbol{\theta} : \sum_{j=1}^{\infty} \theta_j^2 j^{2\beta} < L_0\right\}, \ \beta > 1/2, \ L_0 > 0 \tag{2.11}$$

with smoothness parameter β and radius L_0. If $\boldsymbol{\theta}_0 \in \Theta_\beta(L_0)$, then one has the following bound

$$\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}_{0j_n}\|_2^2 = \sum_{j=j_n+1}^{\infty} \theta_{0j}^2 j^{2\beta} j^{-2\beta} \leq L_0 j_n^{-2\beta}. \tag{2.12}$$

Donoho and Johnstone (1998) give the global (*i.e.* under the l^2 loss) minimax rate $n^{-\beta/(2\beta+1)}$ attached to the Sobolev class of smoothness β. We show that under an additional condition between K, $V_{m,0}$ and l^2, the upper bound ϵ_n on the rate of contraction can be chosen equal to the optimal rate, up to a $\log n$ term.

Proposition 2.6. *Let L_0 denote a positive fixed radius, and $\beta_2 \geq \beta_1 > 1/2$. If for n large enough, there exists a positive constant C_0 such that*

$$\sup_{\beta_1 \leq \beta \leq \beta_2} \sup_{\boldsymbol{\theta}_0 \in \Theta_\beta(L_0)} K\left(p_0^{(n)}, p_{0j_n}^{(n)}\right) \leq C_0 n \left\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}_{0j_n}\right\|_2^2, \text{ and}$$
$$\sup_{\beta_1 \leq \beta \leq \beta_2} \sup_{\boldsymbol{\theta}_0 \in \Theta_\beta(L_0)} V_{m,0}\left(p_0^{(n)}, p_{0j_n}^{(n)}\right) \leq C_0^m n^{m/2} \left\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}_{0j_n}\right\|_2^m, \tag{2.13}$$

and if Conditions $\boldsymbol{A_2}$ - $\boldsymbol{A_5}$ hold with constants independent of $\boldsymbol{\theta}_0$ in the set $\cup_{\beta_1 \leq \beta \leq \beta_2} \Theta_\beta(L_0)$, then for M sufficiently large,

$$\sup_{\beta_1 \leq \beta \leq \beta_2} \sup_{\boldsymbol{\theta}_0 \in \Theta_\beta(L_0)} \mathsf{E}_0^{(n)} \Pi\left(\boldsymbol{\theta} : \ d_n^2(\boldsymbol{\theta}, \boldsymbol{\theta}_0) \geq M \frac{\log n}{L(n)} \epsilon_n^2(\beta) | X^n\right) \underset{n\to\infty}{\longrightarrow} 0,$$

with

$$\epsilon_n(\beta) = \epsilon_0 \left(\frac{\log n}{n}\right)^{\frac{\beta}{2\beta+1}},$$

and ϵ_0 depending on L_0, C_0 and the constants involved in the assumptions, but not depending on β.

Remark 2.7. *In the standard case where d_n is the l^2 norm, ϵ_n is the* optimal rate of contraction*, up to a $\log n$ term (which is quite common in Bayesian nonparametric computations).*

Proof. Let $\beta \in [\beta_1, \beta_2]$ and $\boldsymbol{\theta}_0 \in \Theta_\beta(L_0)$. Then $\boldsymbol{\theta}_0$ satisfies Equation (2.12), and Condition (2.13) implies that

$$K\left(p_0^{(n)}, p_{0j_n}^{(n)}\right) \leq C_0 L_0 n j_n^{-2\beta}, \ V_{m,0}\left(p_0^{(n)}, p_{0j_n}^{(n)}\right) \leq C_0 L_0^m n^{m/2} j_n^{-m\beta}.$$

For given $\boldsymbol{\theta}_0$ and β, the result of Theorem 2.2 holds if Condition $\boldsymbol{A_1}$ is satisfied. This is the case if we choose $\epsilon_n(\beta, \boldsymbol{\theta}_0) \geq C_0 L_0 j_n^{-\beta}$, provided that the bounds in Conditions $\boldsymbol{A_2}$ - $\boldsymbol{A_5}$ and in Equation (2.13) are uniform. Combined with $j_n = \lfloor j_0 n \epsilon_n^2 / \log n \rfloor$, it gives as a tight choice $\epsilon_n(\beta, \boldsymbol{\theta}_0) = \epsilon_0(\beta, \boldsymbol{\theta}_0)(\log n / n)^{\beta/(2\beta+1)}$ with $\epsilon_0(\beta, \boldsymbol{\theta}_0) \leq (L_0 C_0 j_0^{-\beta})^{1/(2\beta+1)}$. So there exists a bound $\epsilon_0 > 0$ such that $\sup_{\beta_1 \leq \beta \leq \beta_2} \sup_{\boldsymbol{\theta}_0 \in \Theta_\beta(L_0)} \epsilon_0(\beta, \boldsymbol{\theta}_0) = \epsilon_0 < \infty$, which concludes the proof. □

2.2.3 Examples

In this section, we apply our results of contraction of Section 2.2.2 to a series of models. The Gaussian white noise example is studied in detail in Section 2.3. We suppose in each model that $\boldsymbol{\theta}_0 \in \Theta_\beta(L_0)$, where $\Theta_\beta(L_0)$ is defined in Equation (2.11).

Throughout, we consider the following prior Π on Θ (or on a curve space $\mathcal{F}$ through the coefficients of the functions in a basis). Let the prior distribution π on k be Poisson with parameter λ, and given k, the prior distribution on θ_j/τ_j, $j = 1, \dots, k$ be standard Gaussian,

$$\begin{aligned} k &\sim \text{Poisson}(\lambda), \\ \frac{\theta_j}{\tau_j} \,|\, k &\sim \mathcal{N}(0,1),\ j = 1, \dots, k, \text{ independently.} \end{aligned} \tag{2.14}$$

It satisfies Equation (2.6) with function $L(k) = \log(k)$ and Equation (2.7) with $\alpha = 2$. Choose then $\tau_j^2 = \tau_0 j^{-2q}$, $\tau_0 > 0$, with $q > 1/2$. It is decreasing and bounded from above by τ_0 so Equation (2.8) is satisfied. Additionally,

$$\min_{j \le k_n} \tau_j = \tau_{k_n} = k_n^{-2q} \ge n^{-H_2}$$

for H_2 large enough, so Equation (2.9) is checked. Since $\boldsymbol{\theta}_0 \in \Theta_\beta(L_0)$,

$$\tau_0^2 \sum_{j=1}^{j_n} \theta_{0j}^2/\tau_j^2 = \sum_{j=1}^{j_n} \theta_{0j}^2 j^{2q} = \sum_{j=1}^{j_n} \theta_{0j}^2 j^{2\beta} j^{2q-2\beta} \le j_n \sum_{j=1}^{j_n} \theta_{0j}^2 j^{2\beta} \le j_n L_0,$$

as soon as $2q - 2\beta \le 1$. Hence by choosing $1/2 < q \le 1$, Equation (2.10) is verified for all $\beta > 1/2$. The prior Π thus satisfies Condition $\boldsymbol{A_5}$.

Since Condition $\boldsymbol{A_5}$ is satisfied, we will show in the three examples that Conditions $\boldsymbol{A_2}$ - $\boldsymbol{A_4}$ and Condition (2.13) hold, thus Proposition 2.6 applies: the posterior distribution attains the optimal rate of contraction, up to a $\log n$ term, that is $\epsilon_n = \epsilon_0 (\log n/n)^{\beta/(2\beta+1)}$, for a distance d_n which is specific to each model. This rate is adaptive in a range of smoothness $[\beta_1, \beta_2]$.

Density

Let us consider the density model in which the density $\boldsymbol{p}$ is unknown, and we observe i.i.d. data

$$X_i \sim \boldsymbol{p}, \quad i = 1, 2, \dots, n,$$

where $\boldsymbol{p}$ belongs to $\mathcal{F}$,

$$\mathcal{F} = \left\{ \boldsymbol{p} \text{ density on } [0,1] : \boldsymbol{p}(0) = \boldsymbol{p}(1) \text{ and } \log \boldsymbol{p} \in L^2(0,1) \right\}.$$

Equality $\boldsymbol{p}(0) = \boldsymbol{p}(1)$ is mainly used for ease of computation. We define the parameter $\boldsymbol{\theta}$ of such a function $\boldsymbol{p}$, and write $\boldsymbol{p} = \boldsymbol{p}_{\boldsymbol{\theta}}$, as the coefficients of $\log \boldsymbol{p}_{\boldsymbol{\theta}}$ in the Fourier basis

$\boldsymbol{\psi} = (\psi_j)_{j\geq 1}$, *i.e.* it can be represented as

$$\log \boldsymbol{p_\theta}(x) = \sum_{j=1}^{\infty} \theta_j \psi_j(x) - c(\boldsymbol{\theta}),$$

where $c(\boldsymbol{\theta})$ is a normalizing constant. We assign a prior to $\boldsymbol{p_\theta}$ by assigning the sieve prior Π of Equation (2.14) to $\boldsymbol{\theta}$.

A natural choice of metric d_n in this model is the Hellinger distance $d_n(\boldsymbol{\theta}, \boldsymbol{\theta'}) = h(\boldsymbol{p_\theta}, \boldsymbol{p_{\theta'}}) = \left(\int \left(\sqrt{\boldsymbol{p_\theta}} - \sqrt{\boldsymbol{p_{\theta'}}} \right)^2 d\mu \right)^{1/2}$. Lemma 2 in Ghosal and van der Vaart (2007) shows the existence of tests satisfying $\boldsymbol{A_4}$ with the Hellinger distance.

Rivoirard and Rousseau (2012b) study this model in detail (Section 4.2.2) in order to derive a Bernstein-von Mises theorem for the density model. They prove that Conditions $\boldsymbol{A_2}$, $\boldsymbol{A_3}$ and (2.13) are valid in this model (see the proof of Condition (C) for $\boldsymbol{A_2}$ and (2.13), and the proof of Condition (B) for $\boldsymbol{A_3}$). With $D_1 = D_2 = 1$, Condition $\boldsymbol{A_3}$ is written $h(\boldsymbol{p_\theta}, \boldsymbol{p_{\theta'}}) \leq D_0 k_n \left\| \boldsymbol{\theta} - \boldsymbol{\theta'} \right\|_{2,k_n}$.

Regression

Consider now the following nonparametric regression model

$$X_i = \boldsymbol{f}(t_i) + \sigma \xi_i, \quad i = 1, \ldots, n,$$

with the regular fixed design $t_i = i/n$ in $[0,1]$, i.i.d. centered Gaussian errors ξ_i with variance σ^2. The unknown σ case is studied in an unpublished paper by Rousseau and Sun. They endow σ with an Inverse Gamma (conjugate) prior. They show that this one dimensional parameter adds an $n \log(\sigma/\sigma_0)$ term in the Kullback-Leibler divergence but does not alter the rates by considering three different cases for σ, either $\sigma < \sigma_0/2$, $\sigma > 3\sigma_0/2$, or $\sigma \in [\sigma_0/2, 3\sigma_0/2]$.

We consider now in more detail the σ known case. Denote $\boldsymbol{\theta}$ the coefficients of a regression function $\boldsymbol{f}$ in the Fourier basis $\boldsymbol{\psi} = (\psi_j)_{j\geq 1}$. So for all $t \in [0,1]$, $\boldsymbol{f}$ can be represented as $\boldsymbol{f}(t) = \sum_{j=1}^{\infty} \theta_j \psi_j(t)$. We assign a prior to $\boldsymbol{f}$ by assigning the sieve prior Π of Equation (2.14) to $\boldsymbol{\theta}$.

Let $\mathbb{P}_t^n = n^{-1} \sum_{i=1}^n \delta_{t_i}$ be the empirical measure of the covariates t_i's, and define the square of the empirical norm by $\|\boldsymbol{f}\|_{\mathbb{P}_t^n}^2 = n^{-1} \sum_{i=1}^n \boldsymbol{f}^2(t_i)$. We use $d_n = \| \cdot \|_{\mathbb{P}_t^n}$.

Let $\boldsymbol{\theta} \in \Theta$ and $\boldsymbol{f}$ the corresponding regression. Basic algebra (see for example Lemma 1.7 in Tsybakov, 2009) provides, for any two j and k,

$$\frac{1}{n}\sum_{i=1}^{n}\psi_j(t_i)\psi_k(t_i) = \delta_{jk},$$

where δ_{jk} stands for Kronecker delta. Hence

$$\|\boldsymbol{f}\|^2_{\mathbb{P}^n_t} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j,k}\theta_j\theta_k\psi_j(t_i)\psi_k(t_i) = \|\boldsymbol{\theta}\|_2^2 = \|\boldsymbol{f}\|_2^2, \tag{2.15}$$

where the last equality is Parseval's. It ensures Condition $\boldsymbol{A_3}$ with $D_0 = D_2 = 1$ and $D_1 = 0$.

The densities $\mathcal{N}(\boldsymbol{f}(t_i), \sigma^2)$ of X_i's are denoted $p_{\boldsymbol{f},i}$, $i = 1, \ldots, n$, and their product $p^{(n)}_{\boldsymbol{f}}$. The quantity $\boldsymbol{f}_{0j_n}$ denotes the truncated version of $\boldsymbol{f}_0$ to its first j_n terms in the Fourier basis.

We have $2K(p_{\boldsymbol{f}_0,i}, p_{\boldsymbol{f},i}) = V_{2,0}(p_{\boldsymbol{f}_0,i}, p_{\boldsymbol{f},i}) = \sigma^{-2}(\boldsymbol{f}_0(t_i) - \boldsymbol{f}(t_i))^2$ and $V_{m,0}(p_{\boldsymbol{f}_0,i}, p_{\boldsymbol{f},i}) = \sigma_m\sigma^{m-2}|\boldsymbol{f}_0(t_i) - \boldsymbol{f}(t_i)|^2$ for $m \geq 2$, where σ_m is the (non centred) $m^{\text{th}}-$moment of a standard Gaussian variable. So using Equation (2.15) we get

$$2K(p^{(n)}_{\boldsymbol{f}_0}, p^{(n)}_{\boldsymbol{f}}) = V_{2,0}(p^{(n)}_{\boldsymbol{f}_0}, p^{(n)}_{\boldsymbol{f}}) = n\sigma^{-2}\|\boldsymbol{f}_0 - \boldsymbol{f}\|^2_{\mathbb{P}^n_t} = n\sigma^{-2}\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}\|_2^2$$

which proves Condition (2.13).

Additionally, both $2\widetilde{K}(p^{(n)}_{\boldsymbol{f}_{0j_n}}, p^{(n)}_{\boldsymbol{f}})$ and $\widetilde{V}_{2,0}(p^{(n)}_{\boldsymbol{f}_{0j_n}}, p^{(n)}_{\boldsymbol{f}})$ are upper bounded by $n\sigma^{-2}(2\|\boldsymbol{f}_{0j_n} - \boldsymbol{f}\|^2_{\mathbb{P}^n_t} + \|\boldsymbol{f}_0 - \boldsymbol{f}_{0j_n}\|^2_{\mathbb{P}^n_t})$. Let $\boldsymbol{\theta} \in \mathcal{A}_n(H_1)$, for a certain $H_1 > 0$. Then, using (2.15) again, the bound is less than

$$n\sigma^{-2}(n^{-H_1} + L_0 j_n^{-2\beta}) \leq Cn\epsilon_n^2$$

for $H_1 > 2\beta/(2\beta+1)$, which ensures Condition $\boldsymbol{A_2}$.

Ghosal and van der Vaart (2007) state in Section 7.7 that tests satisfying $\boldsymbol{A_4}$ exist with $d_n = \|\cdot\|_{\mathbb{P}^n_t}$.

Nonlinear AR(1) model

As a nonindependent illustration, we consider the following Markov chain: the nonlinear autoregression model whose observations $X^n = (X_1, \ldots, X_n)$ come from a stationary time series $X_t, t \in \mathbb{Z}$, such that

$$X_i = \boldsymbol{f}(X_{i-1}) + \xi_i, \quad i = 1, 2, \ldots, n,$$

where the function $\boldsymbol{f}$ is unknown and the residuals ξ_i are standard Gaussian and independent of $(X_1, \dots, X_{i-1})$. We suppose that X_0 is drawn in the stationary distribution.

Suppose that regression functions $\boldsymbol{f}$ are in $L_2(\mathbb{R})$, and uniformly bounded by a constant M_1 (a bound growing with n could also be considered here). We use Hermite functions $\boldsymbol{\psi} = (\psi_j)_{j\geq 1}$ as an orthonormal basis of $\mathbb{R}$, such that for all $x \in \mathbb{R}$, $\boldsymbol{f}(x) = \boldsymbol{f}_{\boldsymbol{\theta}}(x) = \sum_{j=1}^{\infty} \theta_j \psi_j(x)$. This basis is uniformly bounded (by Cramér's inequality). Consider the sieve prior Π in its truncated version (2.3) for $\boldsymbol{\theta}$, with radius r_1 a (possibly large) constant independent of k and n.

We show that Conditions $\boldsymbol{A_1}$-$\boldsymbol{A_4}$ are satisfied, along the lines of Ghosal and van der Vaart (2007) Sections 4 and 7.4. Denote $p_{\boldsymbol{\theta}}(y|x) = \varphi(y - \boldsymbol{f}_{\boldsymbol{\theta}}(x))$ the transition density of the chain, where $\varphi(\,\cdot\,)$ is the standard normal density distribution, and where reference measures relative to x and y are denoted respectively by ν and μ. Define $r(y) = \frac{1}{2}(\varphi(y - M_1) + \varphi(y + M_1))$, and set $d\nu = r d\mu$. Then Ghosal and van der Vaart (2007) show that the chain $(X_i)_{1\leq i\leq n}$ has a unique stationary distribution and prove the existence of tests satisfying $\boldsymbol{A_4}$ relative to the Hellinger semidistance d whose square is given by

$$d^2(\boldsymbol{\theta}, \boldsymbol{\theta'}) = \int\int \left(\sqrt{p_{\boldsymbol{\theta}}(y|x)} - \sqrt{p_{\boldsymbol{\theta'}}(y|x)}\right)^2 d\mu(y) d\nu(x).$$

They show that d is bounded by $\|\cdot\|_2$ (which proves Condition $\boldsymbol{A_3}$) and that

$$K(p_0, p_{\boldsymbol{\theta}}) = V_{2,0}(p_0, p_{\boldsymbol{\theta}}) \lesssim \|\boldsymbol{\theta}_0 - \boldsymbol{\theta}\|_2^2.$$

Thus Equation (2.13) holds. Condition $\boldsymbol{A_2}$ follows from inequalities $\widetilde{K}(p_{0j_n}, p_{\boldsymbol{\theta}}) \lesssim \sum_{j=1}^{j_n} |\theta_{0j} - \theta_j|$ and $\widetilde{V}_{2,0}(p_{0j_n}, p_{\boldsymbol{\theta}}) \lesssim \|\boldsymbol{\theta}_{0j_n} - \boldsymbol{\theta}\|_{2,j_n}^2$ for $\boldsymbol{\theta} \in \Theta_{j_n}$.

2.3 Application to the white noise model

Consider the Gaussian white noise model

$$dX^n(t) = \boldsymbol{f}_0(t)dt + \frac{1}{\sqrt{n}} dW(t), \quad 0 \leq t \leq 1, \tag{2.16}$$

in which we observe processes $X^n(t)$, where $\boldsymbol{f}_0$ is the unknown function of interest belonging to $L^2(0,1)$, $W(t)$ is a standard Brownian motion, and n is the sample size. We assume that $\boldsymbol{f}_0$ lies in a Sobolev ball, $\Theta_\beta(L_0)$, see (2.11). Brown and Low (1996) show that this model is asymptotically equivalent to the nonparametric regression (assuming $\beta > 1/2$). It can be written as the equivalent infinite normal mean model using the

decomposition in a Fourier basis $\boldsymbol{\psi} = (\psi_j)_{j\geq 1}$ of $L^2(0,1)$,

$$X_j^n = \theta_{0j} + \frac{1}{\sqrt{n}}\xi_j,\ j = 1, 2, \ldots \tag{2.17}$$

where $X_j^n = \int_0^1 \psi_j(t)\, dX^n(t)$ are the observations, $\theta_{0j} = \int_0^1 \psi_j(t)\, \boldsymbol{f}_0(t)dt$ the Fourier coefficients of $\boldsymbol{f}_0$, and $\xi_j = \int_0^1 \psi_j(t)dW(t)$ are independent standard Gaussian random variables. The function $\boldsymbol{f}_0$ and the parameter $\boldsymbol{\theta}_0$ are linked through the relation in $L^2(0,1)$, $\boldsymbol{f}_0 = \sum_{j=1}^{\infty} \theta_{0j}\psi_j$.

In addition to results in concentration, we are interested in comparing the risk of an estimate $\hat{\boldsymbol{f}}_n$ corresponding to basis coefficients $\hat{\boldsymbol{\theta}}_n$, under two different losses: the global L^2 loss (if expressed on curves $\boldsymbol{f}$, or l^2 loss if expressed on $\boldsymbol{\theta}$),

$$R_n^{L^2}(\boldsymbol{\theta}_0) = \mathsf{E}_0^{(n)} \left\| \hat{\boldsymbol{f}}_n - \boldsymbol{f}_0 \right\|_2^2 = \mathsf{E}_0^{(n)} \sum_{j=1}^{\infty} \left(\hat{\theta}_{nj} - \theta_{0j} \right)^2,$$

and the local loss at point $t \in [0,1]$,

$$R_n^{\text{loc}}(\boldsymbol{\theta}_0, t) = \mathsf{E}_0^{(n)} \left(\hat{\boldsymbol{f}}_n(t) - \boldsymbol{f}_0(t) \right)^2 = \mathsf{E}_0^{(n)} \left(\sum_{j=1}^{\infty} a_j \left(\hat{\theta}_{nj} - \theta_{0j} \right) \right)^2,$$

with $a_j = \psi_j(t)$. Note that the difference between global and local risks expressions in basis coefficients comes from the parenthesis position with respect to the square: respectively the sum of squares and the square of a sum.

We show that sieve priors allow to construct adaptive estimate in global risk. However, the same estimate does not perform as well under the pointwise loss, which illustrates the result of Cai et al. (2007). We provide a lower bound for the pointwise rate.

2.3.1 Adaptation under global loss

Consider the global l^2 loss on $\boldsymbol{\theta}_0$. The likelihood ratio is given by

$$\frac{p_0^{(n)}}{p_{\boldsymbol{\theta}}^{(n)}}(X^n) = \exp\left(n\langle \boldsymbol{\theta}_0 - \boldsymbol{\theta}, X^n \rangle - \frac{n}{2} \|\boldsymbol{\theta}_0\|_2^2 + \frac{n}{2} \|\boldsymbol{\theta}\|_2^2 \right),$$

where $\langle ., . \rangle$ denotes the l^2 scalar product. We choose here the l^2 distance as $d_n(\boldsymbol{\theta}, \boldsymbol{\theta}') = \left\| \boldsymbol{\theta} - \boldsymbol{\theta}' \right\|_2$. Let us check that Conditions $\boldsymbol{A_2}$ - $\boldsymbol{A_4}$ and Condition (2.13) hold.

The choice of d_n ensures Condition $\boldsymbol{A_3}$ with $D_0 = D_2 = 1$ and $D_1 = 0$. The test statistic of $\boldsymbol{\theta}_0$ against $\boldsymbol{\theta}_1$ associated with the likelihood ratio is $\phi_n(\boldsymbol{\theta}_1) = \mathsf{I}(2\langle \boldsymbol{\theta}_1 - \boldsymbol{\theta}_0, X^n \rangle > \|\boldsymbol{\theta}_1\|_2^2 - \|\boldsymbol{\theta}_0\|_2^2)$. With Lemma 5 of Ghosal and van der Vaart (2007) we have that

$\mathsf{E}_0^{(n)}(\phi_n(\boldsymbol{\theta}_1)) \leq e^{-n\|\boldsymbol{\theta}_1-\boldsymbol{\theta}_0\|_2^2/4}$ and $\mathsf{E}_{\boldsymbol{\theta}}^{(n)}(1-\phi_n(\boldsymbol{\theta}_1)) \leq e^{-n\|\boldsymbol{\theta}_1-\boldsymbol{\theta}_0\|_2^2/4}$ for $\boldsymbol{\theta}$ such that $\|\boldsymbol{\theta}-\boldsymbol{\theta}_1\|_2 \leq \|\boldsymbol{\theta}_1-\boldsymbol{\theta}_0\|_2/4$. It provides a test as in Condition $\boldsymbol{A_4}$ with $c_1 = \zeta = 1/4$.

Moreover, following Lemma 6 of Ghosal and van der Vaart (2007) we have

$$K\left(p_0^{(n)}, p_{\boldsymbol{\theta}}^{(n)}\right) = n\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|_2^2/2 \text{ and } V_{2,0}\left(p_0^{(n)}, p_{\boldsymbol{\theta}}^{(n)}\right) = n\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|_2^2.$$

For $m \geq 2$, we have

$$\begin{aligned} V_{m,0}\left(p_0^{(n)}, p_{\boldsymbol{\theta}}^{(n)}\right) &= \int p_0^{(n)} \left|\log\left(p_0^{(n)}/p_{\boldsymbol{\theta}}^{(n)}\right) - K\left(p_0^{(n)}, p_{\boldsymbol{\theta}}^{(n)}\right)\right|^m d\mu \\ &= n^m \int p_0^{(n)} |\langle \boldsymbol{\theta}_0-\boldsymbol{\theta}, X^n-\boldsymbol{\theta}_0\rangle|^m d\mu \\ &\leq n^m \|\boldsymbol{\theta}_0-\boldsymbol{\theta}\|_2^m \int p_0^{(n)} \|X^n-\boldsymbol{\theta}_0\|_2^m d\mu. \end{aligned}$$

The centred $m^{\text{th}}-$moment of the Gaussian variable X^n is proportional to $n^{-m/2}$, so $V_{m,0}\left(p_0^{(n)}, p_{\boldsymbol{\theta}}^{(n)}\right) \lesssim n^{m/2}\|\boldsymbol{\theta}_0-\boldsymbol{\theta}\|_2^m$, and Condition (2.13) is satisfied. The same calculation shows that Condition $\boldsymbol{A_2'}$ is satisfied: for all $\boldsymbol{\theta} \in \Theta_{j_n}$, $\widetilde{K}\left(p_{0j_n}^{(n)}, p_{\boldsymbol{\theta}}^{(n)}\right) = \frac{n}{2}\|\boldsymbol{\theta}_{0j_n}-\boldsymbol{\theta}\|_{2,j_n}^2$ and $\widetilde{V}_{m,0}\left(p_{0j_n}^{(n)}, p_{\boldsymbol{\theta}}^{(n)}\right) \lesssim n^{m/2}\|\boldsymbol{\theta}_{0j_n}-\boldsymbol{\theta}\|_{2,j_n}^m$.

Conditions $\boldsymbol{A_2}$ - $\boldsymbol{A_4}$ and Condition (2.13) hold, if moreover $\boldsymbol{A_4}$ is satisfied, then by Proposition 2.6, the procedure is adaptive, which is expressed in the following proposition.

Proposition 2.8. *Under the prior* Π *defined in Equations (2.14), the global* l^2 *rate of posterior contraction is optimal adaptive for the Gaussian white noise model, i.e. for* M *large enough and* $\beta_2 \geq \beta_1 > 1/2$

$$\sup_{\beta_1\leq\beta\leq\beta_2} \sup_{\boldsymbol{\theta}_0\in\Theta_\beta(L_0)} \mathsf{E}_0^{(n)}\Pi\left(\boldsymbol{\theta} : \|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|_2^2 \geq M\frac{\log n}{L(n)}\epsilon_n^2(\beta)|X^n\right) \underset{n\to\infty}{\longrightarrow} 0,$$

with $\epsilon_n(\beta) = \epsilon_0\left(\frac{\log n}{n}\right)^{\frac{\beta}{2\beta+1}}$.

The distance here is not bounded, so Corollary 2.3 does not hold. For deriving a risk rate, we need a more subtle result than Theorem 2.2 that we can obtain when considering sets $\mathcal{S}_{n,j}(M) = \left\{\boldsymbol{\theta} : M\frac{\log n}{L(n)}(j+1)\epsilon_n^2 \geq \|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|_2^2 \geq M\frac{\log n}{L(n)}j\epsilon_n^2\right\}$, $j = 1, 2, \ldots$ instead of $\mathcal{S}_n(M) = \left\{\boldsymbol{\theta} : \|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|_2^2 \geq M\frac{\log n}{L(n)}\epsilon_n^2\right\}$. Then the bound of the expected posterior mass of $\mathcal{S}_{n,j}(M)$ becomes

$$\mathsf{E}_0^{(n)}\Pi(\mathcal{S}_{n,j}(M)|X^n) \leq c_7\left(nj\epsilon_n^2\right)^{-m/2} \tag{2.18}$$

for a fixed constant c_7. Hence we obtain the following rate of convergence in risk.

Proposition 2.9. *Under Condition (2.13) with $m \geq 5$, the expected posterior risk given $\boldsymbol{\theta}_0$ and Π converges at least at the same rate ϵ_n*

$$\mathcal{R}_n^{L^2}(\boldsymbol{\theta}_0) = \mathsf{E}_0^{(n)}\Pi\left[\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_2^2 \,|X^n\right] = \mathcal{O}\left(\epsilon_n^2\right),$$

for any $\boldsymbol{\theta}_0$. So the procedure is risk adaptive as well (up to a $\log(n)$ term).

Proof. We have

$$\begin{aligned}\mathcal{R}_n^{L^2}(\boldsymbol{\theta}_0) \leq & \quad \mathsf{E}_0^{(n)}\Pi\left[\left(\mathsf{I}(\boldsymbol{\theta} \notin \mathcal{S}_n(M)) + \sum_{j\geq 1}\mathsf{I}(\boldsymbol{\theta} \in \mathcal{S}_{n,j}(M))\right)\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_2^2 \,|X^n\right] \\ \leq & \quad M\frac{\log n}{L(n)}\epsilon_n^2\left(1 + \sum_{j=1}^{\infty}(j+1)\mathsf{E}_0^{(n)}\Pi(\mathcal{S}_{n,j}(M)|X^n)\right).\end{aligned}$$

Due to (2.18), the last sum in j converges as soon as $m \geq 5$. This is possible in the white noise setting because the conditions are satisfied whatever m. So $\mathcal{R}_n^{L^2}(\boldsymbol{\theta}_0) = \mathcal{O}\left(\epsilon_n^2\right)$. □

We have shown that conditional to the existence of a sieve prior for the white noise model satisfying $\boldsymbol{A_5}$ (*cf.* Section 2.2.3), the procedure has minimax rates (up to a $\log(n)$ term) both in contraction and in risk. We now study the asymptotic behaviour of the posterior under the local loss function.

2.3.2 Lower bound under pointwise loss

The previous section derives rates of convergence under the global loss. Here, under the pointwise loss, we show that the risk deteriorates as a power n factor compared to the benchmark minimax pointwise risk $n^{-(2\beta-1)/2\beta}$ (note the difference with the global minimax rate $n^{-2\beta/(2\beta+1)}$, both given for risks on squares). We use the sieve prior defined as a conditional Gaussian prior in Equation (2.14). Denote by $\hat{\boldsymbol{\theta}}_n$ the Bayes estimate of $\boldsymbol{\theta}$ (the posterior mean). Then the following proposition gives a lower bound on the risk (pointwise square error) under a pointwise loss:

Proposition 2.10. *If the point t is such that $a_j = \psi_j(t) = 1$ for all j $(t = 0)$, then for all $\beta \geq q$, for all $L_0 > 0$, a lower bound on the risk rate under pointwise loss is given by*

$$\sup_{\boldsymbol{\theta}_0 \in \Theta_\beta(L_0)} R_n^{\mathrm{loc}}(\boldsymbol{\theta_0}, t) \gtrsim n^{-\frac{2\beta-1}{2\beta+1}} / \log^2 n.$$

Proof. See the Appendix. □

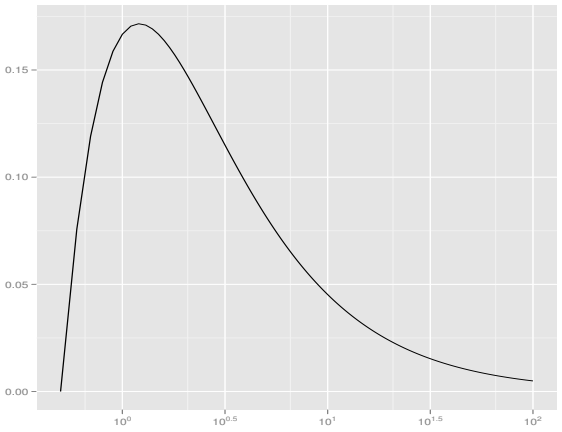

FIGURE 2.2: Variation of the exponent of the penalty in a log scale for β between 1/2 and 100; it is maximum for $\beta = (1+\sqrt{2})/2$

Cai et al. (2007) show that a global optimal estimator cannot be pointwise optimal. The sieve prior leads to an (almost up to a $\log n$ term) optimal global risk and Proposition 2.10 shows that the pointwise risk associated to the posterior mean $\hat{\boldsymbol{\theta}}_n$ is suboptimal with a power of n penalty, whose exponent is

$$\frac{2\beta-1}{2\beta} - \frac{2\beta-1}{2\beta+1} = \frac{2\beta-1}{2\beta(2\beta+1)}.$$

The maximal penalty is for $\beta = (1+\sqrt{2})/2$, and it vanishes as β tends to 1/2 and $+\infty$ (see the Figure 2.2). Abramovich et al. (2007a) also derive such a power n penalty on the maximum local risk of a globally optimal Bayesian estimate, as well as on the reverse case (maximum global risk of a locally optimal Bayesian estimate).

Remark 2.11. *This result is not anecdotal and illustrates the fact that the Bayesian approach is well suited for loss functions that are related to the Kullback-Leibler divergence (i.e. often the l^2 loss). The pointwise loss does not satisfy this since it corresponds to a non smooth linear functional of $\boldsymbol{\theta}$. This possible suboptimality of the posterior distribution of some non smooth functional of the parameter has already been noticed in various other cases, see for instance Rivoirard and Rousseau (2012b) or Rousseau and Kruijer (2011). The question of the existence of a fully Bayesian adaptive procedure to estimate $\boldsymbol{f}_0(t) = \sum_{j=1}^{\infty} a_j \theta_{0j}$ remains an open question.*

Acknowledgements

We would like to thank the referees for their valuable comments which have helped to improve the manuscript.

Bibliography

Abramovich, F., Amato, U., and Angelini, C. (2004). On optimality of Bayesian wavelet estimators. *Scand. J. Stat.*, 31(2):217–234. 49

Abramovich, F., Angelini, C., and De Canditiis, D. (2007a). Pointwise optimality of Bayesian wavelet estimators. *Ann. Inst. Statist. Math.*, 59(3):425–434. 49, 64

Abramovich, F., Grinshtein, V., and Pensky, M. (2007b). On optimality of Bayesian testimation in the normal means problem. *Ann. Statist.*, 35(5):2261–2286. 55

Abramovich, F., Grinshtein, V., Petsa, A., and Sapatinas, T. (2010). On Bayesian testimation and its application to wavelet thresholding. *Biometrika*, 97(1):181–198. 55

Abramovich, F., Sapatinas, T., and Silverman, B. (1998). Wavelet thresholding via a Bayesian approach. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 60(4):725–749. 46

Arbel, J., Gayraud, G., and Rousseau, J. (2013). Bayesian optimal adaptive estimation using a sieve prior. *Scandinavian Journal of Statistics*. 46

Babenko, A. and Belitser, E. (2009). On the posterior pointwise convergence rate of a Gaussian signal under a conjugate prior. *Statist. Probab. Lett.*, 79(5):670–675. 53

Babenko, A. and Belitser, E. (2010). Oracle convergence rate of posterior under projection prior and Bayesian model selection. *Math. Methods Statist.*, 19(3):219–245. 53

Barron, A., Schervish, M., and Wasserman, L. (1999). The consistency of posterior distributions in nonparametric problems. *Ann. Statist.*, 17(2):536–561. 46

Belitser, E. and Ghosal, S. (2003). Adaptive Bayesian inference on the mean of an infinite-dimensional normal distribution. *Ann. Statist.*, 31(2):536–559. 49

Brown, L. and Low, M. (1996). Asymptotic Equivalence of Nonparametric Regression and White Noise. *Ann. Statist.*, 24(6):2384–2398. 60

Cai, T., Low, M., and Zhao, L. (2007). Trade-offs between global and local risks in nonparametric function estimation. *Bernoulli*, 13(1):1–19. 49, 61, 64

De Jonge, R. and van Zanten, J. (2010). Adaptive nonparametric Bayesian inference using location-scale mixture priors. *Ann. Statist.*, 38(6):3300–3320. 49

Donoho, D. and Johnstone, I. (1998). Minimax estimation via wavelet shrinkage. *Ann. Statist.*, 26(3):879–921. 55

Ghosal, S., Ghosh, J. K., and van der Vaart, A. W. (2000). Convergence rates of posterior distributions. *Ann. Statist.*, 28(2):500–531. 46, 48, 51

Ghosal, S., Lember, J., and van der Vaart, A. W. (2008). Nonparametric Bayesian model selection and averaging. *Electron. J. Stat.*, 2:63–89. 49

Ghosal, S. and van der Vaart, A. W. (2007). Convergence rates of posterior distributions for noniid observations. *Ann. Statist.*, 35(1):697–723. 47, 48, 51, 53, 58, 59, 60, 61, 62, 71

Huang, T. (2004). Convergence rates for posterior distributions and adaptive estimation. *Ann. Statist.*, 32(4):1556–1593. 49

Kruijer, W., Rousseau, J., and van der Vaart, A. W. (2010). Adaptive Bayesian density estimation with location-scale mixtures. *Electron. J. Stat.*, 4:1225–1257. 49

Rivoirard, V. and Rousseau, J. (2012a). Bernstein-von Mises theorem for linear functionals of the density. *Ann. Statist.*, 40(3):1489–1523. 49, 54

Rivoirard, V. and Rousseau, J. (2012b). Posterior concentration rates for infinite-dimensional exponential families. *Bayesian Anal.*, 7(2):311–334. 55, 58, 64

Robert, C. (2007). *The Bayesian choice: from decision-theoretic foundations to computational implementation*. Springer Verlag. 50

Rousseau, J. (2010). Rates of convergence for the posterior distributions of mixtures of betas and adaptive nonparametric estimation of the density. *Ann. Statist.*, 38(1):146–180. 49

Rousseau, J., Chopin, N., and Liseo, B. (2012). Bayesian nonparametric estimation of the spectral density of a long or intermediate memory Gaussian process. *Ann. Statist.*, 40(2):964–995. 46

Rousseau, J. and Kruijer, W. (2011). Adaptive Bayesian Estimation of a spectral density. *Preprint*. 49, 64

Scricciolo, C. (2006). Convergence rates for Bayesian density estimation of infinite-dimensional exponential families. *Ann. Statist.*, 34(6):2897–2920. 49

Shen, X. and Wasserman, L. (2001). Rates of convergence of posterior distributions. *Ann. Statist.*, 29(3):687–714. 47, 48, 51

Tsybakov, A. (2009). *Introduction to nonparametric estimation*. Springer Verlag. 59

van der Vaart, A. W. and van Zanten, J. (2008). Rates of contraction of posterior distributions based on Gaussian process priors. *Ann. Statist.*, 36(3):1435–1463. 46

van der Vaart, A. W. and van Zanten, J. (2009). Adaptive Bayesian estimation using a Gaussian random field with inverse Gamma bandwidth. *The Annals of Statistics*, 37(5B):2655–2675. 49

Zhao, L. (1993). *Frequentist and Bayesian aspects of some nonparametric estimation problems*. PhD thesis, Ph. D. thesis. Cornell University. 47

Zhao, L. (2000). Bayesian aspects of some nonparametric problems. *Ann. Statist.*, 28(2):532–552. 47, 72

2.4 Technical lemmas and proofs

2.4.1 Technical lemmas

Set $\mathcal{S}_n(M) = \{\boldsymbol{\theta} : d_n^2(\boldsymbol{\theta}, \boldsymbol{\theta}_0) \geq M \frac{\log n}{L(n)} \epsilon_n^2\}$ and recall that $\Theta_{k_n}(Q) = \{\boldsymbol{\theta} \in \Theta_{k_n} : \|\boldsymbol{\theta}\|_{2,k_n} \leq n^Q\}$, $Q > 0$. We begin with three technical lemmas.

Lemma 2.12. *If Conditions* $\boldsymbol{A_3}$ *and* $\boldsymbol{A_4}$ *hold, then there exists a test* $\boldsymbol{\phi}_n$ *such that for* M *large enough, there exists a constant* c_2 *such that*

$$\mathsf{E}_0^{(n)}(\boldsymbol{\phi}_n) \leq e^{-c_2 M \frac{\log n}{L(n)} n\epsilon_n^2} \quad \textit{and} \quad \mathsf{E}_{\boldsymbol{\theta}}^{(n)}(1 - \boldsymbol{\phi}_n) \leq e^{-c_2 M \frac{\log n}{L(n)} n\epsilon_n^2},$$

for all $\boldsymbol{\theta} \in \mathcal{S}_n(M) \cap \Theta_{k_n}(Q)$.

Proof. Set $r_n = \left(\sqrt{M \frac{\log n}{L(n)}} \frac{\zeta \epsilon_n}{D_0 k_n^{D_1}}\right)^{1/D_2}$. The set $\mathcal{S}_n(M) \cap \Theta_{k_n}(Q)$ is compact relative to the l^2 norm. Let a covering of this set by l^2 balls of radius r_n and centre $\boldsymbol{\theta}^{(i)}$. Its number of elements is $\eta_n \lesssim (Cn^Q/r_n)^{k_n} \lesssim \exp(Ck_n \log n) \lesssim \exp(C \frac{\log n}{L(n)} n\epsilon_n^2)$ due to relation (2.5).

For each centre $\boldsymbol{\theta}^{(i)} \in \mathcal{S}_n(M) \cap \Theta_{k_n}(Q)$, there exists a test $\phi_n(\boldsymbol{\theta}^{(i)})$ satisfying Condition $\boldsymbol{A_4}$. We define the test $\boldsymbol{\phi}_n = \max_i \phi_n(\boldsymbol{\theta}^{(i)})$ which satisfies

$$\mathsf{E}_0^{(n)}(\boldsymbol{\phi}_n) \leq \eta_n e^{-c_1 M \frac{\log n}{L(n)} n\epsilon_n^2} \leq e^{C \frac{\log n}{L(n)} n\epsilon_n^2 - c_1 M \frac{\log n}{L(n)} n\epsilon_n^2} \leq e^{-c_2 M \frac{\log n}{L(n)} n\epsilon_n^2},$$

for M large enough and a constant c_2.

Here, Condition $\boldsymbol{A_3}$ allows to switch from the coverage in term of the l^2 distance to a covering expressed in term of d_n: each $\boldsymbol{\theta} \in \mathcal{S}_n(M) \cap \Theta_{k_n}(Q)$ which lies in a l^2 ball of centre $\boldsymbol{\theta}^{(i)}$ and of radius r_n in the covering of size η_n also lies in a d_n ball of adequate radius

$$d_n(\boldsymbol{\theta}, \boldsymbol{\theta}^{(i)}) \leq D_0 k_n^{D_1} \|\boldsymbol{\theta} - \boldsymbol{\theta}^{(i)}\|_2^{D_2} \leq D_0 k_n^{D_1} r_n^{D_2} = \zeta \epsilon_n \sqrt{M \frac{\log n}{L(n)}}.$$

Then there exists a constant c_2 (the minimum with the previous one)

$$\sup_{\boldsymbol{\theta}\in\mathcal{S}_n(M)\cap\Theta_{k_n}(Q)} \mathsf{E}_{\boldsymbol{\theta}}^{(n)}\left(1-\boldsymbol{\phi}_n\right) \leq e^{-c_2 M \frac{\log n}{L(n)} n\epsilon^2},$$

hence the result follows. □

Lemma 2.13. *Under Condition* $\boldsymbol{A_5}$*, for any constant* $c_6 > 0$*, there exist positive constants* Q*,* C *and* M_0 *such that*

$$\Pi(\Theta_{k_n}^c(Q)) \leq Ce^{-c_6 n\epsilon_n^2}, \tag{2.19}$$

where M_0 *is introduced in the definition (2.5) of* k_n*, and* $\Theta_{k_n}^c(Q)$*, the complementary of* $\Theta_{k_n}(Q)$*, is taken in* Θ*.*

Proof. $\Theta_{k_n}^c(Q)$ is written by $\Theta_{k_n}^c(Q) = \{\boldsymbol{\theta}\in\Theta : \|\boldsymbol{\theta}\|_{2,k_n} > n^Q \text{ or } \exists j > k_n \text{ s.t. } \theta_j \neq 0\}$, so its prior mass is less than $\pi(k > k_n) + \sum_{k\leq k_n} \pi_k \Pi_k(\boldsymbol{\theta}\in\Theta_k : \|\boldsymbol{\theta}\|_{2,k} > n^Q)$, where the last sum is less than $\Pi_{k_n}(\boldsymbol{\theta}\in\Theta_{k_n} : \|\boldsymbol{\theta}\|_{2,k_n} > n^Q)$ because its terms are increasing.

The prior mass of sieves that exceed k_n is controlled by Equation (2.6). We have

$$\pi\left(k \geq k_n\right) \leq \sum_{j\geq k_n} e^{-bjL(j)} \leq \sum_{j\geq k_n} e^{-bjL(k_n)} \leq Ce^{-bk_nL(k_n)}.$$

Since L is a slow varying function, we have $k_nL(k_n) \gtrsim j_n \log(n) \gtrsim n\epsilon_n^2$. Hence $\pi\left(k \geq k_n\right) \leq Ce^{-c_6 n\epsilon_n^2}$ for a constant c_6 as large as needed since it is determined by constant M_0 in Equation (2.5).

Then by the second part of Condition (2.7), $\Pi_{k_n}\left(\boldsymbol{\theta}\in\Theta_{k_n} : \|\boldsymbol{\theta}\|_{2,k_n} > n^Q\right)$ is less than

$$\begin{aligned} & \int_{\|\boldsymbol{\theta}\|_{2,k_n}>n^Q} \prod_{j=1}^{k_n} g(\theta_j/\tau_j)/\tau_j d\theta_j, \\ \leq\ & (G_3 n^{H_2})^{k_n} \int_{\|\boldsymbol{\theta}\|_{2,k_n}>n^Q} \exp(-G_4 \sum_{j=1}^{k_n} |\theta_j|^\alpha/\tau_j^\alpha)\, d\theta_i, \end{aligned} \tag{2.20}$$

by using the lower bound on the τ_j's of Equation (2.9).

If $\alpha \geq 2$, then applying Hölder inequality, one obtains

$$n^{2Q} \leq \|\boldsymbol{\theta}\|_{2,k_n}^2 \leq \|\boldsymbol{\theta}\|_{\alpha,k_n}^2 k_n^{1-2/\alpha},$$

which leads to

$$\|\boldsymbol{\theta}\|_{\alpha,k_n}^\alpha \geq k_n^{1-\alpha/2} n^{Q\alpha}.$$

If $\alpha < 2$, then a classical result states that the l^α norm $\|\,.\,\|_\alpha$ is larger than the l^2 norm $\|\,.\,\|_2$, *i.e.*

$$\|\boldsymbol{\theta}\|_{\alpha,k_n}^\alpha \geq \|\boldsymbol{\theta}\|_{2,k_n}^\alpha \geq n^{Q\alpha}.$$

Eventually the upper bound τ_0 on the τ_j's of Equation (2.8) provides

$$\sum_{j=1}^{k_n} |\theta_j|^\alpha / \tau_j^\alpha \geq \tau_0^{-\alpha} n^{Q\alpha} \min(k_n^{1-\alpha/2}, 1).$$

The integral in the right-hand side of (2.20) is bounded by

$$\exp(-\frac{G_4}{2}\|\boldsymbol{\theta}\|_{2,k_n}^\alpha/\tau_0^\alpha) \int_{\Theta_{k_n}} \exp(-\frac{G_4}{2}\sum_{j=1}^{k_n} |\theta_j|^\alpha/\tau_j^\alpha)\, d\theta_i.$$

The last integral is bounded by C^{k_n}, so

$$\Pi_{k_n}\left(\boldsymbol{\theta} \in \Theta_{k_n} : \|\boldsymbol{\theta}\|_{2,k_n} > n^Q\right) \leq C^{k_n \log n} \exp(-\frac{G_4}{2}\tau_0^{-\alpha} n^{Q\alpha} \min(k_n^{1-\alpha/2}, 1)).$$

The right-hand side of the last inequality can be made smaller than $Ce^{-c_6 n\epsilon_n^2}$ for any constant C and c_6 provided that Q is chosen large enough. This entails result (2.19).

In the truncated case (2.3), we note that if $\sum_{j=1}^{k_n} |\theta_j| \leq r_1$, then $\sum_{j=1}^{k_n} \theta_j^2 \leq r_1^2$, so that for n large enough, $\Pi(\Theta_{k_n}^c(Q)) = \pi(k \geq k_n)$, and the rest of the proof is similar. □

Lemma 2.14. *Under Conditions* $\boldsymbol{A_1}$, $\boldsymbol{A_2}$ *and* $\boldsymbol{A_5}$, *there exists* $c_4 > 0$ *such that*

$$\Pi(\mathcal{B}_n(m)) \geq e^{-c_4 n\epsilon_n^2}.$$

Proof. Let $\boldsymbol{\theta} \in \mathcal{A}_n(H_1)$. For n large enough, Conditions $\boldsymbol{A_1}$ and $\boldsymbol{A_2}$ imply that

$$K(p_0^{(n)}, p_{\boldsymbol{\theta}}^{(n)}) \leq K(p_0^{(n)}, p_{0j_n}^{(n)}) + \widetilde{K}(p_{0j_n}^{(n)}, p_{\boldsymbol{\theta}}^{(n)}) \leq 2n\epsilon_n^2,$$

and

$$\begin{aligned} V_{m,0}(p_0^{(n)}, p_{\boldsymbol{\theta}}^{(n)}) = & \int p_0^{(n)} \left|\log(p_0^{(n)}/p_{0j_n}^{(n)}) - K(p_0^{(n)}, p_{0j_n}^{(n)}) + \right. \\ & \left. \log(p_{0j_n}^{(n)}/p_{\boldsymbol{\theta}}^{(n)}) - \int p_0^{(n)} \log(p_{0j_n}^{(n)}/p_{\boldsymbol{\theta}}^{(n)}) d\mu\right|^m d\mu \\ \leq & \quad 2^m (V_{m,0}(p_0^{(n)}, p_{0j_n}^{(n)}) + \widetilde{V}_{m,0}(p_{0j_n}^{(n)}, p_{\boldsymbol{\theta}}^{(n)})) \leq 2^{m+1} \left(n\epsilon_n^2\right)^{\frac{m}{2}}, \end{aligned}$$

which yields $\mathcal{A}_n(H_1) \subset \mathcal{B}_n(m)$ so that a lower bound for $\Pi(\mathcal{B}_n(m))$ is given by $\Pi(\mathcal{A}_n(H_1))$. Note that for $H_0 > H_1$, then

$$\mathcal{A}_n(H_0) \subset \mathcal{A}_n(H_1) \subset \mathcal{B}_n(m). \tag{2.21}$$

We have

$$\Pi(\mathcal{A}_n(H_1)) = \sum_{k=1}^{\infty} \pi(k)\Pi_k(\mathcal{A}_n(H_1)) \geq \pi(j_n)\Pi_{j_n}(\mathcal{A}_n(H_1)).$$

By the first part of Condition (2.6) we have

$$\pi(j_n) \geq e^{-j_n L(j_n)} \geq e^{-\frac{c_4}{2} n\epsilon_n^2}, \tag{2.22}$$

for c_4 large enough. Now by the first part of Condition (2.7) and by Condition (2.8)

$$\begin{aligned} \Pi_{j_n}(\mathcal{A}_n(H_1)) = & \int_{\|\boldsymbol{\theta}-\boldsymbol{\theta}_{0j_n}\|_{2,j_n} \leq n^{-H_1}} \prod_{j=1}^{j_n} g(\theta_j/\tau_j)/\tau_j d\theta_j \qquad (2.23) \\ \geq & (G_1/\tau_0)^{j_n} \int_{\|\boldsymbol{\theta}-\boldsymbol{\theta}_{0j_n}\|_{2,j_n} \leq n^{-H_1}} \exp(-G_2 \sum_{j=1}^{j_n} |\theta_j|^\alpha/\tau_j^\alpha) d\theta_j. \end{aligned}$$

We can bound above $\tau_j^{-\alpha}$ by $n^{\alpha H_2}$ by Equation (2.9) as $j \leq j_n \leq k_n$. We write $|\theta_j|^\alpha \leq 2^\alpha \left(|\theta_{0j}|^\alpha + |\theta_j - \theta_{0j}|^\alpha\right)$. First, Equation (2.10) gives

$$\sum_{j=1}^{j_n} |\theta_{0j}|^\alpha / \tau_j^\alpha \leq C j_n \log n.$$

Then, if $\alpha \geq 2$

$$\sum_{j=1}^{j_n} |\theta_j - \theta_{0j}|^\alpha \leq \|\boldsymbol{\theta} - \boldsymbol{\theta}_{0j_n}\|_{2,j_n}^\alpha \leq n^{-\alpha H_1},$$

and if $\alpha < 2$ then Hölder's inequality provides

$$\sum_{j=1}^{j_n} |\theta_j - \theta_{0j}|^\alpha \leq \|\boldsymbol{\theta} - \boldsymbol{\theta}_{0j_n}\|_{2,j_n}^\alpha \, j_n^{1-\alpha/2} \leq n^{-\alpha H_1} j_n^{1-\alpha/2}.$$

In both cases we have

$$\sum_{j=1}^{j_n} |\theta_j|^\alpha / \tau_j^\alpha \leq 2^\alpha (C j_n \log n + n^{\alpha(H_2 - H_1)} j_n^{1-\alpha/2}),$$

so choosing $H_2 \leq H_1$ ensures to bound the latter by $j_n \log n$. Last, the integral of the ball in dimension j_n, centered around $\boldsymbol{\theta}_{0j_n}$, and of radius n^{-H_1}, is at least equal to $e^{-C j_n \log n}$, for some given positive constant C.

Noting that $j_n = \lfloor j_0 n\epsilon_n^2/\log(n)\rfloor$ and choosing H_1 large enough, which is possible by Equation (2.21), ensures the existence of $c_4 > 0$ such that $\Pi_{j_n}(\mathcal{A}_n(H_1)) \geq e^{-\frac{c_4}{2}n\epsilon_n^2}$. Combining this with (2.22) allows to conclude.

In the truncated case (2.3), we can first choose r_1 larger than $2\sum_{j=1}^{j_n}|\theta_{0j}|$. If $\boldsymbol{\theta} \in \mathcal{A}_n(H_1)$, then $\sum_{j=1}^{j_n}|\theta_j| \leq \sum_{j=1}^{j_n}(|\theta_j - \theta_{0j}| + |\theta_{0j}|) \leq \sqrt{j_n}n^{-H_1} + r_1/2 \leq r_1$ for n and H_1 large enough. So the expression of integral (2.23) is still valid. □

2.4.2 Proof of Theorem 2.2

Proof. (of **Theorem 2.2**)

Express the quantity of interest $\Pi(\mathcal{S}_n(M)|X^n)$ in terms of N_n, $\widetilde{N_n}$ and D_n defined as follows

$$\frac{\int_{\mathcal{S}_n(M)\cap\Theta_{k_n}(Q)} p_{\boldsymbol{\theta}}^{(n)}/p_{\boldsymbol{\theta}_0}^{(n)} d\Pi(\boldsymbol{\theta}) + \int_{\mathcal{S}_n(M)\cap\Theta_{k_n}^c(Q)} p_{\boldsymbol{\theta}}^{(n)}/p_{\boldsymbol{\theta}_0}^{(n)} d\Pi(\boldsymbol{\theta})}{\int_\Theta p_{\boldsymbol{\theta}}^{(n)}/p_{\boldsymbol{\theta}_0}^{(n)} d\Pi(\boldsymbol{\theta})} := \frac{N_n + \widetilde{N_n}}{D_n}.$$

Denote $\rho_n(c_3) = \exp(-(c_3+1)n\epsilon_n^2)\Pi(\mathcal{B}_n(m))$ for $c_3 > 0$. Introduce $\boldsymbol{\phi}_n$ the test statistic of Lemma 2.12, and take the expectation of the posterior mass of $\mathcal{S}_n(M)$ as follows

$$\begin{aligned}
&\mathsf{E}_0^{(n)}\left(\frac{N_n + \widetilde{N_n}}{D_n}(\boldsymbol{\phi}_n + 1 - \boldsymbol{\phi}_n)(\mathsf{I}(D_n \leq \rho_n(c_3)) + \mathsf{I}(D_n > \rho_n(c_3)))\right) \\
&\leq \mathsf{E}_0^{(n)}(\boldsymbol{\phi}_n) + \mathsf{E}_0^{(n)}\left(\frac{N_n + \widetilde{N_n}}{D_n}(1 - \boldsymbol{\phi}_n)(\mathsf{I}(D_n \leq \rho_n(c_3)) + \mathsf{I}(D_n > \rho_n(c_3)))\right) \\
&\leq \mathsf{E}_0^{(n)}(\boldsymbol{\phi}_n) + \mathbf{p}_0^{(n)}(D_n \leq \rho_n(c_3)) + \frac{\mathsf{E}_0^{(n)}(N_n(1-\boldsymbol{\phi}_n)) + \mathsf{E}_0^{(n)}(\widetilde{N_n})}{\rho_n(c_3)}. \qquad (2.24)
\end{aligned}$$

Lemma 10 in Ghosal and van der Vaart (2007) gives $\mathbf{p}_0^{(n)}(D_n \leq \rho_n(c_3)) \lesssim (n\epsilon_n^2)^{-m/2}$ for every $c_3 > 0$.

Fubini's theorem entails that $\mathsf{E}_0^{(n)}(N_n(1-\boldsymbol{\phi}_n)) \leq \sup_{\mathcal{S}_n(M)\cap\Theta_{k_n}(Q)} \mathsf{E}_{\boldsymbol{\theta}}^{(n)}(1-\boldsymbol{\phi}_n)$. Along with $\mathsf{E}_0^{(n)}(\boldsymbol{\phi}_n)$, it is upper bounded in Lemma 2.12 by $e^{-c_2 M\frac{\log n}{L(n)}n\epsilon_n^2}$.

Lemma 2.13 implies that $\mathsf{E}_0^{(n)}(\widetilde{N_n}) \leq \Pi(\Theta_{k_n}^c(Q)) \leq e^{-c_6 n\epsilon_n^2}$ and Lemma 2.14 yields $\Pi_n(\mathcal{B}_n(m)) \geq e^{-c_4 n\epsilon_n^2}$. Constants c_3 and c_4 are fixed, so we can choose M, M_0 and Q large enough for c_6 to be sufficiently large (see proof of Lemma 2.13), such that $\min(M\frac{\log n}{L(n)}c_2, c_6) > c_3 + c_4 + 1$. It implies that the third term in Equation (2.24) is bounded above by $e^{-c_5 n\epsilon_n^2}$ for some positive c_5. Finally,

$$\mathsf{E}_0^{(n)}\Pi(\mathcal{S}_n(M)|X^n) = \mathcal{O}\left((n\epsilon_n^2)^{-m/2}\right) \underset{n\to\infty}{\longrightarrow} 0,$$

since $n\epsilon_n^2 \underset{n\to\infty}{\longrightarrow} \infty$. □

2.4.3 Proof of Proposition 2.10

The proof of the lower bound in the local risk case uses the next lemma, whose proof follows from Cauchy-Schwarz' inequality.

Lemma 2.15. *If* $\mathsf{E}(B_n^2) = o(\mathsf{E}(A_n^2))$, *then* $\mathsf{E}((A_n + B_n)^2) = \mathsf{E}(A_n^2)(1 + o(1))$.

Proof. (of **Proposition 2.10**)

The coordinates of $\hat{\boldsymbol{\theta}}_n$ are $\hat{\theta}_{nj} = \Pi(\theta_j|X^n) = \sum_{k=1}^{\infty} \pi(k|X^n)\widetilde{\theta}_{nj}(k)$, with $\widetilde{\theta}_{nj}(k) = \tau_j^2/(\tau_j^2 + \frac{1}{n})X_j^n$ if $k \geq j$, and $\widetilde{\theta}_{nj}(k) = 0$ otherwise (see Zhao, 2000).

Denote $u_j(X^n) = \sum_{k\geq j} \pi(k|X^n) = \pi(k \geq j|X^n)$, so that $\hat{\theta}_{nj} = u_j(X^n)\tau_j^2/(\tau_j^2 + \frac{1}{n})X_j^n$. Denote $K_n = n^{1/(2\beta+1)}$ and $J_n = n^{1/2\beta}$. Most of the posterior mass on k is concentrated before K_n, in the sense that there exists a constant c such that

$$\mathsf{E}_0^{(n)}\left(u_{K_n}(X^n)\right) \lesssim \exp\left(-cK_n\right). \tag{2.25}$$

This follows from the exponential inequality

$$P_{\boldsymbol{\theta}_0}^{(n)}[u_{K_n}(X^n) > \exp(-cK_n)] \lesssim \exp(-cK_n),$$

which is obtained by classic arguments in line with Theorem 2.2: writing the posterior quantity $u_{K_n}(X^n)$ as a ratio N_n/D_n, and then using Fubini's theorem, Chebyshev's inequality and an upper bound on $\pi(k > K_n)$.

Due to Relation (2.17), we split in three the sum in the risk

$$R_n^{\mathrm{loc}}(\boldsymbol{\theta}_0, t) = \mathsf{E}_0^{(n)}\left(\sum_{i=1}^{\infty} a_i[(1 - u_i(X^n)\tfrac{\tau_i^2}{\tau_i^2+\frac{1}{n}})\theta_{0i} - u_i(X^n)\tfrac{\tau_i^2}{\tau_i^2+\frac{1}{n}}\tfrac{\xi_i}{\sqrt{n}}]\right)^2$$

by centring the stochastic term X_i^n and writing $1 - u_i(X^n)\frac{\tau_i^2}{\tau_i^2+\frac{1}{n}} = \frac{1}{n}\frac{1}{\tau_i^2+\frac{1}{n}} + \frac{\tau_i^2}{\tau_i^2+\frac{1}{n}}(1 - u_i(X^n))$. The idea of the proof is to show that there is a leading term in the sum, and to apply Lemma 2.15.

Let $R_1 = \left(\sum_{i=1}^{\infty} a_i \frac{1}{n\tau_i^2+1}\theta_{0i}\right)^2$, $R_2 = \mathsf{E}_0^{(n)}\left(\sum_{i=1}^{\infty} a_i \frac{\tau_i^2}{\tau_i^2+\frac{1}{n}}(1-u_i(X^n))\theta_{0i}\right)^2$ and $R_3 = \mathsf{E}_0^{(n)}\left(\sum_{i=1}^{\infty} a_i \frac{\tau_i^2}{\tau_i^2+\frac{1}{n}}u_i(X^n)\frac{\xi_i}{\sqrt{n}}\right)^2$. By using Cauchy-Schwarz' inequality

$$\begin{aligned} R_1 = \left(\sum_{i=1}^{\infty} a_i \frac{1}{n\tau_i^2+1}\theta_{0i}\right)^2 = & \left(\sum_{i=1}^{\infty} a_i \frac{i^{-\beta}}{n\tau_i^2+1}\theta_{0i} i^{\beta}\right)^2 \\ \lesssim & L_0 \sum_{i=1}^{\infty} \frac{i^{-2\beta}}{(ni^{-2q}+1)^2}, \end{aligned}$$

because the a_i's are bounded. If $2\beta - 4q > 1$, then we can write

$$R_1 \lesssim \frac{1}{n^2}\sum_{i=1}^{\infty} i^{-2\beta+4q} \lesssim \frac{1}{n^2},$$

and if $2\beta - 4q \leq 1$, then comparing to an integral provides

$$\begin{aligned} R_1 \lesssim & \int_1^{\infty} \frac{x^{-2\beta}}{(nx^{-2q}+1)^2}dx \\ \lesssim & \left(n^{1/2q}\right)^{1-2\beta}\int_{n^{-1/2q}}^{\infty} \frac{y^{-2\beta}}{(y^{-2q}+1)^2}dy \\ \lesssim & n^{-\frac{2\beta-1}{2q}} \lesssim n^{-\frac{2\beta-1}{2\beta}}, \end{aligned}$$

where the last inequality holds because q is chosen such that $q \leq \beta$. Then $R_1 = \mathcal{O}(n^{-(2\beta-1)/2\beta})$.

For $k = 2, 3$, denote $R_k(b_n, c_n)$ the partial sum of R_k from $j = b_n$ to c_n. Then $R_2(1, J_n)$ is the larger term in the decomposition, and is treated at the end of the section. The upper part $R_2(J_n, \infty)$ is easily bounded by

$$R_2(J_n, \infty) \lesssim \left(\sum_{i=J_n}^{\infty} |\theta_{0i}|\, i^{\beta} i^{-\beta}\right)^2 \lesssim J_n^{-2\beta+1} = \mathcal{O}\left(n^{-\frac{2\beta-1}{2\beta}}\right).$$

We split $R_3(1, J_n)$ in two parts $R_{3,1}(1, J_n)$ and $R_{3,2}(1, J_n)$ by writing $u_i(X^n) = u_{J_n}(X^n) + \pi(i \leq k < J_n | X^n)$ for all $i \leq J_n$:

$$\begin{aligned} nR_3(1, J_n) \lesssim & \mathsf{E}_0^{(n)}\left(\sum_{j=1}^{J_n} \pi(j|X^n)\sum_{i=1}^{j} a_i \frac{\tau_i^2}{\tau_i^2+\frac{1}{n}}\xi_i\right)^2 \\ & + \mathsf{E}_0^{(n)}\left(u_{J_n}(X^n)\sum_{i=1}^{J_n} a_i \frac{\tau_i^2}{\tau_i^2+\frac{1}{n}}\xi_i\right)^2 \\ := & R_{3,1}(1, J_n) + R_{3,2}(1, J_n). \end{aligned}$$

Let $\Gamma_{jn}(X^n) = \sum_{i=1}^{j} a_i \frac{\tau_i^2}{\tau_i^2+\frac{1}{n}}\xi_i$. We have $\sum_{j=1}^{J_n}\pi(j|X^n) \leq 1$ so we can apply Jensen's inequality,

$$\begin{aligned} R_{3,1}(1,J_n) \leq & \quad \mathsf{E}_0^{(n)}\left(\sum_{j=1}^{J_n}\pi(j|X^n)\Gamma_{jn}(X^n)^2\right) \\ \leq & \quad \mathsf{E}_0^{(n)}\max_{j\leq J_n}\left\{\Gamma_{jn}(X^n)^2\right\}. \end{aligned}$$

Noting that $(\Gamma_{jn}(X^n))_{1\leq j\leq J_n}$ is a martingale, we get using Doob's inequality

$$R_{3,1}(1,J_n) \leq \mathsf{E}_0^{(n)}\Gamma_{J_n n}(X^n)^2 = \sum_{i=1}^{J_n}\left(a_i\frac{\tau_i^2}{\tau_i^2+\frac{1}{n}}\right)^2 \lesssim J_n.$$

The second term $R_{3,2}(1,J_n)$ can be upper bounded in the same way as for $R_3(J_n,\infty)$ in Equation (2.26) below by noting that

$$R_{3,2}(1,J_n) \lesssim \mathsf{E}_0^{(n)}\left[u_{J_n}(X^n)^2\left(\sum_{i=K_n}^{\infty}\frac{\tau_i^2}{\tau_i^2+\frac{1}{n}}|\xi_i|\right)^2\right].$$

For the upper part $R_3(J_n,\infty)$, we use the bound (2.25) on $\mathsf{E}_0^{(n)}\left(u_{K_n}(X^n)\right)$,

$$\begin{aligned} nR_3(J_n,\infty) \lesssim & \quad \mathsf{E}_0^{(n)}\left(\sum_{i=K_n}^{\infty}\frac{\tau_i^2}{\tau_i^2+\frac{1}{n}}u_i(X^n)|\xi_i|\right)^2 \\ \lesssim & \quad \mathsf{E}_0^{(n)}\left[u_{K_n}(X^n)^2\left(\sum_{i=K_n}^{\infty}\frac{\tau_i^2}{\tau_i^2+\frac{1}{n}}|\xi_i|\right)^2\right] \qquad (2.26) \\ \lesssim & \quad \left[\mathsf{E}_0^{(n)}u_{K_n}(X^n)^4\right]^{1/2}\left[\mathsf{E}_0^{(n)}\left(\sum_{i=K_n}^{\infty}\frac{\tau_i^2}{\tau_i^2+\frac{1}{n}}|\xi_i|\right)^4\right]^{1/2} \\ \lesssim & \quad \left[\mathsf{E}_0^{(n)}u_{K_n}(X^n)\right]^{1/2}\left[\left(\sum_{i=K_n}^{\infty}\frac{\tau_i^2}{\tau_i^2+\frac{1}{n}}\right)^4\right]^{1/2} \\ \lesssim & \quad e^{-c_2K_n/2}n^{1/q}, \end{aligned}$$

where we bound the different moments of $|\xi_i|$ by a unique constant and then use $\sum_{i=K_n}^{\infty}\tau_i^2/(\tau_i^2+\frac{1}{n}) = \mathcal{O}(n^{1/2q})$. Then $R_3 = \mathcal{O}(n^{-(2\beta-1)/2\beta})$.

To sum up, $R_2(1,J_n)$ is the only remaining term. We build an example where it is of greater order than $n^{-(2\beta-1)/2\beta}$. Let $\boldsymbol{\theta}_0$ be defined by its coordinates $\theta_{0i} = i^{-\beta-1/2}\left(\log(i+1)\right)^{-1}$ such that the series $\sum_i \theta_{0i}^2 i^{2\beta}$ converge, so $\boldsymbol{\theta}_0$ belongs to the Sobolev ball of smoothness β. It is assumed that $a_i = \psi_i(t) = 1$, so all terms in the sum $R_2(1,J_n)$ are *positive*,

hence

$$R_2(1, J_n) \geq \frac{1}{4}\mathsf{E}_0^{(n)}\left(\sum_{i=K_n}^{J_n}(1-u_i(X^n))\theta_{0i}\right)^2,$$

noting that for $i \leq J_n$, we have $n\tau_i^2 \geq n^{1-q/\beta} \geq 1$ because $q \leq \beta$ and $n \geq 1$, so $\tau_i^2/(\tau_i^2 + \frac{1}{n}) \geq 1/2$. Moreover, $u_i(X^n)$ decreases with i, so

$$R_2(1, J_n) \geq \frac{1}{4}\mathsf{E}_0^{(n)}\left((1-u_{K_n}(X^n))^2\right)\left(\sum_{i=K_n}^{J_n}\theta_{0i}\right)^2,$$

where $\mathsf{E}_0^{(n)}\left((1-u_{K_n}(X^n))^2\right)$ is lower bounded by a positive constant for n large enough. Comparing the series $\sum_{i=K_n}^{J_n}\theta_{0i}$ to an integral shows that it is bounded from below by $K_n^{-\beta+1/2}/\log n$. We obtain by using Lemma 2.15 that $R_n^{\mathrm{loc}}(\boldsymbol{\theta}_0, t) = R_2(1, J_n)(1+o(1)) \gtrsim n^{-\frac{2\beta-1}{2\beta+1}}/\log^2 n$, which ends the proof. □

Chapter 3

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

On introduit dans ce chapitre un modèle bayésien non-paramétrique pour étudier de manière probabiliste des données d'espèces par site, c'est à dire des données de population pour lesquelles les individus observés par site par site appartiennent à différentes espèces. Ces données peuvent être représentées par une matrice constituée du nombre d'occurrences de chaque espèce sur chaque site. Notre but est d'étudier l'impact de facteurs, ou variables explicatives, additionnels, tels que des variables environnementales, sur la structure des données, et en particulier sur la diversité. A cet effet, on introduit de la dépendance a priori selon les variables explicatives, et on montre que cela améliore l'inférence a posteriori. On utilise une version dépendante de la distribution GEM, qui représente la distribution des poids du processus de Dirichlet, de la même manière que sont dfinis les processus de Dirichlet dépendants. La loi a priori est définie à partir de la construction *stick-breaking*, dans laquelle on obtient les poids en transformant un processus gaussien, et la dépendance découle de la fonction de variance-covariance de ce dernier. On explicite des propriétés de distribution du modèle, telle que sa fonction de probabilité de partition échangeable jointe. On décrit un algorithme de Monte-Carlo par chaîne de Markov pour l'échantillonnage a posteriori, ainsi que l'échantillonnage de la loi prédictive pour des facteurs inobservés. Les deux algorithmes sont illustrés sur les données simulées et sur des données d'expériences réalisées par des prélèvements dans le sol en Antarctique.

Authors

- Julyan Arbel (Université Paris-Dauphine, CREST, Paris)
- Judith Rousseau (ENSAE, Université Paris-Dauphine, CREST, Paris)
- Kerrie L. Mengersen (Mathematical Sciences, Queensland University of Technology, Brisbane)

Status

Manuscript Arbel et al. (2013a) under preparation.

Abstract

We introduce 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 in distinct species. These data can be represented as a frequency matrix giving the number of times each species is observed in each site. Our aim is to study the impact of additional factors (covariates), for instance environmental factors, on the data structure, and in particular on the diversity. To that purpose, we introduce dependence a priori across the covariates, and show that it improves posterior inference. We use a dependent version of the GEM distribution, which is 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, where the weights are obtained by transforming a Gaussian process, and the dependence stems from the covariance function of the latter. Some distributional properties of the model are derived, such as its joint exchangeable partition probability function. A Markov chain Monte Carlo algorithm for posterior sampling is described, along with the sampling scheme of the predictive distribution for unobserved factors. Both samplers are illustrated on simulated data and on a real data set obtained in experiments conducted in Antarctica soil.

Keywords: Bayesian nonparametrics, Dependent model, Gaussian processes, GEM distribution, Stick-breaking representation.

3.1 Introduction

In this paper we define a dependent random probability measure for the study of species given by sites. Random probability measures are widely used, for instance in Bayesian

nonparametrics as prior distributions on measures, and are also of interest as species sampling models. Their dependent extensions, with respect to a factor, or a covariate, like time, position, *etc*, have been more and more studied recently, roughly under three possible constructions. First some are based on the Chinese Restaurant process, for instance Caron et al. (2006), and are oriented towards in-line data collection and fast implementation in that case. Then others use completely random measures, for example Lijoi et al. (2013a,b), whose analytical tractability allows to study their distributional properties. Eventually, many strategies make use of the stick-breaking representation after the seminal paper MacEachern (1999), for instance (Dunson et al., 2007, Dunson and Park, 2008, Chung and Dunson, 2009, Griffin and Steel, 2006).

As we can see, applications vary, and are mainly tailored by the type of construction. When it comes to the study of species observed by site, the literature is scarce. It is often useful to consider the influence of environmental and soil factors that can influence the abundance patterns. An obstacle in the field of species sampling stems from the need to track species across the factor when individuals are not the same. Arises the question which is tackled in the present work, namely

> How to model species proportions at different sites, indexed by a covariate, for example an environmental factor, and be able to interpret the impact of the latter on the population, for example on its diversity, or on particular species?

We define a dependent version of the GEM distribution (which is the distribution of the weights in a Dirichlet process) for modelling relative proportions. Dependence is introduced via (the covariance function of) Gaussian processes, which allow to define dependent Beta random variable by inverse cumulative distribution functions transforms.

The effect of the factor is measured in terms of population indices like diversity, which is a notion of interest in species data.

The appropriateness of the model is assessed by the study of some of its distributional properties. The exchangeable partition probability function of the model is given for the simple bivariate case. The prior dependence induced at the diversity level is also examined.

The running application used throughout the paper concerns an ecotoxicological data set of abundance data of microbes known as Operational Taxonomic Units (OTUs, or microbes, see Schloss and Handelsman, 2005) measured at different sites in Antarctica soil. The covariate that is used is one of the important environmental insults in the Antarctic region, namely fuel spills as measured by Total Petroleum Hydrocarbon (TPH). As an

indication of the structure of the data, Figure 3.1 shows three diversity indices computed on this data set according to the contaminant.

The rest of the paper is organized as follows. Section 3.2 discusses diversity measures that are used in order to characterise a species population, and their estimation. Section 4.3 reviews classical models in Bayesian nonparametrics and proposes a dependent GEM model for species-by-site count data, whose posterior sampling is developed in Section 3.4. Some of its distributional properties are studied in Section 3.5. Applications to simulated and real data obtained from the Antarctica soil biodiversity study are given in Section 3.6.

3.2 Diversity

In studies oriented towards species sampling and abundance measures, diversity is often a notion of interest. The question of measuring diversity arises in many fields, *eg* 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. Diversity is sometimes described as the crude measure of the number of observed species in a sample. This is also called species *richness*, and was studied with a Bayesian approach by Hill (1979), Boender and Kan (1987), and later in a Bayesian nonparametric setting by Gnedin and Pitman (2006) in the case of α–Gibbs type priors. Diversity is also defined as an index which measures the proximity of a discrete distribution with the uniform distribution. For a discussion between different diversity indices, see for example Cerquetti (2012).

One such index that is predominant in ecology and which will be retained here is the Shannon index H_{Shan}. For a discrete probability distribution $\mathbf{p}$, one defines

$$H_{\mathsf{Shan}}(\mathbf{p}) = -\sum_j p_j \log p_j. \tag{3.1}$$

Other diversity indices include the Simpson index

$$H_{\mathsf{Simp}}(\mathbf{p}) = 1 - \sum_j p_j^2, \tag{3.2}$$

and the generalized diversity index which was proposed by Good (1953) in the form of

$$H_{\mathsf{Good},\alpha,\beta}(\mathbf{p}) = -\sum_j p_j^\alpha \log^\beta p_j, \tag{3.3}$$

for non-negative integer values of α and β.

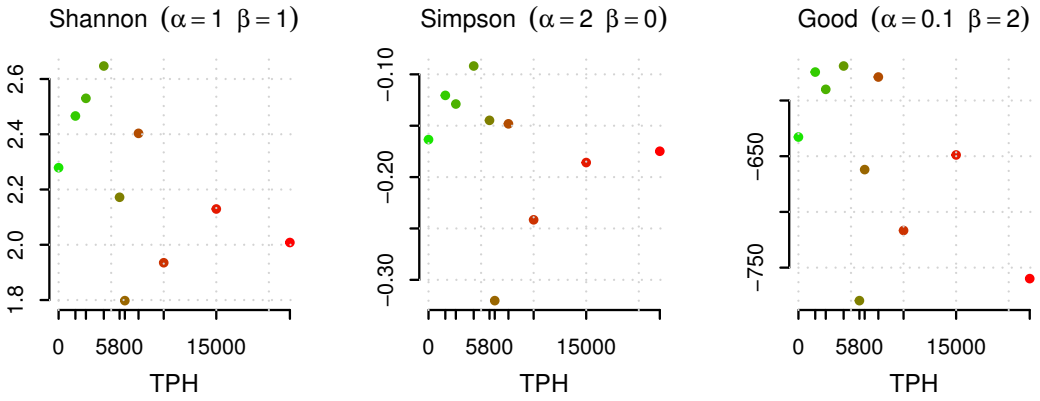

FIGURE 3.1: **Empirical diversity indices** for the data set of Section 3.6.2, indexed by a contaminant factor called TPH. *From left to right*: Shannon, Simpson (minus 1) and Good indices, for α and β parameters as indicated. The x-axis represents the TPH contaminant.

Diversity estimation has been a problem for a long time. Consider a sample Y_i in a discrete distribution $\mathbf{p} = (p_1, p_2, \ldots)$ such that $\mathsf{P}(Y_i = j) = p_j$. The most straightforward estimator is obtained by plugging-in the empirical distribution. Denote by n_j the number of observations equal to j, and by N the total number of observations. Then $\hat{\mathbf{p}}^{\mathsf{emp}} = (\hat{p}_1^{\mathsf{emp}}, \hat{p}_2^{\mathsf{emp}}, \ldots)$ where $\hat{p}_j^{\mathsf{emp}} = n_j/N$. For instance for the Shannon index

$$\hat{H}_{\mathsf{Shan}}{}^{\mathsf{emp}} = -\sum_j \hat{p}_j^{\mathsf{emp}} \log \hat{p}_j^{\mathsf{emp}}. \tag{3.4}$$

Figure 3.1 shows the empirical Shannon index estimator (3.4) and analogous estimators for the Simpson and Good indices computed on the ecotoxicological data set of Section 3.6.2 for varying values of the covariate. The empirical estimator is also the maximum likelihood estimator in the multinomial model. It is known that it is a biased estimator (see *eg* Gill and Joanes, 1979), with a bias of $-(J-1)/(2N)$ in the case of a population of size N with a finite number of species J. It also exhibits the following undesirable property with small sample sizes: if in a sample of a population with J species, all the individuals belong to the same species, then $\hat{H}^{\mathsf{emp}} = 0$.

Gill and Joanes (1979) examine a Bayes estimate which avoids that problem. They use a Dirichlet prior distribution on $\mathbf{p}$, with parameter $\boldsymbol{\alpha} = (\alpha, \ldots, \alpha)$, such that the posterior mean of the p_j's are $\hat{p}_j^{\mathsf{B}} = \frac{n_j+\alpha}{N+\alpha J}$. They deduce a diversity estimate by plug–in

$$\hat{H}^{\mathsf{B}} = -\sum_j \hat{p}_j^{\mathsf{B}} \log \hat{p}_j^{\mathsf{B}}. \tag{3.5}$$

It can be seen that the $\boldsymbol{\alpha}$ parameter smoothes, or flattens, the estimation of the p_j's by preventing them to be 0.

In our work we also adopt a Bayesian approach, but unlike Gill and Joanes (1979), we use Markov chain Monte Calro (thereafter MCMC) to sample directly the posterior distribution of the index and hence do not need to resort to a plug–in estimator. Moreover, our diversity estimation problem does not consist of a pointwise estimate, but a multivariate estimate with as many entries as the number of sites indexed by a covariate X. In order to construct better estimates of the diversity than unrelated ones as in Figure 3.1, we now turn to define a nonparametric Bayes prior on $\mathbf{p}$ that allows to model the dependence of the data with the covariate X across different population, and still retains the desirable flattening property of (3.5).

3.3 Models

3.3.1 Sampling model

We describe here the notation and sampling process of covariate dependent species-by-site count data. Each unique covariate value is indexed by i and is denoted by X_i. Recall that is may correspond to a single site or to a collection of sites with the same covariate value. For the sake of simplicity, we may still speak of *site* i. Individual observations at site i are taxa, or species, indexed by natural numbers $j = 1, 2, \ldots$. The total number of observed species is denoted by J. No hypothesis is made on the unknown total number of species in the population of interest, which might be infinite. Observe $(X_i, \mathbf{Y}_i^{N_i})_{i=1,\ldots,I}$ 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 value X_i. Species j abundance at site i is denoted by N_{ij}, *i.e.* the number of times when $Y_{n,i}) = j$ with respect to n index. Relative abundance satisfy $\sum_j N_{ij} = N_i$.

We model the relative frequencies or abundances $\mathbf{p} = (p(X_i))_i = (p_j(X_i))_{i,\, j=1,2,\ldots}$ by the following. For $i = 1 \ldots I$ and $n = 1 \ldots N_i$:

$$Y_{n,i} \,|\, \mathbf{p}(X_i), X_i \overset{\text{ind}}{\sim} \sum_{j=1}^{\infty} p_j(X_i)\delta_j. \tag{3.6}$$

Note that given the independence assumption in the model in Equation (3.6), it is not necessary to assume that the covariate values X_i are all distinct. The case $X_i = X_j$ for $i \neq j$ is equivalent to considering a single covariate X_i which collapses together observations $\mathbf{Y}_i^{N_i}$ and $\mathbf{Y}_j^{N_j}$. We also denote by $\mathbf{p}_i$ the relative frequencies at site i, $\mathbf{p}_i = (p_j(X_i))_{j=1,2,\ldots}$.

Remark 3.1. *The covariate X is continuous, so in addition to inferring the matrix* $\mathbf{p}$*, we are interested in the inference of the whole paths* $(\mathbf{p}_j(X), X \in \mathcal{X})$*, or* $\mathbf{p}_j(X_*)$ *for a reasonable covariate value* X_**, and for any species* j*.*

Remark 3.2. *Note that we want that sub-models close in the covariate space share close parameters,* i.e. *that the* $p_j(X_i)$*'s have dependence through the* X_i*'s.*

The probability of $Y_{n,i}$ is $f(Y_{n,i}|\mathbf{p}_i, X_i) = \prod_{j=1}^{J} p_j(X_i)^{\mathbb{I}(Y_{n,i}=j)}$. The likelihood of data at site i is given by $L_i(\mathbf{Y}_i^{N_i}|\mathbf{p}_i, X_i) = \prod_{j=1}^{J} p_j(X_i)^{N_{ij}}$, so the likelihood of the model is:

$$L(\mathbf{Y}|\mathbf{p}, X) = \prod_{i=1}^{I}\prod_{j=1}^{J} p_j(X_i)^{N_{ij}}. \tag{3.7}$$

Before we turn to the description of our Baysian nonparametric model, let mention a Bayesian parametric approach to the problem of estimating relative proportions by Holmes et al. (2012) using Dirichlet multinomial mixtures. A natural limitation to these techniques is that the number of species has to be fixed. In addition, we are interested in borrowing information across different sites, which cannot be guaranteed when one uses such independent priors by site.

3.3.2 Dependent GEM distribution

Modelling dependence in the Bayesian nonparametric has played an important in the recent literature. Strategies are diverse, and depend on the application in sight. Some are based on the Chinese Restaurant process, for instance Caron et al. (2006), and are oriented towards in-line data collection and fast implementation in that case. Others use completely random measures, for example Lijoi et al. (2013a,b), whose analytical tractability allows to study their distributional properties. Eventually, many strategies make use of the stick-breaking representation after the seminal paper MacEachern (1999), for instance (Dunson et al., 2007, Dunson and Park, 2008, Chung and Dunson, 2009, Griffin and Steel, 2006). We also follow this line with the precise motivation of our research question in mind, that is to say to be able to interpret the impact of a factor on the population. For this purpose, the stick-breaking construction is well adapted, since each weight can correspond to a species. This is the reason why we choose to index species by integers. This choice is discussed in more details in Remark 3.3. We now describe the GEM distribution, and show how it is extended to incorporate dependence.

A Dirichlet process (DP) is a distribution on probability measures. Its law can be written as the law of the following random probability measure G, known as the stick-breaking

representation of the DP. Let $M > 0$ and G_0 be a probability measure on a space Θ:

$$G = \sum_{j=1}^{\infty} p_j \delta_{\theta_j}, \tag{3.8}$$

$$p_j = V_j \prod_{l<j} (1 - V_l), \text{ with } V_j \overset{\text{iid}}{\sim} Beta(1, M) \text{ and } \theta_j \overset{\text{iid}}{\sim} G_0 \tag{3.9}$$

mutually independently, where δ_{θ_j} stands for the Dirac point mass at θ_j. We write $G \sim \mathsf{DP}(M, G_0)$. The prior induced on $\mathbf{p}_i = (p_j(X_i))_{j=1,2,\ldots}$ by Equations (3.9) is called the Griffiths-Engen-McCloskey prior, abbreviated GEM: $\mathbf{p}_i \sim \mathsf{GEM}(M)$ (see Pitman, 2006).

The motivation for the GEM distribution is explained by Figure 4.4. It shows draws of $(p_j)_{j=1\ldots J}$ in the $\mathsf{GEM}(M)$ prior with various precision parameters M and the observed proportions $(p_{ij})_{j=1\ldots J}$ at different sites i in the real data set under study. The similarity between the graphs is an argument in favour of the use of the $\mathsf{GEM}(M)$ prior for modelling the $\mathbf{p}_i$'s.

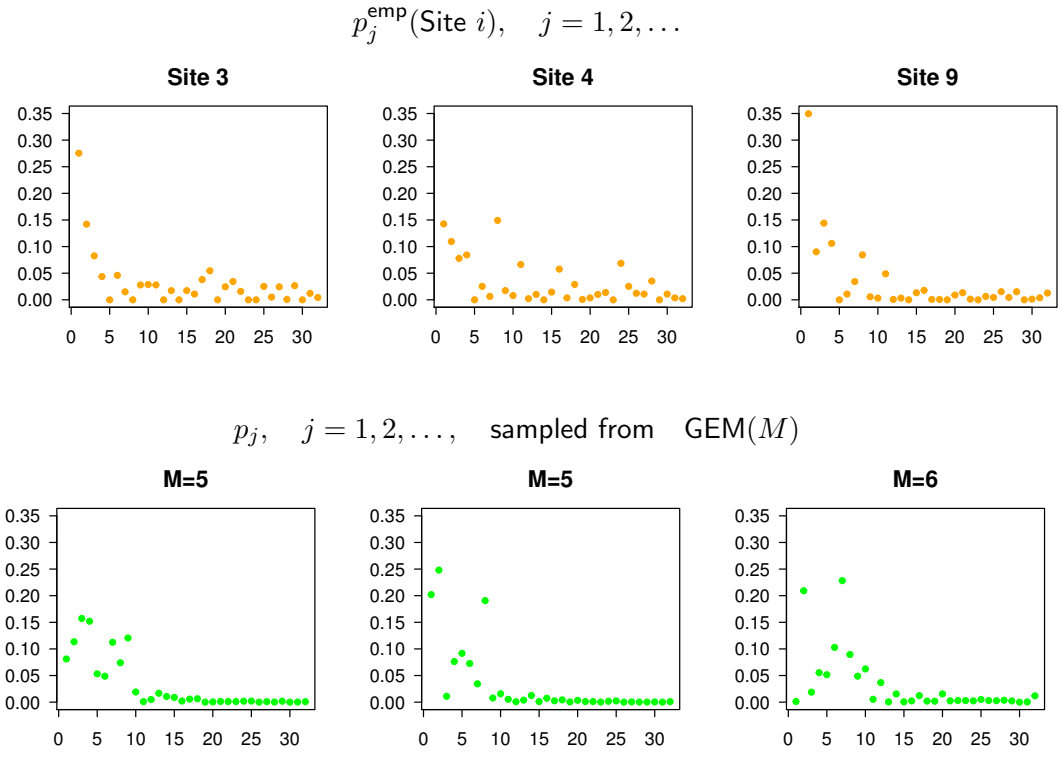

FIGURE 3.2: **Comparison of proportions in raw data and in the prior.** *Top*: proportions $p_j^{\mathsf{emp}}(\mathsf{Site}\ i)$ observed in the data at three sites. *Bottom*: proportions $(p_j)_j$ sampled from the Griffiths-Engen-McCloskey distribution. The x-axis represents the species index $j = 1, 2, \ldots$.

Remark 3.3. *It is important to note that the data in Figure 4.4 are ordered by decreasing overall abundance,* i.e. *species $j = 1$ is the largest species in the whole data set (when all sites are collapsed together). The variations of sampling across the sites explain why the species are not strictly ordered when considered site by site. Since the* GEM(M) *prior on $\mathbf{p}_i$ is "stochastically ordered" (see Pitman, 2006), it puts more mass on the larger species. It makes sense to sort the data in that way and to use a prior with a natural stochastic order on $\mathbf{p}$ since the data under study naturally present large and small species. Figure 4.4 shows that the population structure in terms of species decrease is similar in the data and in the prior.*

As an aside, we first consider the simple case of I independent GEM models at each site i of covariate X_i.

$$\text{for } i = 1, \dots, I, \ \mathbf{p}(X_i) \overset{\text{iid}}{\sim} \mathsf{GEM}(M). \tag{3.10}$$

This is not a satisfactory modelling since no dependence is incorporated in this way, but it is interesting for its posterior in closed-form. It is efficient to estimate this model in term of the $\mathbf{V}$ parameters as the prior is conjugate when written in $\mathbf{V}$, although the parameter of interest remains $\mathbf{p}$. Indeed, using (3.7) and (3.9), the likelihood is

$$L(Y|\mathbf{V}, X) = \prod_{i=1}^{I} \prod_{j=1}^{J} V_j(X_i)^{N_{ij}} (1 - V_j(X_i))^{\bar{N}_{i,j+1}}, \tag{3.11}$$

where $\bar{N}_{i,j+1} = \sum_{l>j} N_{il}$, and the posterior is Beta:

$$\pi(\mathbf{V}|Y) = \prod_{i=1}^{I} \prod_{j=1}^{\infty} \text{Be}(V_j(X_i)|1 + N_{ij}, M + \bar{N}_{i,j+1}). \tag{3.12}$$

In this case it is easy to sample from the posterior. The precision parameter M is endowed with a Gamma prior distribution $\mathsf{Ga}(a_M, b_M)$ which leads to the following posterior conditional

$$M|\mathbf{V} \sim \mathsf{Ga}(a_M + J, b_M - \sum_{j=1}^{J} \log(1 - V_{ij})). \tag{3.13}$$

The model is estimated by a simple Gibbs sampler, which leads to the results shown in Figure 3.5 on the real data set of Section 3.6.2. A comparison between the posterior mean and the maximum likelihood estimate of the relative proportions parameters p_{ij} is deferred to Appendix 3.8.2.

We now turn to the dependent GEM in itself. The seminal paper by MacEachern (1999) extended the classical DP to dependent Dirichlet processes which allow the weights p_j and/or the clusters θ_j to vary with a predictor X, according to stochastic processes

$p_j(X)$ and $\theta_j(X)$. We use the same construction to extend the GEM distribution to the following dependent version, abbreviated $\mathsf{Dep}-\mathsf{GEM}(M)$:

$$p_j(X) = V_j(X)\prod_{l<j}(1-V_l(X)) \text{ with } V_j(X) \overset{\text{iid}}{\sim} Beta(1,M). \tag{3.14}$$

We use the same model as before, that is Equation (3.6), with the $\mathsf{Dep}-\mathsf{GEM}$ prior. The likelihood given in Equation (3.11) is factorized across the species $j = 1 \dots J$. By independence in (3.14), the prior is also factorized on $\mathbf{V}$ through j: $\mathbf{V}_j|X \overset{\text{iid}}{\sim} \pi(\mathbf{V}_j|X)$, so the following factorized posterior is obtained:

$$\pi(\mathbf{V}|Y,X) \propto \prod_{j=1}^{\infty} \pi(\mathbf{V}_j|X) \times L_j(Y|\mathbf{V}_j,X), \tag{3.15}$$

where $L_j(Y|\mathbf{V}_j,X) = \prod_{i=1}^{I} V_j(X_i)^{N_{ij}}(1-V_j(X_i))^{\bar{N}_{i,j+1}}$. Note that for $j > J$, $L_j(Y|\mathbf{V}_j,X) = 1$, so that the posterior updating concerns only J components of observed species. Using a factorized prior across the species does not prevent the introduction of dependence across the sites as the $V_j(X_i)$ can be correlated across i in the joint prior $\pi(\mathbf{V}_j|X)$. The factorized expression of the posterior in Equation (4.8) is convenient as it permits to sample independently across the species j. It reduces the initial problem of estimation in dimension $I \times J$ to J estimation problems of dimension I, which is more efficient with respect to the curse of dimensionality.

The construction of the $\mathsf{Dep}-\mathsf{GEM}$ prior requires to be able to construct a prior on the vector $\mathbf{V}_j$ which is marginally Beta, *i.e.* which meets the definition in (3.14) for each $\mathbf{V}_(X_i)$ and still exhibits dependence across i. We turn in the next section to describe the construction of a prior on $\mathbf{V}_j$, where the dependence is introduced through Gaussian processes. The model is illustrated by a graphical representation in Figure 3.3.

Remark 3.4. *Since the processes* $\mathbf{V}_j$ *are i.i.d. (across* j*), we denote by* $\mathbf{V}$ *a generic* $\mathbf{V}_j$*.*

3.3.3 Dependence through Gaussian processes

This section proposes a construction of a prior on $\mathbf{V} = (V(X_1), \dots, V(X_I))$ which meets the Beta marginal requirement of Equation (3.9). It is built by transforming a Gaussian process (GP) $\mathcal{Z}$ on the covariate space $\mathcal{X}$ with the inverse cumulative distribution function (CDF) transform as follows[1]. Denote by $Z \sim \mathsf{N}(0,\sigma^2)$ a Gaussian random variable,

[1] We use instead a Gaussian random field if $\mathcal{X}$ is multidimensional. This extension is straightforward.

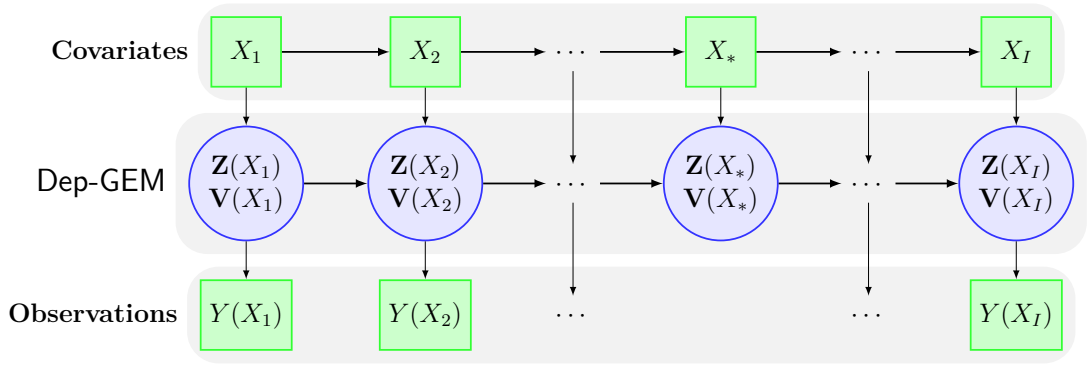

FIGURE 3.3: **Graphical model representation for the Dep − GEM model.** Squares represent observed data, *i.e.* covariates X_i and observations $Y(X_i) = (Y_1(X_i), \ldots, Y_{N_i}(X_i))$, and circles represent parameters for the Dep − GEM model.

by Φ_{σ_Z} its CDF and by F_M a $\mathsf{Beta}(1, M)$ CDF. Then:

$$\Phi_{\sigma_Z}(Z) \sim U(0,1) \text{ and } V = F_M^{-1} \circ \Phi_{\sigma_Z}(Z) \sim \mathsf{Beta}(1, M), \tag{3.16}$$

with $F_M(V) = 1-(1-V)^M$ and $F_M^{-1}(U) = 1-(1-U)^{1/M}$. Denote by $g_{\sigma,M} = F_M^{-1} \circ \Phi_{\sigma_Z}$. The Gaussian process $\mathcal{Z}$ has Gaussian marginals, hence by applying the transform $g_{\sigma,M}$ to the marginals $\mathcal{Z}(X_i)$ with the right standard deviation σ, one obtains marginal Beta random variables $V(X_i) \sim \mathsf{Beta}(1, M)$ which are dependent. The Gaussian process is used as a prior probability distribution over functions. It is fully specified by a mean function m and a covariance function K. We choose to use a centred GP, *i.e.* $m = 0$. The covariance function of $\mathcal{Z}$ is denoted by K and defined by

$$K(X_i, X_j) = \mathsf{Cov}(\mathcal{Z}(X_i), \mathcal{Z}(X_j)). \tag{3.17}$$

We control the overall variance of $\mathcal{Z}$ by a positive pre-factor $\sigma_{\mathbf{Z}}^2$ and write $K = \sigma_{\mathbf{Z}}^2 \tilde{K}$ where $\tilde{K}$ is normalized in the sense that $\tilde{K}(X_i, X_i) = 1$ for all i. We work with the following covariance functions $\tilde{K}_\lambda(X_1, X_2)$:

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)	$\big(1 + (X_1 - X_2)^2/(2\lambda^2)\big)^{-1}$

The Squared Exponential kernel is known to be smooth compared with other popular choices like the Ornstein-Uhlenbeck covariance function. The parameter λ is called the length-scale of the process $\mathcal{Z}$. It tunes how far apart two points X_1 and X_2 have to be

for the process to change significantly. The shorter λ is, the rougher are the paths of the process $\mathcal{Z}$. We adopt the same technique as van der Vaart and van Zanten (2009) who deal with λ by making it random with an inverse Gamma prior distribution (by using a Gamma distribution on a rescaling factor called a bandwidth). They obtain adaptive minimax-optimal posterior contraction rates, which indicates that the length-scale parameter λ correctly adapts to the path smoothness. Gibbs (1997) derived a covariance function where the length-scale $\lambda(X)$ is a (positive) function of X. This case is not studied here, although it could result in interesting behaviour, as noted in Rasmussen and Williams (2006).

We have a set of I points $X = (X_1, \dots, X_I)$ in the covariate space $\mathcal{X}$. So instead of dealing with the whole process $\mathcal{Z}$, we deal with its values at X denoted by $\mathbf{Z} = (Z_1, \dots, Z_I) = (\mathcal{Z}(X_1), \dots, \mathcal{Z}(X_I))$. The vector $\mathbf{Z}$ is a multivariate Gaussian whose covariance matrix $K(X, \lambda, \sigma_{\mathbf{Z}}) = (\sigma^2_{\mathbf{Z}} \tilde{K}_\lambda(X_i, X_j))_{ij}$ is a Gram matrix whose entries are given by Equation (3.17). This prior distribution is

$$\log \pi(\mathbf{Z}|X, \lambda) = \frac{1}{2}\mathbf{Z}^\top K^{-1}(X, \lambda, \sigma_{\mathbf{Z}})\mathbf{Z} - \frac{1}{2}|K(X, \lambda, \sigma_{\mathbf{Z}})| - \frac{I}{2}\log 2\pi,$$

or, written in terms of $\sigma^2_{\mathbf{Z}}$ and $\tilde{K}_\lambda = (\tilde{K}_\lambda(X_i, X_j))_{ij}$:

$$\pi(\mathbf{Z}|\sigma_{\mathbf{Z}}, \lambda, X) \propto \sigma_{\mathbf{Z}}^{-I} |\tilde{K}_\lambda|^{-1/2} \exp\Big(- \frac{\mathbf{Z}^\top \tilde{K}_\lambda^{-1} \mathbf{Z}}{2\sigma^2_{\mathbf{Z}}} \Big).$$

It is convenient to estimate the model in terms of $\mathbf{Z}$, and then to use the transform $\mathbf{V} = g_{\sigma_{\mathbf{Z}}, M}(\mathbf{Z})$. Following Remark 3.4, note that $\mathbf{Z}$ corresponds to a given species j (and should be written $\mathbf{Z}_j$). The likelihood contribution for species j only is

$$L(Y|\mathbf{Z}, X) = L(Y|g^{-1}_{\sigma_{\mathbf{Z}}}(\mathbf{V}), X) = \prod_{i=1}^{I} g_{\sigma_{\mathbf{Z}}, M}(Z_i)^{N_{ij}} (1 - g_{\sigma_{\mathbf{Z}}, M}(Z_i))^{\bar{N}_{i,j+1}}.$$

The hyperparameters are the standard deviation $\sigma_{\mathbf{Z}}$, the length-scale λ and the precision parameter M of the GEM distribution. We use the following hyperpriors:

$$\sigma^2_{\mathbf{Z}} \sim \mathsf{IG}(a_{\mathbf{Z}}, b_{\mathbf{Z}}), \ \lambda \sim \mathsf{IG}(a_\lambda, b_\lambda), \text{ and } M \sim \mathsf{Ga}(a_M, b_M).$$

In absence of dependence, these are common choices since they are conjugate priors. On top of conjugacy, recall that the IG for λ also proves to lead to good convergence results. For the parameters of the hyperpriors, we let $a_{\mathbf{Z}} = b_{\mathbf{Z}} = 1$, $\eta_\lambda = 1$, $a_\lambda = b_\lambda = 1$ and $a_M = b_M = 1$.

The posterior distribution is then:

$$\pi(\mathbf{Z}, \lambda, \sigma_{\mathbf{Z}}, M|Y, X) \propto L(Y|\mathbf{Z}, X)\pi(\mathbf{Z}|X, \lambda, \sigma_{\mathbf{Z}})\pi(\sigma_{\mathbf{Z}})\pi(\lambda)\pi(M). \tag{3.18}$$

3.4 Posterior computation and inference

Here, we describe how to design an MCMC algorithm for the model. Up to a transformation, it is equivalent to sample the parameters in terms of Gaussian vectors $\mathbf{Z}$ or Beta breaks $\mathbf{V}$. We equally denote by π the prior for both parameters. For each of these, we make use of the factorized form of the posterior in Equation (4.8) in order to break the posterior sampling into J independent sampling schemes. It remains a multivariate sampling in terms of the I sites, but avoids a very high dimensional scheme of size $J \times I$.

3.4.1 MCMC algorithm

We use a MCMC algorithm comprising Gibbs and Metropolis-Hastings steps for sampling the posterior distribution of $(\mathbf{Z}, \sigma_{\mathbf{Z}}, \lambda, M)$, which proceeds by sequentially updating each of the parameters $\mathbf{Z}$, $\sigma_{\mathbf{Z}}$, λ and M in its conditional distribution as described in Algorithm 1. Each conditional is sampled by a Metropolis-Hastings (MH) steps. Denote by $P_{\boldsymbol{\theta}}(\cdot)$ the target distribution (full conditional), and by $Q_{\boldsymbol{\theta}}(\cdot\,|\,\boldsymbol{\theta})$ the proposal for a generic parameter $\boldsymbol{\theta}$. The variance of the latter proposal, denoted by $\sigma^2_{Q_{\boldsymbol{\theta}}}$, is tuned during a burn-in period. A generic Metropolis-Hastings step is described in Algorithm 2.

Algorithm 1 Dep − GEM algorithm

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 2 MH algorithm

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

The full conditionals and target distributions are now fully described:

1. Conditional for $\mathbf{Z}$: Metropolis algorithm with Gaussian jumps $\mathbf{Z}' \sim Q_{\mathbf{Z}}(\cdot\,|\,\mathbf{Z}) = \mathsf{N}_I(\mathbf{Z}, \sigma^2_{Q_{\mathbf{Z}}}\tilde{K}_\lambda)$. To use a covariance matrix proportional on the one of the prior, $\tilde{K}_\lambda$, leads to improve the convergence of the algorithm compared to an homoscedastic one. The target distribution is

$$P_{\mathbf{Z}}(\mathbf{Z}) \propto L(Y|\mathbf{Z}, X, \sigma_{\mathbf{Z}}, M)\pi(\mathbf{Z}|X, K(X, \lambda, \sigma_{\mathbf{Z}})).$$

2. Conditional for $\sigma_{\mathbf{Z}}$: Metropolis-Hastings algorithm with a truncated to 0 Gaussian proposal $\sigma_{\mathbf{Z}}' \sim Q_{\sigma_{\mathbf{Z}}}(\cdot \,|\, \sigma_{\mathbf{Z}}) = \mathsf{N}_{\text{0-trunc}}(\sigma_{\mathbf{Z}}, \sigma^2_{Q_{\sigma_{\mathbf{Z}}}})$, and target distribution $P_{\sigma_{\mathbf{Z}}}(\sigma_{\mathbf{Z}}) \propto L(Y|\mathbf{Z}, X, \sigma_{\mathbf{Z}}, M)\sigma_{\mathbf{Z}}^{-I-a_{\mathbf{Z}}/2} \exp\Big(-\frac{\mathbf{Z}^\top \tilde{K}_\lambda^{-1}\mathbf{Z} - 2b_{\mathbf{Z}}}{2\sigma^2_{\mathbf{Z}}}\Big)$.

3. Conditional for λ: Metropolis-Hastings algorithm with a truncated to 0 Gaussian proposal $\lambda' \sim Q_\lambda(\cdot \,|\, \lambda) = \mathsf{N}_{\text{0-trunc}}(\lambda, \sigma^2_{Q_\lambda})$, and target distribution $P_\lambda(\lambda) \propto \pi(\mathbf{Z}|X, K(X, \lambda, \sigma_{\mathbf{Z}}))\pi(\lambda)$.

4. Conditional for M: Metropolis algorithm with a truncated to 0 Gaussian proposal $M' \sim Q_M(\cdot \,|\, M) = \mathsf{N}_{\text{0-trunc}}(M, \sigma^2_{Q_M})$, and target distribution $P_M(M) \propto M^{A_M - 1}\exp(-b_M M)\prod_{i=1}^{I} g_{\sigma_{\mathbf{Z}}, M}(Z_i)^{N_{ij}}(1 - g_{\sigma_{\mathbf{Z}}, M}(Z_i))^{\bar{N}_{i,j+1}}$.

Remark 3.5. *The dimensionality of the* MCMC *algorithm described above equals the number of covariate (or block of covariates). Large dimensions can be an obstacle to the use of traditional methods (mainly for matrix inversion). A direction that has not been investigated could be to replace* MCMC *algorithms with afster approximations, of the type of* INLA *for example, see Rue et al. (2009).*

3.4.2 Predictive distribution

Up to now we have considered the vector $\mathbf{Z}$, which is the evaluation of the GP $\mathcal{Z}$ at the observed covariates X. We are now interested in new outputs, called test outputs, $\mathbf{Z}_*$, associated with (non observed) test covariates X_*. An appealing feature of the use of GP is the possibility to easily derive the predictive distribution of $\mathbf{Z}_*$, which is achieved as follows. The joint distribution of the vector outputs $(\mathbf{Z}, \mathbf{Z}_*)$ according to the prior is

$$\begin{pmatrix} \mathbf{Z} \\ \mathbf{Z}_* \end{pmatrix} \sim \mathsf{N}_{I+I_*}\left[\mathbf{0}, \begin{pmatrix} K(X,X) & K(X,X_*) \\ K(X_*,X) & K(X_*,X_*) \end{pmatrix}\right], \tag{3.19}$$

where the covariance matrices $K(X,X)$, $K(X,X_*) = K(X_*,X)^\top$ and $K(X_*,X_*)$ (resp. $I\times I$, $I\times J$ and $J\times J$ matrices) are defined by their entries according to Equation (3.17). The conditional density of $\mathbf{Z}_*$ given $\mathbf{Z}$ given by Rasmussen and Williams (2006) is the following Gaussian:

$$\begin{aligned} &\mathbf{Z}_* \,|\, X_*, X, \mathbf{Z} \sim \mathsf{N}_{I_*}(m_*(\mathbf{Z}), K_*), \text{ with } m_*(\mathbf{Z}) = K(X_*,X)K(X,X)^{-1}\mathbf{Z}, \\ &\text{and } K_* = K(X_*,X_*) - K(X_*,X)K(X,X)^{-1}K(X,X_*). \end{aligned} \tag{3.20}$$

The predictive distribution of $\mathbf{Z}_*$ is obtained by integrating out $\mathbf{Z}$ in the conditional distribution (3.20) according to the posterior distribution $\pi(\mathbf{Z}|Y,X)$:

$$\pi(\mathbf{Z}_* \,|\, X_*, Y) = \int \pi(\mathbf{Z}_* \,|\, X_*, X, \mathbf{Z})\pi(\mathbf{Z}|Y,X)\mathrm{d}\mathbf{Z}. \tag{3.21}$$

From a practical point of view, this is done with no particular computational burden. Generally speaking, simulation of a predictive distribution of the form of (3.21) is undertaken as follows:

Algorithm 3 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 $\pi(\mathbf{Z}_* \mid X_*, X, \mathbf{Z})$

Hence when a sample of $\mathbf{Z}$ from the posterior distribution $\pi(\mathbf{Z}|Y, X)$ is available, one obtains a sample from the predictive distribution by sampling in the multivariate normal distribution (3.20). One matrix, $K(X, X)$, has to be inverted, but that computation is already done for the MCMC sampler. The variance K_* of (3.20) is to be computed once. Then it is efficient to draw a sample of the desired size from the centred normal $\mathsf{N}(0, K_*)$, and then add the means $m_*(\mathbf{Z})$ for $\mathbf{Z}$ in the posterior sample. We can obtain the predictive distribution of any $\mathbf{Z}_*$ associated with any test covariates X_*, hence the posterior distribution of $\mathbf{V}_*$, and in turn of $\mathbf{p}_*$. This allows prediction in the whole space $\mathcal{X}$.

3.5 Distributional properties

The purpose of this section is to present some general distributional properties of the $\mathsf{Dep}-\mathsf{GEM}$ prior in terms of dependence and predictive rule.

The main trick for the following computations is the *conditional independence* between the samples at two sites, $\mathbf{Y}_1^n = (Y_1(x_1), \ldots, Y_n(x_1))$ and $\mathbf{Y}_2^m = (Y_1(x_2), \ldots, Y_m(x_2))$, given the process $\mathbf{p} \sim \mathsf{Dep}-\mathsf{GEM}(M)$.

First, denote by $c_M = c_M(|x_1 - x_2|)$ the dependence factor between the process at two covariate points x_1 and x_2 defined by:

$$c_M(|x_1 - x_2|) = (M+1)^2 \mathsf{E}(V(x_1)V(x_2)).$$

We identify two extreme cases denoted by:

- (I): *independence*, $V(x_1) \perp\!\!\!\perp V(x_2)$ (*eg* $|x_1 - x_2| \to \infty$), then $c_M = 1$.
- (E): *equality*, $x_1 = x_2$, *i.e.* $V(x_1) = V(x_2)$ in distribution, then $c_M = 2(M+1)/(M+2) = 1 + M/(M+2)$,

3.5.1 Joint law of the first picks

Proposition 3.6. *The joint law for the first picks $Y_1(x_1)$ and $Y_1(x_2)$ at two sites of covariate x_1 and x_2 is:*

$$\mathsf{P}(Y_1(x_1)=j, Y_1(x_2)=k) = (M+1-c_M)M^{|j-k|-1}\frac{(M^2-1+c_M)^{(j\wedge k)-1}}{(M+1)^{j+k}}. \tag{3.22}$$

Proof.

$$\begin{aligned}\mathsf{P}(Y_1(x_1)=j, Y_1(x_2)=k) &= \mathsf{E}\big(\mathsf{P}(Y_1(x_1)=j, Y_1(x_2)=k \,|\, \mathbf{p}(x_1), \mathbf{p}(x_2))\big),\\ &= \mathsf{E}(p_j(x_1)p_k(x_2)) \quad \text{by conditional independence.}\end{aligned}$$

Suppose without loss of generality that $j \geq k$, then the last quantity can be decomposed into the following product of four groups of terms

$$\begin{aligned}&\mathsf{E}(V_j(x_1)) \quad \prod_{k<l<j}\mathsf{E}(\bar{V}_l(x_1))E(\bar{V}_k(x_1)V_k(x_2)) \quad \prod_{l<k}\mathsf{E}(\bar{V}_l(x_1)\bar{V}_l(x_2))\\ &= \frac{1}{M+1}\quad \Big(\frac{M}{M+1}\Big)^{j-k-1}\Big(\frac{1}{M+1}-\frac{c_M}{(M+1)^2}\Big)\quad \Big(1-\frac{2}{M+1}+\frac{c_M}{(M+1)^2}\Big)^{k-1}.\end{aligned}$$

which sum up to (3.22). □

Equation (3.22) reduces to $M^{2(j-1)}/(M+1)^{2j}$ in the (I) case, which is the square of the law of the first pick, $\mathsf{P}(Y_1(x_1)=j)=\mathsf{E}(p_j)$. Surprisingly, it does not reduce to $M^{j-1}/(M+1)^j$ in the (E) case, but to $\frac{M}{(M+1)(M+2)^j}$. In particular, the probability that both species are the jth one is:

$$\mathsf{P}(Y_1(x_1)=Y_1(x_2)=j) = (M+1-c_M)\frac{(M^2-1+c_M)^{j-1}}{M(M+1)^{2j}},$$

thus by summing over all positive j

$$\mathsf{P}(Y_1(x_1)=Y_1(x_2)) = \frac{M+1-c_M}{M(2M+2-c_M)}. \tag{3.23}$$

We can see that in the (I) case, Equation (3.23) reduces the probability that two draws at the same site x_1 belong to the same species:

$$\mathsf{P}(Y_1(x_1)=Y_2(x_1)) = \frac{1}{2M+1},$$

obtained by summing all squares of $M^{j-1}/(M+1)^j$.

3.5.2 Joint EPPF

Let two samples $\mathbf{Y}_1^n = (Y_1(x_1), \ldots, Y_n(x_1))$ and $\mathbf{Y}_2^m = (Y_1(x_2), \ldots, Y_m(x_2))$ partition $[n] = \{1, 2, \ldots, n\}$ and $[m] = \{1, 2, \ldots, m\}$ in $\mathbf{N}_1^n = (n_1, \ldots, n_{k_1})$ and $\mathbf{N}_2^m = (m_1, \ldots, m_{k_2})$. The following proposition gives the joint EPPF of $(\mathbf{Y}_1^n, \mathbf{Y}_2^m)$.

Proposition 3.7. *Let observations $(\mathbf{Y}_1^n, \mathbf{Y}_2^m)$ come from a model with* Dep – GEM *distribution. The probability distribution of the joint partition of $(\mathbf{Y}_1^n, \mathbf{Y}_2^m)$ when written in terms of $(\mathbf{N}_1^n, \mathbf{N}_2^m, k_1, k_2)$ is*

$$p(\mathbf{N}_1^n, \mathbf{N}_2^m, k_1, k_2) = p(\mathbf{N}_2^{m_{k_1+1}^{k_2}}, k_2 - k_1) \prod_{i=1}^{k_1} B_{x_1,x_2}(n_i - 1, m_i - 1, n_{i+1}^{k_1}, m_{i+1}^{k_2}),$$

where $B_{x_1,x_2}(n, m, n', m') = \mathsf{E}\big[V(x_1)^n V(x_2)^m (1 - V(x_1))^{n'} (1 - V(x_2))^{m'}\big]$ and we use the generic notation $m_i^k = m_i + \cdots + m_k$. We denote by $\mathbf{N}_2^{m_{k_1+1}^{k_2}}$ the partition of $m_{k_1+1}^{k_2}$ observations at x_2 unobserved at x_1 in $k_2 - k_1$ species.

Interestingly, the dependence at the partition level arises only through species that are shared at both covariates x_1 and x_2, with the B terms that are moments of dependent Beta random variables. The contribution of unshared species is their own marginal EPPF.

Proof. Using expression (1.14) for the EPPF (see Pitman, 1995), one obtains by conditional independence:

$$\begin{aligned}
&\mathsf{P}\big(\mathbf{N}_1^n, \mathbf{N}_2^m, k_1, k_2\big) \\
&= \mathsf{E}\big[\mathsf{P}\big(\mathbf{N}_1^n, \mathbf{N}_2^m, k_1, k_2 \,|\, \mathbf{p}(x_1), \mathbf{p}(x_2)\big)\big] \\
&= \mathsf{E}\Big[\prod_{i=1}^{k_1} V_i(x_1)^{n_i - 1}(1 - V_i(x_1))^{n_{i+1} + \cdots + n_{k_1}} \prod_{i=1}^{k_2} V_i(x_2)^{m_i - 1}(1 - V_i(x_2))^{m_{i+1} + \cdots + n_{k_2}}\Big].
\end{aligned}$$

The B terms come by independence across i for shared species. The remaining part of the product for the last $k_2 - k_1$ terms at x_2 boils down to the EPPF for $\mathbf{N}_2^{m_{k_1+1}^{k_2}}$. □

See for instance Müller et al. (2011), Lijoi et al. (2013a), Kolossiatis et al. (2013), Sporysheva and Petrone (2013) for other examples of joint EPPF in the case of different dependent processes (based on normalized completely random measures).

[!! Develop the link with these results]

3.5.3 Dependence at the diversity level

We denote a size-biased permutation of $\mathbf{p}$ by $\tilde{\mathbf{p}} = (\tilde{p}_1, \tilde{p}_2, \ldots)$. The first element $\tilde{p}_1$ is called the size-biased pick.

The following result is proved by Pitman (2006), as formula (2.23)

$$\mathsf{E}\Big(\sum f(p_j)\Big) = \mathsf{E}\Big(\sum f(\tilde{p}_j)\Big) = \mathsf{E}\left(\frac{f(\tilde{p}_1)}{\tilde{p}_1}\right). \tag{3.24}$$

Hence the distribution of $\tilde{p}_1$ encodes much information about $\mathbf{p}$. It suffices to compute the expectation of any transform of the form $\sum f(p_j)$, for example the generalized diversity index given in Equation (3.3). When it comes to compute its variance, one need to resort to an extension of this result, as is obtained by Archer et al. (2013)

$$\mathsf{E}\Big(\sum_{i\neq j} f(p_i, p_j)\Big) = \mathsf{E}\left(\frac{f(\tilde{p}_1, \tilde{p}_2)}{\tilde{p}_1\tilde{p}_2}(1-\tilde{p}_1)\right). \tag{3.25}$$

In the case of a $\mathsf{GEM}(1, M)$ prior on $\mathbf{p}$, the prior expectation of Simpson diversity H is found by Cerquetti (2012)

$$\mathsf{E}(H) = \frac{M}{1+M}$$

The result for the Shannon diversity index is given an unpublished work by Cerquetti

$$\mathsf{E}(H) = \psi(M+1) - \psi(1), \tag{3.26}$$

where ψ is the digamma function, *i.e.* the derivative of the log of the gamma function.

Now, we turn to the question of measuring the dependence at the diversity level. We know that the DDP introduces some dependence across the $p_j(X_i)$ in varying X_i. What dependence is induced in $H(X_i)$? In order to answer to this question, one needs the following propositon

Proposition 3.8. *Let* $\mathbf{p}$ *be governed by a* $\mathsf{GEM}(M)$ *prior. Denote by* $\tilde{p}_1(X_1)$ *and* $\tilde{p}_1(X_2)$ *the first size-biased pick respectively at* X_1 *and* X_2*. Then for an arbitrary measurable function* f*, then the following holds*

$$\mathsf{E}\Big(\sum_j f(p_j(X_1)) \sum_j f(p_j(X_2))\Big) = \mathsf{E}\Big(\frac{f(\tilde{p}_1(X_1))}{\tilde{p}_1(X_1)} \frac{f(\tilde{p}_1(X_2))}{\tilde{p}_1(X_2)}\Big). \tag{3.27}$$

Proof.

$$\begin{aligned}
&\mathsf{E}\Big(\sum_j f(p_j(X_1))\sum_j f(p_j(X_2))\Big)\\
&=\mathsf{E}\Big(\sum_j \frac{f(p_j(X_1))}{p_j(X_1)}p_j(X_1)\sum_j \frac{f(p_j(X_2))}{p_j(X_2)}p_j(X_2)\Big)\\
&=\mathsf{E}\left[\mathsf{E}\Big(\frac{f(\tilde{p}_1(X_1))}{\tilde{p}_1(X_1)}|\mathbf{p}(X_1)\Big)\mathsf{E}\Big(\frac{f(\tilde{p}_1(X_2))}{\tilde{p}_1(X_2)}|\mathbf{p}(X_2)\Big)\right]
\end{aligned}$$

and by independance of the first pick with respect to $\mathbf{p}$ at a different covariate location

$$\begin{aligned}
&=\mathsf{E}\left[\mathsf{E}\Big(\frac{f(\tilde{p}_1(X_1))}{\tilde{p}_1(X_1)}|\mathbf{p}(X_1),\mathbf{p}(X_2)\Big)\mathsf{E}\Big(\frac{f(\tilde{p}_1(X_2))}{\tilde{p}_1(X_2)}|\mathbf{p}(X_1),\mathbf{p}(X_2)\Big)\right]\\
&=\mathsf{E}\left[\frac{f(\tilde{p}_1(X_1))}{\tilde{p}_1(X_1)}\frac{f(\tilde{p}_1(X_2))}{\tilde{p}_1(X_2)}\right].
\end{aligned}$$

□

The following result on the dependence between diversity indices is a direct application of Proposition 3.8 and Equation (4.7).

Proposition 3.9. *The covariance between $H(X_1)$ and $H(X_2)$ induced by the* GEM *is controlled by the distribution of the first couple of beta breaks $(V(X_1), V(X_2))$ and the precision parameter M only*

$$\mathsf{Cov}(H(X_1), H(X_2)) = \mathsf{E}(\log V_1(X_1)\log V_2(X_2)) - (\psi(M+1)-\psi(1))^2.$$

[!! Explain and speak about asymptotics in M]

3.6 Applications to the estimation of diversity

We now apply the model to the estimation of diversity as described in Section 3.2. It is a one dimensional output for which the fit of the model is easy to assess.

3.6.1 Simulated data

We begin by assessing the model on simulated data. We use a synthetic model where the true relative proportions depend on a covariate X in the following way:

$$p_j(X) \propto j^{-\frac{3+\cos(X)}{2}},\ j=1,\dots,J,\ J=50. \tag{3.28}$$

We use the six covariate values $X = 0, 1, 2, \ldots, 5$. We sample data of size $N = 100$, $N = 250$ and $N = 1000$ from the distribution given by Equation (3.28) for each X. Since there are J species, this correspond to an average abundance respectively of 2, 5 and 20 by species. We run the model described in Section 3.3.3 using three kinds of Gaussian processes: Ornstein-Uhlenbeck (OU), Squared Exponential (SE) and Rational Quadratic (RQ), see Section 3.3.3. Algorithm 1 is run for 100 000 iterations with a burn-in of 20 000 iterations. The graphs of Figure 3.4 are estimates of the diversity at observed covariates (triangles) along with the predictive estimate between observed covariates. The gray shade indicates a 95% credible interval of the predictive distribution. The color dots represent the empirical diversity on observations, while the coloured line represents the diversity for the true distribution. The graphs below show a convergence of the estimates towards the true diversity as the data size grows. Note that there is not a clear difference between the series of results for the three Gaussian processes. This can be attributed to the prior on λ which acts as expected and adapts rightly to the smoothness of the path.

In addition to studying the model through plots, we now turn to examine numerically the fit of the $\mathsf{Dep}-\mathsf{GEM}$ model (with the Squared Exponential GP only) and of the independent model of Equation (3.10). To this end, we define the sum of squared errors (SSE) for data $\mathbf{Y}^{(N)}$ between the true Shannon index H_{Shan} and an estimator $\hat{H}$ by

$$\mathsf{SSE}(\mathbf{Y}^{(N)}) = \sum_{X} \left(\hat{H}(X, \mathbf{Y}^{(N)}) - H_{\mathsf{Shan}}(X)\right)^2. \tag{3.29}$$

For a thorough comparison that accounts for sampling variability, we use the mean sum of squared errors $\mathsf{MSSE} = \mathsf{E}\big(\mathsf{SSE}(\mathbf{Y}^{(N)})\big)$, where expectation is with respect to the distribution of $\mathbf{Y}^{(N)}$. To that purpose we iterate above estimation 100 times (*i.e.* draw data $\mathbf{Y}^{(N)}$ from the distribution of Equation (3.28) 100 times for each N), and also estimate the independent GEM model. For both models we run an MCMC algorithm of 5 000 iterations with a burn-in of 2 000 iterations, from Algorithm 1 for the dependent model, and from a Gibbs sampler for the second model. This last sampler is based on the beta posterior of Equation (3.12) conditional on the precision parameter M, and on the full conditional of M of Equation (3.13). The results are presented in Table 3.1, and show a global improvement with the $\mathsf{Dep}-\mathsf{GEM}$ model over the independent GEM model.

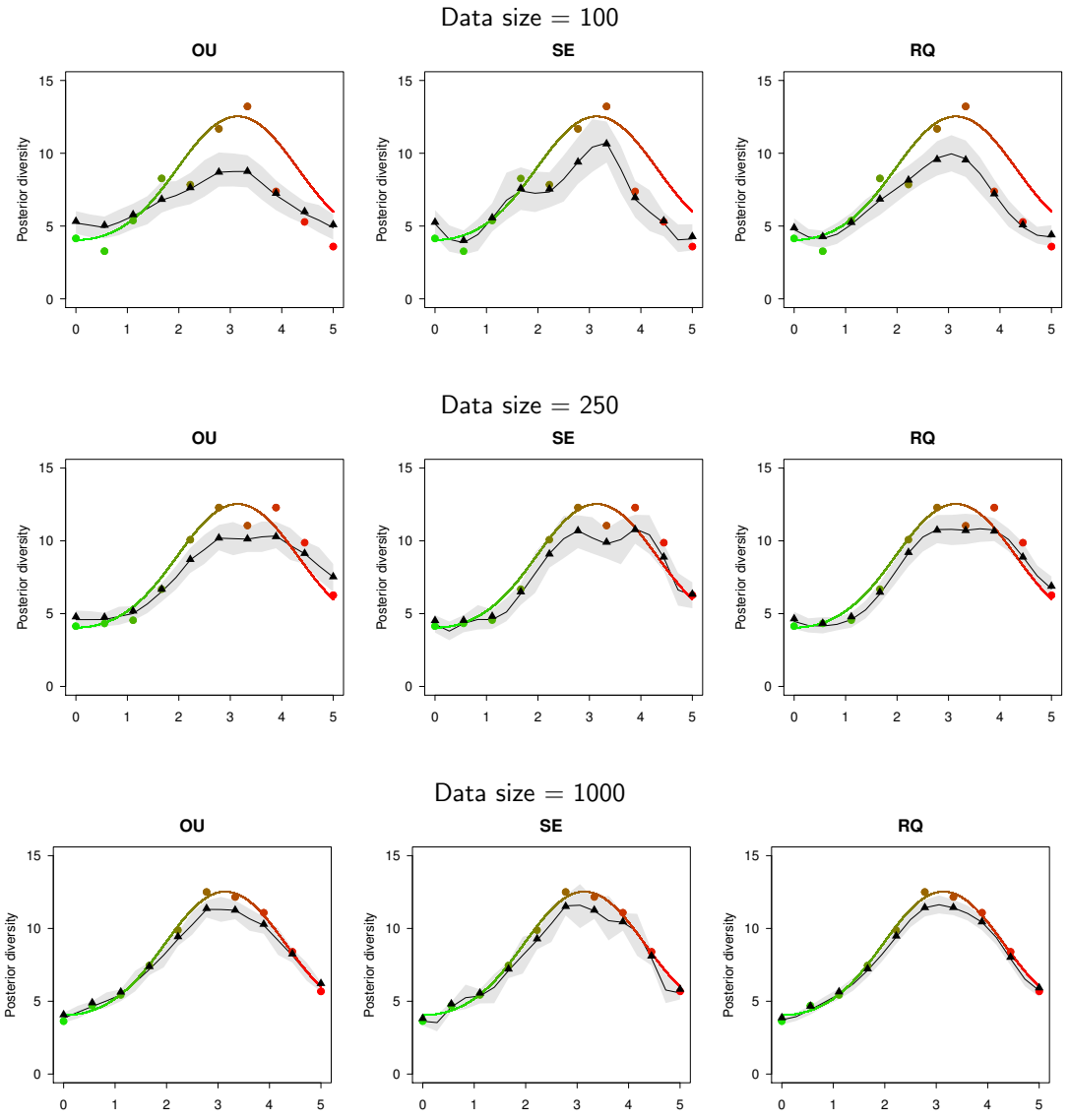

FIGURE 3.4: **Posterior estimation of the Shannon diversity index** in the simulated example (3.28) (100 000 replications). Black triangles are the estimates at observed covariates, the black curve is the estimated predictive, and the gray shade represents a 95% credible interval for the predictive distribution. Colour dots represent the Shannon index in simulated data. Resp. 100, 250 and 1 000 observations were simulated from (3.28) in the first, second and third line. The x-axis represents the TPH contaminant.

3.6.2 Microbial data

We now test the model on real microbial data. The data set consists of measurements of abundance of Operational Taxonomic Units (OTUs, see Schloss and Handelsman, 2005), conducted at sites in Casey, Antarctica. OTU measurements are paired with a contaminant factor. Although a continuous variable, TPH has the same value for several sites. When this is the case, we choose to collapse the sites together by adding the abundances. This data set is studied in more details in Arbel et al. (2013b).

	MSSE	
N	Dep – GEM	⫫ GEM
100	84.3	91.8
250	32.7	36.3
1000	13.9	15.1

TABLE 3.1: **MSSE between true and estimated Shannon diversity index** in example (3.28) in the Dep – GEM model with Sqaured Exponential GP *(first column)* and the independent GEM model *(second column)*, for varying data size N *(in rows)*. Expectation with the data distribution is computed by averaging over 100 data sets simulated from example (3.28) (5 000 replications) for each N.

The comparison between the diversity estimated in the independent GEM and in the Dep – GEM models shows that some smoothing operates in the second case. In the latter, the diversity is also available as a path in X, not only at observed predictor values X_i's, but in the whole space $\mathcal{X}$ equal to the real line here.

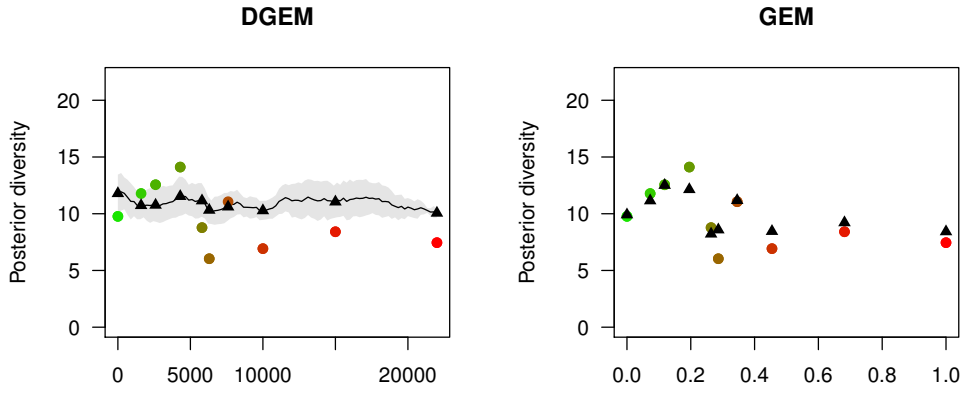

FIGURE 3.5: **Comparison between the dependent and independent model estimations.** *Left*: Dep – GEM model estimates (50 000 replications). *Right*: GEM model estimates (50 000 replications). Black triangles: Posterior mean of the Shannon diversity index. Color dots: Shannon diversity in raw data.

3.7 Discussion

We have presented a Bayesian nonparametric dependent model for species data, based on the distribution of the weights of a DDP, named Dep – GEM distribution, which is constructed thanks to Gaussian processes.

A fundamental advantage of our approach based on the stick-breaking is that it brings a lot of flexibility when it comes to define the dependence structure. It is indeed defined by the kernel of a Gaussian process, compared to a single parameter in many approaches. Such a large flexibility can somehow be criticized, however, when we deal with nonparametric Bayes priors, we think that the more flexibility the better since the different features can be learnt with increasing data. On the other hand, there are examples in the literature where the dependence structure is defined with less parameters, *eg* Caron et al. (2006). If it is much less flexible, the counterpart advantage is the readily availability of marginal posterior sampling schemes (see for instance the construction by Caron et al. (2006) based on the Pólya urn).

In terms of model fit, we have shown that the $\mathsf{Dep}-\mathsf{GEM}$ model improves estimation compared to an independent GEM model. This was conducted by computing the mean sum of squared errors on a simulated example were the fit of the model can be compared to the true sampling process.

There are computational limitations to the use of this model. The estimation can deal with large rows data (large number of observations) since the complexity grows linearly with the number of different observed species J, however the number of unique covariate values I is the dimensionality of the estimation problem and represents its limiting factor. As mentioned before, one could consider to use INLA approximations in the case of prohibitively large I.

The $\mathsf{Dep}-\mathsf{GEM}$ model is tested in the present paper on univariate factors only. An interesting extension concerns multivariate factors, that is to say factors $\mathbf{X} = (X^1, \ldots, X^k) \in \mathbf{R}^k$ (instead of $\mathbf{X} = X \in \mathbf{R}$). To that purpose, all the methodology presented in Section 4.3 and in Section 3.4 remains valid, but some additional care should be taken in defining the Gaussian process $\mathcal{Z}^k : \mathbf{R}^k \to \mathbf{R}$ (instead of $\mathcal{Z} : \mathbf{R} \to \mathbf{R}$) and the k–dimensional array $\mathbf{Z}^k$ (instead of the vector $\mathbf{Z}$). Applications are promising, such as testing joint effects in dynamical models (time $\times$ contaminant factors), in spatial models (position $\times$ contaminant factors), *etc*. This will be the subject of future investigations.

3.8 Appendicies

3.8.1 One-to-one relation between p and V

Note first that there is a one-to-one relation between $\mathbf{p}$ and $\mathbf{V}$ which allows to estimate the model in terms of $\mathbf{p}$ or of $\mathbf{V}$ equivalently. The inverse transform of the stick-breaking

relation of Equation (3.9) is

$$V_j = \frac{p_j}{\left(1 - \sum_{l<j} p_l\right)},$$

since $\sum_{l<j} p_l + \prod_{l<j}(1 - V_l) = 1$.

3.8.2 Posterior mean and maximum likelihood estimates

Here we compare the posterior mean and maximum likelihood estimators of the relative proportions p_{ij} under the independent GEM model of Equation (3.10). In the case of a fixed precision parameter M, the posterior means of V_{ij} and p_{ij} from (3.12) have the following closed-form expressions:

$$\hat{V}_{ij}^{\mathsf{B}} = \frac{1 + N_{ij}}{1 + M + \bar{N}_{i,j}},$$

where $\bar{N}_{i,j} = \sum_{l \geq j} N_{il}$. By independence of the V_{ij}'s and by using Equations (3.9):

$$\hat{p}_{ij}^{\mathsf{B}} = \frac{1 + N_{ij}}{1 + M + \bar{N}_{i,j}} \frac{M + \bar{N}_{i,j}}{1 + M + \bar{N}_{i,j-1}} \cdots \frac{M + \bar{N}_{i,2}}{1 + M + \bar{N}_{i,1}},$$

where $\bar{N}_{i,1}$ boils down to N_i. We can assess how close the posterior mean $\hat{p}_{ij}$ of p_{ij} is from the maximum likelihood estimate $\hat{p}_{ij}^{\mathsf{emp}} = N_{ij}/N_i$ by the following first order approximation (based on the hypothesis of large abundances N_{ij}):

$$\hat{p}_{ij}^{\mathsf{B}} = \frac{1 + N_{ij}}{1 + M + N_i} \prod_{l=2}^{j} \frac{M + \bar{N}_{i,l}}{1 + M + \bar{N}_{i,l}} \approx \frac{1 + N_{ij}}{1 + M + N_i} \left(1 - \sum_{l=2}^{j} \frac{1}{M + \bar{N}_{i,l}}\right),$$

$$\hat{p}_{ij}^{\mathsf{B}} \approx \hat{p}_{ij}^{\mathsf{emp}} \left(1 + \frac{1}{N_{ij}} - \frac{1 + M}{N_i} - \sum_{l=2}^{j} \frac{1}{M + \bar{N}_{i,l}}\right).$$

Acknowledgements

The authors thank Nicolas Chopin for suggesting to work with Gaussian processes, and Annalisa Cerquetti for helpful comments on diversity measures.

Bibliography

Arbel, J., Mengersen, K., and Rousseau, J. (2013a). Bayesian nonparametric dependent models for the study of diversity in species data. *Manuscript under preparation*. 78

Arbel, J., Mengersen, K., Rousseau, J., Raymond, B., and King, C. (2013b). Ecotoxicological data study of diversity using a dependent Bayesian nonparametric model. *Manuscript under preparation*. 97

Archer, E., Park, I. M., and Pillow, J. (2013). Bayesian Entropy Estimation for Countable Discrete Distributions. http://arxiv.org/abs/1302.0328. 94

Boender, C. and Kan, A. R. (1987). A multinomial Bayesian approach to the estimation of population and vocabulary size. *Biometrika*, 74(4):849–856. 80

Caron, F., Davy, M., Doucet, A., Duflos, E., and Vanheeghe, P. (2006). Bayesian inference for dynamic models with Dirichlet process mixtures. In *Proc. International Conference on Information Fusion*. Citeseer. 79, 83, 99

Cerquetti, A. (2012). Bayesian nonparametric estimation of Simpson's evenness index under $\alpha-$Gibbs priors. *arXiv preprint arXiv:1203.1666*. 80, 94

Chung, Y. and Dunson, D. (2009). The local Dirichlet process. *Annals of the Institute of Statistical Mathematics*, pages 1–22. 79, 83

Dunson, D. and Park, J. (2008). Kernel stick-breaking processes. *Biometrika*, 95(2):307. 79, 83

Dunson, D., Pillai, N., and Park, J. (2007). Bayesian density regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(2):163–183. 79, 83

Gibbs, M. N. (1997). *Bayesian Gaussian processes for regression and classification*. PhD thesis, Citeseer. 88

Gill, C. A. and Joanes, D. N. (1979). Bayesian estimation of Shannon's index of diversity. *Biometrika*, 66(1):81–85. 81, 82

Gnedin, A. and Pitman, J. (2006). Exchangeable Gibbs partitions and Stirling triangles. *Journal of Mathematical Sciences*, 138(3):5674–5685. 80

Good, I. J. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika*, 40(3-4):237–264. 80

Griffin, J. and Steel, M. (2006). Order-based dependent Dirichlet processes. *Journal of the American statistical Association*, 101(473):179–194. 79, 83

Hill, B. M. (1979). Posterior moments of the number of species in a finite population and the posterior probability of finding a new species. *Journal of the American Statistical Association*, 74(367):668–673. 80

Holmes, I., Harris, K., and Quince, C. (2012). Dirichlet Multinomial Mixtures: Generative Models for Microbial Metagenomics. *PloS one*, 7(2):e30126. 83

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. 93

Lijoi, A., Nipoti, B., and Prünster, I. (2013a). Bayesian inference with dependent normalized completely random measures. *Bernoulli*, to appear. 79, 83, 93

Lijoi, A., Nipoti, B., and Prünster, I. (2013b). Dependent mixture models: clustering and borrowing information. Technical report, University of Pavia, Department of Economics and Management. 79, 83

MacEachern, S. (1999). Dependent nonparametric processes. *ASA Proceedings of the Section on Bayesian Statistical Science*, pages 50–55. 77, 78, 79, 83, 85

Müller, P., Quintana, F., and Rosner, G. L. (2011). A product partition model with regression on covariates. *Journal of Computational and Graphical Statistics*, 20(1). 93

Pitman, J. (1995). Exchangeable and partially exchangeable random partitions. *Probability Theory and Related Fields*, 102(2):145–158. 93

Pitman, J. (2006). *Combinatorial stochastic processes*, volume 1875. Springer-Verlag. 84, 85, 94

Rasmussen, C. E. and Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning*. MIT Press. 88, 90

Rue, H., Martino, S., and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the royal statistical society: Series b (statistical methodology)*, 71(2):319–392. 90

Schloss, P. and Handelsman, J. (2005). Introducing DOTUR, a computer program for defining operational taxonomic units and estimating species richness. *Applied and environmental microbiology*, 71(3):1501–1506. 79, 97

Sporysheva, P. and Petrone, S. (2013). Bivariate Species Sampling Models. *Manuscript under preparation*. 93

van der Vaart, A. W. and van Zanten, J. (2009). Adaptive Bayesian estimation using a Gaussian random field with inverse Gamma bandwidth. *The Annals of Statistics*, 37(5B):2655–2675. 88

Chapter 4

Ecotoxicological data study of diversity using a dependent Bayesian nonparametric model

On étudie dans ce chapitre le modèle bayésien non-paramétrique du Chapitre 3 dans une perspective plus appliquée, avec comme domaine d'application l'écotoxicologie. Ici, les espèces sont des microbes, et le facteur est une variable de contamination environnementale importante appelée Hydrocarbure de Pétrole Total. On étudie son impact sur les données sous des angles différents: en terme de diversité de Shannon, de clustering (en des groupes de microbes qui réagissent de manière similaire au contaminant), et en terme de décroissance de la proportion relative des espèces (l'estimation de quantités appelées IC_{50}, qui correspondent au niveau de contamination pour lequel la proportion relative est divisée par deux par rapport à la proportion relative à une valeur de contamination de référence). Ce modèle, étudié sur des données microbiennes mesurées dans le sol en Antarctique, est applicable plus généralement à de nombreux autres problèmes dans lesquels la structure des données et les questions inférentielles sont similaires.

Authors

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

Status

Manuscript Arbel et al. (2013b) under preparation.

Abstract

We study the Bayesian nonparametric model of Chapter 3 from a more applied perspective, with a focus in the field of ecotoxicology. Here, the species are microbes, and the factor is an important environmental contaminant called Total Petroleum Hydrocarbon. Its impact on the data is studied from different points of view: in term of Shannon diversity, of clustering (into groups with similar behaviour with respect to the contaminant), and of species relative proportion decrease (estimation of a quantity called IC_{50}, the covariate level for which the relative proportion is divided by 2 compared to the relative proportion at a given covariate value). The model, which is studied on soil microbial data collected in Antarctica, is broadly applicable to a range of other problems with the same data structure and inferential requirements.

keywords Antarctica, Dependent models, Fuel spills, Griffiths-Engen-McCloskey distribution, Shannon diversity index, Species abundance, Species-by-site data, Soil biodiversity, Total Petroleum Hydrocarbon (TPH)

4.1 Introduction

An understanding of the environmental processes that affect ecosystems is of fundamental importance for their management and conservation. Ecotoxicology is primarily concerned with predicting the effects of toxic substances on the biological components of the ecosystem. This information is critical to the derivation of environmental quality

guidelines. These include trigger values or contaminant thresholds, which when exceeded prompt remediation and/or clean-up activities, and remediation targets which define an acceptable level of ecosystem recovery and restoration, and once reached, enable site sign off as no longer posing environmental risk. In remote, high latitude environments such as Antarctica, where field work is logistically difficult and expensive, and where terrestrial ecosystems are comprised of relatively few species and simple food webs, appropriate modelling tools can be valuable as an alternative to traditional toxicity tests using local biota to generate sensitivity data to predict effects on the ecosystem.

While ecotoxicological assessments aim to predict the effects of contaminants on an ecosystem, monitoring and characterizing the state of an entire ecosystem is rarely practicable. Hence toxicity tests are generally conducted on single species (populations), or groups of species (communities), as indicators of the overall system state. Community assessments may provide more representative and relevant information that incorporates complex interactions between species compared with simple single species tests.

Modelling the responses of species or community to contamination gradients is conceptually very similar to the broader goal of modelling species responses to environmental conditions, which is an area of long-standing study in the ecological sciences. Conventional observational studies have generally dealt with a relatively small number of species, and the modelling methods have been developed accordingly. Thus, while methods for single-species modelling are relatively mature and diverse (*eg* Elith et al., 2006), community modelling methods are less well established. One approach to community modelling is to model single species in an independent fashion, and then assemble the individual model predictions into a composite prediction of the community (*eg* Ellis et al., 2011). However, such approaches typically struggle with rare species, which are difficult to model with confidence because of their sparse observations. Appropriate propagation of the uncertainty in individual species models into the composite predictions can also be difficult. An alternative approach, which has become more common in recent years, is to simultaneously model the response of the community as a whole. This can take the form of multi-response modelling of multiple species (*eg* Dunstan et al., 2011, Foster and Dunstan, 2010, Wang et al., 2012), or modelling of univariate summaries of multi-species responses, such as species richness or compositional dissimilarity (*eg* Ferrier and Guisan, 2006, Ferrier et al., 2007) or rank abundance distributions (Foster and Dunstan, 2010). The development of community modelling methods has, at least in part, been driven by the emergence of high-throughput microbial and similar studies, which can provide information on tens of thousands of species simultaneously, many of which can be extremely sparse.

Modelling species abundance patterns is crucial in ecology. The species under study might be diverse, such as microbes (as in this paper), alleles for genes, or animals (*eg* birds in MacArthur, 1957), *etc*. One way of describing these patterns is to model the relative probabilities p_j of each species j. The multinomial-type methodologies for modelling species data are very popular in the ecotoxicological literature (see *eg* Bohlin et al., 2009, Fordyce et al., 2011, De'ath, 2012, Holmes et al., 2012) and the genomic literature, see for example Dunson and Xing (2009) who use a Dirichlet process mixture of product multinomial distributions. One of the main reasons of the success of the multinomial specification is its attractiveness as an intuitive modelling of the species relative proportions. The multinomial distribution, which generalizes the binomial distribution when there are more than two species, is an intuitive framework for species sampling when the sampling process consists of independent observations of a fixed number of species. This distribution gives the probability of any observed combination of theses species conditional to parameters which are the species relative proportions.

It is also particularly successful among ecologists since the species relative proportions are precisely the parameters of interest when the focus is on ecological indices such as species richness, diversity, and evenness (the literature on diversity is extensive, see *eg* Hill, 1973, Patil and Taillie, 1982, Foster and Dunstan, 2010, Colwell et al., 2012, De'ath, 2012). A model on the relative probabilities $\mathbf{p} = (p_1, p_2, \ldots)$ allows inference about most of these indices. Although we focus on this modelling question in the context of diversity in ecology, the same question arises in other areas of science such as biology, engineering, physics, chemistry, economics, health and medicine (see Borges and Roditi, 1998, Havrda and Charvát, 1967, Kaniadakis et al., 2005), and in more mathematical fields such as probability theory and mathematics (Donnelly and Grimmett, 1993). Diversity itself can be defined in a number of ways. Most simply, it refers to a crude measure of the number of species in a sample (see Hill, 1979, Boender and Kan, 1987). A common definition of diversity that is predominant in ecology and which will be used here is the Shannon index defined by

$$H_{\mathsf{Shan}}(\mathbf{p}) = -\sum_j p_j \log p_j. \tag{4.1}$$

Other diversity indices include the Simpson index $H_{\mathsf{Simp}}(\mathbf{p}) = 1 - \sum_j p_j^2$ and the generalized diversity index which was proposed by Good (1953) in the form of $H_{\mathsf{Good},\alpha,\beta}(\mathbf{p}) = -\sum_j p_j^\alpha \log^\beta p_j$ for non-negative integer values of α and β. In this study, focus will be restricted to the Shannon index H_{Shan}, although the procedures could be equally applied to these other measures.

We will describe the approach of Holmes et al. (2012) and present an extension proposed in Arbel et al. (2013a) based on a Bayesian nonparametric model which incorporates dependence with respect to a factor. Improvements of the methodology of Holmes et al.

(2012) mainly consists (i) in allowing for an unknown a priori number of species, thanks to the nonparametric formulation, and (ii) in accounting for additional factors, typically contamination, thanks to the dependence in the model. This brings estimates which are more efficient, both computationally and inferentially, and which allow assessment of the response of species, for instance in terms of diversity, to contamination.

The dataset under study was collected at Casey station in Antarctica. The data comprise counts of a large number (of the order of 1 800) species of microbe species (called OTU) collected at 60 sites, resulting in a very sparse matrix. As is usual practice, the species were combined into 32 groups and the counts of species aggregated within each of these groups. Environmental and soil covariates that are likely to influence the abundance pattern are also available. In the present study, focus is restricted to one of these covariates, namely Total Petroleum Hydrocarbon (TPH), which is a measure of fuel spills. This is one of the most important environmental contaminants in Antarctica.

We will use the following notation in order to describe the data. Let i denote the measurement index, $i \in \{1, \ldots, I\}$, I =10, X_i the TPH level, j the OTU index, $j \in \{1, \ldots, J\}$, J =32 and $N_{i,j}$ its abundance. Total OTU j abundance is denoted N_j, total OTU abundance for site i is denoted N_i (this ambiguous notation is for the sake of simplicity), and overall abundance is denoted N. Bold characters denote vectors or matrices of observations or parameters, *eg* $\boldsymbol{N}_i = (N_{i1}, \ldots, N_{iR})$, $\boldsymbol{N}_j = (N_{1j}, \ldots, N_{Ij})$; $\boldsymbol{N}$ stands for the whole $I \times J$ matrix of observations. The same notations with p instead of N relate to proportions, or average proportions, instead of abundance.

The rest of the paper is organized as follows. Section 4.2 provides a descriptive study of the data. Section 4.3 defines the Bayesian model for species-by-site count data and compares it with other models in the literature. Results are given in Section 4.4, dealing first with the estimation of a univariate output by TPH (OTU diversity), then with classification (of OTUs in terms of response with TPH patterns) and finally with the estimation of model patterns (IC_{50} and IC_{25}).

4.2 Data analysis

In this section we briefly describe the dataset. Data consist of measurements of abundance of Operational Taxonomic Units (OTUs, or microbes, see Schloss and Handelsman, 2005), conducted at sites in Casey, Antarctica.

OTU measurements are paired with covariates mainly of two kinds: first geographical parameters of the place (latitude, longitude, elevation, slope), transect number, soil measurements (water, chemical elements concentrations, TPH). The study will focus on

the effect of TPH level[1] (for Total Petroleum Hydrocarbon, a measure of compounds of hydrocarbons that are found in crude oil) on OTUs abundance. Although a continuous variable, the same value of TPH was recorded for several sites. Given this, we choose to collapse the sites together by adding the abundances, which results in the following 10 unique TPH values (expressed in thousands) 0, 1.6, 2.6, 4.3, 5.8, 6.3, 7.6, 10, 15, 22.

We focus on the subset of data where OTUs' abundance exceeds 40 over all measurements (32 of them). We have studied other subsets of data with either more of less OTUs, with no substantial change in the estimation.

We show in Table 4.1 a subsample of the data at the three lowest and three highest TPH sites for a few OTUs.

	Site	TPH	OTU03964	OTU00527	OTU00396	OTU03930	OTU05279	OTU03882
1	1	0	2	331	197	0	0	0
2	22653	1600	187	73	32	85	1	3
3	22654	2600	384	64	48	198	115	61
8	22645	10000	640	11	8	290	70	97
9	22650	15000	551	17	21	142	227	167
10	22644	22000	379	0	7	226	443	364

TABLE 4.1: Subsample of the data. The rows are sites, the columns species, and cells give the count data. The name of the OTU species is given in the header line.

The abundance data is illustrated on Figures 4.1 and 4.2. On Figure 4.1 OTUs are sorted by decreasing overall abundance. The top part shows total abundance by OTU, while the bottom part gives the abundance per site, where abundance is expressed in natural scale on the left, and in log scale on the right. Figure 4.2 shows abundance N_{ij} site by site, with increasing TPH. Note that the y scales are different across plots. Colours indicate the TPH level of the corresponding site. It goes from green for the lowest TPH measure, $X = 0$, to red for the highest, $X = 22 \times 10^3$.

Although sorted by total abundance, Figure 4.2 indicates that (i) OTUs abundance site by site does not follow the same sorting, and (ii) the relative proportions are very diverse. This remark is also supported by the graphs on top of Figure 4.1 were we see that few OTUs share most of the observations. This leads to look at the number of OTUs which gather a given proportion of the total abundance at each site. This is illustrated on Figure 4.3 for a proportion of 90%. This quantity decreases with TPH, which means that the relative mass of the biggest OTUs increases with TPH. As a consequence, the

[1] It is expressed in mg/kg, so the TPH measure is per mass unit.

diversity as measured by this index deteriorates with TPH. The left side of Figure 4.3 illustrates the Shannon diversity index.

Taking account of theses remarks leads to consider the GEM distribution as a good candidate prior for the distribution of the relative frequencies of OTUs $\mathbf{p}$, as explained later in Section 4.3.

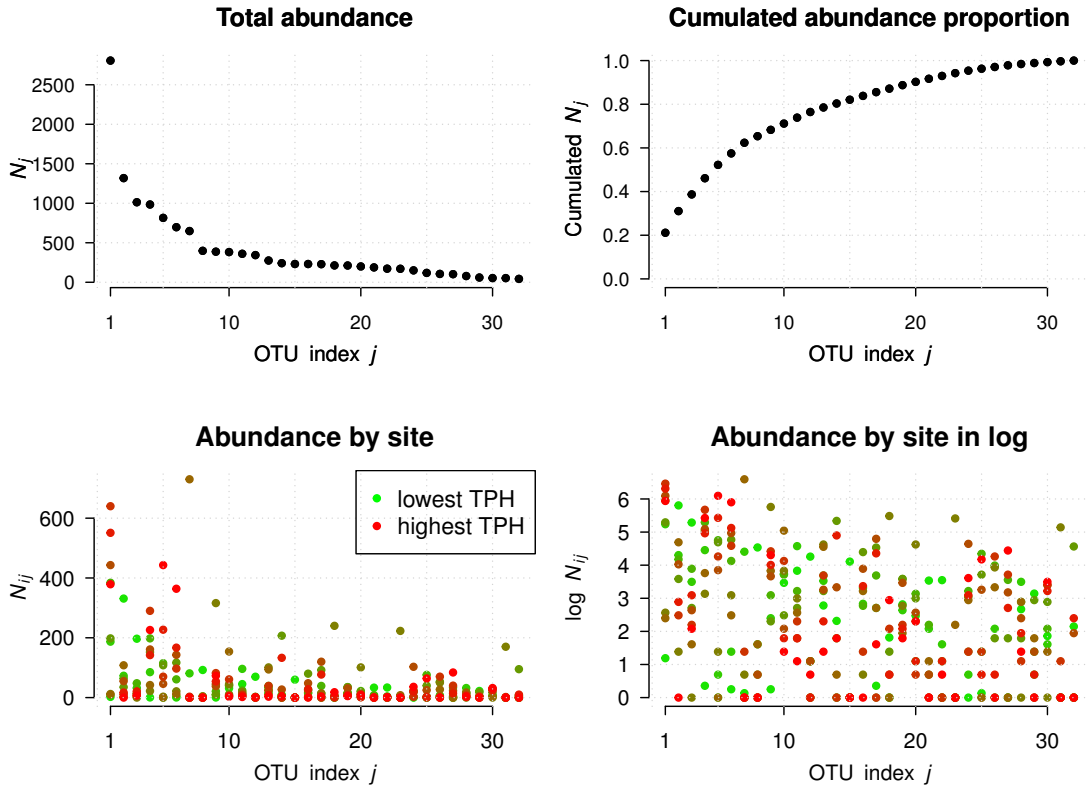

FIGURE 4.1: **OTUs abundance.** The x-axis represents the species index $j = 1, 2, \ldots$. Throughout the paper, the colors represent the TPH level, from green (minimum TPH, $X_1 = 0$) to red (maximum TPH, $X_{10} = 22 \times 10^3$).

4.3 Bayesian model for species–by–site data

4.3.1 Sampling model

We describe here the notation and sampling process of covariate dependent species-by-site count data. Each unique covariate value is indexed by i and is denoted by X_i. Recall that is may correspond to a single site or to a collection of sites with the same covariate value. For the sake of simplicity, we may still speak of *site* i. Individual observations at site i are taxa, or species, indexed by natural numbers $j = 1, 2, \ldots$. The total number of observed species is denoted by J. No hypothesis is made on the unknown total number

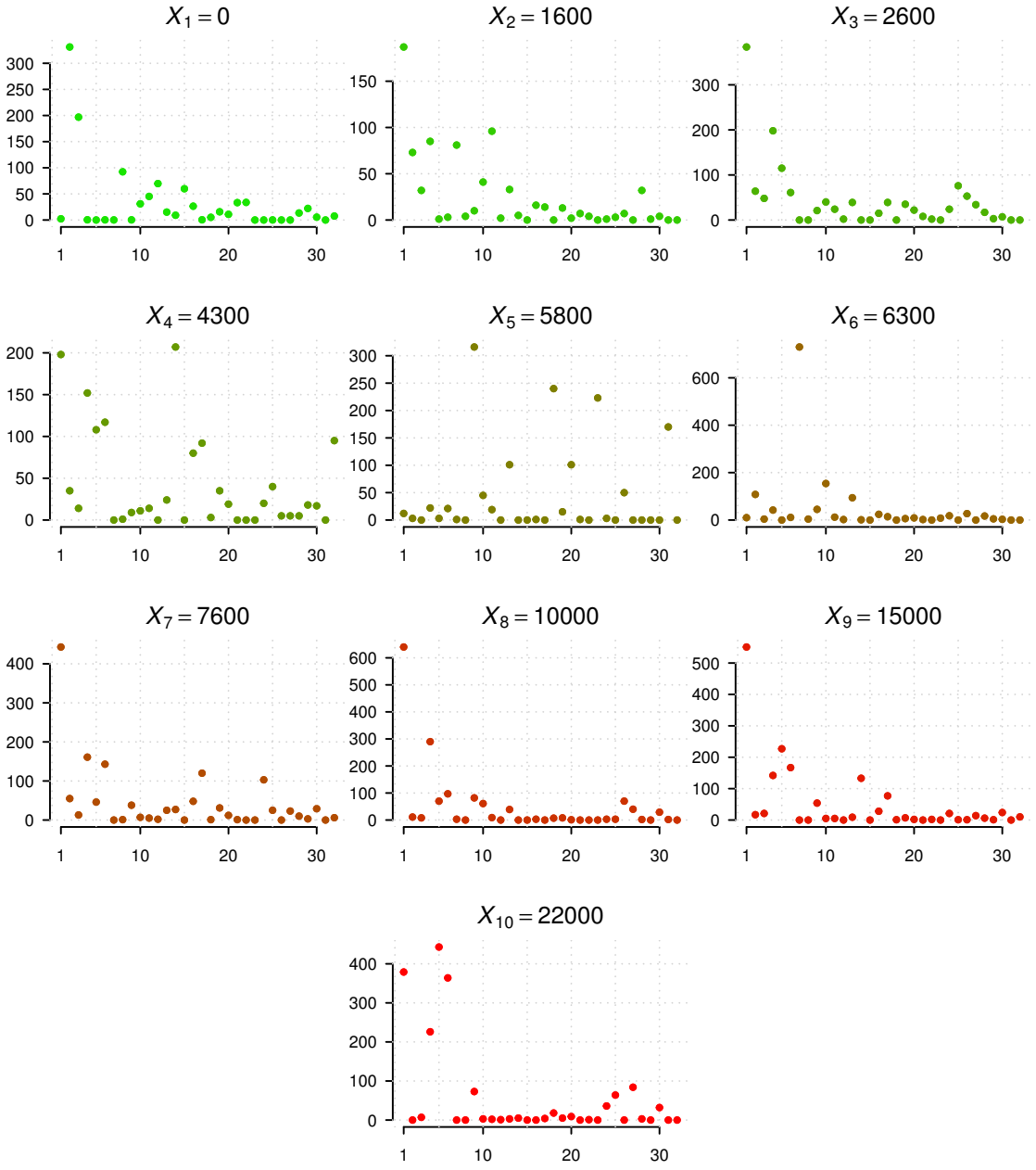

FIGURE 4.2: **OTUs abundance per site.** Each plot represent a site sorted by TPH. The x-axis represents the species index $j = 1, 2, \ldots$.

of species in the population of interest, which might be infinite. Observe $(X_i, \mathbf{Y}_i^{N_i})_{i=1,\ldots,I}$ 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 value X_i. Species j abundance at site i is denoted by N_{ij}, *i.e.* the number of times when $Y_{n,i}) = j$ with respect to n index. Relative abundance satisfy $\sum_j N_{ij} = N_i$.

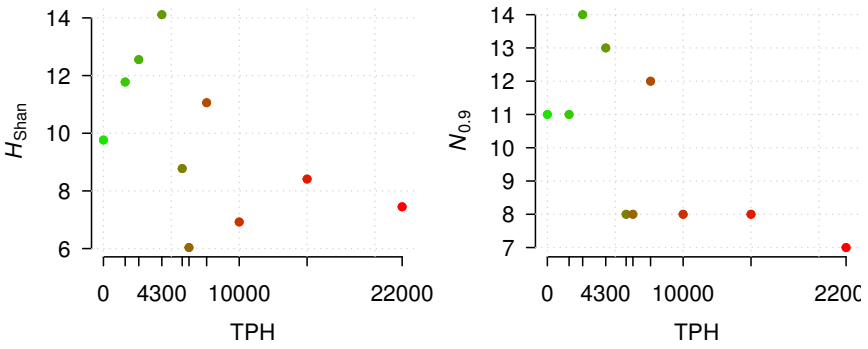

FIGURE 4.3: **Diversity indices.** *Left*: Shannon diversity index H_{Shan}. *Right*: Number of OTUs with cumulated mass at least 90%, denoted by $N_{0.9}$. The x-axis represents the TPH contaminant.

We model the relative frequencies or abundances $\mathbf{p} = (p(X_i))_i = (p_j(X_i))_{i,\,j=1,2,\ldots}$ by the following. For $i = 1 \ldots I$ and $n = 1 \ldots N_i$:

$$\text{for } i = 1 \ldots I,\ n = 1 \ldots N_i, \quad Y_{n,i} \,|\, \mathbf{p}(X_i) \overset{\text{ind}}{\sim} \sum_{j=1}^{\infty} p_j(X_i)\delta_j, \tag{4.2}$$

where $\mathbf{p}(X_i) = (p_j(X_i))_{j=1\ldots J}$ are the parameters of interest. The point mass δ_j means that the probability that observation $Y_{n,i}$ is of species j equals $p_j(X_i)$. The probability of $Y_{n,i}$ is $f(Y_{n,i}|\mathbf{p}(X_i), X_i) = \prod_{j=1}^{J} p_j(X_i)^{\mathsf{I}(Y_{n,i}=j)}$. The likelihood of data at site i is given by $L_i(\mathbf{Y}_i^{N_i}|\mathbf{p}(X_i), X_i) = \prod_{j=1}^{J} p_j(X_i)^{N_{ij}}$, so the likelihood of the model is:

$$L(Y|\mathbf{p}, X) = \prod_{i=1}^{I}\prod_{j=1}^{J} p_j(X_i)^{N_{ij}}. \tag{4.3}$$

Most of the models considered in the literature, such as Holmes et al. (2012), put independent priors on the by-site vectors $\mathbf{p}(X_i)$'s. The main contribution of this article is to use a dependent prior on the $\mathbf{p}(X_i)$'s across the sites i. As we are dealing with continuous covariates X, this approach is arguably more appropriate than discrete prior specifications on $\mathbf{p}(X_i)$ across i.

In Holmes et al. (2012), the total number of species is assumed to be known and denoted by J_{T}, so in this case, the observational model given in Equation (4.2) is equivalent to the following *multinomial model* at the species level:

$$\boldsymbol{N}_i = (N_{i1}, \ldots, N_{iJ_{\mathsf{T}}})|X \sim \mathcal{M}(N_i, p_1(X_i), \ldots, p_{J_{\mathsf{T}}}(X_i)). \tag{4.4}$$

The assumption of a fixed number of species, J_T, and a fixed number of observations by sites, N_i, is a limitation of this model in applications compared to model (4.2).

The vectors $\mathbf{p}(X_i) = (p_1(X_i), \ldots, p_{J_\mathsf{T}}(X_i))$ are first modelled by Holmes et al. (2012) by the Dirichlet distribution $\mathbf{p}(X_i) \sim \mathsf{Dir}(\mathbf{p}(X_i)|\boldsymbol{\alpha})$ with $\boldsymbol{\alpha} = (\alpha, \ldots, \alpha)$. It is a natural choice as it is conjugate to the multinomial model of Equation (4.4), with the following posterior distribution:

$$\mathbf{p}(X_i)|\mathbf{N}_i \sim \mathsf{Dir}(\mathbf{p}(X_i)|\boldsymbol{\alpha} + \mathbf{N}).$$

In the simple case where the hyperparameters are fixed, the posterior mean of the proportions p_{ij} has a closed form expression

$$\hat{p}_{ij}^{\mathsf{B}} = \frac{\alpha + N_{ij}}{\alpha J_\mathsf{T} + N_i},$$

and the marginal posterior variances are given by

$$(\hat{\sigma}_{ij}^{\mathsf{B}})^2 = \frac{(\alpha + N_{ij})(\alpha(J_\mathsf{T} - 1) + N_i - N_{ij})}{(\alpha J_\mathsf{T} + N_i)^2(\alpha J_\mathsf{T} + N_i + 1)}.$$

As noted above, extra flexibility is provided by Holmes et al. (2012) by considering the following mixture of Dirichlet distributions:

$$\mathbf{p}(X_i) \sim \sum_{k=1}^{K} \pi_k \mathsf{Dir}(\mathbf{p}(X_i)|\boldsymbol{\alpha}_k).$$

This prior is convenient for clustering across sites: each site vector $\mathbf{p}(X_i)$ is assumed to derive from a single component k of the mixture. However, in the case of a continuous covariate X, it fails to describe a continuous dependence of the prior with respect to X. There is no clear argument in favour of the use of a mixture: it creates indeed a prior partition of the covariate space $\mathcal{X}$ into an *ad hoc* number K of sets, each associated with a mixture component $k = 1, \ldots, K$. An alternative which is adopted here is to relax this $0-1$ assumption and extend this idea by building a model which evolves smoothly with the covariate X: a slight change in X_i induces a slight change in $\mathbf{p}(X_i)$.

4.3.2 Dependent GEM distribution

The dependent Bayesian nonparametric model which is used here is defined in detail in Chapter 3. We summarize here its most important features.

The Dirichlet process (DP) is very popular in Bayesian nonparametric as a prior distribution on probability measures. It was shown by Sethuraman (1994) that its distribution can be written as the distribution of an infinite mixture of point masses. In our case,

we want to model the relative probabilities $\mathbf{p}$, hence only the weights of the mixture are of interest. Their distribution is called the GEM distribution, after Griffiths-Engen-McCloskey. For taking account of environmental factor like TPH, we favour an extension of the DP called the dependent Dirichlet process (DDP), first proposed by MacEachern (1999). We call $\mathsf{Dep}-\mathsf{GEM}$ the distribution of the weights in the DDP. The full Bayesian model is

$$\begin{aligned} &Y_{n,i} \,|\, \mathbf{p}(X_i), X_i \overset{\text{ind}}{\sim} \sum_{j=1}^{\infty} p_j(X_i)\delta_j, \quad \text{for } i = 1 \dots I,\ n = 1 \dots N_i \\ &\mathbf{p} \sim \mathsf{Dep}-\mathsf{GEM}(M). \end{aligned} \tag{4.5}$$

The $\mathsf{Dep}-\mathsf{GEM}$ distribution in Equation (4.5) implies that the weights $p_j(X_i)$ can be written by

$$p_j(X_i) = V_j(X_i) \prod_{l<j} (1 - V_l(X_i)) \text{ with } V_j(X_i) \overset{\text{iid}}{\sim} Beta(1, M). \tag{4.6}$$

This construction of generic probability weights p_j's is called the *stick-breaking* for the following analogy of breaking a stick. Start with a stick of length 1, break it at a random V_1, define $p_1 = V_1$ and do the same for the remaining stick of size $1 - V_1$: break it at a random V_2, define $p_1 = V_2(1-V_1)$ and start again for the stick of length $(1-V_1)(1-V_2)$, and iterate *ad infinitum*. The remaining length $\prod_{l<j}(1 - V_l)$ goes to zero when j goes to infinity, hence the p_j's sum to 1.

The parameter M is called the *precision parameter* of the prior and is now discussed. The motivation of using this model is explained by Figure 4.4 which shows draws of $(p_j)_{j=1\dots J}$ from the prior distribution with various precision parameters M. The similarity with the graphs the empirical OTU frequencies in Figure 4.2 is an argument in favour of the use of the model: the range of empirical proportions patterns is apparently covered by the prior samples conditional to a large enough range of values for M. This can be ensured by the use of a random prior distribution on M.

Figure 4.4 makes it apparent that the precision parameter M controls the level of the diversity in the prior. For small M, only the first species share most of the weights, whereas in the limiting case $M \to \infty$, the weights tend to be uniformly distributed. Cerquetti showed in an unpublished work that the prior expectation of the Shannon index H_{Shan} under the GEM prior is given by

$$\mathsf{E}(H_{\mathsf{Shan}}) = \psi(M + 1) - \psi(1), \tag{4.7}$$

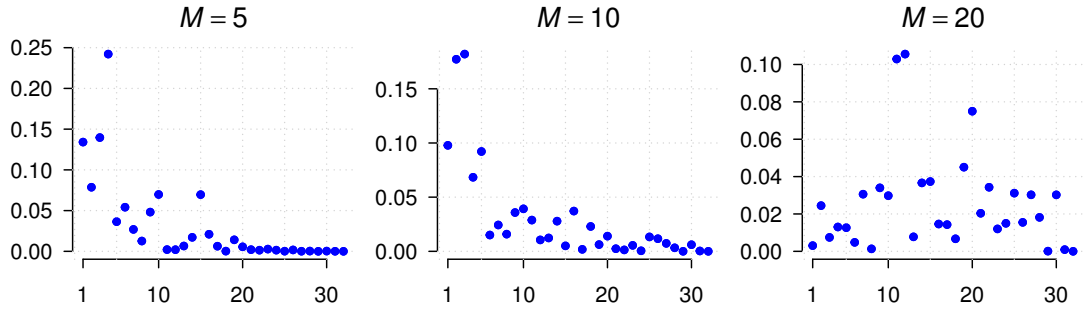

FIGURE 4.4: **Proportions p_j sampled from the Griffiths-Engen-McCloskey distribution.** *From left to right*: precision parameter $M = 5, 10, 20$ (mind the different y axis scaling). The x-axis represents the species index $j = 1, 2, \ldots$.

where ψ is the digamma function, a monotonically increasing function (for similar derivations, see Cerquetti, 2012). This makes clearer the effect of M on the prior diversity, and is illustrated on Figure 1.2.

The likelihood given in Equation (4.3) is factorized across the species $j = 1 \ldots J$. By independence in Equation (4.6), it is also the case of the prior induced on $\mathbf{V}$ through j by the $\mathsf{Dep} - \mathsf{GEM}$ prior, which can thus be written as $\pi(\mathbf{V}|X) = \prod_{j=1}^{\infty} \pi(\mathbf{V}_j|X)$. Thus the following factorized posterior is obtained:

$$\pi(\mathbf{V}|Y, X) \propto \prod_{j=1}^{\infty} \pi(\mathbf{V}_j|X) \times L_j(Y|\mathbf{V}_j, X). \tag{4.8}$$

where $L_j(Y|\mathbf{V}_j, X) = \prod_{i=1}^{I} V_j(X_i)^{N_{ij}}(1 - V_j(X_i))^{\bar{N}_{i,j+1}}$. Note that the latter is equal to 1 for $j > J$ since by definition in this case $N_{ij} = \bar{N}_{i,j+1} = 0$. Using a factorized prior across the species does still allows for dependence across the sites as the $V_j(X_i)$ can be correlated across i in the joint prior $\pi(\mathbf{V}_j|X)$. The factorized expression of the posterior in Equation (4.8) is convenient as it permits sampling independently across the species j. This ameliorates the original problem of estimation of dimension $I \times J$ to J estimation problems in dimension I, which is more efficient with respect to the curse of dimensionality.

The estimation of the model is described in detail in Chapter 3. The $\mathsf{Dep} - \mathsf{GEM}$ prior is constructed with Gaussian processes whose covariance matrices allow a very flexible modelling of the dependence. Posterior sampling is done by a Gibbs algorithm, with Metropolis–Hastings step for non-conjugate conditionals. The model allows estimation of the probability $\mathbf{p}(X)$ also for covariate values which are unobserved. This is achieved by computing the predictive distribution of the Gaussian process. A graphical representation of the model is given in Figure 3.3 of Chapter 3.

4.4 Results

4.4.1 Diversity

We begin with studying the diversity which is a one dimensional output for which the fit of the model is easy to assess.

The comparison between the diversity obtained in the Dep – GEM and in the GEM models in Figure 4.5 shows that some smoothing operates in the first case. This is also shown by a numerical comparison in a simulated example in Table 3.1 of Section 3.6.1 by computing mean sum of squared errors.

An addition advantage in the Dep – GEM modelling is that the diversity is available as a path in X, not only at observed predictor values X_i's, but in the whole space $\mathcal{X}$ equal to the real line here.

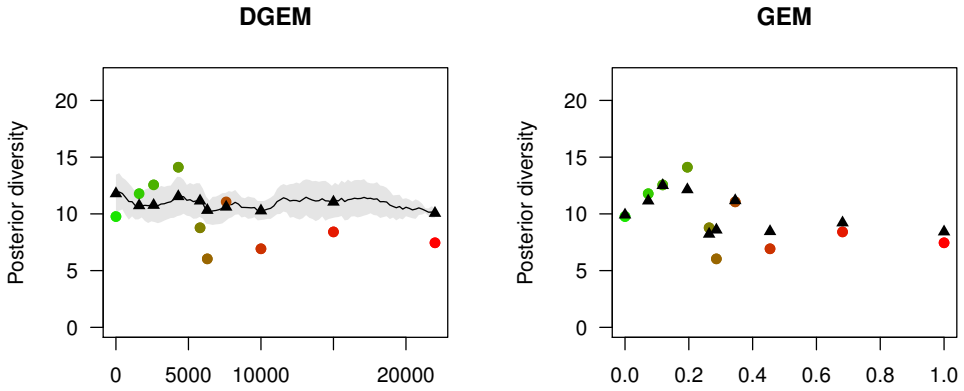

FIGURE 4.5: **Estimation results.** *Left*: Dep – GEM model estimates (100 000 replications). *Right*: GEM model estimates (100 000 replications). Black triangles: Posterior mean of the Shannon diversity index. Gray band: credible interval of the predictive estimate for the diversity index. Color dots: Shannon diversity index in raw data. The x-axis represents the TPH contaminant.

4.4.2 Clustering

The aim of this section is to cluster OTUs with respect to different types of response to TPH exposure in two groups (typically, increasing / decreasing with TPH). We first describe the methodology of the clustering, which uses a posterior sample from any model. Then we compare the clusterings obtained in different models: the Dep – GEM model introduced in Section 4.3.2 which depends on the TPH, and the GEM model, as a model that does not depend on a covariate.

The clustering methodology is based on the following posterior mass for each OTU j

$$m_j = \mathsf{P}\Big(\underset{X<X^{\mathsf{med}}}{\text{mean}}\, p_j(X) > \underset{X\geq X^{\mathsf{med}}}{\text{mean}}\, p_j(X) \,|\, Y \Big),$$

where $X^{\mathsf{med}} = 5\ 800$ is the median of the observed covariates $X_1, \ldots, X_I$. We compare m_j to $1/2$ in order to decide of the clustering:

- if $m_j < 1/2$, then OTU j is increasing with TPH: $+$ group,
- if $m_j > 1/2$, then OTU j is decreasing with TPH: $-$ group.

The clustering results are summed up in Table 4.2. A comparison to existing ecological knowledge about OTUs shows that most of the OTUs within the groupings make biological sense.

	Name	Abundance	Data	DGEM	GEM
1	OTU03964	2806	+	+	+
2	OTU03930	1318	+	+	+
3	OTU05279	1013	+	+	+
4	OTU03882	984	+	–	–
5	OTU03284	815	+	–	+
6	OTU00527	697	–	+	+
7	OTU04061	648	–	–	–
8	OTU03908	398	+	+	+
9	OTU03906	387	–	–	–
10	OTU01369	382	–	–	–
11	OTU05289	360	+	+	–
12	OTU00396	344	–	+	+
13	OTU05005	275	–	+	–
14	OTU02085	241	–	+	–
15	OTU04082	231	–	+	+
16	OTU05400	231	–	–	–
17	OTU01748	229	+	–	–
18	OTU05308	213	–	–	+
19	OTU03907	212	–	–	–

20	OTU05292	200	+	–	–
21	OTU00560	188	–	–	+
22	OTU05403	172	–	–	–
23	OTU00619	171	–	+	–
24	OTU05291	150	+	–	–
25	OTU04854	119	–	–	–
26	OTU00060	105	–	–	+
27	OTU00672	102	–	–	–
28	OTU04855	79	–	+	–
29	OTU04145	60	–	+	+
30	OTU04370	53	–	–	+
31	OTU04856	52	–	–	–
32	OTU01493	43	–	+	–

TABLE 4.2: Clustering of the 32 largest OTUs in 2 groups (+ means increasing with TPH, – means decreasing with TPH) according to the models (Data: row data, $\mathsf{Dep}-\mathsf{GEM}$: dependent model, GEM: independent model)

4.4.3 Estimation of abundance patterns: IC_{25} and IC_{50}

We define IC_{50} as the TPH level for which the OTU relative proportion is divided by 2 compared to its abundance at low TPH level (TPH=0). This holds for the OTU subgroup with a decreasing pattern with TPH, *i.e.* tagged by "–" in the $\mathsf{Dep}-\mathsf{GEM}$ model column of Table 4.2. There is a range of approaches to estimating IC_{50}. Here we consider a simple approach that utilises logistic regression.

We make use of the posterior sample $(p_j^k(X_i))_{ij,k=1\ldots K}$ obtained in the estimation of the $\mathsf{Dep}-\mathsf{GEM}$ model. For each OTU j and for $k = 1, \ldots, K$, we fit the following logistic regression :

$$p_j^k(X_i) = \text{logit}(a_j^k X_i + b_j^k + \epsilon_{ij}^k),$$

with Gaussian errors ϵ_{ij}^k. This allows estimating curves $X \to \hat{p}_j^k(X)$ for each OTU j and for each sample k. As a validation of the clustering procedure of Section 4.4.2, the regression curves behave accordingly to the group the OTU belong to. Hence, now focusing on the OTU j decreasing with TPH, we obtain estimated $\hat{p}_j^k(X)$ curves which

are decreasing, and one can define a posterior sample of IC_{50} and IC_{25} by the TPH values X solving

$$\hat{p}_j^k(X = \mathsf{IC}_{50}) = \frac{\hat{p}_j^k(X=0)}{2} \quad \text{and} \quad \hat{p}_j^k(X = \mathsf{IC}_{25}) = \frac{\hat{p}_j^k(X=0)}{4}.$$

The posterior distributions of both IC_{50} and IC_{25} are illustrated in Figure 4.6 (for the group with a decreasing pattern). The estimation of IC_{50} is more precise than the estimation of IC_{25} because the latter is concerned with a lighter part of the tail of the $p_j(X)$ curves. These distributions are relevant from an ecological point of view, where one needs to interpret the effect of soil contamination on the relative proportions of microbes. However, one limitation of the interpretation of IC_{50} and IC_{25} is there strong sensitivity to the group of OTUs that are accounted for the estimation. Indeed, limiting the OTUs to the most decreasing with TPH ones drastically reduces the estimates of IC_{50} and IC_{25}.

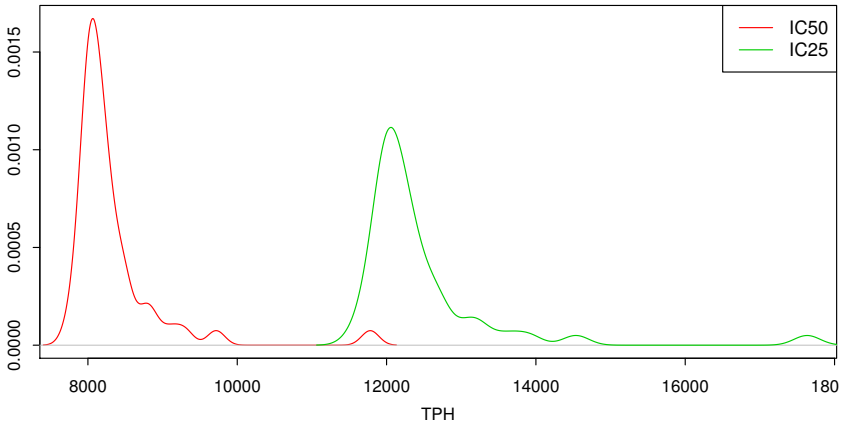

FIGURE 4.6: **Posterior distributions of IC_{50} and IC_{25}.** The x-axis represents the TPH contaminant.

4.5 Discussion

In this study we proposed a dependent Bayesian model for taking account of an environmental factor in population data problems. We applied it to clustering (Table 4.2) and to estimating population quantities such as the diversity (Figure 4.5), and IC_{50} and IC_{25} (Figure 4.6). Compared with existing methods, this allows incorporating the dependence with respect to TPH in the inferential procedure.

There are many ways in which this work can be extended. Here we discuss three examples. First, additional factors to TPH which include other types of environmental factors about soil composition, geographical factors, etc, could be utilised in the model. The

steps for this extension are sketched in Section 3.7. Second, one could consider other approaches to clustering, for instance it would be interesting to compare the inherent clustering of some Bayesian nonparametric procedures like the Hierarchical Dirichlet process by Teh et al. (2006). Last, the methodology used for estimating IC_{50} and IC_{25} is very sensitive to the group of OTUs on which it is based, and could be compared to other approaches.

There are several limitations of the approach, both inferential and computational. From the inference point of view, one can argue that the dependence that we introduce is fixed in a sense by the a prior dependence structure, and can not be learnt by the model. Computationally, the model is difficult to estimate on very large data sets, even if extremely sparse as is the case of the original ecotoxicological data set studied here. Indeed, the sparsity of the data is not favourable utilised in the posterior computation. However, in terms of diversity, most of the information is driven by the largest OTUs, hence working on a subsample of the data is satisfactory in this respect.

This study brings some new elements to the understanding of the effects of contaminants on ecosystems. The grouping is a valuable information for knowing which OTUs are to be studied for deriving environmental quality guidelines, while the IC_{50} and IC_{25} estimates allow fixing thresholds in these guidelines.

The range of suitable applications of the model presented here is not limited to ecology. Census data represents an appealing extension, where species are any categorical variable indexed by integers (nationality, age, for instance, or pairs of such variables), and the covariate is any continuous variable, such as time. French renovated census, which is available annually since 2004, could be the subject of future work.

Bibliography

Arbel, J., Mengersen, K., and Rousseau, J. (2013a). Bayesian nonparametric dependent models for the study of diversity in species data. *Manuscript under preparation*. 106

Arbel, J., Mengersen, K., Rousseau, J., Raymond, B., and King, C. (2013b). Ecotoxicological data study of diversity using a dependent Bayesian nonparametric model. *Manuscript under preparation*. 104

Boender, C. and Kan, A. R. (1987). A multinomial Bayesian approach to the estimation of population and vocabulary size. *Biometrika*, 74(4):849–856. 106

Bohlin, J., Skjerve, E., and Ussery, D. (2009). Analysis of genomic signatures in prokaryotes using multinomial regression and hierarchical clustering. *BMC Genomics*, 10(1):487. 106

Borges, E. P. and Roditi, I. (1998). A family of nonextensive entropies. *Physics Letters A*, 246(5):399–402. 106

Cerquetti, A. (2012). Bayesian nonparametric estimation of Simpson's evenness index under $\alpha-$Gibbs priors. *arXiv preprint arXiv:1203.1666*. 114

Colwell, R. K., Chao, A., Gotelli, N. J., Lin, S.-Y., Mao, C. X., Chazdon, R. L., and Longino, J. T. (2012). Models and estimators linking individual-based and sample-based rarefaction, extrapolation and comparison of assemblages. *Journal of Plant Ecology*, 5:321. 106

De'ath, G. (2012). The multinomial diversity model: linking shannon diversity to multiple predictors. *Ecology*, page in press. doi: 10.1890/11-2155.1. 106

Donnelly, P. and Grimmett, G. (1993). On the asymptotic distribution of large prime factors. *Journal of the London Mathematical Society*, 2(3):395–404. 106

Dunson, D. B. and Xing, C. (2009). Nonparametric bayes modeling of multivariate categorical data. *Journal of the American Statistical Association*, 104(487). 106

Dunstan, P. K., Foster, S. D., and Darnell, R. (2011). Model based grouping of species across environmental gradients. *Ecological Modelling*, 222(4):955–963. 105

Elith, J., Graham, C. H., Anderson, R. P., Dudk, M., Ferrier, S., Guisan, A., Hijmans, R. J., Huettmann, F., Leathwick, J. R., Lehmann, A., Li, J., Lohmann, L. G., Loiselle, B. A., Manion, G., Moritz, C., Nakamura, M., Nakazawa, Y., Overton, J. M. M., Townsend Peterson, A., Phillips, S. J., Richardson, K., Scachetti-Pereira, R., Schapire, R. E., Sobern, J., Williams, S., Wisz, M. S., and Zimmermann, N. E. (2006). Novel

methods improve prediction of species distributions from occurrence data. *Ecography*, 29(2):129151. 105

Ellis, N., Smith, S. J., and Pitcher, C. R. (2011). Gradient forests: calculating importance gradients on physical predictors. *Ecology*, 93(1):156–168. doi: 10.1890/11-0252.1. 105

Ferrier, S. and Guisan, A. (2006). Spatial modelling of biodiversity at the community level. *Journal of Applied Ecology*, 43(3):393–404. 105

Ferrier, S., Manion, G., Elith, J., and Richardson, K. (2007). Using generalized dissimilarity modelling to analyse and predict patterns of beta diversity in regional biodiversity assessment. *Diversity and Distributions*, 13:252–264. 105

Fordyce, J. A., Gompert, Z., Forister, M. L., and Nice, C. C. (2011). A hierarchical bayesian approach to ecological count data: a flexible tool for ecologists. *PLoS ONE*, 6(11):e26785. 106

Foster, S. D. and Dunstan, P. K. (2010). The analysis of biodiversity using rank abundance distributions. *Biometrics*, 66(1):186–195. 105, 106

Good, I. J. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika*, 40(3-4):237–264. 106

Havrda, J. and Charvát, F. (1967). Quantification method of classification processes. concept of structural a-entropy. *Kybernetika*, 3(1):30–35. 106

Hill, B. M. (1979). Posterior moments of the number of species in a finite population and the posterior probability of finding a new species. *Journal of the American Statistical Association*, 74(367):668–673. 106

Hill, M. O. (1973). Diversity and evenness: a unifying notation and its consequences. *Ecology*, 54(2):427–432. 106

Holmes, I., Harris, K., and Quince, C. (2012). Dirichlet Multinomial Mixtures: Generative Models for Microbial Metagenomics. *PloS one*, 7(2):e30126. 106, 111, 112

Kaniadakis, G., Lissia, M., and Scarfone, A. (2005). Two-parameter deformations of logarithm, exponential, and entropy: a consistent framework for generalized statistical mechanics. *Physical Review E*, 71(4):046128. 106

MacArthur, R. H. (1957). On the relative abundance of bird species. *Proceedings of the National Academy of Sciences of the United States of America*, 43(3):293. 106

MacEachern, S. (1999). Dependent nonparametric processes. *ASA Proceedings of the Section on Bayesian Statistical Science*, pages 50–55. 113

Patil, G. and Taillie, C. (1982). Diversity as a concept and its measurement. *Journal of the American statistical Association*, 77(379):548–561. 106

Schloss, P. and Handelsman, J. (2005). Introducing DOTUR, a computer program for defining operational taxonomic units and estimating species richness. *Applied and environmental microbiology*, 71(3):1501–1506. 107

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica*, 4:639–650. 112

Teh, Y., Jordan, M., Beal, M., and Blei, D. (2006). Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581. 119

Wang, Y., Naumann, U., Wright, S. T., and Warton, D. I. (2012). mvabund an r package for model-based analysis of multivariate abundance data. *Methods in Ecology and Evolution*, 3(3):471–474. 105

Chapter 5

Future directions

Ce chapitre présente des travaux en lien avec la statistique bayésienne non-paramétrique qui ont été commencés pendant la thèse et seront poursuivis par la suite. Le premier projet concerne l'estimation adaptative spatiale. On se place dans le cas où la régularité du paramètre varie dans l'espace de définition de ce dernier. On recherche des lois a priori qui sont adaptatives optimales dans toutes les régions de l'espace, c'est à dire dont la loi a posteriori converge avec une vitesse associée à la régularité de la région de l'espace considérée. Le second projet concerne l'estimation de densité dans un cadre multivarié dans lequel les variables ou composantes des observations ne sont pas directement comparables. Le cas de grandeurs physiques avec des unités de mesure différentes, tel que l'espace des phases constitué des variables de position et de moment, représentent un exemple typique. On montre par des simulations que l'estimation de la densité à partir de mélanges de processus de Dirichlet ont des propriétés d'invariance qui font de ces mélanges une solution adaptée à ce type de problèmes.

5.1 Spatially adaptive estimation

Chapter 1 and Chapter 2 have emphasized a growing literature about estimation procedures that deliver adaptive rates of contraction (see *eg* de Jonge and van Zanten, 2010, 2012, Shen and Ghosal, 2012, Scricciolo, 2012). An additional direction with rather limited results concerns the so-called *local* or *spatial* adaptation of rates of contraction. We explain what this means.

A well defined adaptation result states according to what criterion adaptation is obtained (minimax rate of contraction) and in which class of smoothness (leading examples of a smoothness classes include Hölder, Besov, Sobolev, etc). A smoothness property can either be local (*eg* Hölder, Besov) or global (*eg* Sobolev):

- the Hölder space $\mathcal{H}_\alpha = \mathcal{H}_\alpha(L, [0,1])$, $0 < \alpha \leq q$, which is the collection of all functions f that have bounded derivatives up to order $\alpha_0 = \lfloor \alpha \rfloor = \max\{z \in \mathbb{Z} : z < \alpha\}$ and such that the α_0-th derivative satisfies the Hölder condition $|f^{(\alpha_0)}(x) - f^{(\alpha_0)}(y)| \leq L|x-y|^{\alpha-\alpha_0}$, for $L > 0$ and $x, y \in [0,1]$.
- the Sobolev space $\Theta_\beta(L_0)$ for univariate problems defined by

$$\Theta_\beta(L_0) = \left\{ \boldsymbol{\theta} : \sum_{j=1}^{\infty} \theta_j^2 j^{2\beta} < L_0 \right\}, \ \beta > 1/2, \ L_0 > 0 \tag{5.1}$$

  with smoothness parameter β and radius L_0.

Local or *spatial* adaptation deals with local smoothness spaces and means that the posterior distribution adapts to the local rate that is monitored by the local smoothness. In particular, in regression or density problems, the roughest part of the curve to be estimated should not affect the rate where the curve is smoother.

An obstacle of such a result of local adaptation is often encountered in the bound associated to Kullback neighborhoods. Kullback divergence is in essence global, hence it prevents from using the general methodology of Ghosal et al. (2000).

Recent tools that are promising in the field of local adaptation include locally adaptive factor processes (Durante et al., 2012) and splines with randomly placed knots (Belitser and Serra, 2013). Our preliminary investigations are concerned with location-scale mixture priors in the spirit of de Jonge and van Zanten (2010), which are priors Π_n on curves $\boldsymbol{\theta} : (0,1) \to \mathbb{R}$ defined as the law of the random curve on $(0,1)$

$$\boldsymbol{\theta}(x) = \sum_{k=1}^{M} Z_k \frac{1}{\sqrt{M}\sigma} p\left(\frac{x - k/M}{\sigma} \right),$$

where M is a positive integer random variable, Z_k, $k = 1, \ldots, M$ are Gaussian random variables, $p : \mathbb{R} \to \mathbb{R}$ is a kernel with appropriate integrability properties. The prior Π_n leads to interesting results in the following nonparametric regression

$$Y_i = \theta_0(x_i) + \epsilon_i,$$

with observations $Y_1, \ldots, Y_n$, unknown regression function $\boldsymbol{\theta} : (0, 1) \to \mathbb{R}$ supposed to be $\alpha-$Hölder, with $\alpha \in (0, 1]$, known design points $x_1, \ldots, x_n \in [0, 1]$, $x_i = i/n$, and i.i.d. ϵ_i, centered Gaussian with fixed variance.

5.2 Density estimation sensitivity to data scaling

Authors

- Julyan Arbel (Université Paris-Dauphine, CREST, Paris)
- Bernardo Nipoti (University of Turin, Collegio Carlo Alberto, Moncalieri, Italie)

Status

Comment of the paper Bayesian Nonparametric Inference–Why and How by Müller and Mitra published in *Bayesian Analysis*, Volume 8, Issue 2, 2013, pages 326–328, Arbel and Nipoti (2013).

Müller and Mitra (2013) deal with the flexibility of Bayesian nonparametric models and show, through some examples, that their use can be advantageous in common inference problems. As for density estimation, the paper describes the Dirichlet process mixtures (DPM) model by means of an application to inference on T-cell diversity, where the observations are counts. The specific nature of the dataset ensures that the scale of the data is not an issue. Nonetheless this is a ubiquitous concern in density estimation problems with observations from continuous distributions. Clearly, it is desirable that the estimates are not significantly affected by a rescaling of the data. A closely related problem refers to the estimation of multidimensional densities in spaces where different axes represent quantities with different physical dimensions. There is not a natural way to define a metric on the product space and scaling constants need to be set in order to relate units along different axes. This scenario arises, for example, with astronomical observations consisting of position and velocity of stars (e.g., Ascasibar and Binney, 2005). Although we are not aware of existing BNP literature where this problem is directly investigated, it is worth mentioning that, as a matter of fact, BNP models have

been used for density estimation in non–commensurable spaces. For example, both Müller et al. (1996) and Hanson (2006) analyse the well-known ozone dataset and, by means of Dirichlet process mixtures and Mixtures of Pólya tree models respectively, deal with the problem of estimating multivariate densities in, e.g., radiation and ozone concentration product space. In the next section we illustrate, through a simulation study, that the flexibility of the DPM model provides a natural answer to the problem of estimating densities in non–commensurable spaces.

We investigate the performance of location-scale DPM models with multivariate normal kernels (introduced in Müller et al., 1996) for density estimation through the following synthetic example. We generate bivariate samples $\mathbf{D}^{(n)} = (\mathbf{X}^{(n)}, \mathbf{Y}^{(n)})$, of size $n \in \{50, 100, 150, 200\}$, from the mixture of two normals:

$$\frac{1}{3}\,\mathsf{N}\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}1 & 0.9\\0.9 & 1\end{bmatrix}\right) + \frac{2}{3}\,\mathsf{N}\left(\begin{bmatrix}5\\3\end{bmatrix}, \begin{bmatrix}0.7 & 0\\0 & 0.7\end{bmatrix}\right). \tag{5.2}$$

The true density f and a scatter plot of 100 observations are shown in Figure 5.1. Then we consider rescaled data $\mathbf{D}_c^{(n)} = (\mathbf{X}^{(n)}, c\mathbf{Y}^{(n)})$ with varying scale parameter c. We use a DPM model to estimate f, conditional on each sample $\mathbf{D}_c^{(n)}$, and we let $\hat{f}_c^{(n)}$ denote the estimated predictive distribution. Simulations are done by using the R package `DPpackage` (see Jara et al., 2011) (10,000 iterations with a 5,000 burn–in period); the prior specification we have set is standard and, importantly, does not take into account the scale of the data. As a first argument in support of the stability of the model with respect to rescaling, we show in Figure 5.1 the estimates obtained for $n = 100$ and two scales, $c = 0.1$ and $c = 10$.

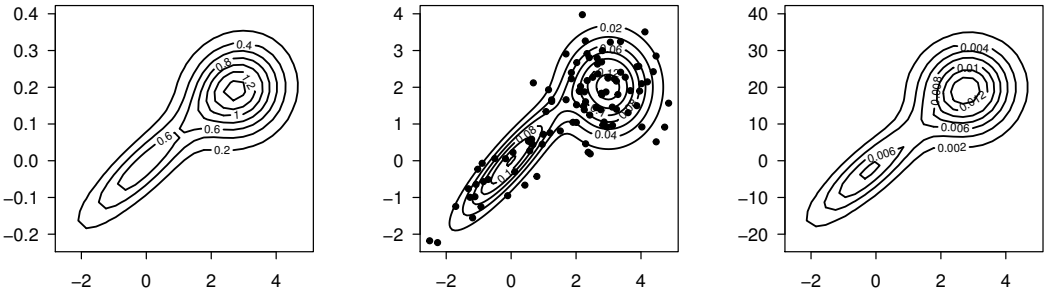

FIGURE 5.1: *(Middle)* Contour of the true density f and scatter plot of 100 observations. *(Left and right)* Contour of the estimates $\hat{f}_c^{(100)}$ for $c = 0.1$ and $c = 10$ respectively.

Additionally, for each n and $c = 10^{3k}$, where $k \in \{-2, \ldots, 2\}$, we summarize in Table 5.1 the fit of the estimate by computing the integrated squared error (ISE) for $\hat{f}_c^{(n)}$ suitably

rescaled, that is

$$\mathsf{ISE}(\mathbf{D}_c^{(n)}) := \int_{\mathbb{R}^2} \left(c\hat{f}_c^{(n)}\left(x, y/c\right) - f(x,y) \right)^2 \mathrm{d}x\mathrm{d}y.$$

It is apparent from Table 5.1 that the fit of $\hat{f}_c^{(n)}$ is not heavily affected by the choice of c. This feature is even more evident when the sample size is large. It is worth stressing that the estimates we got are pretty stable even when the model is tested on data severely rescaled (e.g. $c = 10^{-6}$ and $c = 10^6$).

$n \backslash c$	10^{-6}	10^{-3}	1	10^3	10^6
50	4.73	4.77	4.87	5.25	5.24
100	2.29	2.27	2.25	2.68	2.65
150	1.90	1.92	1.93	2.17	2.35
200	1.07	1.07	1.06	1.13	1.17

TABLE 5.1: $10^3 \times \mathsf{ISE}(\mathbf{D}_c^{(n)})$ for varying data size n *(in rows)* and scale c *(in columns)*.

This toy example suggests that the flexibility of DPM models makes them good candidates for dealing with a whole range of density estimation problems for which there is not a univocal scaling of the data.

The flexibility suggested here is a motivation for future investigation, such as proving an invariance property, or characterizing the dependence of estimation with respect to the scale factor c.

Bibliography

Arbel, J. and Nipoti, B. (2013). Comment on Müller and Mitra (2013). *Bayesian Analysis*, 8(2):326–328. 125

Ascasibar, Y. and Binney, J. (2005). Numerical estimation of densities. *Monthly Notices of the Royal Astronomical Society*, 356(3):872–882. 125

Belitser, E. and Serra, P. (2013). Adaptive Priors based on Splines with Random Knots. *arXiv preprint arXiv:1303.3365*. 124

de Jonge, R. and van Zanten, J. (2010). Adaptive nonparametric Bayesian inference using location-scale mixture priors. *Ann. Statist.*, 38(6):3300–3320. 124

de Jonge, R. and van Zanten, J. (2012). Adaptive estimation of multivariate functions using conditionally Gaussian tensor-product spline priors. *Electronic Journal of Statistics*, 6:1984–2001. 124

Durante, D., Scarpa, B., and Dunson, D. B. (2012). Locally adaptive Bayesian covariance regression. *arXiv preprint arXiv:1210.2022*. 124

Ghosal, S., Ghosh, J. K., and van der Vaart, A. W. (2000). Convergence rates of posterior distributions. *Ann. Statist.*, 28(2):500–531. 124

Hanson, T. (2006). Inference for mixtures of finite Polya tree models. *Journal of the American Statistical Association*, 101(476). 126

Jara, A., Hanson, T., Quintana, F., Müller, P., and Rosner, G. (2011). DPpackage: Bayesian non-and semi-parametric modelling in R. *Journal of statistical software*, 40(5):1. 126

Müller, P., Erkanli, A., and West, M. (1996). Bayesian curve fitting using multivariate normal mixtures. *Biometrika*, 83(1):67–79. 126

Müller, P. and Mitra, R. (2013). Bayesian Nonparametric Inference–Why and How. *Bayesian Analysis*, 8(2):323–356. 125, 127

Scricciolo, C. (2012). Adaptive Bayesian density estimation using Pitman-Yor or normalized inverse-Gaussian process kernel mixtures. *arXiv preprint arXiv:1210.8094*. 124

Shen, W. and Ghosal, S. (2012). Adaptive Bayesian procedures using random series prior. *Preprint*. 124

Chapter 6

Appendix

6.1 Results on the beta distribution

The probability density function of the beta distribution, for $0 \leq v \leq 1$, and shape parameters α and β is

$$f(v;\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} v^{\alpha-1}(1-v)^{\beta-1}.$$

Let $V \sim \mathsf{Beta}(\alpha,\beta)$. The kth moment of V is given by

$$\mathsf{E}(V^k) = \frac{\alpha^{(k)}}{(\alpha+\beta)^{(k)}}.$$

The following quantity is useful when computing the EPPF

$$\mathsf{E}\big(V^k(1-V)^j\big) = \frac{\alpha^{(k)}\beta^{(j)}}{(\alpha+\beta)^{(k+j)}}. \tag{6.1}$$

List of Figures

Notations

Common distributions & random probability measures

Beta	Beta distribution
Exp	Exponential distribution
Ga	Gamma distribution
IG	Inverse–gamma distribution
Dir	Dirichlet distribution
N	Gaussian distribution
Unif	Uniform distribution

Divergences & metrics

d	generic semimetric
h	Hellinger metric
KL	Kullback Leibler divergence
V	Csiszár divergence

Other abbreviations

BNP	Bayesian nonparametric
MCMC	Markov chain Monte Calro
RPM	random probability measure