On the frequentist properties of Bayesian nonparametric methods

Judith Rousseau,[1,2]

[1]CEREMADE, Université Paris Dauphine, Paris 75016, France; email : rousseau@ceremade.dauphine.fr
[2] Laboratoire de Statistique, CREST-ENSAE , Malakoff 92245, France; email : rousseau@ensae.fr

Xxxx. Xxx. Xxx. Xxx. YYYY. AA:1–24

This article's doi:
10.1146/((please add article doi))

Keywords

Asymptotics; Bayesian nonparametrics; Bernstein von Mises; posterior concentration; Empirical Bayes.

Abstract

In this paper, I will review the main results on the asymptotic properties of the posterior distribution in nonparametric or large dimensional models. In particular I will explain how posterior concentration rates can be derived and what we learn from such analysis in terms of impact of the prior distribution in large dimensional models. These results concern fully Bayes and empirical Bayes procedures. I will also describe some of the results that have been obtained recently in semi-parametric models, focusing mainly on the Bernstein - von Mises property. Although these results are theoretical in nature, they shed light on some subtle behaviours of the prior models and sharpen our understanding of the family of functionals that can be well estimated, for a given prior model.

Contents

1. INTRODUCTION

Some five years ago, it was said at a Bayesian nonparametric workshop that the field was now growing so fast that it was not possible to keep up with all the evolutions and new findings. And indeed, Bayesian nonparametrics has grown to be a major field in Bayesian statistics with applications in a large number of fields within biostatistics, physics, economy, social sciences, computational biology, comuter vision and language processing. There is now a collection of textbooks on general Bayesian nonparametric models , such as (Dey et al., 1998, Ghosh and Ramamoorthi, 2003, Hjort et al., 2010) or even on some specific aspects of Bayesian nonparametric, see for instance (Rasmussen and Williams, 2006) for Machine learning and Gaussian processes.

With the elaboration of more sophisticated models, the need to understand their theoretical properties becomes crucial. Theoretical studies on Bayesian nonparametric or large dimensional models can be split - typically - into two parts: asymptotic frequentist properties and probabilistic properties of the random process defining the prior and/or the posterior distribution. In this paper, I will mainly describe the advances that have been obtained on the asymptotic frequentist properties of Bayesian nonparametric procedures.

When opposing Bayesian to frequentist statistics, one is merely opposing the methods of validation, since, at least from a frequentist view - point there is not a frequentist methods but all sorts of different " algorithms", say, and the question is on how to evaluate them. Interestingly, Bayesian statistics form a global approach in that it provides a generic methodology to make inference, together with inherent evaluation tools. This coherency sometimes lead (Bayesian) statisticians to question the need for understanding their (asymptotic) frequentist properties. I will not enter this dispute, however I will try along the way to explain why it is helpful to understand the asymptotic frequentist properties of Bayesian procedures, in particular in complex or large dimensional models, when intuition and subjective inputs cannot be fully invoked.

Although strictly speaking nonparametric designates infinite dimensional parameters, I will also discuss high dimensional models since they share common features with nonparametric models.

Bayesian nonparametric modelling was probably intiated from de Finetti's representa-

tion of infinite exchangeable sequences, see (de Finetti, 1937), which states that any infinite exchangeable sequence $(Y_i, i \in \mathbb{N})$ has a distribution which can be represented as:

$$Y_i | P \overset{iid}{\sim} P, \quad i \in \mathbb{N}, \quad P \sim \Pi, \tag{1}$$

so that the de Finetti measure Π can be understood as a prior distribution on P. Nowadays, more complex structures are modelled and used in practice.

Consider a statistical model associated to a set of observations $Y^n \in \mathcal{Y}^{(n)} \sim P_\theta, \theta \in \Theta$ where n denotes a measure of information of the data Y^n. In the exchangeable model (1) for instance Θ designates the set of probabilities on $\mathcal{Y}^1$, or the set of probability densities on $\mathcal{Y}^1$ if we restrict our attention to dominated models. In regression or classification models of Y on X, Θ may denote the set of regression functions, or the set of conditional distributions or densities given X. Generally speaking Θ can have a very complex structure, be high or infinite dimensional. Hence, in such cases the influence of the prior is strong and does not entirely vanish asymptotically. It is then interesting to understand the types of implicit assumptions which are made by the choice of a specific prior and also within a family of priors which are the hyperparameters whose influence does not disappear as the number of observations increases. In some applications, hyperparameters are determined based on prior knowledge, as in (Yau et al., 2011), in others they are chosen based on the data as in (van de Wiel et al., 2013); in the latter case the approach is called empirical Bayes. In both cases it is important to assess the influence of these choices. From a theoretical view-point subjective priors and data dependent priors do not present the same difficulties, in Section **??** I describe the asymptotic behaviour of posterior distributions associated to priors that do not depend on the data while in Section 4 empirical Bayes posteriors are considered.

Before describing theoretical properties of Bayesian nonparametric procedures, I will recall in Section 2 the two main categories of Bayesian nonparametric prior models : namely those based on Dirichlet processes or its extensions and those based on Gaussian process priors. In Section 3 then the main results on posterior consistency and posterior concentration rates are presented, Section 4 treats the recent results on empirical Bayes procedures and Section 5 briefly describes advances in Semi-parametric models.

2. COMMON BAYESIAN NONPARAMETRIC MODELS

We do not intend to cover the whole spectrum of Bayesian nonparametric, but in this section we will review two important families of processes that are used in Bayesian nonparametric modelling.

2.1. Around the Dirichlet process

The most celebrated process used in prior modelling is the Dirichlet process prior DP, introduced by (Ferguson, 1974). The Dirichlet process can be characterized in many ways. It is parameterized by a mass $M > 0$ and and probability measure G_0 on a space $\mathcal{X}$. An explicit construction of its distribution is known as the stick - breaking representation, it is

due to (Sethuraman, 1994) and is given by

$$\begin{aligned} G &= \sum_{j=1}^{\infty} p_j \delta_{\theta_j}, \\ p_j &= V_j \prod_{l<j}(1-V_l), \text{ with } V_j \overset{iid}{\sim} Beta(1,M) \text{ and } \theta_j \overset{iid}{\sim} G_0, \end{aligned} \tag{2}$$

mutually independently, where δ_{θ_j} stands for the Dirac point mass at θ_j. We write $G \sim DP(M, G_0)$. The Dirichlet process has various other representations which makes it a very useful process, see for instance (Ghosh and Ramamoorthi, 2003, Lijoi and Prünster, 2010). Most often, the Dirichlet process is not used alone in the prior modelling. It is commonly used combined with some kernel f_θ in a mixture model :

$$Y_i|\theta_i \sim f_{\theta_i}, \quad \theta_i|P \overset{iid}{\sim} P, \quad P \sim DP(M, G_0). \tag{3}$$

The above type of model is a powerful tool to estimate the density of Y_i but it can also be considered for clustering given the discrete nature of the Dirichlet process. All sorts of variations around the mixture model (3) can be considered. For instance in (Kyung and Casella, 2010) the authors model the distribution of the random effects in a random effect model using a Dirichlet process. To go beyond exchangeable data, hierarchical Dirichlet processes, dependent Dirichlet processes, infinite hiden Markov models have been constructed, see (Hjort et al., 2010) for descriptions of these extensions.

Also, extensions of the Dirichlet process (2) have been constructed based either on the Sethuraman representation or one of its other representations : normalized completely random measure, Polya urn representation, see (Lijoi and Prünster, 2010), or as a special case of Polya trees, (Lavine, 1992).

2.2. Around Gaussian processes

Gaussian processes form another class of very popular processes used in prior modelling in Bayesian nonparametrics. Bayesian modelling via Gaussian processes has strong connections with machine learning approaches as described in (Rasmussen and Williams, 2006). They are used to model curves. Roughly speaking a zero mean Gaussian process can be viewed as a set of random variables on a probability space $(\Omega, \mathcal{B}, P)$, $(W_t, t \in T)$ for some set T, with finite dimensional marginals following multivariate Gaussian distributions. It is caracterized by a covariance kernel $K(s,t)$, $s,t \in T$. The behaviour of the Gaussian process is therefore driven by the choice of the Kernel. The most well known kernels are the exponential kernel $K_a(s,t) = e^{-a\|t-s\|^2}$, the Matérn Kernel $K_{\nu,a}(s,t) = 2^{1-\nu}(\sqrt{2\nu}\|s-t\|/a)^\nu K_\nu(\sqrt{2\nu}\|s-t\|/a)/\Gamma(\nu)$ where K_ν is the Bessel function, $a, \nu > 0$, and the Brownion motion kernel $K(s,t) = s \wedge t$. The first two refer to stationary Gaussian processes with T a normed space, while the Brownion motion is non stationary and sits on $T \subset \mathbb{R}^+$. These three classes of kernels are associated to very different behaviour of the process, the curves $(W_t, t \in T)$ drawn from these distributions have in particular different smoothness properties. The exponential kernel leads to infinitely differentiable curves, contrarywise to the Matérn or the Brownion motion. A key feature in understanding the behaviour of the Gaussian process associated to a given kernel, is its Reproducing Kernel Hilbert space (RKHS) $\mathbb{H}$. Roughly speaking the RKHS is a Hilbert space and is the closure in $L_2(\Omega, \mathcal{B}, P)$ of the functions $t \to E[W_t \sum_i \alpha_i W_{s_i}]$, see (van der Vaart and van Zanten, 2008a) for a review on the subject.

There are many other ways to construct probabilities on curves, in a similar spirit to Gaussian processes. Indeed Gaussian processes, under weak conditions, can be decomposed as $\sum_j Z_j \lambda_j e_j$ where $(e_j)_j$ form an orthonormal basis, $\lambda_j > 0$ and the $Z_j \overset{iid}{\sim} \mathcal{N}(0,1)$. Other types of projections on linear spaces can be considered using wavelets, splines, Legendre polynomials to name but a few. The prior is then typically formed by (1) choosing the dimension of the space from some distribution on $\mathbb{N}$ and (2) given the dimension of the space, drawing the coefficients of the projection on the space from some specific distribution.

We now describe the tools which have been developped to study the asymptotic behaviour of the posterior distribution on large or infinite dimensional spaces.

3. ASYMPTOTIC PROPERTIES OF THE POSTERIOR DISTRIBUTION

3.1. Notations and setup

Herafter we consider a Bayesian model $(\mathcal{Y}^{(n)}, P_\theta, \theta \in \Theta)$ where $(\Theta, \mathcal{A})$ is the parameter space which is possibly infinite dimensional and $\mathcal{A}$ is its $\sigma-$ field, with a prior probability Π on Θ. We assume that the model is dominated by some measure ν on $\mathcal{Y}^n$ and we write f_θ the density of P_θ with respect to ν and $\ell_n(\theta) = \log f_\theta(Y^n)$ the log-likelihood. Then the posterior distribution can be represented as, for all $B \in \mathcal{A}$,

$$\Pi(B|Y^n) = \frac{\int_B f_\theta(Y^n) d\Pi(\theta)}{\int_\Theta f_\theta(Y^n) d\Pi(\theta)}. \tag{4}$$

Hereafter θ_0 denotes the true value of the parameter, as we are now focusing on the frequentist properties of $\Pi(.|Y^n)$. For all $\theta \in \Theta$, E_θ and V_θ denote respectively expectation and variance with respect to P_θ.

3.2. Posterior consistency

Consider a Bayesian model as described in Section 3.1 with a prior probability Π on Θ, we say that the posterior distribution is consistent with respect to a loss function $d(.,.)$ on $\Theta_0 \subset \Theta$ if for all $\theta_0 \in \Theta_0$,

$$\forall \epsilon > 0; \Pi\left(d(\theta, \theta_0) < \epsilon | Y^n\right) \to 1 \quad P_{\theta_0} \quad \text{a.s. as } n \text{ goes to } +\infty \tag{5}$$

In other words posterior consistency means that the posterior distribution concentrates around the true parameter θ_0, in terms of the loss $d(.,.)$. Posterior concentration is a minimal requirement, in particular in the context of large dimensional models where it is not possible to construct fully subjective priors. Moreover, even from a subjective Bayes point of view, posterior consistency is important since it is the necessary and sufficient condition for the asymptotic merging of 2 posterior distributions associated to 2 different priors as the information in the data, n, goes to infinity, see (Diaconis and Freedman, 1986). Although not all priors lead to posterior consistency, posterior consistency has been verified in a large number of models and of prior distributions. This was initiated by the work of (Schwartz, 1965) in the case of density estimation and extended by (Barron, 1988) for generic models.

From (Schwartz, 1965) and (Barron, 1988), posterior consistency at θ_0 under the loss $d(.,.)$ is achieved if for all $\theta \in \Theta$ there exists $D(\theta_0; \theta)$ (typically the Kullback-Leibler divergence) such that

$$\text{limsup}_n n^{-1} \left(\ell_n(\theta_0) - \ell_n(\theta)\right) = D(\theta_0; \theta), \quad P_{\theta_0} \quad \text{a.s.}$$

and under the following conditions on the model and the prior: for all $\epsilon > 0$,

- (i) Kullback-Leibler condition:

$$\Pi(S_\epsilon) > 0, \quad S_\epsilon = \{\theta; D(\theta_0; \theta) < \epsilon\}$$

- (ii) Testing condition : there exist $\Theta_n \subset \Theta$, $r, \delta > 0$ and a sequence of test functions $\phi_n \in [0, 1]$ such that

$$\Pi(\Theta_n^c) \leq e^{-nr}, \quad E_{\theta_0}(\phi_n) \leq e^{-n\delta}, \quad \sup_{\theta \in \Theta_n, d(\theta_0, \theta) > \epsilon} E_\theta(1 - \phi_n) \leq e^{-n\delta}$$

Condition (i) ensures that the prior mass puts positive (enough) mass on neighbourhoods (here Kullback-Leibler neighbourhoods) of the true distribution while condition (ii) means that we are able to construct (exponential) tests separating the true parameter θ_0 to points that are far away from it in the $d(.,.)$ metric (or pseudo-metric). In other words ϕ_n is a statistical test for the problem $H_0 : \theta = \theta_0$ versus $H_1 : d(\theta_0, \theta) > \epsilon$ and $\theta \in \Theta_n$. Condition (ii) then says that the tests ϕ_n need to have exponentially decreasing first and second type errors. Roughly speaking, when such tests exist then $\ell_n(\theta) - \ell_n(\theta_0) < -n\delta$ for some $\delta > 0$ for all $\theta \in \Theta_n$, with large probability. Since Θ is not necessarily compact, the existence of tests with exponential decay is seldom verified over Θ, but it is enough to construct the tests ϕ_n on a sequence of subsets Θ_n increasing towards Θ and having high prior probability. They are typically established by splitting Θ_n into a number say N_n of subsets $\Theta_{i,n}$ (balls for instance) and to construct for each subset $\Theta_{i,n}$ a test function for the problem $H_0 : \theta = \theta_0$ versus $H_1 : \theta \in \Theta_{i,n}$ and to define ϕ_n as the maximum of such tests. We explain in more details in Section 3.3.1 how these tests are used to derive posterior concentration.

Consistency has been established in a variety of models, as in density estimation for i.i.d data, see (Schwartz, 1965, Barron et al., 1999, Lijoi et al., 2005) for generic results, regression function estimation (Choi and Schervish, 2007, Ghosal and Roy, 2006), or in models with non independent data (Ghosal and Tang, 2006, Rousseau et al., 2012, Vernet, 2014) and recently these results have been extended to data dependent priors (empirical Bayes) in (Petrone et al., 2014). The above papers present general conditions on the true parameter and on the prior distribution to ensure posterior consistency. Specific prior models have also been studied in the literature. However to understand better the impact of the prior distribution on the analysis in complex models it is enlightening to study posterior concentration rates.

3.3. Posterior concentration rates

Posterior concentration (or contraction) rates are defined by:

Definition 1. *The posterior distribution concentrates at rate ϵ_n at θ_0 if there exists $M > 0$ such that*

$$E_{\theta_0}(\Pi(d(\theta, \theta_0) < M\epsilon_n | Y^n)) = 1 + o(1) \tag{6}$$

Posterior concentration or contraction rates are therefore a more precised version of posterior consistency since they provide an upper bound on the rate at which the posterior distribution shrinks towards the true parameter θ_0. Typically ϵ_n depends on caracteristics of θ_0 and on properties of the prior distribution Π.

So why is it interesting to study posterior concentration rates ? From a frequentist point of view, (6) typically implies that Bayesian estimates such as the posterior mean or

the posterior median have a frequentist risk of order ϵ_n, see (Ghosal et al., 2000). It is also interesting for understanding the behaviour of credible balls with respect to $d(.,.)$, i.e. credible sets defined as

$$C_\alpha = \{\theta; d(\theta, \hat{\theta}) \leq z_\alpha\}, \quad \Pi\left(\theta \in C_\alpha | Y^n\right) \geq 1 - \alpha$$

so that z_α is the $1-\alpha$-th quantile of the posterior distribution of $d(\theta, \hat{\theta})$ and $\hat{\theta}$ is some given estimator, like the posterior mean of θ. Indeed as explained in (Hoffman et al., 2013), if $E_{\theta_0}\left(d(\theta_0, \hat{\theta})\right) = O(\epsilon_n)$ and if (6) is satisfied, then $E_{\theta_0}\left(|C_\alpha|\right) = O(\epsilon_n)$ with $|C_\alpha|$ denoting the size (radius) of C_α and

$$\int_\Theta P_\theta(\theta \in C_\alpha) d\Pi(\theta) \geq 1 - \alpha.$$

Hence the credible region is not necessarily a honest confidence region, but on average it is a confidence region with coverage $1 - \alpha$.

Finally, deriving the posterior concentration rates is enlightning about the way the prior distribution acts, which is particularly important in high dimensional models; we now explain how these posterior concentration rate can be derived. We illustrate in Section 3.3.2 in two families of examples why the study of posterior concentration rates shed some light on the impact of the prior.

3.3.1. Conditions and results. Similarly to posterior consistency, posterior concentration rates are obtained by verifying the following types of conditions, see (Ghosal et al., 2000, Ghosal and van der Vaart, 2007a):

- (i) Kullback-Leibler condition: There exists $c_1 > 0$

$$\Pi(\tilde{S}_{\epsilon_n}) \geq e^{-c_1 n \epsilon_n^2} \quad \tilde{S}_{\epsilon_n} = \{\theta; \mathrm{KL}_n(\theta_0, \theta) \leq n\epsilon_n^2; V_2(\theta_0, \theta) \leq n\epsilon_n^2\}$$

  where

$$\mathrm{KL}_n(\theta_0, \theta) = E_{\theta_0}\left(\ell_n(\theta_0) - \ell_n(\theta)\right); \quad V_2(\theta_0, \theta) = V_{\theta_0}\left(\ell_n(\theta_0) - \ell_n(\theta)\right)$$

- (ii) Testing condition : there exist $\Theta_n \subset \Theta$ and a sequence of test functions $\phi_n \in [0, 1]$ such that

$$\Pi(\Theta_n^c) = o(e^{-(c_1+2)n\epsilon_n^2})$$
$$E_{\theta_0}\left(\phi_n\right) = o(1), \quad \sup_{\theta \in \Theta_n, d(\theta_0, \theta) > M\epsilon_n} E_\theta(1 - \phi_n) = o(e^{-(c_1+2)n\epsilon_n^2})$$

Roughly speaking the argument follows from the following decomposition : write $B_{\epsilon_n}(\theta_0) = \{\theta; d(\theta, \theta_0) \leq M\epsilon_n\}$ then

$$\Pi\left(B_{\epsilon_n}^c(\theta_0) | Y^n\right) = \frac{\int_{B_{\epsilon_n}^c(\theta_0)} e^{\ell_n(\theta) - \ell_n(\theta_0)} d\Pi(\theta)}{\int_\Theta e^{\ell_n(\theta) - \ell_n(\theta_0)} d\Pi(\theta)} := \frac{N_n}{D_n},$$

since $N_n/D_n \leq 1$ and $0 \leq \phi_n \leq 1$, we can write

$$\begin{aligned}
E_{\theta_0}\left(\Pi\left(B^c_{\epsilon_n}(\theta_0)|\, Y^n\right)\right) &\leq E_{\theta_0}\left(\phi_n\right) + P_{\theta_0}\left(D_n < \frac{e^{-(c_1+2)n\epsilon_n^2}}{2}\right) + 2e^{(c_1+2)n\epsilon_n^2} E_{\theta_0}(N_n(1-\phi_n)) \\
&\leq E_{\theta_0}\left(\phi_n\right) + P_{\theta_0}\left(D_n < \frac{e^{-(c_1+2)n\epsilon_n^2}}{2}\right) + 2e^{(c_1+2)n\epsilon_n^2} E_{\theta_0}\left(\int_{\Theta_n^c(\theta_0)} e^{\ell_n(\theta)-\ell_n(\theta_0)} d\Pi(\theta)\right) \\
&\quad + 2e^{(c_1+2)n\epsilon_n^2} E_{\theta_0}\left(\int_{B^c_{\epsilon_n}(\theta_0)\cap\Theta_n} (1-\phi_n) e^{\ell_n(\theta)-\ell_n(\theta_0)} d\Pi(\theta)\right) \\
&= E_{\theta_0}\left(\phi_n\right) + P_{\theta_0}\left(D_n < \frac{e^{-(c_1+2)n\epsilon_n^2}}{2}\right) + 2e^{(c_1+2)n\epsilon_n^2} \Pi(\Theta_n^c) \\
&\quad + 2e^{(c_1+2)n\epsilon_n^2} \int_{B^c_{\epsilon_n(\theta_0)}\cap\Theta_n} E_\theta\left(1-\phi_n\right) d\Pi(\theta).
\end{aligned}$$

The Kullback-Leibler condition allows to control $P_{\theta_0}\left(D_n < \frac{e^{-(c_1+2)n\epsilon_n^2}}{2}\right)$ by first bounding from below

$$\begin{aligned}
D_n &\geq \int_{\tilde{S}_{\epsilon_n}} e^{\ell_n(\theta)-\ell_n(\theta_0)} \mathbb{1}_{\ell_n(\theta)-\ell_n(\theta_0)\geq -2n\epsilon_n^2} d\Pi(\theta) \\
&\geq e^{-2n\epsilon_n^2} \Pi\left(\tilde{S}_{\epsilon_n} \cap \{\ell_n(\theta)-\ell_n(\theta_0) \geq -2n\epsilon_n^2\}\right)
\end{aligned}$$

and then using Markov inequality twice,

$$\begin{aligned}
&P_{\theta_0}\left(\Pi\left(\tilde{S}_{\epsilon_n} \cap \{\ell_n(\theta)-\ell_n(\theta_0) \geq -2n\epsilon_n^2\}^c\right) > \frac{\Pi(\tilde{S}_{\epsilon_n})}{2}\right) \\
\leq\ & \frac{2\int_{\tilde{S}_{\epsilon_n}} P_{\theta_0}\left(\ell_n(\theta)-\ell_n(\theta_0) < -2n\epsilon_n^2\right) d\Pi(\theta)}{\Pi(\tilde{S}_{\epsilon_n})} \leq \frac{2}{n\epsilon_n^2}.
\end{aligned}$$

There exist in the literature variations around this decomposition and the above conditions but the ideas are all along these lines.

Following the frequentist literature, one typically caraterizes the concentration rates in terms of a few features of the true parameter. For instance, in the case of curve estimation, it is common practice to either assume some smoothness property of the curve, like Hölder, Sobolev or Besov regularity or some shape constraints such as monotonicity or convexity. The obtained rates tend to be uniform over some functional classes or some collections of functional classes.

There is a growing literature on the field and large classes or prior distributions and models have been studied using the above approach. In the context of density estimation for i.i.d random variables, the reknown Dirichlet process mixture models have been studied by (Ghosal and van der Vaart, 2007b, Kruijer et al., 2010, Scricciolo, 2014, Shen et al., 2013, Canale and de Blasi, 2013) among others in the case of Gaussian mixtures, by (Ghosal, 2001, Rousseau, 2010) for Beta mixtures. Log-linear, log-spline, log-Gaussian process priors have been also considered by (Ghosal et al., 2000, Rivoirard and Rousseau, 2012b, van der Vaart and van Zanten, 2008b, van der Vaart and van Zanten, 2009) to name but a few. In (van der Vaart and van Zanten, 2008b, van der Vaart and van Zanten, 2009)

posterior concentration rates have been derived for general models, when the prior on the unknown curve is constructed using a Gaussian process prior. Their results have been recently extended to a multivariate setup where both anisotropy and dimension reduction are incorporated in the prior model by (Bhattacharya et al., 2014a). Various other sampling models have been studied in the literature, following the above approach, such as inhomogeneous and Aalen point processes (Belitser et al., 2012, Donnet et al., 2014a), regression models (de Jonge and van Zanten, 2010), Gaussian times series (Rousseau et al., 2012) to name but a few.

In (van der Vaart and van Zanten, 2008b), the authors develop a very elegant strategy to verify conditions (i) and (ii) and thus determine posterior concentration rates in the context of Gaussian process priors, whatever the sampling model. Their approach makes use of the reproducing kernel Hilbert space $\mathbb{H}$ (RKHS) associated to a zero-mean Gaussian process W, viewed as a Borel map in a Banach space $(\mathbb{B}, \|.\|)$. More precisely, when the losses $\mathrm{KL}(\theta_0, \theta)$, $V_{\theta_0}(\theta_0, \theta)$ and $d(\theta_0, \theta)$ can be related (locally bounded typically) to the norm $\|\theta - \theta_0\|$ of $\mathbb{B}$, then ϵ_n defined in (i) and (ii) can be bounded by the solution to

$$n\epsilon^2 = \phi_{\theta_0}(\epsilon) := \inf_{h \in \mathbb{H}; \|h - \theta_0\| \leq \epsilon} \|h\|_{\mathbb{H}}^2 - \log P\left(\|W\| \leq \epsilon\right), \tag{7}$$

where $\|h\|_{\mathbb{H}}$ is the RKHS norm. They apply their results to the context of density, non linear regression, classification and white noise model. Other families of prior models have been studied in a generic way, i.e. somehow irrespective of the sampling model. For instance (Arbel et al., 2013) propose general conditions for prior distributions on some parameter $\theta \in \ell_2 = \{\theta = (\theta_i)_{i \in \mathbb{N}}, \sum_i \theta_i^2 < +\infty\}$ defined by

$$\Pi(d\theta) = \sum_{k=1}^{\infty} P(k) \pi(d\theta | k) \mathbb{1}_{\theta \in \mathbb{R}^k}, \tag{8}$$

where the conditional distribution of θ given k, $\pi(.|k)$ has the form

$$\theta_j \overset{iid}{\sim} g(./\tau_j)\tau_j^{-1}, \quad \text{if } j \leq k, \quad \theta_j = 0, \quad \text{if } j > k$$

In other words, under the prior distribution, θ is truncated according to a distribution P on $\mathbb{N}$, and given a truncation level k, the k non-null components of θ are independent.

3.3.2. What do the two conditions (i) and (ii) tell us about the impact of the prior distribution?. In the case of Gaussian process prior models for instance, (7) shows that posterior concentration rates are caracterized by the smoothness of the true curve θ_0 and the smoothness of the Gaussian process itself, i.e. by its RKHS. Indeed, small ball probabilities $\log P\left(\|W\| \leq \epsilon\right)$ depend on the RKHS and the smoother the RKHS the larger $-\log P\left(\|W\| \leq \epsilon\right)$, while $\inf_{h \in \mathbb{H}; \|h - \theta_0\| \leq \epsilon} \|h\|_{\mathbb{H}}^2$ indicates how well θ_0 can be approximated by elements of the RKHS $\mathbb{H}$. Hence, if θ_0 is not smooth enough compared to the elements in the RKHS $\mathbb{H}$ the latter term will be large and the posterior distribution will tend to have a large bias while $-\log P\left(\|W\| \leq \epsilon\right)$ can be viewed as a measure of variance or spread.

Although (7) gives only an upper bound on the posterior concentration rate, some lower bounds have been derived in the literature showing that it is often a sharp upper bound, see (Castillo, 2008). These results have shown that Gaussian processes are not as flexible as one might have hoped and that the behaviour of the posterior distribution is highly dependent on the covariance kernel $K(.,.)$ which in turns determines the RKHS $\mathbb{H}$, since its influence does not disappear asymptotically to first order.

This is not only true for Gaussian processes. Generally speaking, the two main conditions (i) and (ii) above shed light on key features in the behaviour of the posterior distributions. First, the prior model needs to be flexible enough to approximate well the true distribution (in terms of Kullback-Leibler divergence). For instance, consider the problem of density estimation for i.i.d. data, and take a prior model based on location mixtures of Gaussian distributions. The posterior concentration rates associated to this type of prior models have been studied by (Ghosal and van der Vaart, 2007b, Kruijer et al., 2010, Scricciolo, 2014, Shen et al., 2013). Let $x \in \mathbb{R}^d$,

$$f_{P,\sigma}(x) = \int_{\mathbb{R}^d} \varphi_\sigma(x-\mu) dP(\mu), \tag{9}$$

where φ_σ denotes the density of a Gaussian random variable in $\mathbb{R}^d$ with mean 0 and variance $\sigma^2 I_d$. The prior is constructed by considering a prior on (P, σ) where P varies in the set of probability distributions on $\mathbb{R}^d$. A popular choice for the prior on P is the Dirichlet process $DP(M, G)$ with mass M and base measure G, as defined in Section 2.1. Smooth densities on $\mathbb{R}$ can be well approximated by mixtures in the form (9). To understand what it means, we construct finite mixtures of Gaussian densities which approximate f, with as small a number of components as possible. Let f be a density which has Hölder (type) smoothness β, it is possible to construct a probability density f_β close to f such that

$$\mathrm{KL}_1(f, f_{F_\beta,\sigma}) = O(\sigma^\beta |\log \sigma|),$$

for all $\beta > 0$, where F_β is the distribution associated to f_β, see (Kruijer et al., 2010, Shen et al., 2013). Then we approximate the continuous mixture by a finite mixture, and it can be proved that $f_{F_\beta,\sigma}$ can be approximated to the order σ^κ for any $\kappa > 0$ by mixtures $f_{P_N,\sigma}$ where P_N has at most $N = O(|\log \sigma| \sigma^{-1})$ supporting points. Controlling N is a crucial step in proving the Kullback- Leibler condition since it provides an upper bound on the number of constraints on the parameter space that are needed to approximate a density f with smoothness β by densities in the form (9). It thus leads to a lower bound on the prior mass of Kullback-Leibler neighbourhoods of f. Choosing σ in the form $\sigma = n^{-1/(2\beta+1)} (\log n)^q$, $q \in \mathbb{R}$, leads to condition (i) with $\epsilon_n^2 \asymp n^{-2\beta/(2\beta+1)} (\log n)^{2q\beta+1}$.

Under the L_1 or the Hellinger loss functions for $d(.,.)$, the tests in condition (ii) are constructed from the tests of (Schwartz, 1965) or (Birgé, 1983) and are controlled bounding from above the entropy (i.e. the logarithm of the number of small balls needed to cover the set) of subsets of finite location mixtures of Gaussian distributions with at most $n^{1/(2\beta+1)} (\log n)^{q+1}$ components. Finally the posterior concentration rates for densities f with smoothness β (in a local Hölder sense, as described in (Kruijer et al., 2010) or in (Shen et al., 2013)) and under some exponential type condition on the tails of f, is bounded by

$$\epsilon_n \lesssim n^{-\beta/(2\beta+1)} (\log n)^\tau$$

for some $\tau \geq 0$, which is the minimax rate of convergence in this functional class, up to a $\log n$ term.

Although deriving conditions (i) and (ii) is quite informative on the way the prior acts, in the case of these nonparametric mixture models the picture is far from being complete. In the case of smooth density estimation, one expects the posterior distribution on the scale σ to concentrate on small values. This would mean that a common variation of prior model

(9), namely the location - scale mixture written as

$$f_{P,\sigma}(x) = \int_{\mathbb{R}^d} \varphi_\sigma(x-\mu) dP(\mu,\sigma), \tag{10}$$

might not be the best suited prior model for estimating a smooth density. Indeed following the above computations we obtain a much too small lower bound on prior mass on Kullback-Leibler neighbourhoods of f, and the obtained posterior concentration rate is suboptimal, see (Canale and de Blasi, 2013). Whether this is an artefact of the proof or a real suboptimal result remains an open question. To be able to answer such a question, one needs to characterize fully neighbourhoods of the true density to obtain not only a lower bound on their mass but also an upper bound. Given the complexity of the geometry of mixture models, the latter is a much more formidable task than the former. Model (10) is however more commonly used than the location mixture (9), and it is often considered as better behaved. This decrepancy between theory and practice has not yet been resolved.

A second crucial aspect of conditions (i) and (ii) is the existence of tests with second type error bounded by $e^{-Cn\epsilon_n^2}$. This condition restricts the choice of loss functions. In particular, (Hoffman et al., 2013) show that if there exist parameters θ which are close for some intrinsic loss (for which the tests of condition (ii) can be constructed, such as the L_2 loss in the white noise or the Hellinger distance in the density models) to θ_0 but not in terms $d(.,.)$, then the testing method above will lead to suboptimal bounds.

Interestingly, the prior based on model (9) does not depend on the true smoothness β of the density f_0, but the posterior adapts to the unknown smoothness β of f_0. This is one of the strengths of the Bayesian methodology, by naturally incorporating hierarchical structures in the prior it often enables to construct posterior distributions having good frequentist properties over not only a functional class, but a collection of functional classes.

3.4. Bayesian nonparametrics : a useful tool to derive adaptive optimal methods

Bayesian methods have become popular in particular because they can easily incorporate hierarchical modelling. In the case of nonparametric models, this is also the case and for most families of priors studied so far, it has been possible to construct hierarchical versions of them so as to obtain good frequentist properties over collections of functional classes.

If the posterior concentration rate (6) is uniformly bounded when the true parameter θ_0 is allowed to vary in a class $\Theta_\beta \subset \Theta$ by the frequentist minimax estimation rate over the same class under the same loss function, for instance $n^{-\beta/(2\beta+1)}$ for a β- Hölder ball in the setup of density estimation under the L_1 loss, then we say that the posterior concentrates at the minimax rate over Θ_β. If for a collection of classes, for instance $\Theta_\beta, \beta \in [\beta_1, \beta_2]$, the posterior concentrates at the minimax rate within each class, then we say that it concentrates at the minimax adaptive rate. Hierarchical modelling of prior distributions naturally leads to minimax adaptive posterior concentration rates.

For instance, in the context of Gaussian process priors, (van der Vaart and van Zanten, 2009) study conditional Gaussian process priors on curves g defined as:

$$\begin{aligned} A &\sim \Pi_A, \\ g(t) = W(At), &\qquad (W(t), t \in \mathbb{R}^+)) \sim GP(0, K), \end{aligned}$$

where $GP(0,K)$ denotes a Gaussian process prior with mean 0 and covariance kernel $K(s,t) = e^{-(s-t)^2}$ and Π_A is a probability on $\mathbb{R}^+$. The authors then show that for various

types of sampling models parametrized by the curve g, the posterior distribution concentrates around the true curve at a rate which is the minimax optimal estimation rate, up to a $\log n$ term, over a collection of Hölder classes, under a suitable prior Π_A . The prior does not depend on the supposed smoothness for the true curve and the posterior therefore leads to minimax adaptive estimators.

This construction has been extended in particular by (Bhattacharya et al., 2014a) to anisotropic multivariate curves.

There is now a large range of results on posterior concentration rates of hierarchical nonparametric prior models where adaptive minimax (up to a $\log n$ term usually) posterior concentration rates have been achieved. For instance the hierarchical prior construction (8) has been proved to lead to adaptive minimax concentration rates over collections of Sobolev or Besov balls for a variety of models, in (Arbel et al., 2013) and for some linear inverse problems in (Ray, 2013, Knapik and Salomond, 2015). The nonparametric location mixture of Gaussian random variables with an inverse Gamma on the scale parameter σ also leads to adaptive minimax concentration rates over collections of locally Hölder classes, as described above.

In the last few years, Bayesian nonparametric adaptive methods have been studied in the literature where adaptation is achieved not only with respect to some smoothness caracteristic but also with respect to sparsity in high dimensional models. These include the sequence model where one observes n independent observations

$$Y_i = \theta_i + \epsilon_i, \quad \epsilon_i \overset{iid}{\sim} \mathcal{N}(0,1), \quad i \leq n$$

which we will consider as an illustrative example of the types of phenomena that occur in high dimensional frameworks, but recall that it is simpler than other models like high dimensional regression or high dimensional graphical models. The most natural way to design a sparsity prior in this context is to first select the set S of non zero coefficients, and then put a prior on $\theta_S = (\theta_i, i \in S)$. In (Castillo and van der Vaart, 2012) posterior concentration rates around θ in terms of the L_q losses, $\|\theta - \theta_0\|_q$ with $1 < q \leq 2$, are derived under some conditions on such priors. They show that considering a family of priors on S defined by first choosing the size $|S| = p$ according to a distribution with exponential tails and then randomly selecting S given its size p leads to minimax adaptive posterior concentration rate $r_n^2 = p \log(n/p)$ under the loss $\|\theta - \theta_0\|_2^2$ uniformly over the set $\ell_0(p) = \{\theta, \|\theta\|_0 \leq p\}$ with $\|x\|_0$ denoting the number of nonzero coefficients in x.

Although the approach described in Section 3.3 for deriving posterior concentration rates is used, (i) and (ii) are not the only steps in their proof. This is due to the complexity of the parameter space. Before using the usual testing and Kullback-Leibler arguments, the authors first prove that the posterior distribution concentrates on sets that have at most Mp nonzero coefficients, for some large but fixed constant M. Then on this reduced parameter space they prove posterior concentration rates following steps (i) and (ii). This is common, if not inevitable, in high dimensional models with sparse parameters, when one needs to learn also the sparsity of the parameter. Interestingly, if the prior on $|S|$ or on $\theta_S|S$ have too light tails, then the authors prove that the posterior concentrates at a subotimal rate for large values of $\|\theta_0\|_2$. In many applications this is often not a crucial issue, since large signals are easily detected and the statistical analysis is typically used to detect small signals.

The above family of sparse priors is appealing in a high dimensional but sparse context but it is difficult to implement and so far has only be implemented for moderately large

dimensional models. Alternative priors have been proposed in the literature with posterior distributions easier to sample from but their asymptotic properties have not been studied. Recently (Bhattacharya et al., 2014b) have proposed a continuous type of shrinkage, closer in spirit to the Lasso, which also achieves optimal minimax adative posterior concentration rate, under the constraint that the true signal is not too large: $\|\theta_0\|_2^2 \leq p(\log n)^4$ where $p = \|\theta_0\|_0$, the number of nonzero components of θ_0.

(Castillo et al., 2015) have extended the results of (Castillo and van der Vaart, 2012) to the case of high dimensional linear regression, with a prior distribution on the sparsity inducing very sparse models.

Other families of high dimensional models have been considered, in particular sparse matrix and graphical models have been studied by (Banerjee and Ghosal, 2015, Bhattacharya and Dunson, 2011, Pati et al., 2014).

3.5. On frequentist coverage of credible regions in large dimensional models

As mentioned above posterior concentration rates are usefull to assess the size of posterior credible bands and their frequentist coverages verify

$$\int_\Theta P_\theta(\theta \in C_\alpha) d\Pi(\theta) = 1 - \alpha \tag{11}$$

if $1-\alpha$ is the Bayesian coverage of the credible band C_α. This does not imply however that C_α is a honest confidence region in a frequentist sense, i.e.

$$\inf_{\theta \in \Theta} P_\theta \left(\theta \in C_\alpha\right) \geq 1 - \alpha \tag{12}$$

In parametric regular models, thanks to the Bernstein - von Mise theorem, (12) is valid on compact subsets of Θ and for standard credible regions like highest posterior density regions, or ellipses around the posterior mean or mode. In nonparametric models, it is expected not to be satisfied and the first results on frequentist coverage of credible regions in infinite dimensional models were negative. (Cox, 1993) and (Freedman, 1999) exhibited negative results in the context of Gaussian models with Gaussian priors, where almost surely under the prior the frequentist coverage of an ℓ_2 credible ball could be arbitrarily close to 0.

Despite these results, the picture is not all negative. As said previously, an attractive feature of Bayesian (hierarchical) approaches is that they - when properly tuned - are adaptive procedures, and up to $\log n$ terms are often minimax adaptive over collections of functional classes, say $\Theta_\beta, \beta \in A$ (in terms of their posterior concentration rate). Consider, for the sake of simplicity the white noise model and Θ_β a Hölder ball with regularity β. If C_α is a ℓ_2 credible band for the parameter θ constructed under a minimax adaptive posterior distribution then it satisfies

$$\sup_{\beta \in A} \epsilon_n(\beta)^{-1} \sup_{\theta \in \Theta_\beta} E_\theta[|C_\alpha|] = O(1),$$

where $\epsilon_n(\beta)$ is the minimax estimation rate over the class Θ_β. In other words its size is adaptive minimax. It is known, see for instance (Cai and Low, 2006) that there does not exist honest confidence regions (i.e. satisfying (12)) which have adaptive size, unless $A \subset [\beta_1, 2\beta_1]$ for some β_1. Hence C_α cannot be an honest confidence region.

Can we find a subset Θ_0 of Θ over which C_α could be considered an honest confidence region ?

Recently, in (Szabó et al., 2013), the authors have answered this question in the special case of the white noise model and under the empirical Bayes posterior described in Section 4.0.1 and based on (Szabò et al., 2013). They find a set of well behaved parameters $\Theta_0 \subset \ell_2$, called the polished tail parameter set, over which (12) is verified. Their results rely heavily on the precise structure of the prior and sampling model, but they give some insight on what can be expected in other types of models.

In (Castillo and Nickl, 2013), conditions for deriving a weak nonparametric Bernstein von Mises theorem are derived in the white noise and density models, which leads to the construction of credible bands with correct asymptotic frequentist coverage. The types of priors considered in (Castillo and Nickl, 2013) are based on expansions of the curve on wavelet bases. The drawback of this approach is that the credible bands are constructed in terms of weighted L_2 norms which are difficult to interpret in practice. An advantage however is that from this result it is possible to derive Bernstein von Mises theorems for smooth functionals of either the signal in the white noise model or the density in the density model. Obtaining refined results such as Berntein von Mises theorems for finite dimensional functionals of the parameter is typically easier than for the whole parameter, it is however not a simple task and positive general results have been obtained only recently and in still a rather restricted framework. This is presented briefly in Section 5.

SUMMARY POINTS

1. Many common Bayesian nonparametric prior models lead to posterior distribution with minimax concentration rates
2. Using hierarchical priors, it is possible and relatively easy to obtain adaptive procedures. Adaptation may be with respect to some smoothness or some sparsity aspects of the parameter.
3. Posterior contraction rates are related to the size of credible balls or regions but not so much on their frequentist coverage. Understanding the frequentist coverage of a credible ball (or band) is more involved and only a few results have been obtained until now. It is becoming however an active area of research.

Empirical Bayes is an alternative to hierarchical Bayes; we describe in the following section some recent advances that have been obtained on the properties of empirical Bayes methods.

4. Empirical Bayes procedures

Traditionnally, empirical Bayes designates frequentist methods, in the context of multiple experiments where each experiment is associated to a specific parameter and where these parameters have a common distribution Π which is estimated using a frequentist estimator such as the maximum likelihood estimator. This was initiated by Robbins, see for instance (Robbins, 1964). However the term empirical Bayes is also used for any Bayesian approaches where the prior is data - dependent. In this section we are going to focus on the latter, which is widely used in practice, since more often than not, at some stage of the construction of the prior, some information coming from the data is used, in a more or less formalized way. It is typically believed that it should be better than an arbitrary choice of the hyperparameters of the prior.

The setup is the following. Consider a family of prior distributions on a parameter $\theta \in \Theta$ indexed by a hyperparameter γ, $(\Pi(\cdot|\gamma), \gamma \in \Gamma)$. A hierarchical approach would consist in constructing a prior on $\gamma \in \Gamma$, while the empirical Bayes approach selects $\gamma = \hat{\gamma}$ based on the data Y^n. There are many ways to choose $\hat{\gamma}$ and the two main categories are: (a) using moment conditions or similar considerations (b) using the maximum marginal likelihood estimator. There are other methods that do not quite enter into these categrories such as cross-validation or other frequentist methods used to select hyperparameters, but they have not been studied in the Bayesian setup so far.

The most common of the two is (a) although it is the less formalized. For instance, in the case of mixtures of Gaussian random variables,

$$f_{P,\sigma}(y) = \int_{\mathbb{R}} \phi_\sigma(y-\mu) dP(\mu), \quad P \sim DP(M\mathcal{N}(m_0, \tau_0^2)), \quad \sigma \sim \pi_\sigma$$

as discussed in Section 3.3, in (Green and Richardson, 2001) and (Richardson and Green, 1997) the authors advocate, based on invariance considerations, the choice of $\hat{m}_0$ as the midrange of the data and $\hat{\tau}_0^2$ as the range of the data. Another possibility would be to use the relation $\mathbb{E}(Y) = m_0$ and $\mathbb{V}(Y) = E_{\pi_\sigma}(\sigma^2) + \tau_0^2$ and choose $\hat{m}_0 = \bar{Y}_n = \frac{1}{n}\sum_{i=1}^n Y_i$ and $\hat{\tau}_0^2 = S_n^2 = \frac{1}{n}\sum_i (Y_i - \bar{Y}_n)^2$.

The maximum marginal likelihood estimator $\hat{\gamma}_{ML}$ of γ is defined, when it exists, as

$$\hat{\gamma}_{ML} = \text{argmax}_\gamma m_n(\gamma), \quad m_n(\gamma) = \int_\Theta f_\theta(Y^n) d\Pi(\theta|\gamma).$$

This approach has been used for instance by (George and Foster, 2000, Cui and George, 2008, Scott and Berger, 2010) in the context of variable selection in regression models, by (Belitser and Levit, 2002, Clyde and George, 2000, Szabò et al., 2013, Knapik et al., 2012) for sequence or white noise models, by (Liu, 1996) for the mass parameter M in a Dirichlet Process mixture prior.

What are the asymptotic properties of such empirical Bayes properties? Can we derive general methods to study these properties, as is done in Section 3.3 for fully Bayes procedures?

The approach presented in Section 3.3 to prove posterior concentration rates for fully Bayesian posteriors uses repeatedly Fubini's argument, which cannot be applied in a context of data dependent prior. Moreover, in infinite dimensional models the prior distributions $\Pi(\cdot|\gamma)$, $\gamma \in \Gamma$, are often singular with respect to one - another, see for instance (Ghosh and Ramamoorthi, 2003) in the case of Dirichlet processes, or (van der Vaart and van Zanten, 2008b), in the case of Gaussian process priors. However, recently in (Donnet et al., 2014b) the authors derive a general theory to study posterior concentration rates in the context of data dependent priors. Also, building on some this theory, together with the results of (Petrone et al., 2014) in the finite dimensional case and of (Knapik et al., 2012, Szabò et al., 2013) in the Gaussian white noise model with Gaussian process priors, we now have a better understanding of the behaviour the marginal maximum likelihood estimator in infinite dimensional models.

4.0.1. Dealing with data dependence on the prior . Consider a family of prior distributions $(\Pi(\cdot|\gamma), \gamma \in \Gamma)$ and a data dependent value $\hat{\gamma}$ and denote the empirical Bayes posterior by $\Pi(\cdot|Y^n, \hat{\gamma})$. In this section we assume that there exists a compact set $\Gamma_0 \subset \Gamma \subset \mathbb{R}^d$ for some $d < +\infty$, such that with probability going to 1,

$$\hat{\gamma} \in \Gamma_0.$$

The aim is to find the smallest possible sequence ϵ_n, such that $\Pi(B^c_{\epsilon_n}(\theta_0)|Y^n, \hat{\gamma}) \to 0$ in probability under P_{θ_0}. To do so, we in fact prove that

$$\sup_{\gamma \in \Gamma_0} \Pi(B^c_{\epsilon_n}(\theta_0)|Y^n, \gamma) \to 0,$$

so that the pre-selection of Γ_0 is important.

For instance, if $\hat{\gamma} = \bar{Y}_n$ or some other moment estimator, then under simple ergodicity conditions on P_{θ_0}, Γ_0 can be choosen in the form $\Gamma_0 = [\mu_0 - \epsilon, \mu_0 + \epsilon]$ where μ_0 is the limite of $\bar{Y}_n$ under P_{θ_0}, either in probability or almost surely.

The second key step in dealing with data dependent prior is to transfer the data-dependence from the prior to the likelihood. To do so, we consider changes of measure $\psi_{\gamma,\gamma'} : \Theta \to \Theta$ such that if $\theta \sim \Pi(\cdot|\gamma)$ then $\psi_{\gamma,\gamma'}(\theta) \sim \Pi(\cdot|\gamma')$. For instance in the case of the Dirichlet process mixture of Gaussian densities (9), using the stick-breaking representation of the Dirichlet process, we can write $\theta = f_{P,\sigma}$ as

$$f_{P,\sigma}(x) = \sum_{j=1}^{\infty} p_j \phi_\sigma(x - \mu_j), \quad \mu_j \overset{iid}{\sim} \mathcal{N}(m_0, \tau_0^2), \quad p_j = V_j \prod_{i<j}(1 - V_i), \quad V_j \overset{iid}{\sim} \mathcal{B}(1, M).$$

and if $\gamma = (m_0, \tau_0^2)$, then for all $\tau_0' > 0$ and $m_0' \in \mathbb{R}$, defining $\gamma' = (m_0', \tau_0')$ and

$$\mu_j^{'} = \mu_j \frac{\tau_0^{'}}{\tau_0} - m_0 \frac{\tau_0^{'}}{\tau_0} + m_0', \quad \forall j \in \mathbb{N}, \qquad P' = \sum_{j=1}^{\infty} p_j \delta_{(\mu_j^{'})}$$

we have $f_{P',\sigma} \sim \Pi(\cdot|\gamma')$, see (Donnet et al., 2014b) for more examples.

The third step is a chaining argument which consists in partitioning of Γ_0 into N_n bins of size u_n small enough and choosing points in each bin $(\gamma_i)_{i \leq N_n}$ so that

$$\sup_{\gamma \in \Gamma_0} \Pi(B^c_{\epsilon_n}(\theta_0)|Y^n, \gamma) \leq \max_i \Pi(B^c_{\epsilon_n}(\theta_0)|Y^n, \gamma_i) + \max_i \sup_{|\gamma - \gamma_i| \leq u_n} \Pi(B^c_{\epsilon_n}(\theta_0)|Y^n, \gamma).$$

The first term can be handled straightforwardly using the approach described in Section 3.3, while the second needs to be slightly adapted by replacing the Kullback-Leibler neighbourhoods by the sets of θ such that

$$\inf_{|\gamma - \gamma_i| \leq u_n} (\ell_n(\psi_{\gamma_i,\gamma}(\theta)) - \ell_n(\theta_0)) > -2n\epsilon_n^2$$

with large enough probability in condition (i) and by replacing the control of the second type error of condition (ii) with respect to P_θ by a second type error with respect to the measure with density $q^\theta_{\gamma_i}(y^n) = \sup_{|\gamma - \gamma_i| < u_n} f_{\psi_{\gamma_i,\gamma}(\theta)}(y^n)$. These arguments lead to a set of conditions to derive posterior concentration rates for empirical Bayes procedure which resemble conditions (i) and (ii) of Section 3.3, see Theorem 1 of (Donnet et al., 2014b). This is applied to Dirichlet process mixtures of Gaussian distributions, log-spline and log-linear prior models for density estimation and to Dirichlet process mixtures of uniforms in the context of Aalen point processes.

The pre-selection of Γ_0, i.e. the asymptotic behaviour of $\hat{\gamma}$ is important. However, Γ_0 need not necessarily be small, specially if the posterior concentration rate associated to a prior $\Pi(.|\gamma)$ does not depend on γ. For instance in the case of the Dirichlet process mixture of Gaussian distributions with $\gamma = (m_0, \tau_0^2)$ with $\hat{m}_0 = \bar{Y}_n$ and $\hat{\tau}_0^2 = R$, with R the range of the data, the posterior concentration around a density f_0 which has β - Hölder smoothness

as described in Section 3.3 is still of the form $n^{-\beta/(2\beta+1)}$ up to a $\log n$ term, although $\Gamma_0 = [\mu_0 - \epsilon, \mu_0 + \epsilon] \times [a, (\log n)^\kappa]$ for some positive constants ϵ, a, κ. If on the contrary the posterior concentration rate depends on γ, then it is crucial to have Γ_0 shrink fast enough around the best possible value.

4.0.2. Maximum marginal likelihood estimators. Maximum marginal likelihood empirical Bayes procedures are typically used in such context with the hope that the data dependent $\hat{\gamma}_{ML}$ will be close enough to optimal values for γ. This is not always the case and subtil phenomena can occur, as has been shown in (Knapik et al., 2012, Szabò et al., 2013). In both papers the authors consider the white noise model, with inverse operator but for the sake of simplicity we pretend it is equal the identity,

$$Y_i = \theta_i + n^{-1/2}\epsilon_i, \quad \epsilon_i \overset{iid}{\sim} \mathcal{N}(0,1), \quad \theta = (\theta_i)_i$$

and a Gaussian process prior on θ :

$$\theta_i \overset{ind}{\sim} \mathcal{N}(0, \tau_i^2), \quad i \in \mathbb{N}.$$

To model smooth curves under the prior distribution, it is common practice to consider τ_i going to infinity with i. The question is how . In (Szabò et al., 2013) the authors consider $\tau_i^2 = \tau_0^2 i^{-2\alpha-1}$, with $\gamma = \tau_0$ and using an explicit expression on the marginal likelihood $m_n(\gamma)$ show that if the true parameter θ_0 has smoothness β , i.e. $\theta_{0,i}^2 \leq L i^{-2\beta-1}$, then the empirical Bayes posterior has suboptimal posterior concentration rate for all $\beta > \alpha + 1/2$ while it achieves the minimax adaptive posterior concentration rate over $\beta < \alpha + 1/2$. Interestingly in (Knapik et al., 2012) the authors consider $\gamma = \alpha$ and they show that the empirical Bayes posterior concentration rate achieves minimax concentration rates for all β in this case. So why is there such a discrepancy between the two types of maximum marginal likelihood estimators?

In (Rousseau and Szabò, 2015), we describe the asymptotic behaviour of the maximum marginal likelihood estimator $\hat{\gamma}$ for a model

$$Y^n \sim f_\theta, \quad \theta \sim \Pi(.|\gamma), \quad \theta \in \Theta$$

where $(\Theta, \|.\|)$ is a Banach space, under some conditions on the model and the prior. In particular it shown that with probability going to 1,

$$\hat{\gamma} \in \Gamma_0 = \{\gamma; \epsilon_n(\gamma) \leq m_n \epsilon_{n,0}\}$$

where m_n is any sequence going to infinity and $\epsilon_n(\gamma)$ is defined by

$$e^{-n\epsilon_n^2(\gamma)} = \Pi\left(\|\theta - \theta_0\| \leq \epsilon_n(\gamma)|\gamma\right), \quad \epsilon_{n,0} = \inf_{\gamma \in \Gamma} \epsilon_n(\gamma). \tag{13}$$

Hence the maximum marginal likelihood estimator is minimizing the rate $\epsilon_n(\gamma)$ and it can be checked that in the setup of (Knapik et al., 2012) the minimizer is optimal and leads to $n^{-\beta/(2\beta+1)}$ up to a $\log n$ term while in the case of (Szabò et al., 2013) the minimizer can be suboptimal when $\beta > \alpha + 1/2$. Interestingly the asymptotic behaviour of the maximum marginal likelihood estimator $\hat{\gamma}$ is driven by the behaviour of $\Pi\left(\|\theta - \theta_0\| \leq \epsilon_n(\gamma)|\gamma\right)$ and not so much by the sampling model. A similar result was obtained in the parametric framework by (Petrone et al., 2014).

SUMMARY POINTS

1. A methodology has been developped to derive posterior concentration rates for data dependent priors. The approach is similar to the theory developped by (Ghosal and van der Vaart, 2007a) in the case of regular priors.
2. This approach can be quite easily applied to moment - type estimators and good frequentist properties of the empirical Bayes posterior have been obtained n this case.
3. Maximum marginal likelihood estimators have subtil behaviours and they can be apprehended by minimizing the set of candidate rates $\epsilon_n(\gamma)$ defined in (13).

5. Semi-parametric models and the Bernstein von Mises Theorem

In this section, I will describe some of the latest developpments that have been obtained so far in semi-parametric Bayesian inference, i.e. when the parameter of interest ψ is a finite dimensional functional of an infinite dimensional parameter, i.e. $\psi = \psi(\theta)$. Semi - parametric models are often consider in a context where $\theta = (\psi, \eta)$ with $\psi \in \mathcal{S} \subset \mathbb{R}^d$ and $\eta \in \mathcal{E}$ an infinite dimensional model. For instance in the non linear regression model η is the regression function and ψ the variance of the noise, although the parameter of interest is typically η in such models. In partially linear models the regression function of a response variable Y on a covariate vector $X = (X_1, X_2)$ is written $f(X) = X_1\beta + g(X_2)$ where g is non linear and some redundancies may exist between X_1 and X_2. In survival analysis, the Cox regression model is also a very common semi-parametric model. They are not however the only cases of semi-parametric problems and one might be interested in some mode general functionals of an infinite dimensional parameter, such as the cumulative distribution function at a given point, the mean of a distribution, the L_2 norm of a square - integrable curve, etc.

There are many semi-parametric models for which it is possible to estimate ψ at the rate $\sqrt{n}$, see for instance (van der Vaart, 1998) for a general theory on regular semi-parametric models. What would be the Bayesian counter-part of this theory? How can we prove that the marginal posterior distribution of ψ concentrates at the rate $\sqrt{n}$? Can we obtain a more precise description of the marginal posterior distribution of ψ? These questions can be answered by studying if the posterior distribution on ψ verifies the Bernstein - von Mises theorem (BvM). It says that asymptotically the marginal posterior distribution of $\sqrt{n}(\psi - \hat{\psi})$ converges (weakly or strongly) to a $\mathcal{N}(0, V_0)$, under P_{θ_0}, where $\sqrt{n}(\hat{\psi} - \psi(\theta_0))$ converges in distribution to $\mathcal{N}(0, V_0)$ under P_{θ_0} and $\hat{\psi}$ is some estimator of ψ.

Such properties have many interesting implications. In particular they allow to construct credible regions for ψ which have correct asymptotic frequentist coverage.

In (Castillo, 2010) and (Bickel and Kleijn, 2012) sufficient conditions are proposed to derive BvM in separated models in the form $\theta = (\psi, \eta)$. In (Rivoirard and Rousseau, 2012a) and in (Castillo and Rousseau, 2013) sufficient conditions to BvM are provided for linear functional of the density and for smooth functionals of the parameter in general models respectively. To explain the main features of these results I will present the arguments as in (Castillo and Rousseau, 2013).

The conditions are based on the following three ingredients:

- (1) Concentration of the posterior : There exists some shrinking neighbourhood A_n of θ_0 such that
$$\pi(A_n|Y^n) = 1 + o_{P_{\theta_0}}(1).$$
- (2) Local asymptotic normality of the likelihood (LAN): locally around θ_0 the log-likelihood can be approximated by a quadratic form. Assuming that a neighbourhood of θ_0 can be embedded into a Hilbert space $\mathbb{H}$, this local approximation takes the form
$$\ell_n(\theta) - \ell_n(\theta_0) = -\frac{n\|\theta - \theta_0\|_L^2}{2} + \sqrt{n}W_n(\theta - \theta_0) + R_n(\theta - \theta_0),$$
where $\|.\|_L$ is the norm of the Hilbert space and $W_n(.)$ is a linear operator on $\mathbb{H}$, such that $W_n(h) \sim \mathcal{N}(0, \|h\|_L^2)$ when $h \in \mathbb{H}$.
- (3) Smoothness of the functional : On A_n, the functional can be linearly approximated: There exists $\psi_1 \in \mathbb{H}$ such that
$$\psi(\theta) - \psi(\theta_0) =< \psi_1, \theta - \theta_0 >_L + o(1)$$

Then under some mild additional conditions BvM is valid if for $t \in \mathbb{R}$ with $|t|$ small enough,
$$\frac{\int_{A_n} f_{\theta - t\psi_1/\sqrt{n}}(Y^n)d\Pi(\theta)}{\int_{A_n} f_\theta(Y^n)d\Pi(\theta)} \to 1, \quad P_{\theta_0}. \tag{14}$$

Condition (14) is the key condition and roughly speaking means that it is possible to construct a change of variable
$$\theta \to \theta - t\psi_1/\sqrt{n}$$
or close enough to it, leaving the prior and A_n almost unchanged. In the cited papers, some examples are studied where BvM is valid for families of smooth functionals, however examples are also provided where it is shown that BvM does not hold. To illustrate this and explain the meaning of (14), let $\theta \in \ell_2$ and a prior on θ constructed as in (8): with $k \sim \pi_k$ and conditionnally on k, $\theta_i \overset{iid}{\sim} g$ for $i \leq k$ and $\theta_i = 0$ otherwise. Assume, for the sake of simplicity that the functional $\psi(\theta)$ is linear and that the LAN norm $\|.\|_L$ is the L_2 norm, as in the white noise model. Thus $\psi(\theta) =< \psi_1, \theta >$ for some $\psi_1 \in \ell_2$. To prove (14), we need to construct a change of variable $\theta \to \theta - t\psi_1/\sqrt{n}$ which makes sense under the prior distribution. If $\theta \in \mathbb{R}^k$, since $\psi_1 \notin \mathbb{R}^k$, $\theta - t\psi_1/\sqrt{n} \notin \mathbb{R}^k$ and the best approximation of $\theta - t\psi_1/\sqrt{n}$ in $\mathbb{R}^k$ is $\theta_k^{'} = \theta - t\psi_{1[k]}/\sqrt{n}$ where $\psi_{1[k]}$ is the vector made of the first k components of ψ_1. The transform $\theta \to \theta_k^{'}$ is a feasible change of variable for the conditional prior distribution given k, whereas $\theta - t\psi_1/\sqrt{n}$ is not. We have
$$\ell_n(\theta - t\psi_1/\sqrt{n}) - \ell_n(\theta - t\psi_{1[k]}/\sqrt{n}) = -\frac{n\|\psi_1 - \psi_{1[k]}\|^2}{2} + \sqrt{n} < \theta_0 - \theta_{0[k]}, \psi_1 - \psi_{1[k]} > - tW_n(\psi_1 - \psi_{1[k]}) + o_p(1)$$

Hence even if under the posterior distribution $k \to +\infty$, the term $\sqrt{n} < \theta_0 - \theta_{0[k]}, \psi_1 - \psi_{1[k]} >$ may not be negligible, so that $\ell_n(\theta - t\psi_1/\sqrt{n}) - \ell_n(\theta - t\psi_{1[k]}/\sqrt{n}) \neq o(1)$.

It thus appears that to be able to prove (14), we need $\Pi(|\sqrt{n} < \theta_0 - \theta_{0[k]}, \psi_1 - \psi_{1[k]} > | > \epsilon|Y^n) \to 0$ for all $\epsilon > 0$ under P_{θ_0}. If $\|\psi_1 - \psi_{1[k]}\|$ decreases slowly to 0 with k the above condition may fail and BvM would thus not be valid. This is in essence what drives the counter-examples considered in (Rivoirard and Rousseau, 2012a, Castillo and Rousseau, 2013, Castillo, 2012).

On the other hand under a prior with a deterministic and increasing $k = k_n$ with k_n large enough, the BvM will hold for a wider range of linear functionals at the expense of a bad posterior concentration rate on the whole parameter.

This example indicates that one cannot expect a posterior distribution on a high dimensional parameter space to be well behaved for all aspects of the parameter, it is thus important to understand which parameters of interest will be well recovered and which will not. Adding higher levels of hierarchy in the prior, here considering a distribution on the truncation level k for instance, intuitively induces greater flexibility and indeed it induces adaptive posterior concentration rates. In the same time it prevents the BvM theorem to be valid for some functionals of the parameter. Therefore, this notion of flexibility should be taken with some care.

Although there have been significant advances in understanding the asymptotic behaviour of semi-parametric models, a lot of open questions remain un-answered. So far BvM or precise statements on the posterior distribution in semi-parametric models have been obtained mainly for parameterizations based on bases expansions. More complex geometries, as in mixture models, have not yet been studied since the task is more formidable in these prior models; they are however models that are commonly used in practice.

SUMMARY POINTS

1. Some general tools have been developped to derive precise description of the posterior distribution in large and infinite dimensional models, such as BvMs or more generally frequentist coverage of Bayesian credible regions.
2. There are now some positive and negative results on coverage of credible regions and on BvM properties of the posterior distributions.
3. In large dimensional models, the prior has a strong impact and not every aspects of the parameter can be well recovered by the posterior distribution. It is therefore important to understand which parts or functionals of the parameter can be correctly estimated, given a specific prior distribution.

FUTURE ISSUES

1. Mixture prior models or more generally non linear models (as opposed to priors on parameters $\theta \in \ell_2$ for instance) are not well understood yet, at least asymptotically. They are however extensively used in practice and seem to behave well in many cases.
2. Extension of the results on frequentist coverage of credible regions to more general models than the Gaussian white noise models with Gaussian priors.
3. Develop further the theory on Bayesian nonparametric tests. This has not been mentioned in the paper, but it is also an important aspect of Bayesian inference and there have been only a few theoretical results on Bayesian nonparametric tests.

ACKNOWLEDGMENTS

I would like to thank Julyan Arbel, Sophie Donnet, Pierre Jacob, Vincent Rivoirard, Sylvia Richardson and Jean-Bernard Salomond for helpful comments on earlier versions of the manuscript, This work is partially funded by the ANR CALIBRATION.

LITERATURE CITED

Arbel et al., 2013.Arbel, J., Gayraud, G., and Rousseau, J. (2013). Bayesian adaptive optimal estimation using a sieve prior. *Scand. J. Statist.*, to appear.

Banerjee and Ghosal, 2015.Banerjee, S. and Ghosal, S. (2015). Bayesian structure learning in graphical models. *Journal of Multivariate Analysis*, 136:14–162.

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

Barron et al., 1999.Barron, A., Schervish, M., and Wasserman, L. (1999). The consistency of posterior distributions in nonparametric problems. *Ann. Statist.*, 27:536–561.

Belitser and Levit, 2002.Belitser, E. and Levit, B. (2002). On the empirical Bayes approach to adaptive filtering in the Gaussian model. *Interfaces and Free Boundaries*.

Belitser et al., 2012.Belitser, E., Serra, P., and van Zanten, J. H. (2012). Estimating the period of a cyclic non-homogeneous Poisson process. *Scand. J. Stat.*, page to appear.

Bhattacharya and Dunson, 2011.Bhattacharya, A. and Dunson, D. B. (2011). Sparse bayesian infinite factor models. *Biometrika*, 98:291–306.

Bhattacharya et al., 2014a.Bhattacharya, A., Pati, D., and Dunson, D. (2014a). Anisotropic function estimation using multi-bandwidth gaussian processes. *Ann. Statist.*, 42:352–381.

Bhattacharya et al., 2014b.Bhattacharya, A., Pati, D., Pillai, N., and Dunson, D. (2014b). Dirichlet-Laplace priors for optimal shrinkage. Technical report.

Bickel and Kleijn, 2012.Bickel, P. J. and Kleijn, B. J. K. (2012). The semiparametric bernsteinvon mises theorem. *Ann. Statist.*, 40:206–237.

Birgé, 1983.Birgé, L. (1983). Approximation dans les espaces mtriques et thorie de l'estimation. *Probability Theory and Related Fields*, 65:181–237.

Cai and Low, 2006.Cai, T. and Low, M. (2006). Adaptative confidence balls. *Ann. Statist.*, 34:202–228.

Canale and de Blasi, 2013.Canale, A. and de Blasi, P. (2013). Posterior consistency of nonparametric location - scale mixtures of multivariate gaussian density estimation. Technical report.

Castillo, 2008.Castillo, I. (2008). Lower bounds for posterior rates with Gaussian process priors. *Electronic Journal of Statistics*, 2:1281–1299.

Castillo, 2010.Castillo, I. (2010). A semiparametric Bernstein-von mises Theorem for Gaussian process priors. *Probability Theory and Related Fields*.

Castillo, 2012.Castillo, I. (2012). Semiparametric Bernstein–von Mises theorem and bias, illustrated with Gaussian process priors. *Sankhya A*, 74(2):194–221.

Castillo and Nickl, 2013.Castillo, I. and Nickl, R. (2013). Nonparametric Bernstein–von Mises theorems in Gaussian white noise. *Ann. Statist.*, 41(4):1999–2028.

Castillo and Rousseau, 2013.Castillo, I. and Rousseau, J. (2013). A general bernstein–von mises theorem in semi-parametric models. Technical report.

Castillo et al., 2015.Castillo, I., Schmidt-Hieber, J., and Van der Vaart, A. (2015). Bayesian linear regression with sparse prior. Technical report.

Castillo and van der Vaart, 2012.Castillo, I. and van der Vaart, A. (2012). Needles and straw in a haystack: Posterior concentration for possibly sparse sequences. *Ann. Statist.*, 40:2069–2101.

Choi and Schervish, 2007.Choi, T. and Schervish, M. (2007). On posterior consistency in nonparametric regression problems. *J. Multivariate Anal.*, In Press.

Clyde and George, 2000.Clyde, M. A. and George, E. I. (2000). Flexible empirical Bayes estimation for wavelets. *J. Royal Statist. Society Series B*, pages 681–698.

Cox, 1993.Cox, D. (1993). An analysis of bayesian inference for nonparametric regression. *Ann. Statist.*, 21:903–923.

Cui and George, 2008.Cui, W. and George, E. I. (2008). Empirical Bayes vs. fully Bayes variable selection. *J. Statist. Plann. Inference*, 138:888–900.

de Finetti, 1937.de Finetti, B. (1937). La prédiction: ses logiques, ses sources prédictives. *Annals de l'institut Henri Poincaré*, 7:1–68.

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

Dey et al., 1998.Dey, D., Müller, P., and Sinha, D., editors (1998). *Practical nonparametric and semiparametric Bayesian statistics*, volume 133. Springer. Lecture Notes in Statistics.

Diaconis and Freedman, 1986.Diaconis, P. and Freedman, D. (1986). On the consistency of Bayes estimates. *Ann. Statist.*, 14:1–26.

Donnet et al., 2014a.Donnet, S., Rivoirard, V., Rousseau, J., and Scricciolo, C. (2014a). Posterior concentration rates for counting processes with Aalen multiplicative intensities. *arXiv:1407.6033v1*.

Donnet et al., 2014b.Donnet, S., Rivoirard, V., Rousseau, J., and Scricciolo, C. (2014b). Posterior concentration rates for empirical Bayes procedures, with applications to Dirichlet Process mixtures. *arXiv:1406.4406v1*.

Ferguson, 1974.Ferguson, T. (1974). Prior distributions in spaces of probability measures. *Ann. Statist.*, 2:615–629.

Freedman, 1999.Freedman, D. (1999). On the Bernstein Von Mises theorem with infinite dimensional parameter. *Ann. Statist.*, 27:1119–1140.

George and Foster, 2000.George, E. I. and Foster, D. P. (2000). Calibration and empirical Bayes variable selection. *Biometrika*, 87:731–747.

Ghosal, 2001.Ghosal, S. (2001). Convergence rates for density estimation with bernstein polynomials. *Ann. Statist.*, 29(5):1264–1280.

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

Ghosal and Roy, 2006.Ghosal, S. and Roy, A. (2006). Posterior consistency of Gaussian process prior for nonparametric binary regression. *Annals of Statistics*, 34:2413–2429.

Ghosal and Tang, 2006.Ghosal, S. and Tang, Y. (2006). Bayesian consistency for Markov processes. *Sankhya*, 68:227–239.

Ghosal and van der Vaart, 2007a.Ghosal, S. and van der Vaart, A. (2007a). Convergence rates of posterior distributions for non iid observations. *Ann. Statist.*, 35(1):192–223.

Ghosal and van der Vaart, 2007b.Ghosal, S. and van der Vaart, A. (2007b). Posterior convergence rates of Dirichlet mixtures at smooth densities. *Ann. Statist.*, 35(2):697–723.

Ghosh and Ramamoorthi, 2003.Ghosh, J. K. and Ramamoorthi, R. V. (2003). *Bayesian Nonparametrics*. Springer-Verlag, New York.

Green and Richardson, 2001.Green, P. and Richardson, S. (2001). Modelling heterogeneity with and without the Dirichlet process. *Scandinavian J. Statistics*, 28(2):355–375.

Hjort et al., 2010.Hjort, N. L., Holmes, C., Müller, P., and Walker, S. G. (2010). *Bayesian Nonparametrics*. Cambridge University Press, Cambridge, UK.

Hoffman et al., 2013.Hoffman, M., Rousseau, J., and Schmidt-Hieber, J. (2013). On adaptive posterior concentration. Technical report.

Knapik and Salomond, 2015.Knapik, B. and Salomond, J. (2015). A general approach to posterior contraction in nonparametric inverse problems. Technical report.

Knapik et al., 2012.Knapik, B. T., Szabó, B. T., van der Vaart, A. W., and van Zanten, J. H. (2012). Bayes procedures for adaptive inference in nonparametric inverse problems. Technical report.

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

Kyung and Casella, 2010.Kyung, G. and Casella, G. (2010). Estimation in dirichlet random effects models. *Ann. Statist.*, 38:979–1009.

Lavine, 1992.Lavine, M. (1992). Some aspects of polya tree distributions for statistical modelling. *Ann. Statist.*, 20:1222–1235.

Lijoi et al., 2005.Lijoi, A., Pruenster, I., and Walker, S. (2005). On consistency of nonparametric normal mixtures for bayesian density estimation. *J. American Statist. Assoc.*, 100:1292–1296.

Lijoi and Prünster, 2010.Lijoi, A. and Prünster, I. (2010). Models beyond the Dirichlet process. *Bayesian nonparametrics*, 28:80.

Liu, 1996.Liu, J. S. (1996). Nonparametric hierarchical bayes via sequential imputation. *Ann. Statist.*, 24:911–930.

Pati et al., 2014.Pati, D.and Bhattacharya, A., Pillai, N., and Dunson, D. (2014). Posterior contraction in sparse bayesian factor models for massive covariance matrices. *Ann. Statist.*, 42:1102–1130.

Petrone et al., 2014.Petrone, S., Rousseau, J., and Scricciolo, C. (2014). Bayes and empirical Bayes: do they merge? *Biometrika*, 101:285–302.

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

Ray, 2013.Ray, K. (2013). Bayesian inverse problems with non-conjugate priors. *Electron. J. Statist.*, 7:2516–2549.

Richardson and Green, 1997.Richardson, S. and Green, P. (1997). On Bayesian analysis of mixtures with an unknown number of components (with discussion). *J. Royal Statist. Society Series B*, 59:731–792.

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

Rivoirard and Rousseau, 2012b.Rivoirard, V. and Rousseau, J. (2012b). Posterior concentration rates for infinite dimensional exponential families. *Bayesian Analysis*, 7:311–334.

Robbins, 1964.Robbins, H. (1964). The empirical Bayes approach to statistical decision problems. *Ann. Mathemat. Statist.*, 35:1–20.

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

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

Rousseau and Szabò, 2015.Rousseau, J. and Szabò, B. T. (2015). Asymptotic behaviour of maximum marginal lidelihood estimators and the associated empirical bayes posterior distributions. Technical report.

Schwartz, 1965.Schwartz, L. (1965). On Bayes procedures. *Z. Warsch. Verw. Gebiete*, 4:10–26.

Scott and Berger, 2010.Scott, J. G. and Berger, J. O. (2010). Bayes and empirical-bayes multiplicity adjustment in the variable-selection problem. *Ann. Statist.*, 38(5):2587–2619.

Scricciolo, 2014.Scricciolo, C. (2014). Adaptive Bayesian density estimation in L^p-metrics with Pitman-Yor or normalized inverse-Gaussian process kernel mixtures. *Bayesian Analysis*, 9:475–520.

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

Shen et al., 2013.Shen, W., Tokdar, S., and Ghosal, S. (2013). Adaptive Bayesian multivariate density estimation with Dirichlet mixtures. *Biometrika*, 100:623–640.

Szabò et al., 2013.Szabò, B. T., van der Vaart, A. W., and van Zanten, J. H. (2013). Empirical bayes scaling of gaussian priors in the white noise model. *Electron. J. Statist.*, 7:991–1018.

Szabó et al., 2013.Szabó, B. T., van der Vaart, A. W., and van Zanten, J. H. (2013). Frequentist

coverage of adaptive nonparametric Bayesian credible sets. Technical report.

van de Wiel et al., 2013.van de Wiel, M., Leday, G., Pardo, L., Rue, H., van der Vaart, A., and Van Wieringen, W. (2013). Bayesian analysis of rna sequencing data by estimating multiple shrinkage priors. *biostatistics*, 14:113–28.

van der Vaart and van Zanten, 2008a.van der Vaart, A. and van Zanten, J. H. (2008a). Reproducing kernel Hilbert spaces of gaussian priors. 3:200–222.

van der Vaart and van Zanten, 2009.van der Vaart, A. and van Zanten, J. H. (2009). Adaptive Bayesian estimation using a Gaussian random field with inverse Gamma bandwidth. *Ann. Statist.*, 37:2655–2675.

van der Vaart, 1998.van der Vaart, A. W. (1998). *Asymptotic statistics*, volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge.

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

Vernet, 2014.Vernet, E. (2014). Posterior consistency for nonparametric hidden markov models with finite state space. Technical report.

Yau et al., 2011.Yau, C., Papaspiliopoulos, O., Roberts, G. O., and Holmes, C. (2011). Bayesian non-parametric hidden Markov models with applications in genomics. *J. Royal Statist. Society Series B*, 73:1–21.