ESAIM: PROCEEDINGS, Vol. ?, 2013, 1-10
Editors: Will be set by the publisher

ON SOME ASPECTS OF THE ASYMPTOTIC PROPERTIES OF BAYESIAN APPROACHES IN NONPARAMETRIC AND SEMIPARAMETRIC MODELS *

Judith Rousseau[1], Jean-Bernard Salomond[2] and Catia Scricciolo[3]

Abstract. In this paper, we review some recent results obtained in the context of Bayesian non and semiparametric models in terms of posterior concentration, Bernstein-von Mises theorems and tests. Then two specific cases are studied in more details. The first concerns tests for monotonicity and the second some asymptotic properties of empirical Bayes procedures.

Résumé. Cet article est un article de revue et présente un certain nombre de résultats récents sur les propriété fréquentistes de procédures bayésiennes non et semiparamétriques. Nous donnons notamment des conditions permettant d'obtenir un théorème de Bernstein - von Mises pour des fonctionnelles de la densité, des résultats sur la consistance de la loi a posteriori lorsque la loi a priori dépend des données et enfin un test de monotonicité dans un modèle de régression nonparamétrique.

Introduction

0.1. Overview

The analysis of Bayesian nonparametric statistics started slowly five decades ago. During the last fifteen years, it became a thriving research field thanks to the availability of massive computational resources, to algorithmic advances and theoretical breakthroughs. These practical and theoretical developments have allowed statisticians to develop (parametric and nonparametric) models of increasing complexity. As biostatistics, machine learning and other data intensive disciplines became hotbeds for Bayesian methods, the necessity to analyse priors on infinite or at least high dimensional spaces became more and more obvious. So became the necessity to understand the limiting behaviour of posterior probabilities. The study of asymptotic properties of Bayesian nonparametric methods methods was initiated by the seminal papers of [Schwartz, 1965,Barron, 1988] then increased significantly after the works of [Barron et al., 1999, Ghosal et al., 2000a]. Since then posterior concentration has been extensively studied in various types of models including nonparametric regression [Ghosal and van der Vaart, 2007], Markov models [Tang and Ghosal, 2007], Gaussian time series [Choudhuri et al., 2004,Rousseau et al., 2012]. In this paper, we present some recent advances in the study of frequentist properties of Bayesian nonparametric inference.

Consider a dominated model $\{f_\theta^n, \theta \in \Theta\}$ where Θ is a measurable, metric space, with metric (or semi metric) $d(.,.)$, a prior probability π on Θ, and a vector X^n of observations distributed from f_θ^n. There is a vast literature on possible constructions of nonparametric priors, i.e. on priors on infinite dimensional spaces, see

* *ANR grant BANHDITS : ANR-2010-BLAN 0113-03*

[1] ENSAE-CREST and CEREMADE, Université Paris-Dauphine, FRANCE

[2] ENSAE-CREST and CEREMADE, Université Paris-Dauphine, FRANCE

[3] Bocconi University, Italy

for instance [Hjort et al., 2009] for a recent review on Bayesian nonparametric methods. The most popular are either based on the Dirichlet process or on Gaussian processes.

The posterior distribution is said to concentrate or contract with rate ϵ_n at $\theta_0 \in \Theta$ if

$$P^\pi\left[\theta; d(\theta_0, \theta) \geq \epsilon_n | X^n\right] = o_{p_{\theta_0}}(1), \tag{1}$$

where for all measurable subset B of Θ

$$P^\pi\left[B|X^n\right] = \frac{\int_B f_\theta^n(X^n) d\pi(\theta)}{\int_\Theta f_\theta^n(X^n) d\pi(\theta)}, \tag{2}$$

is the posterior probability of B. When ϵ_n corresponds to the minimax estimation rate for estimating θ under the loss $d(.,.)$ in a given class $\mathcal{C} \subset \Theta$ with $\theta_0 \in \mathcal{C}$ we say that ϵ_n is the minimax concentration rate and the posterior is said to concentrate at an adaptive minimax rate. The posterior is said to concentrate at an adaptive minimax rate if the concentration rate is the (adaptive) minimax rate over a collection of sets C_α, $\alpha \in \mathcal{I}$. Minimax and adaptive minimax concentration rates have been obtained in various cases. For instance, in the case of density estimation, nonparametric mixtures of Beta distributions and nonparametric mixtures of Gaussian distributions lead to adaptive minimax concentration rates over collections of Hölder classes up to a $\log n$ term, see [Rousseau, 2010] and [Kruijer et al., 2010]. General minimax adaptive concentration rates have been obtained by [van der Vaart and van Zanten, 2009] using hierarchical Gaussian process priors. Recent extensions of both works to anisotropic multivariate functional classes have been derived by [Shen et al., 2012, Bhattacharya et al., 2012]. A review of the existing results on posterior concentration rates is given in Section 1.1.

In the semiparametric framework the so-called Bernstein-von Mises property has also been recently investigated by [Castillo, 2012b, Castillo, 2012a, Rivoirard and Rousseau, 2012, Bickel and Kleijn, 2012, Bontemps, 2011, Leahu, 2011]. The posterior distribution of a quantity of interest $\psi(\theta)$ is said to verify the Bernstein-von Mises property if it is asymptotically Gaussian and satisfies

$$P^\pi\left[v_n(\psi(\theta) - \hat{\psi}) \leq z \middle| X^n\right] = \Phi_V(z) + o_{p_{\theta_0}}(1), \tag{3}$$

where v_n is a positive sequence going to ∞, Φ_V denotes the cumulative distribution function of a Gaussian random vector with covariance matrix V, and if under P_{θ_0}, $v_n(\hat{\psi} - \psi(\theta_0))$ is asymptotically Gaussian with mean 0 and covariance matrix V. The Bernstein-von Mises property implies in particular that credible regions, such as High Probability Density (HPD) regions or equal tail intervals are also asymptotically confidence regions with the same levels. Such results are described in Section 1.2 in the context of functionals of a curve.

Moreover, it is common practice to replace some hyperparameters entering in the definition of the prior distribution by some quantities that are data dependent. Such an approach is called empirical Bayes. The theory described above cannot be applied in the context of data dependent priors. It is in fact much more complicated to determine conditions on the model and the data dependent prior that ensure even consistency of the posterior. Recently, [Petrone et al., 2012] have studied the asymptotic behaviour of empirical Bayes approaches both in nonparametric and parametric models. In section 2 we review their result.

Another aspect of Bayesian nonparametric inference has also been recently investigated, namely the problems of tests or model choice when at least one of the hypotheses is nonparametric. Bayesian tests are often based on the so-called Bayes factor, defined in the following way: let $\Theta_0, \Theta_1 \subset \Theta$, π_0, π_1 be prior probabilities over Θ_0 and Θ_1 respectively and consider the problem of testing $\theta \in \Theta_0$ versus $\theta \in \Theta_1$. Then the Bayes factor is defined as

$$B_{0/1} = \frac{\int_{\Theta_0} f_\theta^n(X^n) d\pi_0(\theta_0)}{\int_{\Theta_1} f_\theta^n(X^n) d\pi_1(\theta_1)}. \tag{4}$$

and is related to the posterior distribution of Θ_0. The test procedure corresponds to rejecting Θ_0 if $B_{0/1}$ is small. A typical threshold is 1. However the Bayes factor gives more information than the mere 0-1 decision. The test procedures associated to the Bayes factor are thus said to be consistent if $B_{0/1}$ converges in probability to infinity under $P^n_{\theta_0}$ for all $\theta_0 \in \Theta_0$ and if it converges in probability to 0 under $P^n_{\theta_0}$ for all $\theta_0 \in \Theta_1$. Goodness-of-fit tests have been studied in terms of their asymptotic properties among others by [Dass and Lee, 2006, R. McVinish, 2009, Rousseau, 2007, Rousseau and Choi, 2012], see also [Ghosal et al., 2008]. When both hypotheses are nonparametric but one is still embedded in the other, the determination of Bayesian test procedures having good frequentist properties is quite difficult in general. In Section 3 we present a test for monotonicity proposed by [Salomond, 2013], which is both simple to implement and has asymptotic optimal frequentist properties.

0.2. Notations

Throughout the paper the data X^n are assumed to be distributed according to a statistical model $(P^n_\theta, \theta \in \Theta)$. The parameter set Θ is endowed with a sigma-field and prior probability distributions on Θ are denoted by π. The posterior associated to π is denoted $P^\pi[.|X^n]$ and is defined by (2). We consider dominated models and f^n_θ designate the density of P^n_θ with respect to a given dominated measure μ. We defined $K_n(\theta_0, \theta)$ the Kullback-Leibler divergence between $f^n_{\theta_0}$ and f^n_θ and $V_p(\theta_0, \theta)$ the p- recentered moment of the log-likelihood ratio:

$$K_n(\theta_0, \theta) = \int f^n_{\theta_0}(x^n)[\log f^n_{\theta_0}(x^n) - \log f^n_\theta(x^n)]d\mu(x^n);$$

$$V_p(\theta_0, \theta) = \int f^n_{\theta_0}(x^n)\left|\log f^n_{\theta_0}(x^n) - \log f^n_\theta(x^n) - K_n(\theta_0, \theta)\right|^p d\mu(x^n), \quad p \geq 2.$$

We also denote by $E^n_\theta[.]$ expectation with respect to P^n_θ, by $l_n(\theta) = \log f^n_\theta(X^n)$ the log - likelihood. The set of square integrable functions on $[0, 1]$ with respect to Lebesgue measure is denoted by $L_2([0, 1])$ and $h(f_1, f_2)$ defines the Hellinger distance between the two-densities f_1 and f_2:

$$h(f_1, f_2)^2 = \int (\sqrt{f_1}(x) - \sqrt{f_2}(x))^2 d\mu(x).$$

1. A review on the asymptotic behaviour of the posterior distribution in semi and non parametric models

1.1. Posterior concentration rates

In this section we describe the generic result of [Ghosal and van der Vaart, 2007] which relates properties of the priors and the model to the posterior concentration rate. We now recall Theorem 3 of [Ghosal and van der Vaart, 2007]

Theorem 1.1. *Let $d(.,.)$ be a semi-metric on Θ, $\epsilon_n > 0$ converging to 0 such that $(n\epsilon_n^2)^{-1} = o(1)$. Let $S_n = \{\theta; K_n(\theta_0, \theta) \leq n\epsilon_n^2, V_p(\theta_0, \theta) \leq (n\epsilon_n^2)^{p/2}\}$, for some $p \geq 2$. If there exists $\Theta_n \subset \Theta$ and a sequence of tests $\phi_n \in [0, 1]$ such that for all $j \geq 1$, for some constant κ*

$$\frac{\pi\left(\theta \in \Theta_n; j\epsilon_n < d(\theta_0, \theta) < 2j\epsilon_n\right)}{\pi(S_n)} \lesssim e^{\kappa j^2 n\epsilon_n^2/2}, \tag{5}$$

for some constant $c > 2$

$$\frac{\pi\left(\Theta_n^c\right)}{\pi(S_n)} \lesssim e^{-cn\epsilon_n^2},$$

and the tests ϕ_n are such that

$$E^n_{\theta_0}\left[\phi_n\right] = o(1) \qquad \sup_{\theta \in \Theta_n; j\epsilon_n < d(\theta_0, \theta) < 2j\epsilon_n} E^n_\theta\left[1 - \phi_n\right] \leq e^{-\kappa j^2 n^2 \epsilon_n^2}, \tag{6}$$

then for all M_n going to infinity,

$$E^n_{\theta_0}\left(P^\pi\left[d(\theta_0,\theta) > M_n\epsilon_n|X^n\right]\right) = o(1). \tag{7}$$

There are essentially two key conditions appearing in this theorem: (1) A lower bound on the prior mass of Kullback-Leibler neighbourhoods of the true distribution and (2) the existence of tests. The lower bound on Kullback-Leibler neighbourhoods of the true distribution expressed in condition (5) requires to develop some approximation theory (depending on both the sampling and the prior model), however it has been studied now in a wide range of models and priors. The existence of tests, as required in condition (6) is discussed below.

The proof of Theorem 1.1 is straightforward and essentially relies on the following control of the posterior probability: let $B_n = \{d(\theta_0,\theta) > M_n\epsilon_n\}$ then

$$P^\pi\left[d(\theta_0,\theta) > M_n\epsilon_n|X^n\right] = \frac{N_n}{D_n} := \frac{\int_{B_n} e^{l_n(\theta)-l_n(\theta_0)}d\pi(\theta)}{\int_\Theta e^{l_n(\theta)-l_n(\theta_0)}d\pi(\theta)}$$

Using the Fubini and Markov inequalities,

$$\begin{aligned} E^n_{\theta_0}\left(P^\pi\left[d(\theta_0,\theta) > M_n\epsilon_n|X^n\right]\right) &\le E^n_{\theta_0}[\phi_n] + P^n_{\theta_0}\left(D_n < e^{-2n\epsilon_n^2}\pi(S_n)/2\right) \\ &\quad + \frac{e^{2n\epsilon_n^2}}{\pi(S_n)}\int_{B_n\cap\Theta_n} E^n_\theta\left[1-\phi_n\right]d\pi(\theta) + \frac{e^{2n\epsilon_n^2}\pi(\Theta_n^c)}{\pi(S_n)}. \end{aligned}$$

Then D_n is bounded from below by

$$\begin{aligned} D_n &\ge \int_{S_n} 1\!\!1_{l_n(\theta)-l_n(\theta_0)>-2n\epsilon_n^2} e^{l_n(\theta)-l_n(\theta_0)}d\pi(\theta) \\ &\ge e^{-2n\epsilon_n^2}\pi\left(S_n\cap\{l_n(\theta)-l_n(\theta_0) > -2n\epsilon_n^2\}\right) \end{aligned}$$

so that

$$\begin{aligned} P^n_{\theta_0}\left(D_n < e^{-2n\epsilon_n^2}\pi(S_n)/2\right) &\le P^n_{\theta_0}\left[\pi\left(S_n\cap\{l_n(\theta)-l_n(\theta_0) < -2n\epsilon_n^2\}\right) > \pi(S_n)/2\right] \\ &\le \frac{2\int_{S_n} P^n_{\theta_0}\left[l_n(\theta)-l_n(\theta_0)\le -2n\epsilon_n^2\right]d\pi(\theta)}{\pi(S_n)}. \end{aligned}$$

In a variety of models, depending on the loss function (or semi-metric) $d(.,.)$, some tests with exponential power as required in condition (6) exist. For instance, in the case of density or conditional density estimation, Hellinger or L_1 tests have been determined by [Birge, 1983]. Other examples of tests can be found in [Ghosal and van der Vaart, 2007] and in [Rousseau et al., 2012].

Condition (6) leads however to some restriction on the types of loss functions or semi-metric that can be considered in this context. For example, consider in the case of the white noise model, which when expanded on an orthonormal basis can be written as

$$X_i = \theta_i + n^{-1/2}\epsilon_i, \quad i\in\mathbb{N}; \quad \epsilon_i \sim \mathcal{N}(0,1) \text{ independently,}$$

where $\theta = (\theta_i)_{i\in\mathbb{N}} \in \ell_2$ the set of sequences on $\mathbb{R}$ satisfying $\sum_{i=0}^\infty \theta_i^2 < +\infty$. The following prior leads to good frequentist properties in terms of the ℓ_2 loss but not in terms of the sup-norm (L_∞) or pointwise loss. The prior is defined as follows: Let $K\sim\pi_K$ where π_K is a distribution on the set of integers, say the Poisson or the Geometric distribution. Then given K, set $\theta_i = 0$ for all $i > K$ and independently draw θ_i for $i\le K$ from the density $g(./\tau_i)/\tau_i$, where g is a density on $\mathbb{R}$ and the sequence $(\tau_i)_{i\in\mathbb{N}}$ are given beforehand. We call this prior the **sieve prior**.

In [Arbel et al., 2013], it is proved that under some mild conditions on g, π_K and the τ_j's, for all $1/2 < \alpha_1 < \alpha_2 < +\infty$ and all $L > 0$, there exists $M > 0$ such that

$$\sup_{\alpha_1 \le \alpha \le \alpha_2} \sup_{\theta_0 \in \Theta_\alpha(L)} E_{\theta_0}\left[P^\pi\left[\|\theta - \theta_0\|_2 > M(n/\log n)^{-\alpha/(2\alpha+1)}\Big| X^n\right]\right] = o(1)$$

where $\Theta_L(\alpha) = \{\theta \in \ell_2; \sum_{j=1}^\infty j^{2\alpha}\theta_j^2 \le L\}$ is the α - Sobolev ball with radius L. Interestingly if one is interested in local features of the signal, say $\psi(\theta) = \sum_{j=1}^\infty \theta_j$, then the posterior is not so well-behaved since for all α

$$\inf_{\theta_0 \in \Theta_\alpha(L)} E_{\theta_0}\left[P^\pi\left[|\psi(\theta) - \psi(\theta_0)| \le M(n/\log n)^{-(\alpha-1/2)/(2\alpha+1)}\Big| X^n\right]\right] = o(1)$$

whereas the minimax estimation rate of $\psi(\theta)$ on $\Theta_\alpha(L)$ is equal to $n^{-(\alpha-1/2)/(2\alpha)}$. This phenomenon is due to the fact that under the ℓ_2 loss the optimal truncation K is of order $K_n = n^{1/(2\alpha+1)}$ whereas it would be equal to $\tilde{K}_n = n^{1/(2\alpha)}$ for the estimation of $\psi(\theta)$. Since the posterior distribution is driven by the likelihood ratio (i.e. driven by the ℓ_2 loss in the white noise model), it concentrates on truncation values of order smaller than $n^{1/(2\alpha+1)}$ and a bias occurs in the estimation of $\psi(\theta)$.

Interestingly in the specific case of point-wise loss functions, such as $(\psi(\theta) - \psi(\theta_0))^2$ or the sup-norm loss function, the study of posterior concentration rates based on the existence of tests is bound to lead to sup-optimal posterior concentration rates, because of the need to obtain an exponentially small second-type error as in condition (6), see for instance [Giné and Nickl, 2012]. Recently, a general theory on the possibility of using the approach proposed in Theorem 1.1 has been proposed in [Hoffmann et al., 2013] based on a lower bound on the posterior concentration rates.

To conclude this section, it thus appears that Bayesian nonparametric procedures have good frequentist properties when the loss function under study is somewhat related to the Kullback-Leibler divergence, or to phrase it differently when tests with exponential second- type error can be constructed. In semiparametric contexts this is not necessarily the case. The prior has to be chosen carefully for the posterior to have good frequentist properties. In the following section we describe more precisely some asymptotic aspects of Bayesian semiparametric approaches, namely the Bernstein-von Mises property.

1.2. **On the semiparametric Bernstein - von Mises Theorem**

As described in the introduction, the semiparametric Bernstein - von Mises theorem corresponds to determination of the asymptotic posterior distribution of some finite dimensional quantity of interest $\psi(\theta)$ in the form (3). There are essentially two types of semiparametric problems. First the case where $\theta = (\psi, \eta)$ where $\psi \in \mathbb{R}^d$ is the parameter of interest and $\eta \in \mathcal{S}$ is an infinite dimensional nuisance parameter. Bernstein - von Mises theorems have been obtained in this framework by [Castillo, 2012b, Castillo, 2012a, Bickel and Kleijn, 2012] in the case of regular models and by [Kruijer and Rousseau, 2012] in a specific non regular model. The second type corresponds to functionals of the whole parameter, such as the cumulative distribution function at a given point in the density model, or linear functionals of a curve, etc. This has been studied in particular in [Rivoirard and Rousseau, 2012]. In this section we present the latter case and more precisely the results obtained in [Rivoirard and Rousseau, 2012], with improvements obtained in [Castillo and Rousseau, 2013], where sufficient conditions are proposed in the framework of smooth linear functionals of the density.

Let $X^n = (X_1, \cdots, X_n)$ be a n sample, with density f on $[0, 1]$ with respect to Lebesgue measure, where f is unknown. Let $\Psi(f) = \int_0^1 \psi(x)f(x)dx$ be any continuous linear functional of the density f. The aim is to determine the asymptotic posterior distribution of $\sqrt{n}(\Psi(f) - \psi(\mathbb{P}_n))$, where $\psi(\mathbb{P}_n) = n^{-1}\sum_{i=1}^n \psi(X_i)$, and to obtain conditions on the prior model so that it is asymptotically Gaussian with mean 0 and variance $V = \int_0^1 \psi^2(x)f(x)dx - \Psi(f)^2$, assuming that ψ is bounded. We then have the following theorem

Theorem 1.2. *Under the following two conditions:*
C1: Posterior concentration *There exist ϵ_n converging to 0 and*

$$A_n \subset \{f; h(f_0, f) \leq \epsilon_n\}, \qquad \textit{with} \qquad P^\pi\left[A_n \middle| X^n\right] = 1 + o_{p_{f_0}}(1)$$

C2: Change of variable *Let $0 < |t|$, $f \in A_n$ and define*

$$\bar{\psi}_{t,f}(x) = \psi(x) + \frac{\sqrt{n}}{t} \log\left(\int_0^1 f(x) e^{-t\psi(x)/\sqrt{n}} dx\right).$$

There exists $t_0 > 0$ such that for all $|t| \leq t_0$ and $f \in A_n$,

$$\frac{\int_{A_n} e^{l_n(f e^{-t\bar{\psi}_{t,f}/\sqrt{n}}) - l_n(f_0)} d\pi(f)}{\int_{A_n} e^{l_n(f) - l_n(f_0)} d\pi(f)} d\pi(f) = 1 + o_{p_{f_0}}(1), \tag{8}$$

we have:

$$\sup_{z \in \mathbb{R}} \left| P^\pi\left[\sqrt{n}(\Psi(f) - \psi(\mathbb{P}_n)) \leq z \middle| X^n\right] - \Phi_V(z)\right| = o_{p_{f_0}}(1). \tag{9}$$

As seen in the previous section, there is now an extensive literature on posterior concentration rates so that the tools described in Section 1.1 can be applied to verify Condition **C1**. The key condition is (8). To verify condition (8), one needs to construct a change of parameter $Tf = fe^{-t\bar{\psi}_{t,f}/\sqrt{n}}$ for all $f \in A_n$ which only slightly alters the prior π and the set A_n. To illustrate the phenomena that can occur consider a simple sieve model on the set of prior densities, similar to the sieve priors described in Section 1.1: for all $\theta \in \ell_2$, define densities on $[0,1]$ in the form

$$f_\theta(x) = \exp\left(\sum_{j=0}^{\infty} \theta_j \phi_j(x) - c(\theta)\right),$$

where $(\phi_j)_{j \in \mathbb{N}}$ is an orthonormal basis on $L_2([0,1])$ satisfying $\psi_0 = 1$ and consider the **sieve prior** defined above on $\theta \in \ell_2$. We assume that $0 < c_0 \leq f_0 = f_{\theta_0} \leq C_0 < +\infty$ where f_0 denotes the true density of the observations, and $\theta_0 \in \ell_2$. It can be proved that if $K \sim \pi_K$ where π_K is either the Poisson or the Geometric distribution then the posterior concentration rate in the Hellinger loss is of order $(n/\log n)^{-\alpha/(2\alpha+1)}$ over Sobolev balls, for all $\alpha > 1/2$ and condition **C1** is satisfied. However to construct the change of parameters $f \to Tf$, we need to make the change of parameters $\theta \to \tilde{T}\theta$ within each submodel Θ_K corresponding to the first K coefficients. To do so, define $\psi_{[K]} = (\psi_{j,[K]}, j \leq K)$ the coefficients of the orthogonal projection of $\bar{\psi}_{t,f_\theta}$ onto the space spanned by $(\phi_j)_{j \leq K}$, with respect to the inner product $< g_1, g_2 >= \int_0^1 g_1(x) g_2(x) f_0(x) dx$. The change of variable is then constructed as follow: for all K in the asymptotic support of the posterior distribution, set $\theta_{t,K} = \theta - t\psi_{[K]}/\sqrt{n}$ for all $\theta \in \mathbb{R}^K \cap A_n$ with $A_n := \{\|\theta - \theta_0\| \leq M(n/\log n^{-\alpha/(2\alpha+1)})\}$. Condition (8) is valid when $\theta_0 \in \Theta_\alpha(L)$, $\alpha > 1/2$, $L > 0$, if

$$\sup_{\theta \in \mathbb{R}^K \cap A_n} \left| l_n(f_\theta e^{-t\bar{\psi}_{t,f_\theta}/\sqrt{n}}) - l_n(f_{\theta_t}) \right| = o_{p_{f_0}}(1)$$

and

$$\sup_{\theta \in \mathbb{R}^K \cap A_n} \left| \sum_{j=0}^{K} \left(\log g((\theta_j - t\psi_{j,[K]}/\sqrt{n})/\tau_j) - \log g(\theta_j/\tau_j)\right) \right| = o(1).$$

The first condition means that the change of parameters $f \to Tf$ can be approximated within each submodel by $\theta \to \theta_t$ and the second one that the prior is not modified asymptotically by this change of parameters. Under some mild conditions on g and $(\tau_j)_{j \geq 0}$, the latter is verified for any $\alpha > 1/2$. For the former to be verified it is necessary to have K large enough, since the difference between both changes of parameters (in the likelihood) is of the same order as the difference $\sqrt{n}(\bar{\psi}_{f_\theta,t} - \sum_{j=0}^{K}\psi_{j,[K]})$. Thus for (8) to be valid some no-bias condition is required, which is true in particular if for all $\epsilon > 0$,

$$P^\pi\left[K; \sum_{j>K}\psi_j^2 < \epsilon n^{1/(2\alpha+1)}(\log n)^{-2\alpha/(2\alpha+1)} \middle| X^n\right] = o_{p_{f_0}}(1).$$

Hence, a bias may appear when K is apriori random as shown in an example proposed by [Rivoirard and Rousseau, 2012]. On the contrary, if $K = K_n = \lfloor n^{1/(2\beta+1)} \rfloor$ for some $\beta > 1/2$ and $\alpha \geq \beta$, then (9) is valid and the conclusion of Theorem 1.2 holds.

In both Sections 1.1 and 1.2, the priors do not depend on the data, however it is common practice to replace some of the hyperparameters defining the prior by some quantity which is data dependent. For instance, in the case of density estimation, under a Dirichlet mixture of Gaussian prior, the prior puts mass 1 on densities of the form

$$f(x) = \sum_{j=1}^{\infty} p_j \varphi((x - m_j)/\sigma)/\sigma,$$

where the $(p_j)_{j\geq 1}$ drawn from the stick-breaking distribution, the m_j's are independent and identically distributed from a Gaussian prior with mean m_0 and variance τ_0 and σ follows an inverse Gamma distribution, see for instance [Ghosh and Ramamoorthi, 2003]. Then it is common practice to center m_0 on the empirical mean and τ_0 either on the empirical variance or on the square of the difference between the largest and the smallest observations. Another typical example where such data dependent prior is used is through the so-called type-II marginal maximum likelihood estimation, see for instance [Berger, 1985, Clyde and George, 2000, Cui and George, 2008]. Surprisingly, there are few studies on generic conditions to obtain posterior consistency under data dependent priors. In the following section we describe the recent work of [Petrone et al., 2012] on the asymptotic behaviour of empirical Bayes procedures, i.e. under data dependent priors. This paper considers both parametric and nonparametric models but, for the sake of conciseness we will present here only the results dealing with posterior consistency.

2. On consistency for empirical Bayes procedures

In this section, we call empirical Bayes procedures, Bayesian approaches associated with data dependent priors. The general setup is the following, let $f_\theta^n, \theta \in \Theta$ be a statistical model and $(\pi_\lambda, \lambda \in \Lambda)$ is a family of prior distributions where Θ can be either finite or infinite dimensional and $\Lambda \subset \mathbb{R}^d$ for some $0 < d < +\infty$. We consider two types of empirical Bayes approaches. First, the marginal maximum likelihood approach, which consists in choosing

$$\hat{\lambda}_n = \text{argsup}_{\lambda\in\Lambda}\int_\Theta f_\theta^n(X^n)d\pi(\theta|\lambda) := \text{argsup}_{\lambda\in\Lambda} m(X^n|\lambda), \tag{10}$$

which is also known under the name type-II maximum likelihood estimator. The second is the plug-in type of empirical Bayes, where $\hat{\lambda}_n$ is explicitely defined, like an empirical moment of some given quantity. It may happen that $\hat{\lambda}_n$ converges to a given value under P_{θ_0}, but it is not required.

It is common belief that the empirical Bayes posterior should be close to some purely Bayesian posterior, however as in [Diaconis and Freedman, 1986] to obtain asymptotic merging between the empirical Bayesian and any other Bayesian posteriors it is necessary that the empirical Bayes posterior is consistent, for some given topology. We say that the posterior is consistent at θ_0 in a given topology if for any neighbourhood U with respect to this topology,

$$P^\pi\left[U|X^n\right] = 1 + o(1),$$

where the convergence above is either with probability 1 under P_{θ_0} or with probability going to 1 under P_{θ_0}. In this section we present some sufficient conditions to obtain posterior consistency for empirical Bayes procedures. First note that in the case of a fully Bayesian approach, i.e. if the prior does not dependent on the data, in the case of independent and identically distributed observations, the posterior is weakly consistent (i.e. with respect to the weak topology) if the so-called Kullback-Leibler property holds, see for instance [Barron, 1988], i.e. if for all $\epsilon > 0$

$$\pi\left[\theta; K_\infty(f_{\theta_0}, f_\theta) < \epsilon\right] > 0$$

where

$$K_\infty(f_{\theta_0}, f_\theta) := \lim_{n\to+\infty} n^{-1}[l_n(\theta_0) - l_n(\theta)]$$

where the limite above is taken as $P_{\theta_0}^\infty$ almost surely. In the case of empirical Bayes posteriors, it is not enough to assume such a condition. We first present the result in the case of the marginal maximum likelihood empirical Bayes approach.

2.1. Maximum marginal likelihood empirical Bayes case

In this section we study the asymptotic behaviour of the empirical Bayes posterior defined as

$$P^{EB}\left[B|X^n\right] := \frac{\int_B f_\theta^n(X^n)d\pi(\theta|\hat{\lambda}_n)}{\int_\Theta f_\theta^n(X^n)d\pi(\theta|\hat{\lambda}_n)}$$

where $\hat{\lambda}_n$ is defined by (10). We then have the following theorem.

Theorem 2.1. *Let $\hat{\lambda}_n$ be the maximum marginal likelihood estimator. Under the following two assumptions:*

- ***C3*** *There exist constants c_1, $c_2 > 0$ such that, for any U neighbourhood of θ_0 (associated to a given topology)*

$$P_{\theta_0}^*\left[\sup_{\theta\in U^c}[l_n(\theta) - l_n(\theta_0)] \geq -c_1 n\epsilon^2\right] \leq c_2(n\epsilon^2)^{-(1+t)}$$

  for some $t > 0$, where $P_{\theta_0}^*$ denotes the outer measure.
- ***C4*** *For each $\theta_0 \in \Theta$, there exists $\lambda_0 \in \Lambda$ such that, for any $\epsilon > 0$,*

$$\pi(K_\infty(f_{\theta_0}, f_\theta) < \epsilon|\lambda_0) > 0.$$

the EB posterior $P^{EB}(\cdot|X^n) = P^\pi\left(.|\hat{\lambda}_n, X^n\right)$ is consistent at θ_0: i.e. for any neighbourhood U of θ_0, $P^\pi(U^c|\hat{\lambda}, X^n) \to 0, \text{a.s.} [P_{\theta_0}^\infty]$.

On the one hand, condition **C4** is the usual Kullback-Leibler condition and it appears in a rather weak form since it needs only be verified for at least one $\lambda_0 \in \Lambda$. Condition **C3** on the other hand is quite demanding. It has been proved however in a series of nonparametric models where maximum likelihood estimation is considered, see for instance [Wong and Shen, 1995]. In [Petrone et al., 2012], a counter example is given where every fully Bayesian posterior where the prior has full support is consistent everywhere but where the empirical Bayes approach based on the marginal maximum likelihood estimator is inconsistent. This illustrates the fact that the Kullback-Leibler condition is usually not sufficient to ensure consistency in empirical Bayes approaches.

2.2. Plug-in case

Another common approach to empirical Bayes is to use a plug-in data dependent value $\hat{\lambda}_n$. We only require that there exists a sequence of compact sets $\mathcal{K}_n \subseteq \Lambda \subseteq \mathbb{R}^\ell$ such that, with P_{θ_0}–probability one, $\hat{\lambda}_n \in \mathcal{K}_n$ when n is large enough. Typically $\hat{\lambda}_n$ can converge to a given value under P_{θ_0} but this is not necessary. We then have the following theorem

Theorem 2.2. *We consider for all λ, $\lambda' \in \mathcal{K}_n$, there exists a measurable transformation $\psi_{\lambda,\lambda'} : \Theta \to \Theta$ such that if $\theta \sim \pi(\cdot \mid \lambda)$ then $\psi_{\lambda,\lambda'}(\theta) \sim \pi(\cdot \mid \lambda')$ and we assume the following two assumptions on the transformations:*

- ***C5*** *For every $\delta > 0$ and $\lambda \in \mathcal{K}_n$, there exists a sequence u_n such that $u_n^{-\ell} > \exp(-cn)$ for some constant $c > 0$ and a set $S \in \mathcal{B}(\Theta)$ such that $\liminf_n \inf_{\lambda \in \mathcal{K}_n} \pi(S|\lambda) > 0$ and*

$$\sum_{n=1}^{\infty} u_n^{-\ell} \sup_{\theta \in S} P_{\theta_0}^n \left\{ \inf_{\|\lambda - \lambda'\| \leq u_n} (l_n(\psi_{\lambda,\,\lambda'}(\theta)) - l_n(\theta_0)) < -n\delta \right\} < \infty;$$

- ***C6*** *For every U, neighbourhood of θ_0, there exist $\eta_0 > c$ and tests $\phi_n : \mathcal{X}^n \to [0,1]$ such that, for all $\lambda \in \mathcal{K}_n$, $\sum_{n=1}^{\infty} E_{\theta_0}^n \{\phi_n(X^n)\} < \infty$, and*

$$\int_{U^c} \int_{\mathcal{X}^n} \{1 - \phi_n(x^n)\} \sup_{\|\lambda - \lambda'\| \leq u_n} f_{\psi_{\lambda,\,\lambda'}(\theta)}^n (x^n) \, \mathrm{d}\mu(x^n) \mathrm{d}\pi(\theta \mid \lambda) \leq e^{-n\eta_0}.$$

Then, for any neighbourhood U of θ_0, $\Pi(U^c \mid \hat{\lambda}_n, X^n) \to 0$ with probability one under P_{θ_0}.

The transformations $\psi_{\lambda,\lambda'}$ allow us to transfer the dependence on the data in the prior through $\hat{\lambda}_n$ into a modification of the likelihood. Since u_n can typically be choosen as small as n^{-b} for any $b > 0$, conditions **C5** and **C6** are rather mild conditions. In [Petrone et al., 2012], they are proved to hold in the nonparametric density model, where the observations are assumed to be independent and identically distributed and where the prior on the density f is a Dirichlet process location mixture of Gaussian distributions:

$$\int_{\mathbb{R}} \varphi_\sigma(x - \mu) dP(\mu), \quad P \sim DP(\alpha \mathcal{N}(\lambda, \tau^2)), \quad \sigma \sim H$$

where H is a Gamma distribution and if $\hat{\lambda}_n$ is the empirical mean of the observations, then the empirical Bayes procedure is consistent for any true positive, continuous and bounded density f_0 satisfying

$$\left| \int_{\mathbb{R}} f_0(x) \log \left(\inf_{|x-t| \leq \delta} f_0(t) \right) dx \right| < \infty, \quad \int_{\mathbb{R}} |x|^{2+\delta} f_0(x) dx < +\infty.$$

These are the same conditions as those considered in [Wu and Ghosal, 2008] in the case where the base measure of the Dirichlet process is not data dependent.

Sections 1.1 and 2 concern various aspects of Bayesian nonparametric or semiparametric estimation procedures. In the following section we describe another aspect of Bayesian nonparametric inference, namely the problem of nonparametric tests. When both hypotheses are nonparametric this is a difficult issue and there is no theoretical result in the Bayesian literature apart from a partial result in [Holmes et al., 2012]. Section 3 deals with the special case of a Bayesian nonparametric test of monotonicity for the regression function.

3. Bayes test for monotonicity

In this Section we consider the test for monotonicity in the Gaussian regression setting. Consider the usual regression model with regular fixed design $z_i = i/n$

$$X_i = f(z_i) + \sigma\epsilon_i, \text{ where } \epsilon_i \overset{iid}{\sim} \mathcal{N}(0,1), \quad \sigma > 0, \tag{11}$$

where f and σ are unknown. Denoting $\mathcal{F}$ the set of monotone non increasing functions, we want to test

$$H_0 : f \in \mathcal{F} \text{ versus } H_1 : f \notin \mathcal{F} \tag{12}$$

Many tests have been proposed in the frequentist literature (see for instance [Ghosal et al., 2000b, Baraud et al., 2005] or more recently [Akakpo et al., 2012]). In this section we describe a simple Bayesian procedure to address the testing problem (12), proposed in [Salomond, 2013]. We consider a prior distribution π on f of the form

$$f = f_{k,\omega} := \sum_{i=1}^{k} \mathbb{1}_{[(i-1)/k,i/k)}\omega_i, \; d\pi(f) = \pi(k)\pi(\omega_1, \dots, \omega_k|k). \tag{13}$$

In this case the Bayes factor can lead to inconsistent results, when f has flats parts and is thus at the boundary of both hypotheses. To tackle this issue, we force the test to accept more easily the null hypothesis while retaining a good asymptotic power. To do so, we consider a thresholded version of (12), namely

$$H_0' : d(f, \mathcal{F}) \leq \tau \text{ versus } H_1' : d(f, \mathcal{F}) > \tau$$

where $d(f,g) = n^{-1}\sum_{i=1}^{n}(f(i/n)-g(i/n))^2$ is the L^2 norm on the design and $d(f,\mathcal{F}) = \inf_{g\in\mathcal{F}} d(f,g)$ and $\tau > 0$ is some threshold derived from prior knowledge on the tolerance we can accept for departure from monotonicity under H_0. In many situations, such a knowledge is not available, [Salomond, 2013] thus poposes an automatic calibration of the threshold such that the test has good asymptotic properties. This idea of test approximation is similar to the one proposed in [Rousseau, 2007] and to approximation of a point null hypothesis by an interval hypothesis testing. When $f = f_{\omega,k}$ as defined in (13) monotonicty can be represented as

$$H(\omega, k) = \max_{j>i}(\omega_j - \omega_i) \leq 0,$$

which corresponds to the sup-norm between regression function f and the set of monotone non increasing functions, when f is piecewise constant. The following theorem gives sufficient conditions on the prior as well as an automatic calibration for the threshold and the prior such that our test achieve good frequentist properties together with good finite sample performances.

Theorem 3.1. *Assume that the prior on σ has a positive density on $\mathbb{R}^+$ with respect to Lebesgue measure and that conditionnally on k, $\omega_1, \cdots, \omega_k$ are independent and identically distributed from an absolutely continuous distribution with respect to Lebesgue measure with positive and continuous density on $\mathbb{R}$. Assume further that π_k satisfies*

$$e^{-C_d kL(k)} \leq \pi_k(k) \leq e^{-C_u kL(k)} \tag{14}$$

where $L(k)$ is either equal to $\log(k)$ or to 1, for some positive constants C_d and C_u. Let $\tau_n^k(M_0) = M_0\sqrt{k\log(n)/n}$, for $M_0 > 0$ and δ_n^π the testing procedure

$$\delta_n^\pi = \mathbb{1}\left\{P^\pi\left(H(\omega,k) > \tau_n^k|X^n\right) > 1/2\right\},$$

then there exist some $M, L > 0$ depending on M_0 such that for all $\alpha \in (0,1]$

$$\begin{aligned} \sup_{f\in\mathcal{F}, \|f\|_\infty \leq L} \mathrm{E}_f^n(\delta_n^\pi) = o(1) \\ \sup_{f, d(f,\mathcal{F})>\rho, f\in\mathcal{H}(\alpha,L)} \mathrm{E}_f^n(1-\delta_n^\pi) = o(1) \end{aligned} \tag{15}$$

for all $\rho > \rho_n(\alpha) = M(n/\log(n))^{-\alpha/(2\alpha+1)}v_n$ where $v_n = 1$ when $L(k) = \log(k)$ and $v_n = \sqrt{\log(n)}$ when $L(k) = 1$.

Note that the case $L(k) = \log k$ is satisfied for instance by any Poisson prior on k, while $L(k) = 1$ corresponds for instance to a Geometric distribution. Here both M and L depend on M_0.

Similarly to the frequentist test proposed in the literature, our testing procedure has good asymptotic properties for Hölder smooth functions under the alternative. Note that neither the prior nor the hyperparameters depend on the regularity α of the regression function under the alternative. Thus our test is adaptive. Interestingly our calibration leads to a separation rate ρ_n that is the minimax separation rate up to a $\log n$ factor.

The separation rate, which is the minimal value ρ such that (15) is still valid, gives some information on the amount of tolerance expected using such a test. This can be seen as a criterion of effectiveness of our threshold.

The proof of Theorem 3.1 is as follows. For each k, we approximate the true regression function f_0 in the submodel $\mathcal{G}_k$ of piecewise constant functions associated with k bins on $\{[0,1/k),\dots[1-1/k,1)\}$ by $f_{\omega^0,k}$ which minimizes the Kullback-Leibler divergence with f_0. This leads to a closed form expression for $\omega^0=(\omega_1^0,\dots,\omega_k^0)$:

$$\omega_i^0 = n_i^{-1}\sum_{j,j/n\in[(i-1)/k,i/k)} f_0(j/n),\ n_i=\mathrm{Card}\,\{j, j/n\in[(i-1)/k,i/k)\} \tag{16}$$

so that $f_{\omega^0,k}$ belongs to $\mathcal{F}$ for all k when $f_0\in\mathcal{F}$. To prove the first part of (15), note that $H(\omega,k)\le 2\max|\omega_i-\omega_i^0|$ if $f_0\in\mathcal{F}$ so that the threshold τ_n^k needs to be as large as the posterior concentration rate of ω to ω^0 in the misspecified model $\mathcal{G}_k$. Then to prove the second part of (15), we bound from below $H(\omega,k)$ by $H(\omega^0,k)-2\max|\omega_i-\omega_i^0|$ which implies a constraint on the separation rate of the test to ensure that uniformly over $d(f_0,\mathcal{F})\ge\rho_n$ and $f\in\mathcal{H}(\alpha,L)$ we have $H(\omega,k)>\tau_n^k$.

Contrarywise to the frequentist test proposed by [Baraud et al., 2005], the least favourable regression function under the null is not the constant function, although conditionnally on k this is the case. In [Salomond, 2013] an extensive discussion on ways to calibrate M_0 is conducted. To illustrate this theoretical result, we present some results from a simulation study based on the following prior:

$$\pi := \begin{cases} k & \sim \mathrm{Geom}(p)\\ \sigma^2|k & \sim IG(\alpha,\beta)\\ \omega|\sigma,k & \sim \mathcal{N}_k\left(m, \frac{\sigma^2}{\mu}I_k\right)\end{cases} \tag{17}$$

for $0<1<p$, $\alpha,\beta,\mu>0$ and $m\in\mathbb{R}^k$. This specific choice has the advantage to allow for exact computations of the posterior and thus the implementation of the testing procedure is straightforward. The test is run for the following nine functions considered also in [Baraud et al., 2005, Akakpo et al., 2012]. The testing procedure in this paper is calibrated so that under constant true regression functions the type I error is approximately of order 0.05.

$$\begin{aligned} f_1(x) =& -15(x-0.5)^3 \mathbb{1}_{x\le 1/2} - 0.3(x-0.5) + e^{-250(x-0.25)^2}\\ f_2(x) =& 0.15x\\ f_3(x) =& 0.2e^{-50(x-0.5)^2}\\ f_4(x) =& -0.5\cos(6\pi x)\\ f_5(x) =& -0.2x + f_3(x)\\ f_6(x) =& -0.2x + f_4(x)\\ f_7(x) =& -(1+x) + 0.25e^{-50(x-0.5)^2}\\ f_8(x) =& -0.5x^2\\ f_9(x) =& 0\end{aligned} \tag{18}$$

For each function, we run the testing procedure for 500 repeated samples, from which we compute empirical type I or type II errors depending on the true function. Functions f_1 to f_7 are not monotone non increasing and the results obtained with the testing procedure proposed in [Salomond, 2013] are very similar to those obtained by [Baraud et al., 2005, Akakpo et al., 2012]. Functions f_8 and f_9 are monotone non increasing and the Type I error is smaller than 0.05. The results are presented in Table 1.

The authors would like to thank the anonymous referee for very helpful suggestions on the presentation of the paper.

Table 1. Percentage of rejection for the simulated examples for 500 samples.

	f_0	σ^2	Barraud et	Akakpo et	Bayes Test, n :			
			al. $n = 100$	al. $n = 100$	100	250	500	1000
H_0	f_1	0.01	99	99	98.0	100.0	100.0	100.0
	f_2	0.01	99	100	98.4	100.0	100.0	100.0
	f_3	0.01	99	98	100.0	100.0	100.0	100.0
	f_4	0.01	100	99	100.0	100.0	100.0	100.0
	f_5	0.004	99	99	100.0	100.0	100.0	100.0
	f_6	0.006	98	99	100.0	100.0	100.0	100.0
	f_7	0.01	76	68	69.1	100.0	100.0	100.0
H_1	f_8	0.01	-	-	3.0	0.8	1.0	2.8
	f_9	0.01	-	-	5.0	2.8	2.2	3.6

References

[Akakpo et al., 2012] Akakpo, N., Balabdaoui, F., and Durot, C. (2012). Testing monotonicity via local least concave majorants.

[Arbel et al., 2013] Arbel, J., Gayraud, G., and Rousseau, J. (2013). Bayesian adaptive optimal estimation using a sieve prior. *Scandinavian journal of statistics*, to appear.

[Baraud et al., 2005] Baraud, Y., Huet, S., and Laurent, B. (2005). Testing convex hypotheses on the mean of a Gaussian vector. Application to testing qualitative hypotheses on a regression function. *Ann. Statist.*, 33(1):214–257.

[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 consitency of posterior distributions in nonparametric problems. *Ann. Statist.*, 27:536–561.

[Berger, 1985] Berger, J. (1985). *Statistical Decision Theory and Bayesian Analysis.* Springer-Verlag, New York, second edition.

[Bhattacharya et al., 2012] Bhattacharya, A., Pati, D., and Dunson, D. (2012). Anisotropic function estimation using multi-bandwidth gaussian processes. Technical report.

[Bickel and Kleijn, 2012] Bickel, P. J. and Kleijn, B. J. K. (2012). The semiparametric Bernstein-von Mises theorem. *Ann. Statist.*, 40:206–237.

[Birge, 1983] Birge, L. (1983). Approximation dans les espaces métriques et théorie de l'estimation. *Z. Wahrsch. Verw. Gebiete*, 65:181–237.

[Bontemps, 2011] Bontemps, D. (2011). Bernstein von Mises theorems for gaussian regressions with increasing number of regressors. *Ann. Statist.*, 39:2557–2584.

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

[Castillo, 2012b] Castillo, I. (2012b). A semiparametric Bernstein-von Mises theorem for Gaussian process priors. *Probab. Theory Related Fields*, 152(1-2):53–99.

[Castillo and Rousseau, 2013] Castillo, I. and Rousseau, J. (2013). A General Bernstein–von Mises Theorem in semiparametric models. *ArXiv e-prints.*

[Choudhuri et al., 2004] Choudhuri, N., Ghosal, S., and Roy, A. (2004). Bayesian Estimation of the Spectral Density of a Time Series. *J. American Statist. Assoc.*, 99(468):1050–1060.

[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.

[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.

[Dass and Lee, 2006] Dass, S. and Lee, J. (2006). A note on the consistency of bayes factors for testing point null versus non-parametric alternatives. *J. Statist. Plann. Inference*, 119:143–152.

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

[Ghosal et al., 2000a] Ghosal, S., Ghosh, J. K., and van der Vaart, A. W. (2000a). Convergence rates of posterior distributions. *Ann. Statist.*, 28(2):500–531.

[Ghosal et al., 2008] Ghosal, S., Lember, J., and van der Vaart, A. (2008). Nonparametric Bayesian model selection and averaging. *Electronic journal of statistics*, 2:63–89.

[Ghosal et al., 2000b] Ghosal, S., Sen, A., and van der Vaart, A. W. (2000b). Testing monotonicity of regression. *Ann. Statist.*, 28(4):1054–1082.

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

[Ghosh and Ramamoorthi, 2003] Ghosh, J. and Ramamoorthi, R. (2003). *Bayesian non parametrics.* Springer-Verlag, New York.

[Giné and Nickl, 2012] Giné, E. and Nickl, R. (2012). Rates of contraction for posterior distributions in L^r-metrics, $1 \leq r \leq \infty$. *ArXiv e-prints.*

[Hjort et al., 2009] Hjort, N., Holmes, C., Müller, P., and Walker, S. (2009). *Bayesian Nonparametrics: Principles and Practice.* Cambridge University Press, Cambridge, UK.

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

[Holmes et al., 2012] Holmes, C., Caron, F., Griffin, J., and Stephens, D. (2012). Two -sample Bayesian nonparametric hypothesis testing. Technical report.

[Kruijer and Rousseau, 2012] Kruijer, W. and Rousseau, J. (2012). Bayesian semi-parametric estimation of the long-memory parameter under fexp priors. Technical report.

[Kruijer et al., 2010] Kruijer, W., Rousseau, J., and van der Vaart, A. (2010). Adaptive bayesian density estimation with location-scale mixtures. *Electronic journal of statistics.*

[Leahu, 2011] Leahu, H. (2011). On the bernstein von mises phenomenon in the Gaussian white nois model.

[Petrone et al., 2012] Petrone, S., Rousseau, J., and Scricciolo, C. (2012). Bayes and empirical bayes: do they merge? Technical report.

[R. McVinish, 2009] R. McVinish, J. Rousseau, K. M. (2009). Bayesian goodness-of-fit testing with mixtures of triangular distributions. *Scandinavian Journ. Statist.*, 36:337–354.

[Rivoirard and Rousseau, 2012] Rivoirard, V. and Rousseau, J. (2012). On the Bernstein Von Mises theorem for linear functionals of the density.

[Rousseau, 2007] Rousseau, J. (2007). Approximating interval hypotheses: p-values and Bayes factors. In Bernardo, J. M., Berger, J. O., Dawid, A. P., and Smith, A. F. M., editors, *Bayesian Statistics 8: Proceedings of the Eigth International Meeting.* Oxford University Press.

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

[Rousseau and Choi, 2012] Rousseau, J. and Choi, T. (2012). Bayes factor consistency in non i.i.d models. Technical report.

[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.

[Salomond, 2013] Salomond, J. (2013). Adaptive bayes tests for monotonicity. Technical report.

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

[Shen et al., 2012] Shen, W., Tokdar, S., and Ghosal, S. (2012). Adaptive bayesian multivariate density estimation with dirichlet mixtures. Technical report.

[Tang and Ghosal, 2007] Tang, Y. and Ghosal, S. (2007). Posterior consistency of dirichlet mixtures for estimating a transition density. *Journal of Statistical Planning and Inference*, 137:1711–1726.

[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. *aos*, 37:2655–2675.

[Wong and Shen, 1995] Wong, W. H. and Shen, X. (1995). Probability inequalities for likelihood ratios and convergence rates of sieves mles. *Ann. Statist.*, 23:339–362.

[Wu and Ghosal, 2008] Wu, Y. and Ghosal, S. (2008). Kullback leibler property of kernel mixture priors in bayesian density estimation. *Electronic Journal of Statistics*, 2:298–331.