Posterior concentration rates for empirical Bayes procedures with applications to Dirichlet process mixtures

Sophie Donnet, Vincent Rivoirard, Judith Rousseau, Catia Scricciolo

Submitted to Bernoulli
arXiv: 1406.4406v1

Posterior concentration rates for empirical Bayes procedures with applications to Dirichlet process mixtures

SOPHIE DONNET[1] VINCENT RIVOIRARD[2] JUDITH ROUSSEAU[3] and CATIA SCRICCIOLO[4]

[1] *UMR MIA-Paris, AgroParisTech, INRA, Université Paris-Saclay, 75005, Paris, France.*
E-mail: sophie.donnet@agroparistech.fr

[2] *CEREMADE, Université Paris-Dauphine, Place du Maréchal de Lattre de Tassigny, 75775 Paris Cedex 16, France. E-mail:* rivoirard@ceremade.dauphine.fr

[3] *CEREMADE, Université Paris-Dauphine, Place du Maréchal de Lattre de Tassigny, 75775 Paris Cedex 16, France*
and
CREST-ENSAE, 3 Avenue Pierre Larousse, 92240 Malakoff, France.
E-mail: rousseau@ceremade.dauphine.fr

[4] *Department of Economics, University of Verona, Via Cantarane 24, 37129 Verona, Italy.*
E-mail: catia.scricciolo@univr.it

We provide conditions on the statistical model and the prior probability law to derive contraction rates of posterior distributions corresponding to data-dependent priors in an empirical Bayes approach for selecting prior hyper-parameter values. We aim at giving conditions in the same spirit as those in the seminal article of Ghosal and van der Vaart [23]. We then apply the result to specific statistical settings: density estimation using Dirichlet process mixtures of Gaussian densities with base measure depending on data-driven chosen hyper-parameter values and intensity function estimation of counting processes obeying the Aalen model. In the former setting, we also derive recovery rates for the related inverse problem of density deconvolution. In the latter, a simulation study for inhomogeneous Poisson processes illustrates the results.

Keywords: Aalen model, counting processes, Dirichlet process mixtures, empirical Bayes, posterior contraction rates.

1. Introduction

In a Bayesian approach to statistical inference, the prior distribution should, in principle, be chosen independently of the data; however, it is not always an easy task to elicit the prior hyper-parameter values and a common practice is to replace them by summaries of the data. The prior is then data-dependent and the approach falls under the umbrella of empirical Bayes methods, as opposed to fully Bayes methods. Consider a statistical

model $(\mathbb{P}_\theta^{(n)} : \theta \in \Theta)$ on a sample space $\mathcal{X}^{(n)}$, together with a family of prior distributions $(\pi(\cdot \mid \gamma) : \gamma \in \Gamma)$ on a parameter space Θ. A Bayesian statistician would either set the hyper-parameter γ to a specific value γ_0 or integrate it out using a probability distribution for it in a hierarchical specification of the prior for θ. Both approaches would lead to prior distributions for θ that do not depend on the data. However, it is often the case that knowledge is not *a priori* available to either fix a value for γ or elicit a prior distribution for it, so that a value for γ can be more easily chosen using the data. Throughout the paper, we will denote by $\hat{\gamma}_n$ a data-driven choice for γ. There are many instances in the literature where an empirical Bayes choice for the prior hyper-parameters is performed, sometimes without explicitly mentioning it. Some examples concerning the parametric case can be found in Ahmed and Reid [2], Berger [6] and Casella [8]. Regarding the nonparametric case, Richardson and Green [36] propose a default empirical Bayes approach to deal with parametric and nonparametric mixtures of Gaussian densities; McAuliffe et al. [33] propose another empirical Bayes approach for Dirichlet process mixtures of Gaussian densities, while in Szabó et al. [50] an empirical Bayes procedure is proposed in the context of the Gaussian white noise model to obtain rate adaptive posterior distributions. There are many other instances of empirical Bayes methods in the literature, especially in applied problems.

Our aim is not to claim that empirical Bayes methods are somehow better than fully Bayes methods, rather to provide tools to study frequentist asymptotic properties of empirical Bayes posterior distributions, given their wide use in practice. Very little is known about the asymptotic behavior of such empirical Bayes posterior distributions in a general framework. It is a common belief that if $\hat{\gamma}_n$ asymptotically converges to some value γ^*, then the empirical Bayes posterior distribution associated with $\hat{\gamma}_n$ is eventually "close" to the fully Bayes posterior associated with γ^*. Results have been obtained in specific statistical settings by Clyde and George [10], Cui and George [11] for wavelets or variable selection, by Szabó et al. [48, 49, 50] for the Gaussian white noise model, by Scricciolo [43] for conditional density estimation, by Sniekers and van der Vaart [47], Serra and Krivobokova [44] for Gaussian regression with Gaussian priors. Recently, Petrone et al. [34] have investigated asymptotic properties of empirical Bayes posterior distributions obtaining general conditions for consistency and, in the parametric case, for strong merging between fully Bayes and maximum marginal likelihood empirical Bayes posterior distributions.

In this article, we are interested in studying the frequentist asymptotic behaviour of empirical Bayes posterior distributions in terms of contraction rates. Let $d(\cdot, \cdot)$ be a loss function on Θ, say a pseudo-metric. For $\theta_0 \in \Theta$ and $\epsilon > 0$, let $U_\epsilon := \{\theta : d(\theta, \theta_0) \leq \epsilon\}$ be a neighborhood of θ_0. The empirical Bayes posterior distribution is said to concentrate at θ_0 with rate ϵ_n relative to d, where ϵ_n is a positive sequence converging to zero, if the empirical Bayes posterior probability of the set U_{ϵ_n} tends to one in $\mathbb{P}_{\theta_0}^{(n)}$-probability. In the case of fully Bayes procedures, there has been so far a vast literature on posterior consistency and contraction rates since the seminal articles of Barron et al. [4] and Ghosal et al. [22]. Following ideas of Schwartz [41], Ghosal et al. [22] in the case of independent and identically distributed (iid) observations and Ghosal and van der Vaart [23] in the

case of non-iid observations have developed an elegant and powerful methodology to assess posterior contraction rates which boils down to lower bounding the prior mass of Kullback-Leibler type neighborhoods of $\mathbb{P}_{\theta_0}^{(n)}$ and to constructing exponentially powerful tests for testing $H_0 : \theta = \theta_0$ against $H_1 : \theta \in \{\theta' : d(\theta', \theta_0) > \epsilon_n\}$. However, this approach cannot be immediately taken to deal with posterior distributions corresponding to data-dependent priors. In this article, we develop a similar methodology for assessing posterior contraction rates in the case where the prior distribution depends on the data through a data-driven choice $\hat{\gamma}_n$ for γ.

In Theorem 1, we provide sufficient conditions for deriving contraction rates of empirical Bayes posterior distributions, in the same spirit as those presented in Theorem 1 of Ghosal and van der Vaart [23]. To our knowledge, this is the first result on posterior contraction rates for data-dependent priors which is neither model nor prior specific. The theorem is then applied to nonparametric mixture models. Two relevant applications are considered: Dirichlet process mixtures of Gaussian densities for the problems of density estimation and density deconvolution in Section 3; Dirichlet process mixtures of uniform densities for estimating intensity functions of counting processes obeying the Aalen model in Section 4. Theorem 1 has also been applied to Gaussian process priors and sieve priors in Rousseau and Szabó [38].

Dirichlet process mixtures (DPM) have been introduced by Ferguson [20] and have proved to be a major tool in Bayesian nonparametrics, see for instance Hjort et al. [28]. Rates of convergence for fully Bayes posterior distributions corresponding to DPM of Gaussian densities have been widely studied: they lead to minimax-optimal, possibly up to a logarithmic factor, estimation procedures over a wide collection of density function classes, see Ghosal and van der Vaart [24, 25], Kruijer et al. [31], Scricciolo [42] and Shen et al. [45]. In Section 3.1, we extend existing results to the case of a Gaussian base measure for the Dirichlet process prior with data-driven chosen mean and variance, as advocated for instance in Richardson and Green [36]. Furthermore, in Section 3.2, due to some new inversion inequalities, we get, as a by-product, empirical Bayes posterior recovery rates for the problem of density deconvolution when the error distribution is either ordinary or super-smooth and the mixing density is modeled as a DPM of normal densities with a Gaussian base measure having data-driven selected mean and variance. The problem of Bayesian density deconvolution when the mixing density is modeled as a DPM of Gaussian densities and the error distribution is super-smooth has been recently studied by Sarkar et al. [40].

In Section 4, we focus on Aalen multiplicative intensity models which constitute a major class of counting processes extensively used in the analysis of data arising from various fields like medicine, biology, finance, insurance and social sciences. General statistical and probabilistic literature on such processes is very huge and we refer the reader to Andersen et al. [3], Daley and Vere-Jones [12, 13] and Karr [29] for a good introduction. In the Bayesian nonparametric setting, practical and methodological contributions have been obtained by Lo [32], Adams et al. [1], Cheng and Yuan [9]. Belitser et al. [5] have been the first ones to investigate the frequentist asymptotic behaviour of posterior distributions for intensity functions of inhomogeneous Poisson processes. In Theorem 3, we

derive rates of convergence for empirical Bayes estimation of monotone non-increasing intensity functions of counting processes satisfying the Aalen multiplicative intensity model using DPM of uniform distributions with a truncated gamma base measure whose scale parameter is data-driven chosen. Numerical illustrations are presented in this context in Section 4.3. Final remarks are exposed in Section 5. Proofs of the results in Sections 3 and 4 are deferred to the Supplementary Material.

Notation and context Let $(\mathcal{X}^{(n)}, \mathcal{A}_n, (\mathbb{P}_\theta^{(n)} : \theta \in \Theta))$ be a sequence of statistical experiments, where $\mathcal{X}^{(n)}$ and Θ are Polish spaces endowed with their Borel σ-fields $\mathcal{A}_n$ and $\mathcal{B}$, respectively. Let $X^{(n)} \in \mathcal{X}^{(n)}$ be the observations. We assume that there exists a σ-finite measure $\mu^{(n)}$ on $\mathcal{X}^{(n)}$ dominating all probability measures $\mathbb{P}_\theta^{(n)}$ for $\theta \in \Theta$. For any $\theta \in \Theta$, let $p_\theta^{(n)} := \mathrm{d}\mathbb{P}_\theta^{(n)}/\mathrm{d}\mu^{(n)}$ and $\ell_n(\theta) := \log p_\theta^{(n)}$ be the log-likelihood. We denote by $\mathbb{E}_\theta^{(n)}[\cdot]$ expected values with respect to $\mathbb{P}_\theta^{(n)}$. We consider a family of prior distributions $(\pi(\cdot \mid \gamma) : \gamma \in \Gamma)$ on Θ, where $\Gamma \subseteq \mathbb{R}^d$, $d \geq 1$. We denote by $\pi(\cdot \mid \gamma, X^{(n)})$ the posterior distribution corresponding to the prior law $\pi(\cdot \mid \gamma)$,

$$\pi(B \mid \gamma, X^{(n)}) = \frac{\int_B e^{\ell_n(\theta)} \pi(\mathrm{d}\theta \mid \gamma)}{\int_\Theta e^{\ell_n(\theta)} \pi(\mathrm{d}\theta \mid \gamma)}, \quad B \in \mathcal{B}.$$

Given $\theta_1, \theta_2 \in \Theta$, let

$$\mathrm{KL}(\theta_1; \theta_2) := \mathbb{E}_{\theta_1}^{(n)}[\ell_n(\theta_1) - \ell_n(\theta_2)]$$

be the Kullback-Leibler divergence of $\mathbb{P}_{\theta_2}^{(n)}$ from $\mathbb{P}_{\theta_1}^{(n)}$. Let $V_k(\theta_1; \theta_2)$ be the re-centered k-th absolute moment of the log-likelihood difference associated with θ_1 and θ_2,

$$V_k(\theta_1; \theta_2) := \mathbb{E}_{\theta_1}^{(n)}[|\ell_n(\theta_1) - \ell_n(\theta_2) - \mathbb{E}_{\theta_1}^{(n)}[\ell_n(\theta_1) - \ell_n(\theta_2)]|^k], \quad k \geq 2.$$

Let θ_0 denote the true parameter value. For any sequence of positive real numbers $\epsilon_n \to 0$ such that $n\epsilon_n^2 \to \infty$ and any real $k \geq 2$, let

$$\bar{B}_{k,n} := \{\theta : \mathrm{KL}(\theta_0; \theta) \leq n\epsilon_n^2, \ V_k(\theta_0; \theta) \leq (n\epsilon_n^2)^{k/2}\} \tag{1.1}$$

be the ϵ_n-Kullback-Leibler type neighborhood of θ_0. The role played by these sets will be clarified in Remark 2. Throughout the text, for any set B, constant $\zeta > 0$ and pseudo-metric d, we denote by $D(\zeta, B, d)$ the ζ-packing number of B by d-balls of radius ζ, namely, the maximal number of points in B such that the distance between every pair is at least ζ. The symbols "$\lesssim$" and "$\gtrsim$" are used to indicate inequalities valid up to constants that are fixed throughout.

2. Empirical Bayes posterior contraction rates

The main result of the article is presented in Section 2.1 as Theorem 1: the key ideas are the identification of a set $\mathcal{K}_n$, whose role is discussed in Section 2.2, such that $\hat{\gamma}_n$ takes values in it with probability tending to one, and the construction of a parameter transformation which allows to transfer data-dependence from the prior distribution to the likelihood. Examples of such transformation are given in Section 2.3.

2.1. Main theorem

Let $\hat{\gamma}_n : \mathcal{X}^{(n)} \to \Gamma$ be a measurable function of the observations and let

$$\pi(\cdot \mid \hat{\gamma}_n, X^{(n)}) := \pi(\cdot \mid \gamma, X^{(n)})|_{\gamma=\hat{\gamma}_n}$$

be the associated empirical Bayes posterior distribution. In this section, we present a theorem providing sufficient conditions to obtain posterior contraction rates for empirical Bayes posteriors. Our aim is to give conditions resembling those usually considered in a fully Bayes approach. We first define usual mathematical objects. We assume that, with probability tending to one, $\hat{\gamma}_n$ takes values in a subset $\mathcal{K}_n$ of Γ,

$$\mathbb{P}_{\theta_0}^{(n)} (\hat{\gamma}_n \in \mathcal{K}_n^c) = o(1). \tag{2.1}$$

For any sequence of positive reals $u_n \to 0$, let $N_n(u_n)$ stand for the u_n-covering number of $\mathcal{K}_n$ relative to the Euclidean distance denoted by $\|\cdot\|$, that is, the minimal number of balls of radius u_n needed to cover $\mathcal{K}_n$. For instance, if $\mathcal{K}_n$ is included in a ball of $\mathbb{R}^d$ of radius R_n, then $N_n(u_n) = O((R_n/u_n)^d)$.

We consider posterior contraction rates relative to a loss function $d(\cdot, \cdot)$ on Θ using the following neighborhoods

$$U_{J_1\epsilon_n} := \{\theta \in \Theta : d(\theta, \theta_0) \leq J_1\epsilon_n\},$$

with J_1 a positive constant. We assume that $d(\cdot, \cdot)$ is a pseudo-metric, although this assumption can be relaxed, see Remark 3. For every integer $j \in \mathbb{N}$, we define

$$S_{n,j} := \{\theta \in \Theta : d(\theta, \theta_0) \in (j\epsilon_n, (j+1)\epsilon_n]\}.$$

In order to obtain posterior contraction rates with data-dependent priors, we express the impact of $\hat{\gamma}_n$ on the prior distribution as follows: for all $\gamma, \gamma' \in \Gamma$, we construct a measurable transformation

$$\psi_{\gamma,\gamma'} : \Theta \to \Theta$$

such that, if $\theta \sim \pi(\cdot \mid \gamma)$, then $\psi_{\gamma,\gamma'}(\theta) \sim \pi(\cdot \mid \gamma')$. Let $e_n(\cdot, \cdot)$ be another pseudo-metric on Θ.

We consider the following assumptions.

[A1] There exists a sequence of positive reals $u_n \to 0$ such that

$$\log N_n(u_n) = o(n\epsilon_n^2). \tag{2.2}$$

There exists a sequence of sets $\tilde{B}_n \subseteq \Theta$ such that, for some constant $C_1 > 0$,

$$\sup_{\gamma \in \mathcal{K}_n} \sup_{\theta \in \tilde{B}_n} \mathbb{P}_{\theta_0}^{(n)} \left(\inf_{\gamma' :\ \|\gamma'-\gamma\| \leq u_n} \ell_n(\psi_{\gamma,\gamma'}(\theta)) - \ell_n(\theta_0) < -C_1 n\epsilon_n^2 \right) = o(N_n(u_n)^{-1}). \tag{2.3}$$

[A2] For every $\gamma \in \mathcal{K}_n$, there exists a sequence of sets $\Theta_n(\gamma) \subseteq \Theta$ so that

$$\sup_{\gamma \in \mathcal{K}_n} \int_{\Theta \setminus \Theta_n(\gamma)} Q_{\theta,\gamma}^{(n)}(\mathcal{X}^{(n)}) \frac{\pi(\mathrm{d}\theta \mid \gamma)}{\pi(\tilde{B}_n \mid \gamma)} = o(N_n(u_n)^{-1} e^{-C_2 n\epsilon_n^2}) \tag{2.4}$$

for some constant $C_2 > C_1$, where $Q_{\theta,\gamma}^{(n)}$ is the measure having density $q_{\theta,\gamma}^{(n)}$ with respect to $\mu^{(n)}$:

$$q_{\theta,\gamma}^{(n)} := \frac{\mathrm{d}Q_{\theta,\gamma}^{(n)}}{\mathrm{d}\mu^{(n)}} := \sup_{\gamma' : \|\gamma' - \gamma\| \leq u_n} e^{\ell_n(\psi_{\gamma,\gamma'}(\theta))}.$$

Also, there exist constants ζ, $K > 0$ such that

- for all j large enough,

$$\sup_{\gamma \in \mathcal{K}_n} \frac{\pi(S_{n,j} \cap \Theta_n(\gamma) \mid \gamma)}{\pi(\tilde{B}_n \mid \gamma)} \leq e^{Knj^2\epsilon_n^2/2}, \tag{2.5}$$

- for all $\epsilon > 0$, $\gamma \in \mathcal{K}_n$ and $\theta \in \Theta_n(\gamma)$ with $d(\theta, \theta_0) > \epsilon$, there exist tests $\phi_n(\theta)$ satisfying

$$\mathbb{E}_{\theta_0}^{(n)}[\phi_n(\theta)] \leq e^{-Kn\epsilon^2} \quad \text{and} \quad \sup_{\theta' : e_n(\theta', \theta) \leq \zeta\epsilon} \int_{\mathcal{X}^{(n)}} [1 - \phi_n(\theta)] \mathrm{d}Q_{\theta',\gamma}^{(n)} \leq e^{-Kn\epsilon^2}, \tag{2.6}$$

- for all j large enough,

$$\log D(\zeta j\epsilon_n, S_{n,j} \cap \Theta_n(\gamma), e_n) \leq K(j+1)^2 n\epsilon_n^2/2, \tag{2.7}$$

- there exists a constant $M > 0$ such that for all $\gamma \in \mathcal{K}_n$,

$$\sup_{\gamma' : \|\gamma' - \gamma\| \leq u_n} \sup_{\theta \in \Theta_n(\gamma)} d(\psi_{\gamma,\gamma'}(\theta), \theta) \leq M\epsilon_n. \tag{2.8}$$

We can now state the main theorem.

Theorem 1. *Let $\theta_0 \in \Theta$. Assume that $\hat{\gamma}_n$ satisfies condition (2.1) and that conditions* [A1] *and* [A2] *are verified for a sequence of positive reals $\epsilon_n \to 0$ such that $n\epsilon_n^2 \to \infty$. Then, for a constant $J_1 > 0$ large enough,*

$$\mathbb{E}_{\theta_0}^{(n)}[\pi(U_{J_1\epsilon_n}^c \mid \hat{\gamma}_n, X^{(n)})] = o(1),$$

where $U_{J_1\epsilon_n}^c$ is the complement of $U_{J_1\epsilon_n}$ in Θ.

Remark 1. We can replace (2.6) and (2.7) with a condition involving the existence of a global test ϕ_n over $S_{n,j}$ satisfying requirements similar to those of equation (2.7) in Ghosal and van der Vaart [23] without modifying the conclusion:

$$\mathbb{E}_{\theta_0}^{(n)}[\phi_n] = o(N_n(u_n)^{-1}) \quad \text{and} \quad \sup_{\gamma \in \mathcal{K}_n} \sup_{\theta \in S_{n,j}} \int_{\mathcal{X}^{(n)}} (1 - \phi_n) \mathrm{d}Q_{\theta,\gamma}^{(n)} \leq e^{-Knj^2\epsilon_n^2}.$$

Note also that, when the loss function $d(\cdot, \cdot)$ is not bounded, it is often the case that getting exponential control on the error rates in the form $e^{-Kn\epsilon_n^2}$ or $e^{-Knj^2\epsilon_n^2}$ is not possible for large values of j. It is then enough to consider a modification $\tilde{d}(\cdot, \cdot)$ of the loss function which affects only the values of θ for which $d(\theta, \theta_0)$ is large and to verify (2.6) and (2.7) for $\tilde{d}(\theta, \theta_0)$ by defining $S_{n,j}$ and the covering number $D(\cdot)$ with respect to $\tilde{d}(\cdot, \cdot)$. See the proof of Theorem 3 as an illustration of this remark.

Remark 2. The requirements of assumption [A2] are similar to those proposed by Ghosal and van der Vaart [23] for deriving contraction rates for fully Bayes posterior distributions, see, for instance, their Theorem 1 and its proof. We need to strengthen some conditions to take into account that we only know that $\hat{\gamma}_n$ lies in a compact set $\mathcal{K}_n$ with high probability by replacing the likelihood $p_\theta^{(n)}$ with $q_{\theta,\gamma}^{(n)}$. Note that in the definition of $q_{\theta,\gamma}^{(n)}$ we can replace the centering point γ of a ball with radius u_n with any fixed point in the ball. This is used, for instance, in the context of DPM of uniform distributions in Section 4. In the applications of Theorem 1, condition [A1] is typically verified by resorting to Lemma 10 in Ghosal and van der Vaart [23] and by considering a set $\tilde{B}_n \subseteq \bar{B}_{k,n}$, with $\bar{B}_{k,n}$ as defined in (1.1). The only difference with the general theorems of Ghosal and van der Vaart [23] lies in the control of the log-likelihood difference $\ell_n(\psi_{\gamma,\gamma'}(\theta)) - \ell_n(\theta_0)$ when $\|\gamma' - \gamma\| \leq u_n$. We thus need that $N_n(u_n) = o((n\epsilon_n^2)^{k/2})$. In nonparametric cases where $n\epsilon_n^2$ is a power of n, the sequence u_n can be chosen very small, as long as k can be chosen large enough, so that controlling the above difference uniformly is not such a drastic condition. In parametric models where at best $n\epsilon_n^2$ is a power of $\log n$, this becomes more involved and u_n needs to be large or $\mathcal{K}_n$ needs to be small enough so that $N_n(u_n)$ can be chosen of the order $O(1)$. In parametric models, it is typically easier to use a more direct control of the ratio $\pi(\theta \mid \gamma)/\pi(\theta \mid \gamma')$ of the prior densities with respect to a common dominating measure. In nonparametric models, this is usually not possible since in most cases no such dominating measure exists.

Remark 3. In Theorem 1, $d(\cdot, \cdot)$ can be replaced by any positive loss function. In this case, condition (2.8) needs to be rephrased: for every $J_2 > 0$, there exists $J_1 > 0$ such that, for all $\gamma, \gamma' \in \mathcal{K}_n$ with $\|\gamma - \gamma'\| \leq u_n$, for every $\theta \in \Theta_n(\gamma)$,

$$d(\psi_{\gamma,\gamma'}(\theta), \theta_0) > J_1\epsilon_n \quad \text{implies} \quad d(\theta, \theta_0) > J_2\epsilon_n. \tag{2.9}$$

2.2. On the role of the set $\mathcal{K}_n$

To prove Theorem 1, it is enough to show that the posterior contraction rate of the empirical Bayes posterior associated with $\hat{\gamma}_n \in \mathcal{K}_n$ is bounded from above by the worst contraction rate over the class of posterior distributions corresponding to the family of priors $(\pi(\cdot \mid \gamma) : \gamma \in \mathcal{K}_n)$:

$$\pi(U_{J_1\epsilon_n}^c \mid \hat{\gamma}_n, X^{(n)}) \leq \sup_{\gamma \in \mathcal{K}_n} \pi(U_{J_1\epsilon_n}^c \mid \gamma, X^{(n)}).$$

In other terms, the impact of $\hat{\gamma}_n$ is summarized through $\mathcal{K}_n$. Hence, it is important to preliminarily figure out which set $\mathcal{K}_n$ could be. In the examples developed in Sections 3 and 4, the hyper-parameter γ has no impact on the posterior contraction rate, at least on a large subset of Γ, so that, as long as $\hat{\gamma}_n$ stays in this range, the posterior contraction rate of the empirical Bayes posterior is the same as that of any prior associated with a fixed γ. In those cases where γ has an influence on posterior contraction rates, determining $\mathcal{K}_n$ is crucial. For instance, Rousseau and Szabó [38] study the asymptotic behaviour of the maximum marginal likelihood estimator and characterize the set $\mathcal{K}_n$; they then apply Theorem 1 to derive contraction rates for certain empirical Bayes posterior distributions. Suppose that the posterior $\pi(\cdot \mid \gamma, X^{(n)})$ converges at rate $\epsilon_n(\gamma) = (n/\log n)^{-\alpha(\gamma)}$, where the mapping $\gamma \mapsto \alpha(\gamma)$ is Lipschitzian, and that $\hat{\gamma}_n$ concentrates on an oracle set $\mathcal{K}_n = \{\gamma : \epsilon_n(\gamma) \leq M_n \epsilon_n^*\}$, where $\epsilon_n^* = \inf_\gamma \epsilon_n(\gamma)$ and M_n is some sequence such that $M_n \to \infty$, then, under the conditions of Theorem 1, we can deduce that the empirical Bayes posterior contraction rate is bounded above by $M_n \epsilon_n^*$. Proving that the empirical Bayes posterior distribution has optimal posterior contraction rate then boils down to proving that $\hat{\gamma}_n$ converges to the oracle set $\mathcal{K}_n$. This is what happens in the context considered by Szabó et al. [50], as explained in Rousseau and Szabó [38].

2.3. On the parameter transformation $\psi_{\gamma,\gamma'}$

A key idea of the proof of Theorem 1 is the construction of a parameter transformation $\psi_{\gamma,\gamma'}$ which allows to transfer data-dependence from the prior to the likelihood as in Petrone et al. [34]. The transformation $\psi_{\gamma,\gamma'}$ can be easily identified in a number of cases. Note that this transformation only depends on the family of prior distributions and not on the sampling model.

For Gaussian process priors in the form

$$\theta_i \overset{\text{ind}}{\sim} \mathcal{N}(0, \tau^2(1+i)^{-(2\alpha+1)}), \quad i \in \mathbb{N},$$

the following ones

$$\psi_{\tau,\tau'}(\theta_i) = \frac{\tau'}{\tau}\theta_i, \quad i \in \mathbb{N},$$
$$\psi_{\alpha,\alpha'}(\theta_i) = (1+i)^{-(\alpha'-\alpha)}\theta_i, \quad \alpha' \geq \alpha, \quad i \in \mathbb{N},$$

are possible transformations, see Rousseau and Szabó [38]. Similar ideas can be used for priors based on splines with independent coefficients.

The transformation $\psi_{\gamma,\gamma'}$ can be constructed also for Polya tree priors based on a specific family of partitions $(\mathcal{T}_k)_{k\geq 1}$ with parameters $\alpha_\epsilon = ck^2$ when $\epsilon \in \{0,1\}^k$. When $\gamma = c$,

$$\psi_{c,c'}(\theta_\epsilon) = G^{-1}_{c'k^2,c'k^2}(G_{ck^2,ck^2}(\theta_\epsilon)), \quad \forall\, \epsilon \in \{0,1\}^k, \quad \forall\, k \geq 1,$$

where $G_{a,b}$ denotes the cumulative distribution function (cdf) of a Beta random variable with parameters (a, b).

In Sections 3 and 4, we apply Theorem 1 to two types of Dirichlet process mixture models: DPM of Gaussian distributions used to model smooth densities and DPM of uniform distributions used to model monotone non-increasing intensity functions in the context of Aalen point processes. In the case of nonparametric mixture models, there exists a general construction of the transformation $\psi_{\gamma,\gamma'}$. Consider a mixture model in the form

$$f(\cdot) = \sum_{j=1}^{K} p_j h_{\theta_j}(\cdot), \quad K \sim \pi_K, \tag{2.10}$$

where, conditionally on K, $p = (p_j)_{j=1}^K \sim \pi_p$ and $\theta_1, \ldots, \theta_K$ are iid with cumulative distribution function G_γ. Dirichlet process mixtures correspond to $\pi_K = \delta_{(+\infty)}$ and to π_p equal to the Griffiths-Engen-McCloskey (GEM) distribution obtained from the stick-breaking construction of the mixing weights, cf. Ghosh and Ramamoorthi [26]. Models in the form (2.10) also cover priors on curves if the $(p_j)_{j=1}^K$ are not restricted to the simplex. Denote by $\pi(\cdot \mid \gamma)$ the prior probability on f induced by (2.10). For all $\gamma, \gamma' \in \Gamma$, if f is represented as in (2.10) and is distributed according to $\pi(\cdot \mid \gamma)$, then

$$f'(\cdot) = \sum_{j=1}^{K} p_j h_{\theta'_j}(\cdot), \quad \text{with} \quad \theta'_j = G_{\gamma'}^{-1}(G_\gamma(\theta_j)),$$

is distributed according to $\pi(\cdot \mid \gamma')$, where $G_{\gamma'}^{-1}$ denotes the generalized inverse of the cdf $G_{\gamma'}$. If $\gamma = M$ is the mass hyper-parameter of a Dirichlet process (DP), a possible transformation from a DP with mass M to a DP with mass M' is through the stick-breaking representation of the weights:

$$\psi_{M,M'}(V_j) = G_{1,M'}^{-1}(G_{1,M}(V_j)), \quad \text{where} \quad p_j = V_j \prod_{i<j}(1 - V_i), \quad j \geq 1.$$

We now give the proof of Theorem 1.

2.4. Proof of Theorem 1

Because $\mathbb{P}_{\theta_0}^{(n)}(\hat{\gamma}_n \in \mathcal{K}_n^c) = o(1)$ by assumption, we have

$$\mathbb{E}_{\theta_0}^{(n)}[\pi(U_{J_1\epsilon_n}^c \mid \hat{\gamma}_n, X^{(n)})] \leq \mathbb{E}_{\theta_0}^{(n)}\left[\sup_{\gamma \in \mathcal{K}_n} \pi(U_{J_1\epsilon_n}^c \mid \gamma, X^{(n)})\right] + o(1).$$

The proof then essentially boils down to controlling $\mathbb{E}_{\theta_0}^{(n)}[\sup_{\gamma\in\mathcal{K}_n} \pi(U_{J_1\epsilon_n}^c \mid \gamma, X^{(n)})]$. We split $\mathcal{K}_n$ into $N_n(u_n)$ balls of radius u_n and denote their centers by $(\gamma_i)_{i=1,\ldots,N_n(u_n)}$. We thus have

$$\mathbb{E}_{\theta_0}^{(n)}[\pi(U_{J_1\epsilon_n}^c \mid \hat{\gamma}_n, X^{(n)}) 1_{\mathcal{K}_n}(\hat{\gamma}_n)] \leq N_n(u_n) \max_i \mathbb{E}_{\theta_0}^{(n)}\left[\rho_n(\gamma_i)\right],$$

where the index i ranges from 1 to $N_n(u_n)$ and

$$\begin{aligned}\rho_n(\gamma_i) &:= \sup_{\gamma:\,\|\gamma-\gamma_i\|\le u_n} \pi(U^c_{J_1\epsilon_n} \mid \gamma,\, X^{(n)}) \\ &= \sup_{\gamma:\,\|\gamma-\gamma_i\|\le u_n} \frac{\int_{U^c_{J_1\epsilon_n}} e^{\ell_n(\theta)-\ell_n(\theta_0)}\pi(\mathrm{d}\theta \mid \gamma)}{\int_\Theta e^{\ell_n(\theta)-\ell_n(\theta_0)}\pi(\mathrm{d}\theta \mid \gamma)} \\ &= \sup_{\gamma:\,\|\gamma-\gamma_i\|\le u_n} \frac{\int_{\psi^{-1}_{\gamma_i,\gamma}(U^c_{J_1\epsilon_n})} e^{\ell_n(\psi_{\gamma_i,\gamma}(\theta))-\ell_n(\theta_0)}\pi(\mathrm{d}\theta \mid \gamma_i)}{\int_\Theta e^{\ell_n(\psi_{\gamma_i,\gamma}(\theta))-\ell_n(\theta_0)}\pi(\mathrm{d}\theta \mid \gamma_i)}.\end{aligned}$$

So, it is enough to prove that $\max_i \mathbb{E}^{(n)}_{\theta_0}[\rho_n(\gamma_i)] = o(N_n(u_n)^{-1})$. We mimic the proof of Lemma 9 of Ghosal and van der Vaart [23]. Let i be fixed. For every j large enough, by condition (2.7), there exist $L_{j,n} \le \exp(K(j+1)^2 n\epsilon_n^2/2)$ balls of centers $\theta_{j,1}, \ldots, \theta_{j,L_{j,n}}$, with radius $\zeta j\epsilon_n$ relative to the e_n-distance, that cover $S_{n,j} \cap \Theta_n(\gamma_i)$. We consider tests $\phi_n(\theta_{j,\ell})$, $\ell = 1, \ldots, L_{j,n}$, satisfying (2.6) with $\epsilon = j\epsilon_n$. By setting

$$\phi_n := \max_{j\ge J_1} \max_{\ell\in\{1,\ldots,L_{j,n}\}} \phi_n(\theta_{j,\ell}),$$

by virtue of conditions (2.6), applied with $\gamma = \gamma_i$, and (2.2), we obtain that, for any $K' < K$,

$$\mathbb{E}^{(n)}_{\theta_0}[\phi_n] \le \sum_{j\ge J_1} L_{j,n} e^{-Kj^2 n\epsilon_n^2} = O(e^{-K' J_1^2 n\epsilon_n^2/2}) = o(N_n(u_n)^{-1}).$$

Moreover, for any $j \ge J_1$, any $\theta \in S_{n,j} \cap \Theta_n(\gamma_i)$ and any $i = 1, \ldots, N_n(u_n)$,

$$\int_{\mathcal{X}^{(n)}} (1-\phi_n)\mathrm{d}Q^{(n)}_{\theta,\gamma_i} \le e^{-Kj^2 n\epsilon_n^2}. \tag{2.11}$$

Since for all i we have $\rho_n(\gamma_i) \le 1$, it follows that

$$\mathbb{E}^{(n)}_{\theta_0}[\rho_n(\gamma_i)] < \mathbb{E}^{(n)}_{\theta_0}[\phi_n] + \mathbb{P}^{(n)}_{\theta_0}(A^c_{n,i}) + \frac{e^{C_2 n\epsilon_n^2}}{\pi(\tilde{B}_n \mid \gamma_i)} C_{n,i}, \tag{2.12}$$

with

$$A_{n,i} = \left\{ \inf_{\gamma:\,\|\gamma-\gamma_i\|\le u_n} \int_\Theta e^{\ell_n(\psi_{\gamma_i,\gamma}(\theta))-\ell_n(\theta_0)}\pi(\mathrm{d}\theta \mid \gamma_i) > e^{-C_2 n\epsilon_n^2}\pi(\tilde{B}_n \mid \gamma_i) \right\} \tag{2.13}$$

and

$$C_{n,i} = \mathbb{E}^{(n)}_{\theta_0}\left[(1-\phi_n) \sup_{\gamma:\,\|\gamma-\gamma_i\|\le u_n} \int_{\psi^{-1}_{\gamma_i,\gamma}(U^c_{J_1\epsilon_n})} e^{\ell_n(\psi_{\gamma_i,\gamma}(\theta))-\ell_n(\theta_0)}\pi(\mathrm{d}\theta \mid \gamma_i) \right].$$

We now study the last two terms in (2.12). Since

$$\begin{aligned} e^{C_1 n\epsilon_n^2} \inf_{\gamma:\,\|\gamma-\gamma_i\|\le u_n} & e^{\ell_n(\psi_{\gamma_i,\gamma}(\theta))-\ell_n(\theta_0)} \\ &\ge \mathbf{1}_{\{\inf_{\gamma:\,\|\gamma-\gamma_i\|\le u_n} \exp(\ell_n(\psi_{\gamma_i,\gamma}(\theta))-\ell_n(\theta_0)) \ge e^{-C_1 n\epsilon_n^2}\}}, \end{aligned}$$

we have

$$\begin{aligned}\mathbb{P}_{\theta_0}^{(n)}(A_{n,i}^c) &\le \mathbb{P}_{\theta_0}^{(n)}\left(\int_{\tilde{B}_n}\inf_{\gamma:\,\|\gamma-\gamma_i\|\le u_n} e^{\ell_n(\psi_{\gamma_i,\gamma}(\theta))-\ell_n(\theta_0)}\frac{\pi(\mathrm{d}\theta\mid\gamma_i)}{\pi(\tilde{B}_n\mid\gamma_i)} \le e^{-C_2 n\epsilon_n^2}\right)\\ &< (1-e^{-(C_2-C_1)n\epsilon_n^2})^{-1}\\ &\quad\times\int_{\tilde{B}_n}\mathbb{P}_{\theta_0}^{(n)}\left(\inf_{\gamma:\,\|\gamma-\gamma_i\|\le u_n}\ell_n(\psi_{\gamma_i,\gamma}(\theta))-\ell_n(\theta_0) < -C_1 n\epsilon_n^2\right)\frac{\pi(\mathrm{d}\theta\mid\gamma_i)}{\pi(\tilde{B}_n\mid\gamma_i)}.\end{aligned}$$

Then, by condition (2.3), $\mathbb{P}_{\theta_0}^{(n)}(A_{n,i}^c) = o(N_n(u_n)^{-1})$. For γ such that $\|\gamma-\gamma_i\| \le u_n$, under condition (2.8), for any $\theta \in \Theta_n(\gamma_i)$,

$$d(\psi_{\gamma_i,\gamma}(\theta),\,\theta_0) \le d(\psi_{\gamma_i,\gamma}(\theta),\,\theta) + d(\theta,\,\theta_0) \le M\epsilon_n + d(\theta,\,\theta_0),$$

then, for every $J_2 > 0$, choosing $J_1 > J_2 + M$ we have

$$\psi_{\gamma_i,\gamma}^{-1}(U_{J_1\epsilon_n}^c) \subset \big(U_{J_2\epsilon_n}^c \cup \Theta_n^c(\gamma_i)\big).$$

Note that this corresponds to (2.9). This leads to

$$\begin{aligned}C_{n,i} &\le \mathbb{E}_{\theta_0}^{(n)}\left[(1-\phi_n)\int_{\psi_{\gamma_i,\gamma}^{-1}(U_{J_1\epsilon_n}^c)}\sup_{\gamma:\,\|\gamma-\gamma_i\|\le u_n} e^{\ell_n(\psi_{\gamma_i,\gamma}(\theta))-\ell_n(\theta_0)}\pi(\mathrm{d}\theta\mid\gamma_i)\right]\\ &\le \int_{U_{J_2\epsilon_n}^c\cup\Theta_n^c(\gamma_i)}\int_{\mathcal{X}^{(n)}}(1-\phi_n)\mathrm{d}Q_{\theta,\gamma_i}^{(n)}\pi(\mathrm{d}\theta\mid\gamma_i)\\ &\le \int_{\Theta_n^c(\gamma_i)}Q_{\theta,\gamma_i}^{(n)}(\mathcal{X}^{(n)})\pi(\mathrm{d}\theta\mid\gamma_i) + \sum_{j\ge J_2}\int_{S_{n,j}\cap\Theta_n(\gamma_i)}\int_{\mathcal{X}^{(n)}}(1-\phi_n)\mathrm{d}Q_{\theta,\gamma_i}^{(n)}\pi(\mathrm{d}\theta\mid\gamma_i).\end{aligned}$$

Using (2.4), (2.5) and (2.11),

$$\begin{aligned}C_{n,i} &\le \sum_{j\ge J_2} e^{-Kj^2 n\epsilon_n^2}\pi(S_{n,j}\cap\Theta_n(\gamma_i)\mid\gamma_i) + o(N_n(u_n)^{-1}e^{-C_2 n\epsilon_n^2}\pi(\tilde{B}_n\mid\gamma_i))\\ &\le \sum_{j\ge J_2} e^{-Kj^2 n\epsilon_n^2/2}\pi(\tilde{B}_n\mid\gamma_i) + o(N_n(u_n)^{-1}e^{-C_2 n\epsilon_n^2}\pi(\tilde{B}_n\mid\gamma_i)),\end{aligned}$$

whence $C_{n,i} = o(N_n(u_n)^{-1}e^{-C_2 n\epsilon_n^2}\pi(\tilde{B}_n\mid\gamma_i))$. Consequently,

$$\max_{i=1,\ldots,N_n(u_n)} e^{C_2 n\epsilon_n^2}C_{n,i}/\pi(\tilde{B}_n\mid\gamma_i) = o(N_n(u_n)^{-1}),$$

which concludes the proof of Theorem 1. □

We now consider two applications of Theorem 1 to DPM models. They present different features: the first one deals with density estimation and considers DPM with smooth (Gaussian) kernels, the second one deals with intensity estimation in Aalen point processes and considers DPM with irregular (uniform) kernels. Estimating Aalen intensity functions has strong connections with density estimation, but it is not identical: as far as the control of data-dependence of the prior is concerned, the main difference lies in the different regularity of the kernels.

3. Adaptive rates for empirical Bayes DPM of Gaussian densities

Let $X^{(n)} = (X_1, \ldots, X_n)$ be n iid observations from a Lebesgue density p_0 on $\mathbb{R}$. Consider the following prior distribution on the class of Lebesgue densities p on the real line:

$$\begin{aligned} p(\cdot) = p_{F,\sigma}(\cdot) &:= \int_{-\infty}^{\infty} \phi_\sigma(\cdot - \theta) \mathrm{d}F(\theta), \\ F \sim \mathrm{DP}(\alpha_{\mathbb{R}} \mathcal{N}(m, s^2)) \quad &\text{independent of} \quad \sigma \sim \mathrm{IG}(\nu_1, \nu_2), \quad \nu_1, \nu_2 > 0, \end{aligned} \tag{3.1}$$

where $\alpha_{\mathbb{R}}$ is a finite positive constant, $\phi_\sigma(\cdot) := \sigma^{-1}\phi(\cdot/\sigma)$, with $\phi(\cdot)$ the density of a standard Gaussian distribution, and $\mathcal{N}(m, s^2)$ denotes a Gaussian distribution with mean m and variance s^2. Set $\gamma = (m, s^2) \in \Gamma \subseteq \mathbb{R} \times \mathbb{R}_+^*$, where $\mathbb{R}_+^*$ denotes the set of strictly positive real numbers, let $\hat{\gamma}_n : \mathbb{R}^n \to \Gamma$ be a measurable function of the observations. Typical choices are $\hat{\gamma}_n = (\bar{X}_n, S_n^2)$, with the sample mean $\bar{X}_n = \sum_{i=1}^n X_i/n$ and the sample variance $S_n^2 = \sum_{i=1}^n (X_i - \bar{X}_n)^2/n$, or $\hat{\gamma}_n = (\bar{X}_n, R_n)$, with the range $R_n = \max_{1\le i\le n} X_i - \min_{1\le i\le n} X_i$, as in Richardson and Green [36]. Let $\mathcal{K}_n \subset \mathbb{R} \times \mathbb{R}_+^*$ be a compact set, independent of the data $X^{(n)}$, such that

$$\mathbb{P}_{p_0}^{(n)} (\hat{\gamma}_n \in \mathcal{K}_n) = 1 + o(1), \tag{3.2}$$

where p_0 denotes the true sampling density. Throughout this section, we assume that p_0 satisfies the following tail condition:

$$p_0(x) \lesssim e^{-c_0|x|^\tau} \quad \text{for } |x| \text{ large enough}, \tag{3.3}$$

with finite constants $c_0, \tau > 0$. Let $\mathbb{E}_{p_0}[X_1] = m_0 \in \mathbb{R}$ and $\mathbb{V}\mathrm{ar}_{p_0}[X_1] = \tau_0^2 \in \mathbb{R}_+^*$. If $\hat{\gamma}_n = (\bar{X}_n, S_n^2)$, then condition (3.2) is satisfied for $\mathcal{K}_n = [m_0 - (\log n)/\sqrt{n}, m_0 + (\log n)/\sqrt{n}] \times [\tau_0^2 - (\log n)/\sqrt{n}, \tau_0^2 + (\log n)/\sqrt{n}]$, while, if $\hat{\gamma}_n = (\bar{X}_n, R_n)$, then $\mathcal{K}_n = [m_0 - (\log n)/\sqrt{n}, m_0 + (\log n)/\sqrt{n}] \times [r_n, 2(2c_0^{-1} \log n)^{1/\tau}]$, with a sequence $r_n \downarrow 0$.

3.1. Empirical Bayes density estimation

Mixtures of Gaussian densities have been extensively studied and used in the Bayesian nonparametric literature. Posterior contraction rates have been first investigated by Ghosal and van der Vaart [24, 25]. Subsequently, following an idea of Rousseau [37], Kruijer et al. [31] have shown that nonparametric location mixtures of Gaussian densities lead to adaptive posterior contraction rates over the full scale of locally Hölder log-densities on $\mathbb{R}$. This result has been extended to super-smooth densities by Scricciolo [42] and to the multivariate case by Shen et al. [45]. The key idea is that, for an ordinary smooth density p_0 with regularity level $\beta > 0$, given $\sigma > 0$ small enough, there exists a finite mixing distribution F^*, with at most $N_\sigma = O(\sigma^{-1}|\log\sigma|^{\rho_2})$ support points in $[-a_\sigma, a_\sigma]$, where $a_\sigma = O(|\log\sigma|^{1/\tau})$, such that the corresponding Gaussian mixture density $p_{F^*,\sigma}$ satisfies

$$\mathbb{P}_{p_0} \log(p_0/p_{F^*,\sigma}) \lesssim \sigma^{2\beta}$$

and (3.4)

$$\mathbb{P}_{p_0}|\log(p_0/p_{F^*,\sigma}) - \mathbb{P}_{p_0}\log(p_0/p_{F^*,\sigma})|^k \lesssim \sigma^{k\beta}, \quad k \geq 2,$$

where we have used the notation $\mathbb{P}_{p_0} f$ to abbreviate $\int f \mathrm{d}\mathbb{P}_{p_0}$; see, for instance, Lemma 4 in Kruijer et al. [31]. In all of the above-mentioned articles, only the case where $k = 2$ has been considered for the second inequality in (3.4), but the extension to any $k > 2$ is straightforward. The regularity assumptions considered in Kruijer et al. [31], Scricciolo [42] and Shen et al. [45] to meet (3.4) are slightly different. For instance, Kruijer et al. [31] assume that $\log p_0$ satisfies some locally Hölder conditions, while Shen et al. [45] consider Hölder-type conditions on p_0 and Scricciolo [42] Sobolev-type assumptions. To avoid taking into account all these special cases, in the ordinary smooth case, we state (3.4) as an assumption. Regarding the super-smooth case, defined for any $\alpha \in (0, 1]$ and any pair of densities p and p_0, the ρ_α-divergence of p from p_0 as

$$\rho_\alpha(p_0; p) := \alpha^{-1}\mathbb{P}_{p_0}[(p_0/p)^\alpha - 1],$$

a counter-part of requirement (3.4) is the following one:

$$\text{for some fixed } \alpha \in (0, 1], \quad \rho_\alpha(p_0; p_{F^*,\sigma}) \lesssim e^{-c_\alpha(1/\sigma)^r}, \tag{3.5}$$

where c_α is a positive constant possibly depending on α and F^* is a distribution on $[-a_\sigma, a_\sigma]$, with $a_\sigma = O(\sigma^{-r/(\tau\wedge 2)})$, having at most $N_\sigma = O((a_\sigma/\sigma)^2)$ support points. Because for any pair of densities p and p_0,

$$\mathbb{P}_{p_0}\log(p_0/p) = \lim_{\beta\to 0^+}\rho_\beta(p_0; p) \leq \rho_\alpha(p_0; p) \quad \text{for every } \alpha \in (0, 1],$$

inequality (3.5) is stronger than the one on the first line of (3.4) and allows to derive an almost sure lower bound on the denominator of the ratio defining the empirical Bayes posterior probability of the set $U^c_{J_1\epsilon_n}$, see Lemma 2 of Shen and Wasserman [46]. Following the trail of Lemma 8 in Scricciolo [42], it can be proved that inequality (3.5) holds for any density p_0 satisfying the monotonicity and tail conditions (b) and (c), respectively, of Section 4.2 in Scricciolo [42], together with the following integrability condition

$$\int_{-\infty}^{\infty}|\hat{p}_0(t)|^2 e^{2(\rho|t|)^r}\mathrm{d}t \leq 2\pi L^2 \quad \text{for some } r \in [1, 2] \text{ and } \rho, L > 0, \tag{3.6}$$

where $\hat{p}_0(t) = \int_{-\infty}^{\infty} e^{itx}p_0(x)\mathrm{d}x$, $t \in \mathbb{R}$, is the characteristic function of p_0. Densities satisfying requirement (3.6) form a large class including relevant statistical examples, like the Gaussian distribution which corresponds to $r = 2$, the Cauchy distribution which corresponds to $r = 1$; symmetric stable laws with $1 \leq r \leq 2$, the Student's-t distribution, distributions with characteristic functions vanishing outside a compact set as well as their mixtures and convolutions. We then have the following theorem, where the pseudo-metric d defining the ball $U_{J_1\epsilon_n}$ centered at p_0, with radius $J_1\epsilon_n$, can equivalently be the Hellinger or the $\mathbb{L}_1$-distance.

Theorem 2. *Consider a prior distribution of the form (3.1), with a data-driven choice $\hat{\gamma}_n$ for γ satisfying condition (3.2), where $\mathcal{K}_n \subseteq [m_1, m_2] \times [a_1, a_2(\log n)^{b_1}]$ for some constants $m_1, m_2 \in \mathbb{R}$, $a_1, a_2 > 0$ and $b_1 \geq 0$. Suppose that p_0 satisfies the tail condition (3.3). Consider either one of the following cases.*

(*i*) ***Ordinary smooth case**. Suppose that the exponent τ appearing in (3.3) is such that $\tau \geq 1$. Assume that, for $\beta > 0$, requirement (3.4) holds with $k > 8(\beta + 1)$. Let*

$$\epsilon_n = n^{-\beta/(2\beta+1)}(\log n)^{a_3}, \quad \text{for some constant } a_3 \geq 1 + [\tau(2 + 1/\beta)]^{-1}.$$

(*ii*) ***Super-smooth case**. Assume that (3.6) holds. Suppose that the exponent τ appearing in (3.3) is such that $\tau > 1$ and $(\tau - 1)r \leq \tau$. Assume further that the monotonicity condition (b) in Section 4.2 of Scricciolo [42] is satisfied. Let*

$$\epsilon_n = n^{-1/2}(\log n)^{a_4}, \quad \text{for some constant } a_4 \geq [1/2 + 1/r + 1/(\tau \wedge 2)].$$

Then, under either case (i) or case (ii), for a sufficiently large constant $J_1 > 0$,

$$\mathbb{E}_{p_0}^{(n)}[\pi(U_{J_1\epsilon_n}^c \mid \hat{\gamma}_n, X^{(n)})] = o(1).$$

In Theorem 2, the constant a_3 is the same as that appearing in the convergence rate of the posterior distribution corresponding to a non data-dependent prior with a fixed γ.

3.2. Empirical Bayes density deconvolution

We now present some results on adaptive recovery rates, relative to the $\mathbb{L}_2$-distance, for empirical Bayes density deconvolution when the error density is either ordinary or super-smooth and the mixing density is modeled as a DPM of Gaussian kernels with data-driven chosen hyper-parameter values for the base measure. The problem of deconvolving a density when the mixing density is modeled as a DPM of Gaussian kernels and the error density is super-smooth has been recently investigated by Sarkar et al. [40]. In a frequentist approach, rates for density deconvolution have been studied by Carroll and Hall [7] and Fan [17, 18, 19]. Consider the model

$$X = Y + \varepsilon,$$

where Y and ε are independent random variables. Let p_Y denote the Lebesgue density on $\mathbb{R}$ of Y and K the Lebesgue density on $\mathbb{R}$ of the error measurement ε. The density of X is then the convolution of K and p_Y, denoted by $p_X(\cdot) = (K * p_Y)(\cdot) = \int_{-\infty}^{\infty} K(\cdot - y)p_Y(y)\mathrm{d}y$. The error density K is assumed to be completely known and its characteristic function $\hat{K}$ to satisfy either

$$|\hat{K}(t)| \gtrsim (1 + t^2)^{-\eta/2}, \quad t \in \mathbb{R}, \qquad \text{(ordinary smooth case)} \tag{3.7}$$

for some $\eta > 0$, or

$$|\hat{K}(t)| \gtrsim e^{-\varrho|t|^{r_1}}, \quad t \in \mathbb{R}, \qquad \text{(super-smooth case)} \tag{3.8}$$

for some constant $\varrho > 0$ and exponent $r_1 > 0$. The density p_Y is modeled as a DPM of Gaussian kernels as in (3.1), with a data-driven choice $\hat{\gamma}_n$ for γ. Assuming data $X^{(n)} = (X_1, \ldots, X_n)$ are iid observations from a density $p_{0X} = K * p_{0Y}$ such that the characteristic function $\hat{p}_{0Y}$ of the true mixing distribution satisfies

$$\int_{-\infty}^{\infty} (1+t^2)^{\beta_1} |\hat{p}_{0Y}(t)|^2 \mathrm{d}t < \infty \quad \text{for some } \beta_1 > 1/2, \tag{3.9}$$

we derive adaptive rates for recovering p_{0Y} using empirically selected prior distributions.

Proposition 1. *Suppose that $\hat{K}$ satisfies either condition (3.7) (ordinary smooth case) or condition (3.8) (super-smooth case) and that $\hat{p}_{0Y}$ satisfies the integrability condition (3.9). Consider a prior for p_Y of the form (3.1), with a data-driven choice $\hat{\gamma}_n$ for γ as in Theorem 2. Suppose that $p_{0X} = K * p_{0Y}$ satisfies the conditions of Theorem 2 stated for p_0. Then, there exists a sufficiently large constant $J_1 > 0$ so that*

$$\mathbb{E}_{p_{0X}}^{(n)} [\pi(\|p_Y - p_{0Y}\|_2 > J_1 v_n \mid \hat{\gamma}_n, X^{(n)})] = o(1),$$

where, for some constant $\kappa_1 > 0$,

$$v_n = \begin{cases} n^{-\beta_1/[2(\beta_1+\eta)+1]} (\log n)^{\kappa_1}, & \text{if } \hat{K} \text{ satisfies (3.7)}, \\ (\log n)^{-\beta_1/r_1}, & \text{if } \hat{K} \text{ satisfies (3.8)}. \end{cases}$$

The obtained rates are minimax-optimal, up to a logarithmic factor, in the ordinary smooth case and minimax-optimal in the super-smooth case. Inspection of the proof of Proposition 1 shows that, since the result is based on inversion inequalities relating the $\mathbb{L}_2$-distance between the true mixing density and the (random) approximating mixing density in an efficient sieve set $\mathcal{S}_n$ to the $\mathbb{L}_2$- or the $\mathbb{L}_1$-distance between the corresponding mixed densities, once adaptive rates are known for the direct problem of fully or empirical Bayes estimation of the true sampling density p_{0X}, the same proof yields adaptive recovery rates for both the fully and the empirical Bayes density deconvolution problems. If compared to the approach followed in Sarkar et al. [40], the present strategy simplifies the derivation of adaptive recovery rates for Bayesian density deconvolution. To our knowledge, the ordinary smooth case is treated here for the first time also for the fully Bayes approach.

4. Application to counting processes with Aalen multiplicative monotone non-increasing intensities

In this section, we illustrate our results for counting processes with Aalen multiplicative intensities. Bayesian nonparametric methods have been so far mainly adopted to explore possible prior distributions on intensity functions with the aim of showing that Bayesian nonparametric inference for inhomogeneous Poisson processes can give satisfactory results in applications, see, *e.g.*, Kottas and Sansó [30]. Results on frequentist asymptotic

properties of posterior distributions, like consistency or rates of convergence, have been first obtained by Belitser et al. [5] for inhomogeneous Poisson processes. In Donnet et al. [15] a general theorem on posterior concentration rates for Aalen processes is proposed and some families of priors are studied. Section 4.2 extends these results to the empirical Bayes setting and to the case of monotone non-increasing intensity functions. Section 4.3 illustrates our procedure on artificial data.

4.1. Notation and setup

Let N be a counting process adapted to a filtration $(\mathcal{G}_t)_t$ with compensator Λ so that $(N_t - \Lambda_t)_t$ is a zero-mean $(\mathcal{G}_t)_t$-martingale. A counting process satisfies the *Aalen multiplicative intensity model* if $\mathrm{d}\Lambda_t = \lambda(t)Y_t\mathrm{d}t$, where λ is a non-negative deterministic function called in the sequel, with slight abuse, the intensity function, and $(Y_t)_t$ is an observable non-negative predictable process. Informally,

$$\mathbb{E}[N[t,\, t+\mathrm{d}t] \mid \mathcal{G}_{t^-}] = \lambda(t)Y_t\mathrm{d}t, \tag{4.1}$$

see Andersen et al. [3], Chapter III. We assume that $\Lambda_t < \infty$ almost surely for every t. We also assume that the processes N and Y both depend on an integer n and we consider estimation of λ (not depending on n) in the asymptotic perspective $n \to \infty$, while T is kept fixed. This model encompasses several particular examples: inhomogeneous Poisson processes, censored data and Markov processes. See Andersen et al. [3] for a general exposition, Donnet et al. [15], Gaïffas and Guilloux [21], Hansen et al. [27] and Reynaud-Bouret [35] for specific studies in various settings.

We denote by λ_0 the true intensity function which we assume to be bounded on $\mathbb{R}_+$. We define $\mu_n(t) := \mathbb{E}^{(n)}_{\lambda_0}[Y_t]$ and $\tilde{\mu}_n(t) := n^{-1}\mu_n(t)$. We assume the existence of a non-random set $\Omega \subseteq [0,\, T]$ such that there are constants m_1, m_2 satisfying

$$m_1 \leq \inf_{t\in\Omega} \tilde{\mu}_n(t) \leq \sup_{t\in\Omega} \tilde{\mu}_n(t) \leq m_2 \quad \text{for every } n \text{ large enough}, \tag{4.2}$$

and there exists $\alpha \in (0,\, 1)$ such that, defined $\Gamma_n := \{\sup_{t\in\Omega} |n^{-1}Y_t - \tilde{\mu}_n(t)| \leq \alpha m_1\} \cap \{\sup_{t\in\Omega^c} Y_t = 0\}$, where Ω^c is the complement of Ω in $[0,\, T]$, then

$$\lim_{n\to\infty} \mathbb{P}^{(n)}_{\lambda_0}(\Gamma_n) = 1. \tag{4.3}$$

Assumption (4.2) implies that, on Γ_n,

$$\forall\, t \in \Omega, \quad (1-\alpha)\tilde{\mu}_n(t) \leq \frac{Y_t}{n} \leq (1+\alpha)\tilde{\mu}_n(t). \tag{4.4}$$

Under mild conditions, assumptions (4.2) and (4.3) are easily satisfied for the three examples mentioned above: inhomogeneous Poisson processes, censored data and Markov processes, see Donnet et al. [15] for a detailed discussion. Recall that the log-likelihood for Aalen processes is

$$\ell_n(\lambda) = \int_0^T \log(\lambda(t))\mathrm{d}N_t - \int_0^T \lambda(t)Y_t\mathrm{d}t.$$

Since N is empty on Ω^c almost surely, we only consider estimation over Ω. So, we set $\mathcal{F} = \left\{\lambda : \Omega \to \mathbb{R}_+ \mid \int_\Omega \lambda(t)\mathrm{d}t < \infty\right\}$ endowed with the $\mathbb{L}_1$-norm: for all λ, $\lambda' \in \mathcal{F}$, let $\|\lambda - \lambda'\|_1 = \int_\Omega |\lambda(t) - \lambda'(t)|\mathrm{d}t$. We assume that $\lambda_0 \in \mathcal{F}$ and, for every $\lambda \in \mathcal{F}$, we write

$$\lambda = M_\lambda \times \bar{\lambda}, \qquad \text{with } M_\lambda = \int_\Omega \lambda(t)\mathrm{d}t \text{ and } \bar{\lambda} \in \mathcal{F}_1, \tag{4.5}$$

where $\mathcal{F}_1 = \{\lambda \in \mathcal{F} : \int_\Omega \lambda(t)\mathrm{d}t = 1\}$. Note that a prior probability measure π on $\mathcal{F}$ can be written as $\pi_M \otimes \pi_1$, where π_M is a probability distribution on $\mathbb{R}_+$ and π_1 is a probability distribution on $\mathcal{F}_1$. This representation will be used in the next section.

4.2. Empirical Bayes concentration rates for monotone non-increasing intensities

In this section, we focus on estimation of monotone non-increasing intensities, which is equivalent to considering $\bar{\lambda}$ monotone non-increasing in the parameterization (4.5). To construct a prior on the set of monotone non-increasing densities on $[0, T]$, we use their representation as mixtures of uniform densities as provided by Williamson [51] and we consider a Dirichlet process prior on the mixing distribution:

$$\bar{\lambda}(\cdot) = \int_0^\infty \frac{\mathbf{1}_{(0,\,\theta)}(\cdot)}{\theta}\mathrm{d}P(\theta), \qquad P \mid A, G_\gamma \sim \mathrm{DP}(AG_\gamma), \tag{4.6}$$

where G_γ is a distribution on $[0, T]$. This prior has been studied by Salomond [39] for estimating monotone non-increasing densities. Here, we extend his results to the case of a monotone non-increasing intensity function of an Aalen process with a data-driven choice $\hat{\gamma}_n$ for γ.

We study the family of distributions G_γ with Lebesgue density g_γ belonging to one of the following families: for $\gamma > 0$ and $a > 1$,

$$g_\gamma(\theta) = \frac{\gamma^a \theta^{a-1}}{G(T\gamma)} e^{-\gamma\theta}\mathbf{1}_{\{0\le\theta\le T\}} \quad \text{or} \quad \left(\frac{1}{\theta} - \frac{1}{T}\right)^{-1} \sim \mathrm{Gamma}(a,\,\gamma), \tag{4.7}$$

where G is the cdf of a Gamma$(a,\,1)$ random variable. We then have the following result, which is an application of Theorem 1. We denote by $\pi(\cdot \mid \gamma,\, N)$ the posterior distribution given the observations of the process N.

Theorem 3. *Let $\bar{\epsilon}_n = (n/\log n)^{-1/3}$. Assume that the prior π_M for the mass M is absolutely continuous with respect to Lebesgue measure, with positive and continuous density on $\mathbb{R}_+$, and has finite Laplace transform in a neighbourhood of 0. Assume that the prior $\pi_1(\cdot \mid \gamma)$ on $\bar{\lambda}$ is a DPM of uniform distributions defined in (4.6), with $A > 0$ and base measure G_γ defined as in (4.7). Let $\hat{\gamma}_n$ be a measurable function of the observations satisfying $\mathbb{P}_{\lambda_0}^{(n)}(\hat{\gamma}_n \in \mathcal{K}) = 1 + o(1)$ for some fixed compact subset $\mathcal{K} \subset (0,\,\infty)$. Assume*

also that (4.2) and (4.3) are satisfied and that, for any $k \geq 2$, there exists $C_{1k} > 0$ such that

$$\mathbb{E}_{\lambda_0}^{(n)}\left[\left(\int_\Omega [Y_t - \mu_n(t)]^2 \mathrm{d}t\right)^k\right] \leq C_{1k} n^k. \tag{4.8}$$

Then, there exists a sufficiently large constant $J_1 > 0$ such that

$$\mathbb{E}_{\lambda_0}^{(n)}[\pi(\lambda : \ \|\lambda - \lambda_0\|_1 > J_1 \bar{\epsilon}_n \mid \hat{\gamma}_n, N)] = o(1)$$

and

$$\sup_{\gamma \in \mathcal{K}} \mathbb{E}_{\lambda_0}^{(n)}[\pi(\lambda : \ \|\lambda - \lambda_0\|_1 > J_1 \bar{\epsilon}_n \mid \gamma, N)] = o(1).$$

The proof of Theorem 3 consists in verifying conditions [A1] and [A2] of Theorem 1 and is based on Theorem 3.1 of Donnet et al. [15]. As observed in Donnet et al. [15], condition (4.8) is quite mild and is satisfied for inhomogeneous Poisson processes, censored data and Markov processes. Notice that the concentration rate $\bar{\epsilon}_n$ of the empirical Bayes posterior distribution is the same as that obtained by Salomond [39] for the fully Bayes posterior. Up to a $(\log n)$-factor, this is the minimax-optimal convergence rate over the class of bounded monotone non-increasing intensities.

Note that in Theorem 3, $\mathcal{K}_n$ is chosen to be fixed and equal to $\mathcal{K}$, which covers a large range of possible choices for $\hat{\gamma}_n$. For instance, in the simulation study of Section 4.3, a moment type estimator has been considered which converges almost surely to a fixed value, so that $\mathcal{K}$ is a fixed interval around such value.

4.3. Numerical illustration

We present an experiment to highlight the impact of an empirical Bayes prior distribution for finite sample sizes in the case of an inhomogeneous Poisson process. Let $(W_i)_{i=1,\ldots,N(T)}$ be the observed points of the process N over $[0, T]$, where $N(T)$ is the observed number of jumps. We assume that $Y_t \equiv n$ (n being known). In this case, the compensator Λ of N is non-random and the larger n, the larger $N(T)$.

Estimation of M_{λ_0} and $\bar{\lambda}_0$ can be done separately, given the factorization in (4.5). Considered a gamma prior distribution on M_λ, that is, $M_\lambda \sim \text{Gamma}\,(a_M, b_M)$, we have $M_\lambda \mid N \sim \text{Gamma}\,(a_M + N(T), b_M + n)$. Nonparametric Bayesian estimation of $\bar{\lambda}_0$ is more involved. However, in the case of DPM of uniform densities as a prior on $\bar{\lambda}$, we can use the same algorithms considered for density estimation. In this section, we restrict ourselves to the case where the base measure of the Dirichlet process is the second alternative in (4.7), *i.e.*, under G_γ, it is $\theta \sim \left[T^{-1} + 1/\text{Gamma}\,(a, \gamma)\right]^{-1}$. It satisfies the assumptions of Theorem 3 and presents computational advantages due to conjugacy. Three hyper-parameters are involved in this prior, namely, the mass A of the Dirichlet process, a and γ. The hyper-parameter A strongly influences the number of classes in the posterior distribution of $\bar{\lambda}$. In order to mitigate its influence on the posterior distribution, we propose to consider a hierarchical approach by putting a gamma prior distribution on A,

thus $A \sim \text{Gamma}\,(a_A,\, b_A)$. In absence of additional information, we set $a_A = b_A = 1/10$, which corresponds to a weakly informative prior. Theorem 3 applies to any $a > 1$. We arbitrarily set $a = 2$; the influence of a is not studied in this article. We compare three strategies for determining γ in our simulation study.

Strategy 1: Empirical Bayes - We propose the following estimator:

$$\hat{\gamma}_n = \Psi^{-1}\left[\overline{W}_{N(T)}\right], \quad \overline{W}_{N(T)} = \frac{1}{N(T)} \sum_{i=1}^{N(T)} W_i, \tag{4.9}$$

where

$$\Psi(\gamma) := \mathbb{E}\left[\overline{W}_{N(T)}\right] = \frac{\gamma^a}{2\Gamma(a)} \int_{1/T}^{\infty} \frac{e^{-\gamma/(\nu - \frac{1}{T})}}{(\nu - \frac{1}{T})^{(a+1)}} \frac{1}{\nu} \mathrm{d}\nu,$$

$\mathbb{E}[\cdot]$ denoting expectation under the marginal distribution of N. Hence, $\hat{\gamma}_n$ converges to $\Psi^{-1}(\mathbb{E}[\overline{W}_{N(T)}])$ as n goes to infinity and $\mathcal{K}_n$ can be chosen as any small, but fixed, compact neighbourhood of $\Psi^{-1}(\mathbb{E}[\overline{W}_{N(T)}]) > 0$.

Strategy 2: Fixed γ - In order to avoid an empirical Bayes prior, one can fix $\gamma = \gamma_0$. To study the impact of a bad choice of γ_0 on the behaviour of the posterior distribution, we choose γ_0 different from the calibrated value $\gamma^* = \Psi^{-1}(\mathbb{E}_{theo})$, with $\mathbb{E}_{theo} = \int_0^T t\bar{\lambda}_0(t)\mathrm{d}t$. We thus consider

$$\gamma_0 = \rho \cdot \Psi^{-1}(\mathbb{E}_{theo}), \qquad \rho \in \{0.01, 30, 100\}.$$

Strategy 3: Hierarchical Bayes - We consider a prior on γ, that is, $\gamma \sim \text{Gamma}\,(a_\gamma,\, b_\gamma)$. In order to make a fair comparison with the empirical Bayes posterior distribution, we center the prior distribution at $\hat{\gamma}_n$. Besides, in the simulation study, we consider two different hierarchical hyper-parameters $(a_\gamma,\, b_\gamma)$ corresponding to two prior variances. More precisely, $(a_\gamma,\, b_\gamma)$ are such that the prior expectation is equal to $\hat{\gamma}_n$ and the prior variance is equal to σ_γ^2, the values of σ_γ being specified in Table 1.

Samples of size 30000 (with a warm-up period of 15000 iterations) are generated from the posterior distribution of $(\bar{\lambda},\, A,\, \gamma) \mid N$ using a Gibbs algorithm, decomposed into two or three steps depending on whether or not a fully Bayes strategy is adopted:

$$[1] \quad \bar{\lambda} \mid A,\, \gamma,\, N \qquad [2] \quad A \mid \bar{\lambda},\, \gamma,\, N \qquad [3]^\dagger \quad \gamma \mid A,\, \bar{\lambda},\, N.$$

Step $[3]^\dagger$ only exists if a fully Bayes strategy (strategy 3) is adopted. We use the algorithm developed by Fall and Barat [16]; details can be found in Donnet et al. [14]. The various strategies for calibrating γ are tested on 3 different intensity functions (non null over $[0,\, T]$, with $T = 8$):

$$\begin{aligned}
\lambda_{0,1}(t) &= [4\, \mathbf{1}_{[0,\,3)}(t) + 2\, \mathbf{1}_{[3,\,8]}(t)], \\
\lambda_{0,2}(t) &= e^{-0.4t}, \\
\lambda_{0,3}(t) &= \left[\cos^{-1}\Phi(t)\mathbf{1}_{[0,\,3)}(t) - \left(\frac{1}{6}\cos^{-1}\Phi(3)t - \frac{3}{2}\cos^{-1}\Phi(3)\right)\mathbf{1}_{[3,\,8]}(t)\right],
\end{aligned}$$

where $\Phi(\cdot)$ is the cdf of the standard normal distribution. For each intensity $\lambda_{0,1}$, $\lambda_{0,2}$ and $\lambda_{0,3}$, we simulate 3 datasets corresponding to $n = 500$, 1000 and 2000, respectively. In what follows, we denote by D_n^i the dataset associated with n and intensity $\lambda_{0,i}$.

To compare the three different strategies used to calibrate γ, several criteria are taken into account: tuning of the hyper-parameters, quality of the estimation, convergence of the MCMC and computational time. In terms of tuning effort on γ, the empirical Bayes and the fixed γ approaches are comparable and significantly simpler than the hierarchical one, which increases the space to be explored by the MCMC algorithm and consequently slows down its convergence. Moreover, setting an hyper-prior distribution on γ requires to choose the parameters of this additional distribution, that is, a_γ and b_γ, and to postpone the problem, even though these second-order hyper-parameters are presumably less influential. In our simulation study, the computational time, for a fixed number of iterations, here equal to $N_{iter} = 30000$, turned out to be also a key point. Indeed, the simulation of $\bar{\lambda}$, conditionally on the other variables, involves an accept-reject (AR) step (see equation (B3) in Donnet et al. [14]). For small values of γ, we observe that the acceptance rate of the AR step drops down dramatically, thus inflating the execution time of the algorithm. The computational times (CpT) are summarized in Table 1, which also provides the number of points for each of the 9 datasets $N(T)$, $\hat{\gamma}_n$ being computed using (4.9), $\gamma^* = \Psi^{-1}(\mathbb{E}_{theo})$, the perturbation factor ρ used in the fixed γ strategy and the standard deviation σ_γ of the prior distribution of γ (the prior mean being $\hat{\gamma}_n$) used in the two hierarchical approaches. The second hierarchical prior distribution (last column of Table 1) corresponds to a prior distribution more concentrated around $\hat{\gamma}_n$. We use the algorithm developed by Fall and Barat [16]; details can be found in Donnet et al. [14]. Note that, as described in Donnet et al. [14] (formula B.5 at the end of the paper), the distribution of $\gamma \mid A, \bar{\lambda}, N$ is a gamma whose parameters are easily calculated. As a consequence, this supplementary step in the MCMC algorithm has a negligible computational cost and does not decrease the acceptance rate of the chain.
On Figures 1, 2 and 3, for each strategy and each dataset we plot the posterior median of $\bar{\lambda}_1$, $\bar{\lambda}_2$ and $\bar{\lambda}_3$, respectively, together with a pointwise credible interval using the 10% and 90% empirical quantiles obtained from the posterior simulation. Table 2 gives the distances between the normalized intensity estimates $\hat{\bar{\lambda}}_i$ and the true $\bar{\lambda}_i$ for each dataset and each prior specification. The estimates and the credible intervals for the second hierarchical distribution were very similar to the ones obtained with the empirical strategy and so were not plotted.

For the function $\lambda_{0,1}$, the 4 strategies lead to the same quality of estimation in terms of loss/distance between the functions of interest. In this case, it is thus interesting to have a look at the computational time in Table 1. We notice that for a small γ_0 or for a diffuse prior distribution on γ, possibly generating small values of γ, the computational time explodes. This phenomenon can be so critical that the user may have to stop the execution and re-launch the algorithm. Moreover, the posterior mean of the number of non-empty components in the mixture computed over the last 10000 iterations is equal to 4.21 for $n = 500$ in the empirical strategy, to 11.42 when γ is arbitrarily fixed, to 6.98 under the hierarchical diffuse prior and to 3.77 with the hierarchical concentrated prior.

			Empirical		γ fixed		Hierarchical		Hierarchical 2	
		$N(T)$	$\hat{\gamma}_n$	CpT	$\rho\Psi^{-1}(E_{theo})$	CpT	σ_γ	CpT	σ_γ	CpT
$\lambda_{0,1}$	D^1_{500}	499	0.0386	523.57	0.01×0.0323	2085.03	0.005	12051.22	0.001	447.75
	D^1_{1000}	1036	0.0372	783.53		1009.58		791.28		773.33
	D^1_{2000}	2007	0.0372	1457.40		1561.64		1477.50		1456.03
$\lambda_{0,2}$	D^2_{500}	505	0.6605	1021.73	100×0.6667	1022.59	0.1	663.54	0.01	1047.42
	D^2_{1000}	978	0.6857	1873.05		1416.40		1207.07		2018.89
	D^2_{2000}	2034	0.6827	4849.80		2236.02		2533.62		4644.55
$\lambda_{0,3}$	D^3_{500}	483	0.4094	782.19	30×0.4302	822.12	0.1	788.14	0.01	788.00
	D^3_{1000}	1058	0.4398	1610.47		2012.96		1559.17		1494.75
	D^3_{2000}	2055	0.4677	3546.57		9256.71		3179.96		2770.83

Table 1. Computational Time (CpT in seconds), hyper-parameters for the different strategies and datasets

		$\lambda_{0,1}$			$\lambda_{0,2}$			$\lambda_{0,3}$		
		D^1_{500}	D^1_{1000}	D^1_{2000}	D^2_{500}	D^2_{1000}	D^2_{2000}	D^3_{500}	D^3_{1000}	D^3_{2000}
$d_{\mathbb{L}_1}$	Empir	0.0246	0.0238	0.0207	0.0921	0.0817	0.0549	0.1382	0.0596	0.0606
	Fixed	0.0161	0.0219	0.0211	0.5381	0.7221	0.6356	0.3114	0.2852	0.2885
	Hierar	0.0132	0.0233	0.0317	0.1082	0.1280	0.0969	0.2154	0.1378	0.1405
	Hiera 2	0.0191	0.0240	0.0208	0.0925	0.0815	0.0552	0.1383	0.0607	0.0724

Table 2. $\mathbb{L}_1$-distances between the estimates and the true densities for all datasets and strategies

In this case, choosing a small value of γ leads to a posterior distribution on mixtures with too many non-empty components. These phenomena tend to disappear when n increases. For $\lambda_{0,2}$ and $\lambda_{0,3}$, a bad choice of γ - here γ too large in strategy 2 - or a not enough informative prior on γ, namely, a hierarchical prior with large variance, has a significant negative impact on the behaviour of the posterior distribution. Contrariwise, the medians of the empirical and informative hierarchical posterior distributions of λ have similar losses, as seen in Table 2.

5. Final remarks

In this article, we stated sufficient conditions for assessing contraction rates of posterior distributions corresponding to data-dependent priors. The proof of Theorem 1 relies on two main ideas:

a) replacing the empirical Bayes posterior probability of the set $U^c_{J_1\epsilon_n}$ by the supremum (with respect to γ) of the posterior probability of $U^c_{J_1\epsilon_n}$ over a set $\mathcal{K}_n$;

b) shifting data-dependence from the prior to the likelihood using a suitable parameter transformation $\psi_{\gamma,\gamma'}$.

We do not claim that all nonparametric data-dependent priors can be handled using Theorem 1, yet, we believe it can be applied to many relevant situations. In Section 2, we have described possible parameter transformations for some families of prior distributions. To

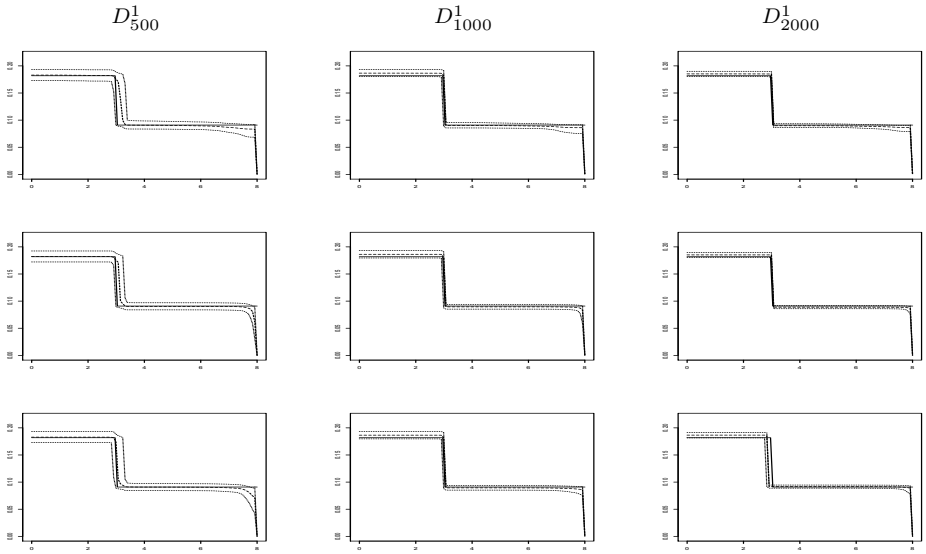

Figure 1. Estimation of $\bar{\lambda}_1$ from D^1_{500} (first column), D^1_{1000} (second column) and D^1_{2000} (third column) using different strategies: empirical prior (row 1), fixed γ (row 2), hierarchical empirical prior (row 3). True density (plain line), estimate (dashed line) and confidence band (dotted lines)

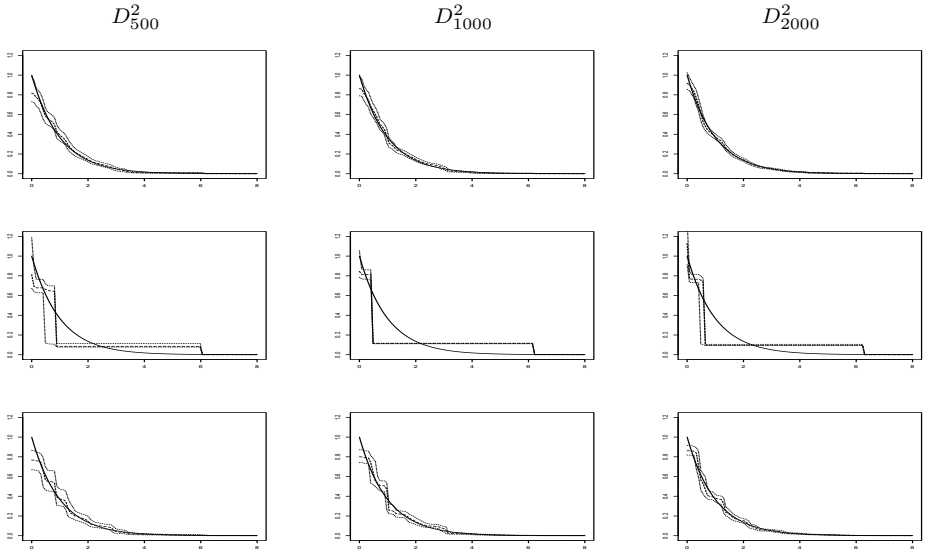

Figure 2. Estimation of $\bar{\lambda}_2$ from D^2_{500} (first column), D^2_{1000} (second column) and D^2_{2000} (third column) using different strategies: empirical prior (row 1), fixed γ (row 2), hierarchical empirical prior (row 3). True density (plain line), estimate (dashed line) and confidence band (dotted lines)

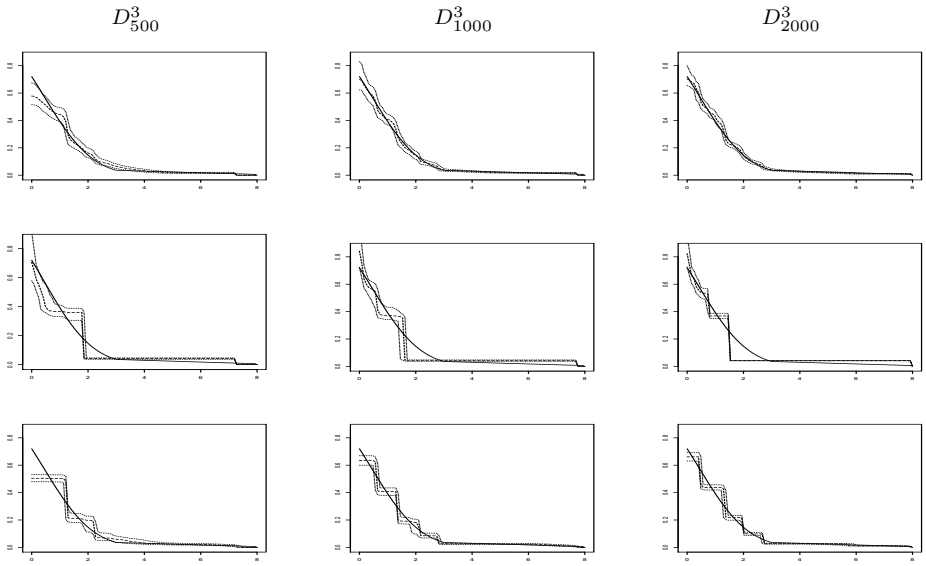

Figure 3. Estimation of $\bar{\lambda}_3$ from D^3_{500} (first column), D^3_{1000} (second column) and D^3_{2000} (third column) using different strategies: empirical prior (row 1), fixed γ (row 2), hierarchical empirical prior (row 3). True density (plain line), estimate (dashed line) and confidence band (dotted lines)

apply Theorem 1 in these cases, it is then necessary to control

$$\inf_{\gamma' \colon \|\gamma'-\gamma\| \leq u_n} \ell_n(\psi_{\gamma,\gamma'}(\theta)) - \ell_n(\theta_0) \quad \text{and} \quad \sup_{\gamma' \colon \|\gamma'-\gamma\| \leq u_n} \ell_n(\psi_{\gamma,\gamma'}(\theta)) - \ell_n(\theta_0).$$

This is typically achieved by bounding above the supremum on the right-hand side of the last display by a well-behaved function of the data, say $m_{\theta,\gamma}$:

$$\sup_{\gamma' \colon \|\gamma'-\gamma\| \leq u_n} \ell_n(\psi_{\gamma,\gamma'}(\theta)) - \ell_n(\theta_0) \leq u_n m_{\theta,\gamma}(X^{(n)}).$$

Similarly for the infimum. This has been illustrated in the examples of Sections 3 and 4.

An important feature of the proposed approach is the identification of a set $\mathcal{K}_n$ satisfying condition (2.1). When $\hat{\gamma}_n$ corresponds to a moment estimator, the set $\mathcal{K}_n$ is easily identified; this is the case in the examples herein considered. When $\hat{\gamma}_n$ is implicitly defined, as it is the case for the maximum marginal likelihood estimator, it is more difficult to characterize $\mathcal{K}_n$ and a preliminary study is needed to be able to apply Theorem 1. This is the approach taken in Rousseau and Szabó [38], where posterior contraction rates are provided for the maximum marginal likelihood estimator following this scheme. The authors provide some examples where minimax-optimal posterior contraction rates are attained and some others where sub-optimal rates are found. This mainly depends on the family of prior distributions. In particular, sub-optimal posterior contraction rates

are obtained for Gaussian priors in the form

$$\theta_i \overset{\text{ind}}{\sim} \mathcal{N}(0, \gamma i^{-(2\alpha+1)}), \quad i \in \mathbb{N},$$

when the true parameter belongs to a Sobolev ball with smoothness $\beta > \alpha + 1/2$.

Although data-dependent prior distributions are commonly used in practice, theoretical properties have been so far considered only for maximum marginal likelihood empirical Bayes procedures when an explicit expression of the marginal likelihood is available. The present contribution is an attempt at filling this gap.

Acknowledgements

This research benefited from the support of the "Chaire Economie et Gestion des Nouvelles Données", under the auspices of the Institut Louis Bachelier, Havas-Media and Paris-Dauphine. The research of Sophie Donnet, Vincent Rivoirard and Judith Rousseau was partly supported by the French Agence Nationale de la Recherche (ANR 2011 BS01 010 01 projet Calibration). Most of the research work of Catia Scricciolo was done when she was affiliated to Bocconi University, whose financial support is gratefully acknowledged. The authors would like to thank the Editor, an Associate Editor and two anonymous referees whose suggestions helped to improve the final presentation of the article.

6. Supplementary material

Supplement to "Posterior concentration rates for empirical Bayes procedures with applications to Dirichlet process mixtures". This supplement contains the proofs of Theorem 2, Proposition 1 and Theorem 3 from the mentioned article.

References

[1] Adams, R. P., Murray, I., and MacKay, D. J. (2009). Tractable nonparametric Bayesian inference in Poisson processes with Gaussian process intensities. *Proceedings of the 26th Annual International Conference on Machine Learning. ACM.*

[2] Ahmed, S. and Reid, N. (2001). *Empirical Bayes and Likelihood Inference.* Lecture Notes in Statistics, Springer.

[3] Andersen, P. K., Borgan, A., Gill, R. D., and Keiding, N. (1993). *Statistical models based on counting processes.* Springer Series in Statistics. Springer-Verlag, New York.

[4] Barron, A., Schervish, M., and Wasserman, L. (1999). The consistency of posterior distributions in nonparametric problems. *Ann. Statist.*, 27(2):536–561.

[5] Belitser, E., Serra, P., and van Zanten, H. (2015). Rate-optimal Bayesian intensity smoothing for inhomogeneous Poisson processes. *J. Statist. Plann. Inference*, 166:24–35.

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

[7] Carroll, R. and Hall, P. (1988). Optimal rates of convergence for deconvolving a density. *J. American Statist. Assoc.*, 83(404):1184–1186.

[8] Casella, G. (1985). An introduction to empirical Bayes data analysis. *The American Statistician*, 39(2):83–87.

[9] Cheng, N. and Yuan, T. (2013). Nonparametric Bayesian lifetime data analysis using Dirichlet process lognormal mixture model. *Naval Res. Logist.*, 60(3):208–221.

[10] Clyde, M. A. and George, E. I. (2000). Flexible empirical Bayes estimation for wavelets. *J. Royal Statist. Society Series B*, 62(4):681–698.

[11] Cui, W. and George, E. I. (2008). Empirical Bayes vs. fully Bayes variable selection. *J. Statist. Plann. Inference*, 138(4):888–900.

[12] Daley, D. J. and Vere-Jones, D. (2003). *An introduction to the theory of point processes. Vol. I.* Probability and its Applications (New York). Springer-Verlag, New York, second edition. Elementary theory and methods.

[13] Daley, D. J. and Vere-Jones, D. (2008). *An introduction to the theory of point processes. Vol. II.* Probability and its Applications (New York). Springer, New York, second edition. General theory and structure.

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

[15] Donnet, S., Rivoirard, V., Rousseau, J., and Scricciolo, C. (2016). Posterior concentration rates for counting processes with Aalen multiplicative intensities. *arXiv:1407.6033v1, to appear in Bayesian Analysis*.

[16] Fall, M. D. and Barat, É. (2012). Gibbs sampling methods for Pitman-Yor mixture models. Technical report.

[17] Fan, J. (1991a). Global behavior of deconvolution kernel estimates. *Statistica Sinica*, 1:541–551.

[18] Fan, J. (1991b). On the optimal rates of convergence for nonparametric deconvolution problems. *Ann. Statist.*, 19(3):1257–1272.

[19] Fan, J. (1992). Deconvolution with supersmooth errors. *Canadian J. Statist.*, 20(2):155–169.

[20] Ferguson, T. (1974). Prior distributions in spaces of probability measures. *Ann. Statist.*, 2(4):615–629.

[21] Gaïffas, S. and Guilloux, A. (2012). High-dimensional additive hazards models and the Lasso. *Electron. J. Statist.*, 6:522–546.

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

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

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

[25] Ghosal, S. and van der Vaart, A. W. (2001). Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities. *Ann.*

Statist., 29(5):1233–1263.

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

[27] Hansen, N. R., Reynaud-Bouret, P., and Rivoirard, V. (2015). Lasso and probabilistic inequalities for multivariate point processes. *Bernoulli*, 21(1):83–143.

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

[29] Karr, A. F. (1991). *Point processes and their statistical inference*, volume 7 of *Probability: Pure and Applied*. Marcel Dekker Inc., New York, second edition.

[30] Kottas, A. and Sansó, B. (2007). Bayesian mixture modeling for spatial Poisson process intensities, with applications to extreme value analysis. *J. Statist. Plann. Inference*, 137(10):3151–3163.

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

[32] Lo, A. Y. (1992). Bayesian inference for Poisson process models with censored data. *J. Nonparametr. Statist.*, 2(1):71–80.

[33] McAuliffe, J. D., Blei, D. M., and Jordan, M. I. (2006). Nonparametric empirical Bayes for the Dirichlet process mixture model. *Statistics and Computing*, 16(1):5–14.

[34] Petrone, S., Rousseau, J., and Scricciolo, C. (2014). Bayes and empirical Bayes: do they merge? *Biometrika*, 101(2):285–302.

[35] Reynaud-Bouret, P. (2006). Penalized projection estimators of the Aalen multiplicative intensity. *Bernoulli*, 12(4):633–661.

[36] 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(4):731–792.

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

[38] Rousseau, J. and Szabó, B. T. (2015). Asymptotic behaviour of the empirical Bayes posteriors associated to maximum marginal likelihood estimator. Technical report.

[39] Salomond, J.-B. (2014). Concentration rate and consistency of the posterior distribution for selected priors under monotonicity constraints. *Electron. J. Statist.*, 8(1):1380–1404.

[40] Sarkar, A., Pati, D., Mallick, B. K., and Carroll, R. J. (2013). Adaptive posterior convergence rates in Bayesian density deconvolution with supersmooth errors. Technical report, arXiv:1308.5427v2.

[41] Schwartz, L. (1965). On Bayes procedures. *Z. Warsch. Verw. Gebiete*, 4(1):10–26.

[42] Scricciolo, C. (2014). Adaptive Bayesian density estimation in L^p-metrics with Pitman-Yor or normalized inverse-Gaussian process kernel mixtures. *Bayesian Analysis*, 9(2):475–520.

[43] Scricciolo, C. (2015). Empirical Bayes conditional density estimation. *STATISTICA*, LXXV(1):37–55.

[44] Serra, P. and Krivobokova, T. (2014). Adaptive empirical Bayesian smoothing splines. Technical report.

[45] Shen, W., Tokdar, S., and Ghosal, S. (2013). Adaptive Bayesian multivariate density estimation with Dirichlet mixtures. *Biometrika*, 100(3):623–640.

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

[47] Sniekers, S. and van der Vaart, A. (2015). Adaptive Bayesian credible sets in regression with a Gaussian process prior. *Electron. J. Statist.*, 9(2):2475–2527.

[48] Szabó, B., van der Vaart, A., and van Zanten, H. (2015a). Honest Bayesian confidence sets for the L^2-norm. *Journal of Statistical Planning and Inference*, 166:36–51. Special Issue on Bayesian Nonparametrics.

[49] Szabó, B., van der Vaart, A. W., and van Zanten, J. H. (2015b). Frequentist coverage of adaptive nonparametric Bayesian credible sets. *Ann. Statist.*, 43(4):1391–1428.

[50] Szabó, 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.

[51] Williamson, R. E. (1956). Multiply monotone functions and their Laplace transforms. *Duke Math. J.*, 23:189–207.