Biometrika (2012), **99**, 1, *pp*. 1–24
© 2012 Biometrika Trust
Printed in Great Britain

Advance Access publication on 31 July 2012

Bayes and empirical Bayes: do they merge?

BY S. PETRONE

Department of Decision Sciences, Bocconi University, Via G. Röntgen 1, 20136 Milano, Italy
sonia.petrone@unibocconi.it

J. ROUSSEAU

CREST-ENSAE and Université Paris Dauphine, 3, Avenue P. Larousse, 92240 Malakoff, France
rousseau@ceremade.dauphine.fr

AND C. SCRICCIOLO

Department of Decision Sciences, Bocconi University, Via G. Röntgen 1, 20136 Milano, Italy
catia.scricciolo@unibocconi.it

SUMMARY

Bayesian inference is attractive for its coherence and good frequentist properties. However, eliciting a honest prior may be difficult and a common practice is to take an empirical Bayes approach using some estimate of the prior hyperparameters. Although not rigorous, the underlying idea is that, for sufficiently large sample size, empirical Bayes should lead to similar inferential answers as a proper Bayesian inference. However, precise mathematical results seem to be missing. In this work, we give rigorous results in terms of merging of Bayesian and empirical Bayes posterior distributions. We study two notions of merging: Bayesian weak merging and frequentist merging in total variation. We also show that, under regularity conditions, empirical Bayes asymptotically gives an oracle selection of the prior hyperparameters. Examples include empirical Bayes density estimation with Dirichlet process mixtures.

Some key words: Bayesian model selection; Bayesian weak merging; Dirichlet process mixtures; Frequentist strong merging; Maximum marginal likelihood estimation; Posterior consistency; Regression with g-priors.

1. INTRODUCTION AND MOTIVATION

The Bayesian approach to inference is appealing in treating uncertainty probabilistically through conditional distributions. If $(X_1, \ldots, X_n)$, conditionally on θ, have joint density $p_\theta^{(n)}$ and θ has prior density $\pi(\theta \mid \lambda)$, then the information on θ, given the data, is expressed through the conditional, or posterior, density $\pi(\theta \mid \lambda, x_1, \ldots, x_n) \propto p_\theta^{(n)}(x_1, \ldots, x_n)\pi(\theta \mid \lambda)$. Although Bayesian procedures are increasingly popular, it is common experience that expressing honest prior information can be difficult and, in practice, one is tempted to use some estimate $\hat{\lambda}_n \equiv \hat{\lambda}_n(x_1, \ldots, x_n)$ of the prior hyperparameter λ and a posterior density $\pi(\cdot \mid \hat{\lambda}_n, x_1, \ldots, x_n)$. This mixed approach is usually referred to as empirical Bayes in the literature, see Lehmann & Casella (1998). The underlying idea is that, when the sample size is large, empirical Bayes should lead to inferential results similar to those of any Bayesian procedure. An empirical Bayesian would then achieve the goal of inference without completely specifying a prior distribution.

An empirical Bayes approach is not justified from a Bayesian point of view, it is, however, attractive as a computationally simpler alternative to a more rigorous, but usually analytically more complex, hierarchical specification of the prior of the kind $\int \pi(\cdot \mid \lambda)h(\lambda)\,\mathrm{d}\lambda$. Thus, for a Bayesian statistician, empirical Bayes is of interest for two reasons: when it is difficult to honestly fix λ, it is expected that a data-driven choice of λ may lead to better inferential results; also, the empirical Bayes posterior could be a simple approximation of a hierarchical posterior distribution. These are possibly the reasons for the wide use of empirical Bayes in practical applications. However, to be rigorously justified, it is necessary (a) to prove whether it is true that empirical Bayes and (hierarchical) Bayes will asymptotically agree and (b) to investigate whether empirical Bayes procedures have some optimality property (versus a fixed choice of λ). To our knowledge, general results on such asymptotic agreement and on optimality properties are missing. The aim of this paper is to provide results in both directions. First, we will give conditions for the asymptotic agreement, or merging, of empirical Bayes and Bayesian solutions, but we will also single out situations wherein empirical Bayes and Bayes diverge and, thus, from a Bayesian viewpoint, require special care. Then, we will show that, in regular parametric cases, the maximum marginal likelihood selection of λ converges to a limit that is optimal, in the sense that it corresponds to an oracle choice of the prior that mostly favors the true model. Thus, for sufficiently large sample size, empirical Bayes would give a solution that is close to the Bayesian oracle and, in this sense, can be expected to exploit information more efficiently than a fixed choice of λ.

Although not rigorously justified, empirical Bayes is used quite often by practitioners and in the literature, see, for instance, George & Foster (2000) in the context of variable selection in regression; Clyde & George (2000) for wavelets shrinkage estimation; Liu (1996) and McAuliffe *et al.* (2006) in Bayesian nonparametric mixture models; Favaro *et al.* (2009) in Bayesian nonparametric inference for species diversity. A systematic comparison of empirical Bayes and Bayesian procedures appears to be less explored. A careful comparison of empirical Bayes and Bayesian variable selection criteria in regression is developed by Cui & George (2008). In this context, a surprising result has been recently highlighted by Scott & Berger (2010) who show an asymptotic discrepancy between empirical Bayes and Bayesian inferences. Empirical Bayes and hierarchical Bayesian procedures for nonparametric curve estimation are studied in Belitser & Levit (2003), Belitser & Enikeeva (2008) and in the recent work by Knapik *et al.* (2012). In these problems, the hyperparameter λ can be endowed with an interpretation as a model index and a direct relationship to the true parameter exists a priori. However, in general, the hyperparameter merely characterizes some aspects of the prior so that there exists no notion of true value of λ; thus, it is not clear which could be a desirable limit for the sequence of $\hat{\lambda}_n$. We will propose a notion of oracle value instead of true value for λ in § 4.

The term empirical Bayes is used with different meanings in the literature. Another common use refers to problems where a prior distribution is introduced but a frequentist interpretation of it is possible, typically in hierarchical models where X_i, conditionally on θ_i, has density p_{θ_i} and the θ_i are a sample from a latent distribution $G(\cdot \mid \lambda)$. Integrating out the θ_i, the X_i are independent and identically distributed according to $\int p_\theta(\cdot)\,\mathrm{d}G(\theta \mid \lambda)$. In these problems, maximum likelihood estimation of λ, i.e., of the latent distribution G, is often referred to as empirical Bayes. A Bayesian approach would assign a prior distribution to λ. In these cases, a comparison between Bayes and empirical Bayes reduces to the interesting, but more standard, comparison between Bayes and maximum likelihood procedures, which, however, is not the primary object of this work.

The first question we address is whether empirical Bayes and Bayesian posterior distributions will asymptotically be close. A relevant counterexample has been recently exhibited by Scott &

Berger (2010) in the case of variable selection in regression models. They consider a Bayesian approach where variable selection is based on an inclusion vector $\gamma = (\gamma_1, \ldots, \gamma_k) \in \{0, 1\}^k$ which selects among k potential regressors and the prior on γ assumes that γ_j are independent Bernoulli random variables with parameter λ, $\pi(\gamma_1, \ldots, \gamma_k \mid \lambda) = \lambda^{k_\gamma}(1-\lambda)^{k-k_\gamma}$, where $k_\gamma = \sum_{j=1}^k \gamma_j$ is the selected number of covariates. In this framework, George & Foster (2000) have shown that an empirical Bayes procedure that estimates the inclusion probability λ from the data by, for example, the maximum marginal likelihood estimator, may be preferable to a Bayesian procedure that uses a fixed value of λ. Scott & Berger (2010) compare this empirical Bayes approach with a hierarchical Bayesian procedure that assigns a prior to λ. Surprisingly, they prove an asymptotic discrepancy between the two procedures. They show that the empirical Bayes posterior distribution on the set of models can be degenerate on the null model ($\gamma = (0, \ldots, 0)$) or on the full model ($\gamma = (1, \ldots, 1)$). This might still lead to interesting pointwise estimates of the model or of the whole parameter, but it is far from being satisfactory in terms of the posterior distribution. We shed light on these phenomena by describing when and why maximum marginal likelihood empirical Bayes procedures will be pathological or, conversely, when and why they will have some oracle property. These results have therefore the practical interest of characterizing, at least in the parametric case, those families of priors which can be jointly used with empirical Bayes procedures and those which instead should be avoided, especially if interest does not merely lie in point estimation, but in more general features of the posterior distribution.

We formalize the asymptotic comparison in terms of merging of empirical Bayes and Bayesian procedures. We consider two notions of merging. First, we study Bayesian weak merging in the sense of Diaconis & Freedman (1986). Then, we study frequentist strong merging in the sense of Ghosh & Ramamoorthi (2003) which compares posterior distributions in terms of total variation distance in the frequentist sense, that is, almost surely with respect to the true probability law $P_0^{(\infty)}$ of $(X_i)_{i\geqslant 1}$. Note that when strong merging holds, if the Bernstein von-Mises theorem holds in the L_1-sense for the Bayesian posterior, then it also holds for the empirical Bayes posterior.

Weak merging of Bayes and empirical Bayes means that any Bayesian is sure that his posterior distribution and the empirical Bayes posterior distribution will eventually be close, in the sense of weak convergence. It is thus a minimal requirement. However, it is not guaranteed and it holds if and only if the empirical Bayes posterior distribution is consistent at the true value θ_0 of the parameter in the frequentist sense. Therefore, it is of interest to provide conditions under which the empirical Bayes posterior is consistent in a general context covering both parametric and nonparametric problems, see § 3.

Simple examples show that, even when consistency and weak merging hold, the empirical Bayes posterior may have unexpected and counterintuitive behaviors. Frequentist strong merging is a way to refine the analysis. Obtaining strong merging of Bayesian posteriors in nonparametric contexts is often impossible since pairs of priors are typically singular. Thus, in tackling this issue, we concentrate on parametric models and on the specific, but important, case of the maximum marginal likelihood estimator $\hat{\lambda}_n$. We find that the behavior of the empirical Bayes posterior is essentially driven by the behavior of the prior at θ_0. Roughly speaking, if $\sup_\lambda \pi(\theta_0 \mid \lambda)$ is achieved at a value λ^* (here unique for simplicity) in the boundary of Λ such that $\pi(\cdot \mid \lambda^*)$ is degenerate at θ_0, then the empirical Bayes posterior will not merge with any Bayesian posterior distribution. We illustrate this behavior in Bayesian regression with g-priors and in model selection. Conversely, if $\sup_\lambda \pi(\theta_0 \mid \lambda) < \infty$, which is the case if λ^* is in the interior of Λ, then $\hat{\lambda}_n$ converges to λ^* and frequentist strong merging holds. The value λ^* can be understood as the

prior oracle since it is the value of the hyperparameters for which the prior mostly favors the truth θ_0. Under this respect, the empirical Bayes posterior achieves some kind of optimality.

2. General context and notation

Let $\mathcal{X}$ and Θ denote the observational space and the parameter space, respectively. In order to cover both parametric and nonparametric problems, we only require that they are complete and separable metric spaces, equipped with their Borel σ-fields $\mathcal{B}(\mathcal{X})$ and $\mathcal{B}(\Theta)$, respectively. Let $(X_i)_{i\geqslant 1}$ be a sequence of random elements, with X_i taking values in $\mathcal{X}$. Suppose that, given θ, the probability measure of the process $(X_i)_{i\geqslant 1}$ is $P_\theta^{(\infty)}$ and, for every $n = 1, 2, \ldots$, denote by $P_\theta^{(n)}$ the joint probability law of $(X_1, \ldots, X_n)$. We consider a dominated collection of probability measures $P_\theta^{(n)}$ with respect to some σ-finite measure $\mu^{(n)}$ and denote the densities by $p_\theta^{(n)}$. In the sequel, we use the short notations $(X_i) = (X_i)_{i\geqslant 1}$, $X_{1:n} = (X_1, \ldots, X_n)$ and $x^\infty = (x_1, x_2, \ldots)$.

Let $\{\Pi(\cdot \mid \lambda) : \lambda \in \Lambda\}$ be a family of prior probability measures on Θ, with $\Lambda \subseteq \mathbb{R}^\ell$, for finite ℓ, endowed with the Borel σ-field induced by the Euclidean distance $\|\lambda - \lambda'\| = \{\sum_{j=1}^\ell (\lambda_j - \lambda'_j)^2\}^{1/2}$ whenever, in a hierarchical Bayesian approach, a prior distribution for λ is considered. For a prior $\Pi(\cdot \mid \lambda)$, we denote by $\Pi(\cdot \mid \lambda, X_{1:n})$ the corresponding posterior distribution of θ, given $X_{1:n}$, which can be computed by the Bayes' rule because the model is dominated. The empirical Bayes approach consists in estimating the hyperparameter λ by $\hat\lambda_n \equiv \hat\lambda_n(X_{1:n})$ and plugging the estimate into the posterior distribution. In general, $\hat\lambda_n$ asymptotically takes values in the closure $\bar\Lambda$ of Λ. If Λ is open and λ_0 is in the boundary $\partial\Lambda$ of Λ, we define $\Pi(\cdot \mid \lambda_0)$ as the σ-additive weak limit of $\Pi(\cdot \mid \lambda)$ as $\lambda \to \lambda_0$, if it exists.

Throughout the paper, we assume that, for each θ, $P_\theta^{(\infty)}$-almost surely, for all large n, the estimator $\hat\lambda_n$ takes values in Λ or in its boundary $\partial\Lambda$, in the latter case assuming that the prior exists as a weak limit (thus ruling out improper priors), and $0 < \int_\Theta p_\theta^{(n)}(X_{1:n})\,\mathrm{d}\Pi(\theta \mid \hat\lambda_n) < \infty$. Then, we say that the empirical Bayes posterior is well defined and it is obtained by plugging $\hat\lambda_n$ into the Bayesian posterior, that is, for every Borel set B,

$$\Pi(B \mid \hat\lambda_n, X_{1:n}) = \frac{\int_B p_\theta^{(n)}(X_{1:n})\,\mathrm{d}\Pi(\theta \mid \hat\lambda_n)}{\int_\Theta p_\theta^{(n)}(X_{1:n})\,\mathrm{d}\Pi(\theta \mid \hat\lambda_n)}.$$

Many types of estimators $\hat\lambda_n$ can be considered: the maximum marginal likelihood estimator, defined as $\hat\lambda_n \in \operatorname{argmax}_{\lambda\in\bar\Lambda} m(X_{1:n} \mid \lambda)$, where $m(X_{1:n} \mid \lambda) = \int_\Theta p_\theta^{(n)}(X_{1:n})\,\mathrm{d}\Pi(\theta \mid \lambda)$, is the most popular. This tacitly assumes that, for every θ, $P_\theta^{(\infty)}$-almost surely, $m(X_{1:n} \mid \lambda)$ has a maximum over $\bar\Lambda$ and we will write $\hat m(X_{1:n}) = m(X_{1:n} \mid \hat\lambda_n)$. We will present general results for the empirical Bayes posterior distribution with any type of estimator $\hat\lambda_n$ as well as specific results for the empirical Bayes posterior with the maximum marginal likelihood estimator.

3. Bayesian weak merging and consistency

3·1. *General results*

Bayesian merging is a natural way to formalize the idea that the empirical Bayes posterior and the Bayesian posterior will asymptotically be close. Blackwell & Dubins (1962) give fundamental results on Bayesian strong merging, which, however, cannot be applied to empirical Bayes since they are based on properties of the probability law of the process (X_i), whereas the

empirical Bayes approach only gives a sequence of posterior distributions, without a properly defined probability law of (X_i). Diaconis & Freedman (1986) give a notion of weak merging that applies even when strong merging does not. Two sequences of probability measures p_n and q_n are said to merge weakly if and only if $|\int f \,\mathrm{d}p_n - \int f \,\mathrm{d}q_n|$goes to 0, for all continuous and bounded functions f. They show that two Bayesian statisticians with different priors merge weakly if and only if one of them has a consistent posterior, in the frequentist sense, at θ, for every θ in Θ. An analogous result holds in the present context: the empirical Bayesian merges weakly with any Bayesian if and only if the empirical Bayes posterior is consistent at θ, for every θ in Θ. The result is herein restricted to the case of exchangeable sequences, thus, given θ, the X_i are independent and identically distributed according to P_θ. Given a prior Π on Θ, we use Π_n to denote the posterior distribution and P_Π for the exchangeable probability law of the process (X_i) defined through Π. Recall that a posterior distribution Π_n is said to be consistent at θ if Π_n converges weakly to a point mass at θ with P_θ^∞-probability one, where P_θ^∞ denotes the infinite product measure on $\mathcal{X}^\infty$. A posterior distribution Π_n is said to be consistent if it is consistent at every $\theta \in \Theta$. The following result is a consequence of Theorem A.1 in Diaconis & Freedman (1986).

PROPOSITION 1. *Let $\theta \mapsto P_\theta$ be one-to-one and such that, for every $B \in \mathcal{B}(\mathcal{X})$, the map $\theta \mapsto P_\theta(B)$ is $\mathcal{B}(\Theta)$-measurable. Let $\Pi(\cdot \mid \hat{\lambda}_n,\, X_{1:n})$ be the empirical Bayes posterior obtained from a family of priors $\{\Pi(\cdot \mid \lambda) : \lambda \in \Lambda\}$ and an estimator $\hat{\lambda}_n$ of λ. Then, for any prior probability measure q on Θ, the empirical Bayes posterior and the Bayesian posterior q_n merge weakly with P_q-probability one if and only if the empirical Bayes posterior is consistent.*

The proof is straightforward since it suffices to note that the proof for the equivalences (i)–(iv) in Theorem A.1 of Diaconis & Freedman (1986) goes through to the present case: in fact, it is based on the properties of the Bayesian posterior q_n, whereas for the empirical Bayes posterior $\Pi(\cdot \mid \hat{\lambda}_n,\, X_{1:n})$ only consistency is required.

Proposition 1 shows that any Bayesian, in particular, any Bayesian with a hierarchical prior $\int_\Lambda \pi(\cdot \mid \lambda) h(\lambda)\,\mathrm{d}\lambda$, can be sure that her estimate with respect to the quadratic loss of any continuous and bounded function f will asymptotically agree with the empirical Bayes estimate if and only if the empirical Bayes posterior is consistent. Thus, even a minimal requirement as weak merging is not guaranteed. It is worth highlighting that consistency refers to the posterior distribution of θ and cannot be referred to the estimator $\hat{\lambda}_n$ since, in our context, there is generally no notion of true value of λ.

Besides its Bayesian motivation in terms of merging, consistency is a fundamental property of autonomous interest from the frequentist point of view. Therefore, we study consistency of empirical Bayes posterior distributions for dependent sequences, beyond the case of exchangeability, covering both parametric and nonparametric cases. Clearly, consistency of empirical Bayes posterior distributions requires more care than consistency of Bayesian posteriors because the prior is data-dependent through $\hat{\lambda}_n$ and one has to control the behavior of the sequence $\hat{\lambda}_n$. We give two results: one for procedures where $\hat{\lambda}_n$ is the maximum marginal likelihood estimator and the other one for procedures where $\hat{\lambda}_n$ is a convenient estimator.

To be more specific, let $(\Theta,\, d)$ be a complete and separable metric space. For parametric models with $\theta \in \mathbb{R}^k$, where k is finite, $d(\theta,\, \theta')$ can be the Euclidean distance $\|\theta - \theta'\| = \{\sum_{j=1}^k (\theta_j - \theta'_j)^2\}^{1/2}$. In nonparametric problems, for example in density estimation, Θ can be some collection of dominated probability measures on $\mathcal{X}$ and one could identify θ with the density $p_\theta^{(n)}$ itself so that natural metrics can be the metric of weak topology or the Hellinger metric which, for independent and identically distributed observations, writes as $d(\theta,\, \theta_0) =$

$\{\int (p_\theta^{1/2} - p_{\theta_0}^{1/2})^2 \, \mathrm{d}\mu\}^{1/2}$. For any $\epsilon > 0$, let $U_\epsilon = \{\theta \in \Theta : \, d(\theta, \theta_0) < \epsilon\}$ denote the open ball centered at θ_0 with radius ϵ. Since $(\Theta, \, d)$ is separable, the definition of consistency can be restated in terms of neighborhoods of θ_0, see, e.g., Ghosh & Ramamoorthi (2003), page 17. Therefore, the empirical Bayes posterior is consistent at θ_0, in the sense of the metric d, if, for any $\epsilon > 0$, $\Pi(U_\epsilon^c \mid \hat{\lambda}_n, \, X_{1:n}) \to 0$ with $P_0^{(\infty)}$-probability one, where $P_0^{(\infty)}$ denotes the probability measure of (X_i) under θ_0. For $\theta \in \Theta$, let $R(p_\theta^{(n)}) = p_\theta^{(n)}(X_{1:n}) / p_{\theta_0}^{(n)}(X_{1:n})$ denote the likelihood ratio. We will use the following assumptions.

Assumption A1. There exists $\lambda \in \Lambda$ such that, for all $\eta > 0$, there is a measurable set $B_{\mathrm{KL}}(\theta_0; \, \eta) \subset \{\theta \in \Theta : \, \liminf_n \{n\eta + \log R(p_\theta^{(n)})\} = +\infty\}$ with $\Pi\{B_{\mathrm{KL}}(\theta_0; \, \eta) \mid \lambda\} > 0$.

Assumption A2. There exists a sequence $\Theta_n \subset \Theta$ such that

(i) the model $p_\theta^{(n)}$ is strongly regular in the sense that there exist constants $c_1, \, c_2 > 0$ such that, for every $\epsilon > 0$,

$$P_0^{(n)} \left\{ \sup_{\theta \in U_\epsilon^c \cap \Theta_n} R(p_\theta^{(n)}) \geqslant e^{-c_1 n \epsilon^2} \right\} \leqslant c_2 (n\epsilon^2)^{-(1+t)}$$

for some $t > 0$. The supremum may not be measurable: in this case, the preceding probability statement is understood in terms of the outer measure, see, e.g., van der Vaart (1998);

(ii) the empirical Bayes posterior probability $\Pi(\Theta_n^c \mid \hat{\lambda}_n, \, X_{1:n}) \to 0$ with $P_0^{(\infty)}$-probability one.

Assumption A1 is the usual Kullback-Leibler prior support condition considered in most results on posterior consistency. It is written in its generic form. In the case where for all $\theta \in \Theta$ $-\log\{R(p_\theta^{(n)})\}/n$ converges almost surely to a deterministic constant, typically the limit of its expectation under $P_{\theta_0}^{(n)}$, say $KL_\infty(\theta)$, then the set $\{\theta \in \Theta : \, \liminf_n \{n\eta + \log R(p_\theta^{(n)})\} = +\infty\}$ is given by $\{\theta; KL_\infty(\theta) < \eta\}$. It can also be proved following Lemma 10 in Ghosal & van der Vaart (2007a) with $\epsilon_n^2 = \epsilon^2$ and $k > 2$. In the parametric case, for independent and identically distributed observations, if the model p_θ is regular, see Johnson (1970), pages 852–853, it is satisfied if θ_0 is in the support of the prior for some $\lambda \in \Lambda$. In the nonparametric case, it has been shown to hold for various families of priors. In the present context, Assumption A1 is used when $\hat{\lambda}_n$ is the maximum marginal likelihood estimator to bound from below $\hat{m}(X_{1:n}) / p_{\theta_0}^{(n)}(X_{1:n})$. For other types of estimators $\hat{\lambda}_n$, a variant of Assumption A1 is considered, see condition (iii) in Proposition 3.

When compared to the assumptions usually considered for posterior consistency, condition (i) of Assumption A2 is quite strong; it is, however, a common assumption in the maximum likelihood estimation literature. It is verified for most parametric models, see, e.g., Lemma 2.3 in Johnson (1970), and for nonparametric models, for instance, Wong & Shen (1995) proved that, for independent and identically distributed observations with density p_θ, if d is the Hellinger metric, then a sufficient condition for (i) of Assumption A2 to hold is that there exist constants $c_3, \, c_4 > 0$ such that, for every $\epsilon > 0$,

$$\int_{\epsilon^2/2^8}^{\sqrt{2}\epsilon} H_{[]}^{1/2}(u/c_3, \, \Theta_n, \, d) \, \mathrm{d}u \leqslant c_4 n^{1/2} \epsilon^2 \tag{3.1}$$

for n large enough, where the function $H_{[]}(\cdot, \, \Theta_n, \, d)$ denotes the Hellinger bracketing metric entropy of Θ_n. Recall that if $\mathcal{F}$ is a set of non-negative, μ-integrable functions and d a metric

on this set, then an ϵ-bracketing (with respect to d) is a set of pairs of μ-integrable functions $(l_1, u_1), \ldots, (l_m, u_m)$ such that (i) for each $f \in \mathcal{F}$ there exists (l_j, u_j) so that $l_j \leqslant f \leqslant u_j$ μ-almost everywhere and (ii) $d(l_j, u_j) \leqslant \epsilon$ for every $j = 1, \ldots, m$. The smallest number of such brackets to cover $\mathcal{F}$ is called the bracketing number and is denoted by $N_{[]}(\epsilon, \mathcal{F}, d)$. The bracketing entropy is defined as $H_{[]}(\epsilon, \mathcal{F}, d) = \log N_{[]}(\epsilon, \mathcal{F}, d)$.

Condition (i) of Assumption A2 is used to prove that, $P_0^{(\infty)}$-almost surely, $\int_{U_\epsilon^c \cap \Theta_n} R(p_\theta^{(n)}) \, \mathrm{d}\Pi(\theta \mid \hat{\lambda}_n)$ is eventually exponentially small. In fact, by the first Borel-Cantelli lemma, condition (i) of Assumption A2 implies that, $P_0^{(\infty)}$-almost surely, $\sup_{\theta \in U_\epsilon^c \cap \Theta_n} R(p_\theta^{(n)}) < e^{-c_1 n \epsilon^2}$ for all large n, whence

$$\int_{U_\epsilon^c \cap \Theta_n} R(p_\theta^{(n)}) \, \mathrm{d}\Pi(\theta \mid \hat{\lambda}_n) \leqslant \sup_{\theta \in U_\epsilon^c \cap \Theta_n} R(p_\theta^{(n)}) < e^{-c_1 n \epsilon^2}. \tag{3.2}$$

Note that the bound in (3.2) is valid for any type of estimator $\hat{\lambda}_n$.

While condition (ii) of Assumption A2 trivially holds for parametric models where $\Theta_n = \Theta$, for nonparametric models, typically, it has to be checked case by case.

We begin to study consistency of the empirical Bayes posterior when $\hat{\lambda}_n$ is the maximum marginal likelihood estimator as defined in § 2.

PROPOSITION 2. *Under Assumptions A1 and A2, the empirical Bayes posterior* $\Pi(\cdot \mid \hat{\lambda}_n, X_{1:n})$*, where* $\hat{\lambda}_n$ *is the maximum marginal likelihood estimator, is consistent at* θ_0*.*

Proof. It suffices to show that, for every $\epsilon > 0$, the posterior probability $\Pi(U_\epsilon^c \mid \hat{\lambda}_n, X_{1:n})$ converges to zero with $P_0^{(\infty)}$-probability one. We write $\Pi(U_\epsilon^c \mid \hat{\lambda}_n, X_{1:n}) = \Pi(U_\epsilon^c \cap \Theta_n \mid \hat{\lambda}_n, X_{1:n}) + \Pi(U_\epsilon^c \cap \Theta_n^c \mid \hat{\lambda}_n, X_{1:n})$, where the second addendum converges to zero $P_0^{(\infty)}$-almost surely by condition (ii) of Assumption A2. As for the first term, we can write

$$\Pi(U_\epsilon^c \cap \Theta_n \mid \hat{\lambda}_n, X_{1:n}) = \frac{\int_{U_\epsilon^c \cap \Theta_n} R(p_\theta^{(n)}) \, \mathrm{d}\Pi(\theta \mid \hat{\lambda}_n)}{\int_\Theta R(p_\theta^{(n)}) \, \mathrm{d}\Pi(\theta \mid \hat{\lambda}_n)} = \frac{N_n}{D_n}.$$

All the following probability statements are understood to hold $P_0^{(\infty)}$-almost surely. By definition of $\hat{m}(X_{1:n})$, we have $D_n \geqslant m(X_{1:n} \mid \lambda)/p_{\theta_0}^{(n)}(X_{1:n}) \equiv D_n(\lambda)$ for all large n, where λ is as required in Assumption A1. Thus, $\Pi(U_\epsilon^c \cap \Theta_n \mid \hat{\lambda}_n, X_{1:n}) \leqslant N_n/D_n(\lambda)$. Under condition (i) of Assumption A2, by (3.2), $N_n < e^{-c_1 n \epsilon^2}$ for all large n. Reasoning as in Lemma 10 of Barron (1988), for any $\eta > 0$, $D_n(\lambda) > e^{-n\eta}$ for all large n. Choosing $0 < \eta < c_1 \epsilon^2$, for $\delta = (c_1 \epsilon^2 - \eta) > 0$, we have $\Pi(U_\epsilon^c \cap \Theta_n \mid \hat{\lambda}_n, X_{1:n}) = N_n/D_n \leqslant N_n/D_n(\lambda) < e^{-n\delta}$ for all large n and the assertion follows. □

In some applications, $\hat{\lambda}_n$ is chosen to be a convenient statistics rather than the maximum marginal likelihood estimator. As shown in Proposition 3 below, knowledge of the behaviour of $\hat{\lambda}_n$ allows to establish consistency for the empirical Bayes posterior without Assumption A2.

PROPOSITION 3. *Assume that*

(i) there exists a compact set $\mathcal{K} \subseteq \Lambda \subseteq \mathbb{R}^\ell$ *such that, with* $P_0^{(\infty)}$*-probability one,* $\hat{\lambda}_n \in \mathcal{K}$ *when* n *is large enough;*

(ii) for all $\lambda, \lambda' \in \mathcal{K}$*, there exists a measurable transformation* $\psi_{\lambda, \lambda'} : \Theta \to \Theta$ *such that if* $\theta \sim \Pi(\cdot \mid \lambda)$ *then* $\psi_{\lambda, \lambda'}(\theta) \sim \Pi(\cdot \mid \lambda')$*;*

(iii) for every $\delta > 0$ and $\lambda \in \mathcal{K}$, there exists a sequence u_n such that $u_n^{-\ell} > e^{-cn}$ for some constant $c > 0$ and a set $S \in \mathcal{B}(\Theta)$ such that $\inf_{\lambda \in \mathcal{K}} \Pi(S|\lambda) > 0$ and

$$\sum_{n=1}^{\infty} u_n^{-\ell} \sup_{\theta \in S} P_0^{(n)} \left\{ \inf_{\|\lambda - \lambda'\| \leqslant u_n} \log R(p^{(n)}_{\psi_{\lambda,\,\lambda'}(\theta)}) < -n\delta \right\} < \infty;$$

(iv) for every $\epsilon > 0$, there exist $\eta_0 > c$ and tests $\phi_n : \mathcal{X}^n \to [0,1]$ such that, for all $\lambda \in \mathcal{K}$, $\sum_{n=1}^{\infty} E_0^{(n)} \{\phi_n(X_{1:n})\} < \infty$, where $E_0^{(n)}$ denotes expectation under $P_0^{(n)}$, and

$$\int_{U_\epsilon^c} \int_{\mathcal{X}^n} \{1 - \phi_n(x_{1:n})\} \sup_{\|\lambda - \lambda'\| \leqslant u_n} p^{(n)}_{\psi_{\lambda,\,\lambda'}(\theta)}(x_{1:n})\, \mathrm{d}\mu^{(n)}(x_{1:n}) \mathrm{d}\Pi(\theta \mid \lambda) \leqslant e^{-n\eta_0}.$$

Then, for any $\epsilon > 0$, $\Pi(U_\epsilon^c \mid \hat{\lambda}_n,\, X_{1:n}) \to 0$ with $P_0^{(\infty)}$-probability one.

The proof of Proposition 3 is reported in the Supplementary Material. Condition (i) is a natural requirement when $\hat{\lambda}_n$ is an explicit estimator as opposed to the maximum marginal likelihood estimator, since $\hat{\lambda}_n$ typically converges to some $\lambda_0 \in \Lambda$ so that $\mathcal{K}$ is a neighborhood of λ_0. Condition (ii) characterizes the action of $\hat{\lambda}_n$ on the prior. By expressing a change from λ to λ' in the prior as a change from θ to $\psi_{\lambda,\,\lambda'}(\theta)$ in the likelihood, the dependence on the data is transferred from the prior to the likelihood. Conditions (ii)–(iv) are illustrated in Example 3 of § 3·2. Note that the same sequence u_n is considered in conditions (iii) and (iv), but this is not necessary. The general idea to show that conditions (iii) and (iv) are satisfied is to bound, for all λ, λ' such that $\|\lambda' - \lambda\| \leqslant u_n$, the density $p^{(n)}_{\psi_{\lambda,\,\lambda'}(\theta)}$ by a density g_θ times a term of the form $e^{o(n)}$.

As previously mentioned, the term empirical Bayes is also used in the literature with a different meaning from the one considered here; yet, the above results could be extended to cover these cases. The term empirical Bayes is often used in hierarchical models where $X_1, \ldots, X_p$ have joint density $\prod_{i=1}^p p_{\theta_i}(x_i)$, conditionally on $\theta_1, \ldots, \theta_p$, and θ_i are a random sample from a latent distribution $G(\cdot \mid \gamma)$ or, in a nonparametric setting, G. Usually, X_i is a sufficient statistics for the n_i observations $X_{i,1}, \ldots, X_{i,n_i}$, with $\sum_{i=1}^p n_i = n$. Since the latent distribution can be regarded as a prior on θ_i, maximum likelihood estimation of γ has been referred to as parametric empirical Bayes (Morris (1983)) or, as nonparametric empirical Bayes (Robbins (1956)) when based on nonparametric maximum likelihood estimation of G. This meaning of empirical Bayes differs from the one we consider in this work because it presumes the existence of a true γ_0 or a true mixing distribution G_0, whereas, in our context, there is generally no true value of the hyperparameter. However, our results apply to Bayesian inference for these problems where a (possibly nonparametric) prior $\Pi(G \mid \lambda)$ is assigned to the latent distribution and one wants an empirical Bayes selection of λ. We illustrate this in Examples 2 and 3 of § 3·2, where interest lies in estimating a mixture density. If, instead, interest lies in a specific parameter θ_i, our results could be extended to study the asymptotic behavior of the empirical Bayes posterior $G(\theta_i \mid \hat{\gamma}_{n,\,p},\, X_1, \ldots, X_p)$ obtained by plugging the maximum marginal likelihood estimate $\hat{\gamma}_{n,\,p} = \operatorname{argmax}_\gamma m(x_1, \ldots, x_p \mid \gamma)$, where $m(x_1, \ldots, x_p \mid \gamma) = \prod_{i=1}^p \int p_\theta(x_i)\, \mathrm{d}G(\theta \mid \gamma)$. Since this estimator exploits information from all experiments, it would not fit into the setting of Proposition 2, but Proposition 3 could be used to study the asymptotic behavior of $G(\theta_i \mid \hat{\gamma}_{n,\,p},\, X_1, \ldots, X_p)$, as $n_i,\, p \to \infty$, based on consistency results for the maximum likelihood estimator $\hat{\gamma}_{n,\,p}$.

3·2. *Examples*

Example 1. Consistency of the Bayesian posterior does not imply consistency of the empirical Bayes posterior, as shown by the following counterexample. Consider Bahadur (1958)'s

example, see also Lehmann & Casella (1998), pages 445–447, Ghosh & Ramamoorthi (2003), pages 29–31. Let X_i be independent and identically distributed random variables with values in $(0, 1]$ and density p_θ indexed by $\theta = 1, 2, \ldots$. For each θ, the density p_θ is constant on $(0,\, 1]$, except on the interval $(a_\theta,\, a_{\theta-1}]$ wherein is equal to $e^{x^{-2}}$. Define $a_0 = 1$ and a_θ by $\int_{a_\theta}^{a_{\theta-1}} (e^{x^{-2}} - C)\,\mathrm{d}x = 1 - C$, where $0 < C < 1$ is a given constant. Since $\int_0^1 e^{x^{-2}}\,\mathrm{d}x = \infty$, the a_θ are uniquely determined and a_θ tends to zero as $\theta \to \infty$. For each $\theta = 1, 2, \ldots$, define

$$p_\theta(x) = \begin{cases} e^{x^{-2}} & (x \in (a_\theta,\, a_{\theta-1}]), \\ C & (x \in (0, 1] \cap (a_\theta,\, a_{\theta-1}]^c), \end{cases}$$

and $p_\theta = 0$ otherwise. The maximum likelihood estimator $\hat\theta_n$ exists and tends to ∞ in probability, regardless of the true value θ_0 of θ. It is, therefore, inconsistent. On the other hand, Θ being countable, by Doob's theorem, any proper prior on Θ leads to a consistent posterior at all $\theta \in \Theta$. However, an empirical Bayes posterior for θ can be degenerate at $\hat\theta_n$, hence inconsistent. This may happen if, for some $\lambda \in \partial\Lambda$, the prior $\Pi(\cdot \mid \lambda)$ is degenerate at $\hat\theta_n$. Such a family of priors can be constructed, for example, by discretizing a Gaussian distribution with parameters μ and τ^2. For $\lambda = (\mu,\, \tau^2)$,

$$\begin{aligned} &\Pi(\{1\} \mid \lambda) = \Phi((-5/2,\, 1/2] \mid \lambda), \\ &\Pi(\{k\} \mid \lambda) = \Phi((k - 3/2,\, k - 1/2] \mid \lambda) + \Phi((-k - 1/2,\, -k - 3/2] \mid \lambda) \quad (k = 2,\, 3,\, \ldots), \end{aligned}$$

where $\Phi(\cdot \mid \mu,\, \tau^2)$ denotes the probability law of a Gaussian distribution with parameters μ and τ^2. For any fixed $\mu = 0,\, 1,\, \ldots$, letting $\tau \to 0$, we have as a limit the Dirac mass at $\mu + 1$ because $\Pi(\{\mu + 1\} \mid \lambda)$ converges to one, while, for any other $k \neq \mu + 1$, the probability $\Pi(\{k\} \mid \lambda)$ converges to zero.

If there exists $\lambda \in \bar\Lambda$ such that $\Pi(\cdot \mid \lambda)$ is degenerate at $\hat\theta_n$, such a λ is the maximum marginal likelihood estimator because $m(X_{1:n} \mid \lambda) \leqslant \prod_{i=1}^n p_{\hat\theta_n}(X_i)$. For the above defined prior, such a λ exists and it is given by $\hat\lambda_n = (\hat\theta_n - 1,\, 0)$, for which $\Pi(\cdot \mid \hat\lambda_n)$ is degenerate at $\hat\theta_n$. Consequently, also the empirical Bayes posterior is degenerate at $\hat\theta_n$, hence inconsistent.

Example 2. Popular Bayesian kernel methods for density estimation assume that X_i, given G, are independently distributed according to $p_G(\cdot) = \int_\Psi K_\psi(\cdot)\,\mathrm{d}G(\psi)$, with G having a nonparametric prior. Note that this model can be written as $X_i \mid \psi_i \sim K_{\psi_i}$ independently and $\psi_i \mid G \sim G$ independently, with p_G obtained by integrating out the ψ_i, as discussed at the end of § 3·1. Most popular models for univariate density estimation use a Gaussian kernel $K_\psi(\cdot) = \phi(\cdot \mid \mu,\, \sigma^2)$, where $\phi(\cdot \mid \mu,\, \sigma^2)$ denotes the Gaussian density with parameters μ and σ^2, and a Dirichlet process prior for G with base measure $\lambda\bar\alpha$, that is, $G \sim \mathrm{DP}(\lambda\bar\alpha)$, where λ is a positive scalar and $\bar\alpha$ is a probability measure on $(-\infty,\, \infty) \times (0,\, \infty)$. The choice of the scale parameter λ has a crucial impact on inference and this has suggested either to treat it as random by assigning to it a hyperprior in a hierarchical Bayesian approach or to select it by empirical Bayes (Liu (1996), McAuliffe *et al.* (2006)), which has computational advantages. In particular, Liu (1996) considers the maximum marginal likelihood estimator of λ for Dirichlet process mixtures of Binomial distributions, but his argument remains valid for more general kernels, see Petrone & Raftery (1997). Liu (1996) shows that the maximum marginal likelihood estimator $\hat\lambda_n$ is the solution of

$$\sum_{i=1}^n \frac{\lambda}{\lambda + i - 1} = E(C_n \mid \lambda,\, X_{1:n}), \tag{3.3}$$

where $E(C_n \mid \lambda,\, X_{1:n})$ is the expected number of occupied clusters under the posterior distribution, given λ. Even if the model is parameterized in the mixing distribution G, Dirichlet process mixtures of Gaussians are usually thought of as priors on spaces of densities over $\mathcal{X}$. If we assume that G and $\bar{\alpha}$ belong to the collection of probability measures $\mathcal{G} = \{G : \text{support}(G) \subseteq A \times [\underline{\sigma},\, \bar{\sigma}]\}$, with $A \subset \mathbb{R}$ a compact interval and $0 < \underline{\sigma} < \bar{\sigma} < \infty$, then, for d the Hellinger distance between any pair of densities $p_G,\, p_{G'}$, from Theorem 3.2 of Ghosal & van der Vaart (2001), the set of densities $\Theta = \{p_G :\, G \in \mathcal{G}\}$ has bracketing Hellinger metric entropy satisfying (3.1), so that condition (i) of Assumption A2 is verified for $\Theta_n = \Theta$. Furthermore, if the true density is itself a mixture of Gaussians p_{G_0}, with $G_0 \in \mathcal{G}$, then the Kullback-Leibler prior support condition, Assumption A1, is satisfied. The existence of a solution for (3.3) implies that the empirical Bayes posterior distribution for the density is well defined and, by Proposition 2, we get Hellinger consistency of the empirical Bayes posterior for the density.

Example 3. Consider again Bayesian density estimation based on Dirichlet process mixtures, where empirical Bayes is now used to choose a hyperparameter of the base measure of the Dirichlet process. We focus on location mixtures of Gaussians, with X_i independently distributed according to $p_{F,\,\sigma}(\cdot) = \int_{-\infty}^{\infty} \phi(\cdot \mid \mu,\, \sigma^2)\, \mathrm{d}F(\mu)$, conditionally on $(F,\, \sigma)$, and $F \sim \mathrm{DP}\{\alpha_{\mathbb{R}}\, N(\lambda,\, \tau^2)\}$, where $\alpha_{\mathbb{R}}$ is a positive constant, $N(\lambda,\, \tau^2)$ denotes the Gaussian distribution with parameters $\lambda,\, \tau^2$, and σ^2 has an inverse-gamma prior $\mathrm{IG}(a,\, b)$, with $a,\, b > 0$. The choice of an inverse-gamma distribution is only for the sake of simplicity, indeed, any prior on σ whose tails ensure posterior consistency of fully Bayes Dirichlet process mixtures would lead to the same result. We consider empirical Bayes selection of λ and a natural candidate is the sample mean $\hat{\lambda}_n = \bar{X}_n$. The empirical Bayes prior for F is $\mathrm{DP}\{\alpha_{\mathbb{R}}\, N(\bar{X}_n,\, \tau^2)\}$. We prove Hellinger consistency of the empirical Bayes posterior for the unknown density of the data using Proposition 3. As in Wu & Ghosal (2008), we assume that the sampling density p_0 is positive, continuous and bounded on $\mathbb{R}$ and satisfies

$$-\int_{-\infty}^{\infty} p_0(x) \log \left\{ \inf_{|t-x|<\delta} p_0(t) \right\} \mathrm{d}x < \infty, \qquad \int_{-\infty}^{\infty} x^{2+2\eta} p_0(x)\, \mathrm{d}x < \infty$$

for $\eta,\, \delta > 0$. Let $m_0 = E_0(X_1)$ be the mean of X_1 under p_0. Consider the compact $\mathcal{K} = [m_0 - 1,\, m_0 + 1]$. Then, with probability one, $\bar{X}_n \in \mathcal{K}$ for all large n and assumption (i) of Proposition 3 is satisfied. Here, Θ is the space of Lebesgue densities on the real line and the prior $\Pi(\cdot \mid \lambda)$ on Θ is induced by $\mathrm{DP}\{\alpha_{\mathbb{R}}\, N(\lambda,\, \tau^2)\} \times \mathrm{IG}(a,\, b)$ via the mapping $(F,\, \sigma) \mapsto p_{F,\,\sigma}$. To construct the transformation required by (ii) of Proposition 3 note that, by the stick-breaking representation of the Dirichlet process, $F = \sum_{j=1}^{\infty} p_j \xi_j$ almost surely, with $\xi_j \sim N(\lambda,\, \tau^2)$ independently. So, for any $\lambda,\, \lambda' \in \mathbb{R}$, we can let $\psi_{\lambda,\,\lambda'}(p_{F,\,\sigma}) = p_{F',\,\sigma}$, where $F' = \sum_{j=1}^{\infty} p_j \xi'_j$, with $\xi'_j = \xi_j - \lambda + \lambda' \sim N(\lambda',\, \tau^2)$ independently, so that $F' \sim \mathrm{DP}\{\alpha_{\mathbb{R}}\, N(\lambda',\, \tau^2)\}$. With abuse of notation, we write $\psi_{\lambda,\,\lambda'}(F,\, \sigma) = (F',\, \sigma)$ in place of $\psi_{\lambda,\,\lambda'}(p_{F,\,\sigma}) = p_{F',\,\sigma}$. Note that $p^{(n)}_{\psi_{\lambda,\,\lambda'}(F,\,\sigma)}(X_{1:n}) = p^{(n)}_{F',\,\sigma}(X_{1:n}) = p^{(n)}_{F,\,\sigma}(X_{1:n} - \lambda' + \lambda)$, where $(X_{1:n} - \lambda' + \lambda)$ stands for $(X_1 - \lambda' + \lambda,\, \ldots,\, X_n - \lambda' + \lambda)$. Conditions (iii) and (iv) of Proposition 3 can be shown to hold by bounding $p^{(n)}_{\psi_{\lambda,\,\lambda'}(F,\,\sigma)}(X_{1:n})$, or equivalently $p^{(n)}_{F,\,\sigma}(X_{1:n} - \lambda' + \lambda)$, from below and above for all $\lambda,\, \lambda'$ such that $|\lambda' - \lambda| \leqslant u_n$ and then using results in Wu & Ghosal (2008) together with the construction of tests as in Ghosal *et al.* (1999). Details are provided in the Supplementary Material.

While being very simple, the following parametric example is illuminating in showing that, even when consistency and weak merging hold, the empirical Bayes posterior may have different behaviours: it may diverge from any Bayesian posterior and underestimate the uncertainty on θ.

Example 4. Consider $X_i \mid \theta \sim N(\theta,\, \sigma^2)$ independently, with σ^2 known. This model satisfies Assumption A2 with $\Theta_n = \Theta$. Let $\theta \sim N(\mu,\, \tau^2)$.

Case 1. If τ^2 is fixed and $\lambda = \mu$ is estimated by the maximum marginal likelihood estimator, then $\hat{\lambda}_n = \bar{X}_n$ and the resulting empirical Bayes posterior is $N\{\bar{X}_n,\, (1/\tau^2 + n/\sigma^2)^{-1}\}$, which has a completely regular density. This sequence of posterior distributions can be seen to be consistent by direct computations. Thus, it merges weakly with any Bayesian posterior.

Case 2. Let us now consider empirical Bayes inference when the prior variance $\lambda = \tau^2$ is estimated by the maximum marginal likelihood estimator, the prior mean μ being fixed. Then, see, e.g., Lehmann & Casella (1998), page 263, $\sigma^2 + n\hat{\tau}_n^2 = \max\{\sigma^2,\, n(\bar{X}_n - \mu)^2\}$ so that $\hat{\tau}_n^2 = (\sigma^2/n)\max\{n(\bar{X}_n - \mu)^2/\sigma^2 - 1,\, 0\}$. The resulting posterior $\Pi(\cdot \mid \hat{\tau}_n^2,\, X_{1:n})$ is Gaussian with mean $\mu_n = (\sigma^2/n)/(\hat{\tau}_n^2 + \sigma^2/n)\mu + \hat{\tau}_n^2/(\hat{\tau}_n^2 + \sigma^2/n)\bar{X}_n$ and variance $(1/\hat{\tau}_n^2 + n/\sigma^2)^{-1}$. A hierarchical Bayesian approach would assign a prior to τ^2 like $1/\tau^2 \sim \text{Gamma}(a,\, b)$. This would lead to a Student's-t prior distribution for θ with flatter tails which may give better frequentist properties, see, e.g., Berger & Robert (1990), Berger & Strawderman (1996). However, the Student's-t prior is no longer conjugate and the empirical Bayes posterior is simpler to compute. Yet, the empirical Bayes posterior is only partially regular, in the sense that $\hat{\tau}_n^2$ can be equal to zero so that $\Pi(\cdot \mid \hat{\tau}_n^2,\, X_{1:n})$ can be degenerate at μ. Simple computations show that the probability that $\hat{\tau}_n = 0$ converges to zero when $\theta_0 \neq \mu$, but it remains strictly positive when $\theta_0 = \mu$. This suggests that, if $\theta_0 \neq \mu$, the hierarchical and the empirical Bayes posterior densities can asymptotically be close; however, if $\theta_0 = \mu$, there is a positive probability that the empirical Bayes and the Bayesian posterior distributions are singular. From a Bayesian perspective, the possible degeneracy of the empirical Bayes posterior is a pathological behaviour.

Case 3. One may object that, although unsatisfactory from a Bayesian viewpoint, the empirical Bayes posterior would be degenerate at the true value θ_0 of θ. However, the case where $\lambda = (\mu,\, \tau^2)$ shows that empirical Bayes may dramatically underestimate the posterior uncertainty. In this case, the maximum marginal likelihood estimator for λ is $\hat{\lambda}_n = (\bar{X}_n,\, 0)$. The posterior is then completely irregular, in the sense that it is always degenerate at $\bar{X}_n$. This is clearly an extreme example, but it is more general than the Gaussian case and applies, in particular, to location-scale family of priors. In fact, if the model $p_\theta^{(n)}$ admits a maximum likelihood estimator $\hat{\theta}_n$ and $\pi(\cdot \mid \lambda)$ is of the form $\tau^{-1}g\{(\cdot - \mu)/\tau\}$, with $\lambda = (\mu,\, \tau)$, for some unimodal density g which is maximum at zero, then $\hat{\lambda}_n = (\hat{\theta}_n,\, 0)$ and the empirical Bayes posterior is a point mass at $\hat{\theta}_n$. This shows that such families of priors should not be jointly used with maximum marginal likelihood empirical Bayes procedures.

Example 5. The counterintuitive behavior of the empirical Bayes posterior shown in Example 4 extends to the regression setting. Consider the canonical Gaussian regression model $Y = 1\alpha + X\beta + \varepsilon$, with $\varepsilon \sim N(0,\, \sigma^2 I_n)$, where $Y = (Y_1,\, \ldots,\, Y_n)^{\mathrm{T}}$ is the response vector, X is the $(n \times k)$-fixed design matrix of full rank k and I_n is the n-dimensional identity matrix. With abuse of notation, we also denote by X the design matrix whose columns have been re-centered so that $1^{\mathrm{T}}X = 0^{\mathrm{T}}$. Assume that $n^{-1}(X^{\mathrm{T}}X)$ converges to a positive definite matrix V as $n \to \infty$. A popular prior for $\theta = (\alpha,\, \beta,\, \sigma^2)$, especially in the variable selection literature, see, e.g., Clyde & George (2000), George & Foster (2000), is

$$\pi(\alpha,\, \sigma^2) \propto \sigma^{-2}, \qquad \beta \mid \sigma^2 \sim N\{0,\, g\sigma^2(X^{\mathrm{T}}X)^{-1}\}, \quad g > 0,$$

which is a modified version of the original Zellner (1986)'s g-prior. Since the choice of g has a crucial impact on the shrinking effect in estimation, data-driven choices of g have been suggested. An empirical Bayes selection of g based on the maximum marginal likelihood gives (see equation (9) in Liang *et al.* (2008)) $\hat{g}_n = \max\{F_n - 1,\, 0\}$, where $F_n = R^2(n-1-k)/\{(1-R^2)k\}$, R^2 being the coefficient of determination. Thus, $\hat{g}_n = 0$ if and only if $F_n \leqslant 1$. Suppose that Y is generated by the model with parameter values α_0, β_0, σ_0^2. It turns out that

$$\begin{cases} \liminf_{n\to\infty} P(\hat{g}_n = 0) = \liminf_{n\to\infty} P(F_n \leqslant 1) > 0 & (\beta_0 = 0), \\ \lim_{n\to\infty} P(\hat{g}_n > 0) = \lim_{n\to\infty} P(F_n > 1) = 1 & (\beta_0 \neq 0). \end{cases} \tag{3.4}$$

Detailed computations to prove (3.4) are provided in the Supplementary Material. When $\beta_0 \neq 0$, the probability that the empirical Bayes posterior is non-degenerate tends to one. However, when $\beta_0 = 0$, the probability that $\hat{g}_n = 0$ does not asymptotically vanish. In this case, with positive probability, the empirical Bayes posterior $\Pi(\cdot \mid \hat{g}_n,\, Y)$ for β is degenerate at β_0, hence singular with respect to the Bayesian posterior $\Pi(\cdot \mid g,\, Y)$ or any hierarchical Bayesian posterior.

4. Frequentist strong merging and asymptotic behavior of $\hat{\lambda}_n$

4·1. *Heuristics*

As illustrated in Examples 4 and 5, stronger forms of merging are needed to explain divergent behaviors between the empirical Bayes posterior, with the maximum marginal likelihood estimator, and Bayesian posterior distributions. In this section, we study frequentist strong merging, in the sense of Ghosh & Ramamoorthi (2003), of empirical Bayes and Bayesian posterior distributions. Two sequences Π_n and q_n of posterior distributions are said to merge strongly if their total variation distance converges to zero $P_0^{(\infty)}$-almost surely. We confine ourselves to the case where $\Theta \subseteq \mathbb{R}^k$, for finite k, and assume that, for every $\lambda \in \Lambda \subseteq \mathbb{R}^\ell$, with finite ℓ, the prior distribution $\Pi(\cdot \mid \lambda)$ has density $\pi(\cdot \mid \lambda)$ with respect to some σ-finite measure ν. Before stating a general result which describes the asymptotic behaviour of the empirical Bayes posterior, we present an informal argument to explain the heuristics behind it. Under usual regularity conditions on the model, the marginal likelihood can be thus approximated:

$$m(X_{1:n} \mid \lambda) = \pi(\theta_0 \mid \lambda) \times \frac{p_{\hat{\theta}_n}^{(n)}(X_{1:n})}{n^{k/2}}\{1 + o_p(1)\}.$$

If we could interchange the maximization and the limit, we would have

$$\operatorname*{argmax}_{\lambda}\, m(X_{1:n} \mid \lambda) = \operatorname*{argmax}_{\lambda}\, \pi(\theta_0 \mid \lambda) + o_p(1).$$

An interesting phenomenon occurs: the maximum marginal likelihood estimate asymptotically maximizes the prior density $\pi(\theta_0 \mid \lambda)$ of the true value θ_0 of the parameter θ. In other words, it selects the most interesting values of the prior hyperparameter λ. We call the set of values of λ maximizing $\pi(\theta_0 \mid \lambda)$ the prior oracle set of hyperparameters and denote it by Λ^*. In terms of strong merging, Λ^* may correspond to unpleasant values if the supremum is achieved for values of λ in the boundary $\partial\Lambda$ for which the prior is a point mass at θ_0. Then, the empirical Bayes posterior is degenerate. This is what happens in Cases 2 and 3 of Example 4 and in Example 5 and, more generally, when $\pi(\cdot \mid \lambda)$ is a location-scale family and λ contains the scale parameter. In such cases, the limit and the maximization cannot be interchanged. We now present these ideas more rigorously.

The map $g:\ \theta \mapsto \sup_{\lambda\in\Lambda}\ \pi(\theta \mid \lambda)$ from Θ to $\mathbb{R}^+$ induces a partition $\{\Theta_0,\ \Theta_0^c\}$ of Θ, with $\Theta_0 = \{\theta \in \Theta:\ g(\theta) < \infty\}$ and $\Theta_0^c = \{\theta \in \Theta:\ g(\theta) = \infty\}$. As illustrated in the above heuristic discussion and proved in § 4·2 and § 4·3 below, if $\theta_0 \in \Theta_0$, then the empirical Bayes posterior is regular, a case which we will refer to as the non-degenerate case; if, instead, $\theta_0 \in \Theta_0^c$, the empirical Bayes posterior is degenerate and fails to merge strongly with any regular Bayesian posterior, a case which we will refer to as the degenerate case.

4·2. *Non-degenerate case*

In the non-degenerate case, we provide sufficient conditions for the empirical Bayes posterior $\Pi(\cdot \mid \hat{\lambda}_n,\ X_{1:n})$, where $\hat{\lambda}_n$ is the maximum marginal likelihood estimator, to merge strongly with any oracle posterior $\Pi(\cdot \mid \lambda^*,\ X_{1:n})$. A consequence of Theorem 1 below is that the empirical Bayes posterior merges strongly with the Bayesian posterior corresponding to any hierarchical prior $\int_\Lambda \pi(\cdot \mid \lambda)h(\lambda)\,\mathrm{d}\lambda$ that is positive and continuous at θ_0; indeed, it merges strongly with the Bayesian posterior corresponding to any prior whose ν-density is positive and continuous at θ_0.

We need to introduce some more notation. For $\theta_0 \in \Theta_0$, we define the prior oracle set of hyperparameters $\Lambda^* = \{\lambda^* \in \Lambda:\ \pi(\theta_0 \mid \lambda^*) = g(\theta_0)\}$. For any pair of ν-densities $\pi,\ \pi'$ on Θ, let $\|\pi - \pi'\|_1$ denote the L_1-distance $\int_\Theta |\pi(\theta) - \pi'(\theta)|\,\mathrm{d}\nu(\theta)$.

THEOREM 1. *Suppose that $\theta_0 \in \Theta_0$. If Assumption A2 is satisfied for $\Theta_n = \Theta$ and*

(*i*) *the map $g:\ \theta \mapsto \sup_{\lambda\in\Lambda}\ \pi(\theta \mid \lambda)$ is positive and continuous at θ_0;*
(*ii*) *there exists a non-empty subset $\tilde{\Lambda}^*$ of Λ^* such that, for every $\lambda^* \in \tilde{\Lambda}^*$, the map $\theta \mapsto \pi(\theta \mid \lambda^*)$ is continuous at θ_0 and, for any $\epsilon,\ \eta > 0$, $\Pi\{U_\epsilon \cap B_{\mathrm{KL}}(\theta_0;\ \eta) \mid \lambda^*\} > 0$;*

then, for every $\lambda^* \in \tilde{\Lambda}^*$,

$$\frac{\hat{m}(X_{1:n})}{m(X_{1:n} \mid \lambda^*)} \to 1 \tag{4.1}$$

with $P_0^{(\infty)}$-probability one. If, in addition to the preceding assumptions,

(*iii*) *$\tilde{\Lambda}^* = \Lambda^*$ is included in the interior of Λ and, for any $\delta > 0$, there exist $\epsilon,\ \eta > 0$ so that*

$$\sup_{\theta\in U_\epsilon}\ \sup_{d(\lambda,\,\Lambda^*)>\delta} \frac{\pi(\theta \mid \lambda)}{g(\theta)} \leqslant 1 - \eta,$$

where $d(\lambda,\ \Lambda^*) = \inf_{\lambda^*\in\Lambda^*} \|\lambda - \lambda^*\|$;

then, $P_0^{(\infty)}$-almost surely,

$$d(\hat{\lambda}_n,\ \Lambda^*) \to 0 \tag{4.2}$$

and, for every $\lambda^* \in \Lambda^*$,

$$\|\pi(\cdot \mid \hat{\lambda}_n,\ X_{1:n}) - \pi(\cdot \mid \lambda^*,\ X_{1:n})\|_1 \to 0. \tag{4.3}$$

The proof of Theorem 1 is reported in the Supplementary Material. Convergence in (4.2) asserts that, if $\theta_0 \in \Theta_0$, then, with $P_0^{(\infty)}$-probability one, the maximum marginal likelihood estimator $\hat{\lambda}_n$ converges to the oracle set of hyperparameters Λ^*, thus, asymptotically giving the best selection of λ. Furthermore, by (4.3), strong merging between the empirical Bayes posterior and any oracle posterior $\Pi(\cdot \mid \lambda^*,\ X_{1:n})$ holds. By an adaptation to dependent data of Theorem 1.3.1 in Ghosh & Ramamoorthi (2003), pages 18–20, if $\pi(\cdot \mid \lambda^*)$ and any other ν-density q on

Θ are positive and continuous at θ_0, then the corresponding Bayesian posterior densities merge strongly. Then, under the assumptions of Theorem 1, by the triangular inequality, the empirical Bayes posterior merges strongly with the Bayesian posterior corresponding to any prior with the above properties.

As noted in § 3·1, Assumption A2 is satisfied for any regular parametric model, including any regular exponential family. Apart from Assumption A2, the conditions of Theorem 1 only concern the family of priors. For example, assume that $\Theta = \mathbb{R}$ and $\pi(\cdot \mid \lambda) = \tau^{-1} g\{(\cdot - \mu)/\tau\}$, where g is a continuously differentiable, positive, unimodal density, with mode at zero, such that there exists $x^* \neq 0$ so that $x^* \{\mathrm{d} \log g(x)/\mathrm{d}x\}|_{x=x^*} = -1$, which, for instance, holds for a Gaussian or a Student's-t density. If $\lambda = \mu$ and τ is fixed as in Case 1 of Example 4, $\lambda^* = \theta_0$ and $g(\theta_0) = \tau^{-1} g(0)$. Hence, (i) and (ii) are satisfied. Similarly, if $\theta_0 \neq \mu$, for $\lambda = \tau$ as in Case 2 of Example 4, then $\tau x^* = \theta_0 - \mu$ and conditions (i) and (ii) are met. Example 5 is another example where, if $\beta_0 \neq 0$, the conditions of Theorem 1 are satisfied.

It is worth noting that Theorem 1 also applies to the semi-parametric framework where the model is parameterized by (θ, ζ), with $\theta \in \Theta \subseteq \mathbb{R}^k$, for finite k, and ζ an infinite-dimensional component taking values in a separable metric space Z, and the prior for (θ, ζ) is of the form

$$\Pi(\mathrm{d}\theta, \mathrm{d}\zeta \mid \lambda) = \pi(\theta \mid \lambda)\, \mathrm{d}\nu(\theta) \times \Pi(\mathrm{d}\zeta \mid \theta).$$

Letting $U_\epsilon = \{\theta \in \Theta : \|\theta - \theta_0\| < \epsilon\}$ and $p_\theta^{(n)}(X_{1:n}) = \int_Z p_{\theta,\zeta}^{(n)}(X_{1:n})\, \mathrm{d}\Pi(\zeta \mid \theta)$, if condition (ii) of Theorem 1 is satisfied for Kullback-Leibler neighbourhoods of the whole parameter (θ_0, ζ_0) and $p_\theta^{(n)}$ satisfies Assumption A2, then (4.1)–(4.3) hold under conditions (i) and (iii) on the marginal prior density $\pi(\cdot \mid \lambda)$ for θ. This is of interest because if the marginal Bayesian posterior distribution for θ, given λ^*, satisfies the Bernstein-von Mises theorem, then also the marginal empirical Bayes posterior for θ satisfies the Bernstein-von Mises theorem. Assumption A2 on the model $p_\theta^{(n)}$ can be verified following the techniques developed in Bickel & Kleijn (2012) and Castillo (2012).

4·3. *Degenerate case and extension to the model choice framework*

Examples 4 and 5 in § 3·2 suggest that strong merging may fail when $g(\theta_0) = \infty$. We generalize this finding and show that such pathological behavior is not so much related to the sampling model $p_\theta^{(n)}$ but rather to the family of priors $\{\Pi(\cdot \mid \lambda) : \lambda \in \Lambda\}$. In the following theorem, we assume the dominating measure ν to be Lebesgue measure on Θ.

THEOREM 2. *Suppose that $\theta_0 \in \Theta_0^c$. If Assumption A2 is satisfied for $\Theta_n = \Theta$ and*

- (i) *there exists $\lambda^* \in \partial\Lambda$ such that $\Pi(\cdot \mid \lambda^*) = \delta_{\theta_0}$;*
- (ii) *with $P_0^{(n)}$-probability tending to one, $\hat{m}(X_{1:n}) \geqslant p_{\theta_0}^{(n)}(X_{1:n})$;*
- (iii) *the model admits a local asymptotic normality expansion in the following form: for every $\epsilon > 0$, there exists a set, with $P_0^{(n)}$-probability tending to one, wherein, uniformly in $\theta \in U_\epsilon$,*

$$l_n(\theta) - l_n(\hat{\theta}_n) \in -\frac{n(\theta - \hat{\theta}_n)^{\mathrm{T}} I(\theta_0)(\theta - \hat{\theta}_n)}{2}(1 \pm \epsilon),$$

where $\hat{\theta}_n$ denotes the maximum likelihood estimator, $l_n(\theta) = \log p_\theta^{(n)}$ and $I(\theta_0)$ is the Fisher information matrix at θ_0;

- (iv) *$l_n(\hat{\theta}_n) - l_n(\theta_0)$ converges in distribution to a χ^2-distribution with k degrees of freedom;*

then the empirical Bayes posterior $\Pi(\cdot \mid \hat{\lambda}_n,\, X_{1:n})$*, with the maximum marginal likelihood estimator, cannot merge strongly with the Bayesian posterior* $\Pi(\cdot \mid \lambda,\, X_{1:n})$*,* $\lambda \in \Lambda$*, corresponding to any prior Lebesgue density* $\pi(\cdot \mid \lambda)$ *which is positive and continuous at* θ_0*.*

Theorem 2 asserts that, under regularity conditions, if $\theta_0 \in \Theta_0^c$ and there exists $\lambda^* \in \partial\Lambda$ so that $\Pi(\cdot \mid \lambda^*) = \delta_{\theta_0}$, then the empirical Bayes posterior cannot merge strongly neither with the Bayesian posterior $\Pi(\cdot \mid \lambda^*,\, X_{1:n})$ nor with the Bayesian posterior corresponding to any prior with density that is positive and continuous at θ_0, in particular, with any smooth hierarchical prior.

Assumption A2 and conditions (iii) and (iv) are verified by regular models. When $\Theta = \mathbb{R}$, priors verifying (i) and (ii) are, for instance, those of the form $\tau^{-1} g\{(\cdot - \mu)/\tau\}$, with $\lambda = (\mu,\, \tau)$ or $\lambda = \tau$, for g satisfying the same conditions as stated in § 4·2. Consider the case where $\theta_0 = \mu$ and $\lambda = \tau$, then $\lambda^* = 0$ and (i) is satisfied. Moreover, for any sequence $(\lambda_p)_{p \geqslant 1}$ converging to zero, $P_0^{(\infty)}$-almost surely, $\hat{m}(X_{1:n}) \geqslant m(X_{1:n} \mid \lambda_p)$ and $m(X_{1:n} \mid \lambda_p)$ converges to $p_{\theta_0}^{(n)}(X_{1:n})$ so that (ii) is satisfied if, for all $x_{1:n}$, $p_\theta^{(n)}(x_{1:n})$ is a continuous and bounded function of θ. This has been illustrated in Cases 2 and 3 of Example 4 and in Example 5.

Theorem 2 is restricted to priors that are absolutely continuous with respect to Lebesgue measure, however, as stated in Proposition 4 below, a similar result holds for a model selection procedure. Consider a general model choice framework with competing models having densities $p_{j,\theta_j}^{(n)}$, with $\theta_j \in \Theta_j$, with respect to some dominating σ-finite measure $\mu^{(n)}$, for $j = 1,\, \ldots,\, J$. Let $\Pi_{\theta|j}$ denote a probability measure on Θ_j and consider a family of prior probability measures $\{\Pi_J(\cdot \mid \lambda) :\, \lambda \in \Lambda\}$ on $\{1,\, \ldots,\, J\}$, each one having probability mass function $\pi_J(\cdot \mid \lambda)$. Denote the marginal likelihood in the model j by $m_j(X_{1:n}) = \int_{\Theta_j} p_{j,\theta_j}^{(n)}(X_{1:n})\, \mathrm{d}\Pi_{\theta|j}(\theta_j)$. We assume that there exists a true parameter $\theta_0 \in \cup_{j=1}^J \Theta_j$ and denote by $j_0 \equiv j(\theta_0)$ the index of the model containing θ_0; in case of nested models, j_0 denotes the index of the smallest model containing θ_0. Denote by $P_0^{(\infty)}$ the probability law of (X_n) corresponding to $(j_0,\, \theta_0)$. Let $\hat{\lambda}_n$ be the maximum marginal likelihood estimator of λ, defined as $\hat{\lambda}_n = \operatorname{argmax}_{\lambda \in \bar{\Lambda}} m(X_{1:n} \mid \lambda)$, where $m(X_{1:n} \mid \lambda) = \sum_{j=1}^J m_j(X_{1:n}) \pi_J(j \mid \lambda)$. Let $\lambda^* = \operatorname{argmax}_{\lambda \in \bar{\Lambda}} \pi_J(j_0 \mid \lambda)$, where, for simplicity, we assume a unique point of maximum. The following result holds.

PROPOSITION 4. *In the above framework, assume that*

(i) for every $j \neq j_0$*, we have* $m_j(X_{1:n})/m_{j_0}(X_{1:n}) \to 0$ *with* $P_0^{(\infty)}$*-probability one;*

(ii) the prior on the model index is such that, if, $P_0^{(\infty)}$*-almost surely, for every sequence* λ_n*,* $\pi_J(j_0 \mid \lambda_n) \to \pi_J(j_0 \mid \lambda^*)$*, then* $\lambda_n \to \lambda^*$ *with* $P_0^{(\infty)}$*-probability one;*

then $\hat{\lambda}_n \to \lambda^*$ *with* $P_0^{(\infty)}$*-probability one.*

Proof. By definition of $\hat{\lambda}_n$,

$$0 \leqslant \frac{m(X_{1:n} \mid \hat{\lambda}_n) - m(X_{1:n} \mid \lambda^*)}{m_{j_0}(X_{1:n})} = \sum_{j=1}^J \frac{m_j(X_{1:n})}{m_{j_0}(X_{1:n})} \{\pi_J(j \mid \hat{\lambda}_n) - \pi_J(j \mid \lambda^*)\}$$
$$\leqslant \pi_J(j_0 \mid \hat{\lambda}_n) - \pi_J(j_0 \mid \lambda^*) + 2 \sum_{j \neq j_0} \frac{m_j(X_{1:n})}{m_{j_0}(X_{1:n})},$$

whence

$$0 \leqslant \frac{m(X_{1:n} \mid \hat{\lambda}_n) - m(X_{1:n} \mid \lambda^*)}{m_{j_0}(X_{1:n})} + \pi_J(j_0 \mid \lambda^*) - \pi_J(j_0 \mid \hat{\lambda}_n) \leqslant 2 \sum_{j \neq j_0} \frac{m_j(X_{1:n})}{m_{j_0}(X_{1:n})},$$

where the last sum converges to zero $P_0^{(\infty)}$-almost surely by assumption (i). Hence, with $P_0^{(\infty)}$-probability one, $\pi_J(j_0 \mid \hat{\lambda}_n) \to \pi_J(j_0 \mid \lambda^*)$ and, by assumption (ii), $\hat{\lambda}_n \to \lambda^*$. □

We tacitly assumed that the prior $\Pi_J(\cdot \mid \lambda)$ is non-degenerate for every λ in the interior of Λ; however, it may be degenerate for λ in the boundary $\partial\Lambda$. If $\lambda^* \in \partial\Lambda$ and $\Pi_J(\cdot \mid \lambda^*)$ is degenerate, it is degenerate at the true model index j_0. By Proposition 4, if follows that the empirical Bayes prior distribution $\Pi_J(\cdot \mid \hat{\lambda}_n)$ is asymptotically a point mass at j_0. It follows that the corresponding empirical Bayes posterior is degenerate at j_0 and cannot merge strongly with any regular Bayesian posterior.

Scott & Berger (2010)'s finding is a special case of this result. We briefly recall their setup. Consider a regression model $X_i = z_i^{\mathrm{T}}\beta + \epsilon_i$, where $\epsilon_i \sim N(0,\, \phi^{-1})$ independently and z_i is the $(k \times 1)$-vector of possible regressors. The aim is to select the best set of covariates among the k candidates. Variable selection is based on an inclusion vector $\gamma = (\gamma_1,\, \ldots,\, \gamma_k) \in \{0,\, 1\}^k$, where $\gamma_j = 1$ if the jth covariate is included. The prior on $\theta = (\beta,\, \phi,\, \gamma)$ is defined as $\pi(\theta \mid \lambda) = \pi(\beta,\, \phi \mid \gamma)\, \pi(\gamma \mid \lambda)$, where $\pi(\beta,\, \phi \mid \gamma)$ is degenerate on a space determined by γ, say of values $(\beta_\gamma,\, \phi)$, where β_γ has dimension $k_\gamma = \sum_{j=1}^k \gamma_j$. Given λ, the γ_j are independent Bernoulli random variables with $\pi(\gamma \mid \lambda) = \lambda^{k_\gamma}\, (1 - \lambda)^{k - k_\gamma}$. Here γ characterizes the model. We let $(\beta_0,\, \phi_0)$ be the true parameter values, β_0 denoting the true k-dimensional vector of regression coefficients with some elements possibly equal to zero, which also gives the vector of indicators $\gamma_0 \equiv \gamma_0(\beta_0,\, \phi_0)$ associated to the true model. Denote by $\beta_{0,\,\gamma}$ the restriction of β_0 to the coefficients present in the model γ. Since each model is regular, $P_0^{(\infty)}$-almost surely,

$$\frac{m_\gamma(X_{1:n})}{p^{(n)}_{\theta_0}} = \frac{c_\gamma \pi(\beta_{0,\,\gamma},\, \phi_0 \mid \gamma)}{n^{(k_\gamma + 1)/2}} \times \frac{p^{(n)}_{\hat{\beta}_\gamma,\, \hat{\phi}}(X_{1:n})}{p^{(n)}_{\beta_{0,\,\gamma},\, \phi_0}(X_{1:n})}\, (1 + o(1))\,,$$

where $\hat{\beta}_\gamma$, $\hat{\phi}$ are the maximum likelihood estimators in the model γ and c_γ is bounded. This implies that, for every $\gamma \neq \gamma_0$, we have $m_\gamma(X_{1:n})/m_{\gamma_0}(X_{1:n})$ converging to zero $P_0^{(\infty)}$-almost surely, thus condition (i) of Proposition 4 holds. It can be easily seen that also (ii) of Proposition 4 is satisfied. Hence, by Proposition 4, the maximum marginal likelihood estimator $\hat{\lambda}_n$ converges almost surely to $\lambda^* = \operatorname{argmax}_\lambda \pi(\gamma_0 \mid \lambda) = k_{\gamma_0}/k$. If $\gamma_0 = (0,\, \ldots,\, 0)$ then $\lambda^* = 0$; if, instead, $\gamma_0 = (1,\, \ldots,\, 1)$, then $\lambda^* = 1$. Both values correspond to degenerate distributions on γ. Thus, in model selection, the discrete nature of the problem does not prevent failing of strong merging between the empirical Bayes posterior and the posterior corresponding to either a hierarchical prior or a prior with a fixed λ.

ACKNOWLEDGEMENTS

We would like to thank the Editor, an Associate Editor and two anonymous referees for thoughtful and constructive comments that helped improving the original manuscript. This work originated from a question by Persi Diaconis. We are grateful to him and Jim Berger for stimulating discussions.

SUPPLEMENTARY MATERIAL

Supplementary material available at *Biometrika* online includes the proofs of Proposition 3 and of Theorems 1 and 2 together with technical derivations on Examples 3 and 5.

REFERENCES

BAHADUR, R. R. (1958). Examples of inconsistency of maximum likelihood estimates. *Sankhya* **20**, 207–210.

BARRON, A. R. (1988). The exponential convergence of posterior probabilities with implications for Bayes estimators of density functions. *University of Illinois at Urbana-Campaign, Technical Report* **7**, April 1988.

BELITSER, E. & ENIKEEVA, F. (2008). Empirical Bayesian test of the smoothness. *Math. Methods Statist.* **17**, 1–18.

BELITSER, E. & LEVIT, B. (2003). On the empirical Bayes approach to adaptive filtering. *Math. Methods Statist.* **12**, 131–154.

BERGER, J. O. & ROBERT, C. (1990). Subjective hierarchical Bayes estimation of a multivariate normal mean: on the frequentist interface. *Ann. Statist.* **18**, 617–651.

BERGER, J. O. & STRAWDERMAN, W. E. (1996). Choice of hierarchical priors: admissibility in estimation of normal means. *Ann. Statist.* **24**, 931–951.

BICKEL, P. J. & KLEIJN, B. J. K. (2012). The semiparametric Bernstein-von Mises theorem. *Ann. Statist.* **40**, 206–237.

BLACKWELL, D. & DUBINS, L. (1962). Merging of opinions with increasing information. *Ann. Math. Stat.* **33**, 882–886.

CASTILLO, I. (2012). A semiparametric Bernstein-von Mises theorem for Gaussian process priors. *Probab. Theory Related Fields*, **152**, 53–99.

CLYDE, M. A. & GEORGE, E. I. (2000). Flexible empirical Bayes estimation for wavelets. *J. R. Statist. Soc.* B **62**, 681–698.

CUI, W. & GEORGE, E. I. (2008). Empirical Bayes vs. fully Bayes variable selection. *J. Statist. Plann. Inference* **138**, 888–900.

DIACONIS, P. & FREEDMAN, D. (1986). On the consistency of Bayes estimates. *Ann. Statist.* **14**, 1–26.

FAVARO, S., LIJOI, A., MENA, R. H. & PRÜNSTER, I. (2009). Bayesian nonparametric inference for species variety with a two parameter Poisson-Dirichlet process prior. *J. R. Statist. Soc.* B, **71**, 993–1008.

GEORGE, E. I. & FOSTER, D. P. (2000). Calibration and empirical Bayes variable selection. *Biometrika* **87**, 731–747.

GHOSAL, S., GHOSH, J. K. & RAMAMOORTHI, R. V. (1999). Posterior consistency of Dirichlet mixtures in density estimation. *Ann. Statist.* **27**, 143-158.

GHOSAL, S. & 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**, 1233–1263.

GHOSAL, S. & VAN DER VAART, A. W. (2007). Convergence rates of posterior distributions for noniid observations. *Ann. Statist.* **35**, 192–223.

GHOSH, J. K. & RAMAMOORTHI, R. V. (2003). *Bayesian Nonparametrics.* New York: Springer-Verlag.

JOHNSON, R. A. (1970). Asymptotic expansions associated with posterior distributions. *Ann. Math. Statist.* **41**, 851–864.

KNAPIK, B. T., SZABÓ, B. T., VAN DER VAART, A. W. & VAN ZANTEN, J. H. (2012). Bayes procedures for adaptive inference in nonparametric inverse problems. `http://arxiv.org/abs/1209.3628`

LEHMANN, E. L. & CASELLA, G. (1998). *Theory of Point Estimation*, 2nd ed. New York: Springer-Verlag.

LIANG, F., PAULO, R., MOLINA, G., CLYDE, M. A. & BERGER, J. O. (2008). Mixtures of g-priors for Bayesian variable selection. *J. Amer. Statist. Assoc.* **103**, 410–423.

LIU, J. S. (1996). Nonparametric hierarchical Bayes via sequential imputation. *Ann. Statist.* **24**, 911–930.

MCAULIFFE, J. D., BLEI, D. M. & JORDAN, M. I. (2006). Nonparametric empirical Bayes for the Dirichlet process mixture model. *Stat. Comput.* **16**, 5–14.

MORRIS, C. N. (1983). Parametric empirical Bayes inference: theory and applications. *J. Amer. Statist. Assoc.* **78**, 47–55.

PETRONE, S. & RAFTERY, A. E. (1997). A note on the Dirichlet process prior in Bayesian nonparametric inference with partial exchangeability. *Statist. Prob. Letters* **36**, 69–83.

ROBBINS, H. (1956). An empirical Bayes approach to statistics. *Proc. Third Berkeley Symp. on Math. Statist. and Prob., Vol. 1*, 157–163. Univ. of Calif. Press, Berkeley and Los Angeles.

SCOTT, J. G. & BERGER, J. O. (2010). Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. *Ann. Statist.* **38**, 2587–2619.

VAN DER VAART, A. W. (1998). *Asymptotic Statistics*, Cambridge University Press.

WONG, W. H. & SHEN, X. (1995). Probability inequalities for likelihood ratios and convergence rates of sieves MLEs. *Ann. Statist.* **23**, 339–362.

WU, Y. & GHOSAL, S. (2008). Kullback Leibler property of kernel mixture priors in Bayesian density estimation. *Electron. J. Statist.* **2**, 298–331.

ZELLNER, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti*, eds. P. K. Goel & A. Zellner, 233–243. North-Holland/Elsevier, Amsterdam.

5. Supplementary Material

5·1. *Proof of Proposition* 3

By assumption (i), on a set with $P_0^{(\infty)}$-probability one, $\hat{\lambda}_n \in \mathcal{K} \subseteq \Lambda \subseteq \mathbb{R}^\ell$, which we cover with balls of radius u_n, the number K_n of such balls being of order $O(u_n^{-\ell})$. We denote by λ_j, $j = 1, \ldots, K_n$, the centers of these balls. For all ϵ, $\delta > 0$,

$$
\begin{aligned}
&P_0^{(n)}\{\Pi(U_\epsilon^c \mid \hat{\lambda}_n,\, X_{1:n}) > \epsilon,\, \hat{\lambda}_n \in \mathcal{K}\} \\
&\qquad \leqslant E_0^{(n)}\{\phi_n(X_{1:n})\} + \frac{1}{\epsilon}\sum_{j=1}^{K_n} E_0^{(n)}\left[\sup_{\|\lambda-\lambda_j\|\leqslant u_n}\{1-\phi_n(X_{1:n})\}\Pi(U_\epsilon^c \mid \lambda,\, X_{1:n})\right] \\
&\qquad \leqslant E_0^{(n)}\{\phi_n(X_{1:n})\} + \frac{1}{\epsilon}\sum_{j=1}^{K_n} P_0^{(n)}\left\{\inf_{\|\lambda-\lambda_j\|\leqslant u_n} D_n(\lambda) < e^{-n\delta}\Pi(S|\lambda_j)/2\right\} \\
&\qquad\qquad + \frac{2}{\epsilon}\sum_{j=1}^{K_n}\frac{e^{n\delta}}{\Pi(S|\lambda_j)} E_0^{(n)}\left[\sup_{\|\lambda-\lambda_j\|\leqslant u_n}\{1-\phi_n(X_{1:n})\}N_n(\lambda)\right],
\end{aligned}
$$

where

$$
D_n(\lambda) = \int_\Theta R(p_\theta^{(n)})\,\mathrm{d}\Pi(\theta \mid \lambda), \qquad N_n(\lambda) = \int_{U_\epsilon^c} R(p_\theta^{(n)})\,\mathrm{d}\Pi(\theta \mid \lambda).
$$

We now study the second and third terms of the above inequality. For all $j = 1, \ldots, K_n$ and $\|\lambda - \lambda_j\| \leqslant u_n$,

$$
\begin{aligned}
D_n(\lambda) &\geqslant \int_S R(p_\theta^{(n)})\,\mathrm{d}\Pi(\theta \mid \lambda) \\
&\geqslant \int_S \inf_{\|\lambda-\lambda_j\|\leqslant u_n} R(p^{(n)}_{\psi_{\lambda,\,\lambda_j}(\theta)})\,\mathrm{d}\Pi(\theta \mid \lambda_j) \\
&\geqslant e^{-n\delta}\int_S \mathbb{1}\left\{\inf_{\|\lambda-\lambda_j\|\leqslant u_n} R(p^{(n)}_{\psi_{\lambda,\,\lambda_j}(\theta)}) \geqslant e^{-n\delta}\right\}\mathrm{d}\Pi(\theta \mid \lambda_j),
\end{aligned}
$$

so that

$$
\begin{aligned}
&P_0^{(n)}\left\{\inf_{\|\lambda-\lambda'\|\leqslant u_n} D_n(\lambda) < e^{-n\delta}\Pi(S|\lambda')/2\right\} \\
&\qquad\qquad \leqslant \frac{2}{\Pi(S|\lambda')}\int_S P_0^{(n)}\left\{\inf_{\|\lambda-\lambda'\|\leqslant u_n} R(p^{(n)}_{\psi_{\lambda,\,\lambda'}(\theta)}) < e^{-n\delta}\right\}\mathrm{d}\Pi(\theta \mid \lambda').
\end{aligned}
$$

Similarly,

$$
\begin{aligned}
&E_0^{(n)}\left[\sup_{\|\lambda-\lambda_j\|\leqslant u_n}\{1-\phi_n(X_{1:n})\}N_n(\lambda)\right] \\
&\qquad \leqslant \int_{U_\epsilon^c}\int_{\mathcal{X}^n}\{1-\phi_n(x_{1:n})\}\sup_{\|\lambda-\lambda_j\|\leqslant u_n} p^{(n)}_{\psi_{\lambda,\,\lambda_j}(\theta)}(x_{1:n})\,\mathrm{d}\mu^{(n)}(x_{1:n})\,\mathrm{d}\Pi(\theta \mid \lambda_j) \leqslant e^{-n\eta_0}
\end{aligned}
$$

by assumption (iv). The proof is completed by combining partial results and using condition (iii).

5·2. *Details for the study of Example* 3

Recall that we consider Bayesian density estimation based on Dirichlet process mixtures, where empirical Bayes is used to choose a hyperparameter of the base measure of the Dirichlet process. We focus on location mixtures of Gaussians, with X_i independently distributed according to $p_{F,\sigma}(\cdot) = \int_{-\infty}^{\infty} \phi(\cdot \mid \mu, \sigma^2)\,\mathrm{d}F(\mu)$, conditionally on (F, σ), and $F \sim \mathrm{DP}\{\alpha_{\mathbb{R}}\, N(\lambda, \tau^2)\}$, $\sigma^2 \sim \mathrm{IG}(a, b)$, with $a, b > 0$. We consider empirical Bayes selection of λ based on $\hat{\lambda}_n = \bar{X}_n$. It is shown in § 3·2 that condition (i) of Proposition 3 is satisfied and that the transformation required in condition (ii) can be defined by exploiting the stick-breaking representation of the Dirichlet process as $\psi_{\lambda,\lambda'}(F, \sigma) = (F', \sigma)$, where $F = \sum_{j=1}^{\infty} p_j \xi_j \sim \mathrm{DP}\{\alpha_{\mathbb{R}}\, N(\lambda, \tau^2)\}$ and $F' = \sum_{j=1}^{\infty} p_j \xi'_j \sim \mathrm{DP}\{\alpha_{\mathbb{R}}\, N(\lambda', \tau^2)\}$. As $\phi(\cdot \mid \xi'_j, \sigma^2) = \phi(\cdot - \lambda' + \lambda \mid \xi_j, \sigma^2)$, we have $p^{(n)}_{\psi_{\lambda,\lambda'}(F,\sigma)}(X_{1:n}) = p^{(n)}_{F,\sigma}(X_{1:n} - \lambda' + \lambda)$, where $(X_{1:n} - \lambda' + \lambda) = (X_1 - \lambda' + \lambda, \ldots, X_n - \lambda' + \lambda)$. Thus, in order to bound $p^{(n)}_{\psi_{\lambda,\lambda'}(F,\sigma)}(X_{1:n})$ from above and below so to meet conditions (iii) and (iv) of Proposition 3, we can bound $p^{(n)}_{F,\sigma}(X_{1:n} - \lambda' + \lambda)$. For every $u_n > 0$,

$$\begin{aligned}\sup_{|\lambda-\lambda'|\leqslant u_n} p^{(n)}_{F,\sigma}(X_{1:n} - \lambda' + \lambda) &\leqslant \prod_{i=1}^{n} \int_{-\infty}^{\infty} \phi(X_i \mid \xi, \sigma^2) e^{u_n |X_i - \xi|/\sigma^2}\,\mathrm{d}F(\xi) \\ &= c^n_{n,\sigma} \prod_{i=1}^{n} \int_{-\infty}^{\infty} g_\sigma(X_i - \xi)\,\mathrm{d}F(\xi),\end{aligned}$$

where $g_\sigma(\cdot)$ is the probability density proportional to $\phi(y \mid 0, \sigma^2) e^{u_n |y|/\sigma^2}$ and $c_{n,\sigma} = \int_{-\infty}^{\infty} \phi(y \mid 0, \sigma^2) e^{u_n |y|/\sigma^2}\,\mathrm{d}y \leqslant e^{u_n^2/(2\sigma^2)}(1 + 2u_n/\sigma)$. Also,

$$\begin{aligned}\inf_{|\lambda-\lambda'|\leqslant u_n} p^{(n)}_{F,\sigma}(X_{1:n} - \lambda' + \lambda) &\geqslant \prod_{i=1}^{n} \int_{-\infty}^{\infty} \phi(X_i \mid \xi, \sigma^2) e^{-u_n |X_i - \xi|/\sigma^2}\,\mathrm{d}F(\xi) \\ &= \tilde{c}^n_{n,\sigma} \prod_{i=1}^{n} \int_{-\infty}^{\infty} \tilde{g}_\sigma(X_i - \xi)\,\mathrm{d}F(\xi),\end{aligned} \tag{5.1}$$

where $\tilde{g}_\sigma(\cdot)$ is the probability density proportional to $\phi(y \mid 0, \sigma^2) e^{-u_n |y|/\sigma^2}$ and $\tilde{c}_{n,\sigma} = \int_{-\infty}^{\infty} \phi(y \mid 0, \sigma^2) e^{-u_n |y|/\sigma^2}\,\mathrm{d}y \geqslant e^{-u_n^2/(2\sigma^2)}(1 - 2u_n/\sigma)$.

For fixed $\epsilon > 0$, choose $0 < \delta < \epsilon^2$ small enough and $a_n = n^{1/2}$. Consider the sieve sets $\mathcal{F}_n = \{(F, \sigma) : F([-a_n, a_n]) > 1 - \delta,\ \sigma \in [\underline{\sigma}_n, \bar{\sigma}_n]\}$ and $\Theta_n = \{p_{F,\sigma} : (F, \sigma) \in \mathcal{F}_n\}$, with $\underline{\sigma}_n = \underline{\sigma}_0 n^{-1/2}$ and $\bar{\sigma}_n = e^{n\bar{\sigma}_0}$, for constants $\underline{\sigma}_0, \bar{\sigma}_0 > 0$. Note that if $n^2 u_n = o(1)$, $c_{n,\sigma} = 1 + o(1)$ over $\mathcal{F}_n$ and over $\tilde{\mathcal{F}}_n = \{(F, \sigma) : F([-a_n, a_n]) = 1,\ \sigma \in [\underline{\sigma}_n, \bar{\sigma}_n]\}$,

$$\sup_{|\lambda-\lambda'|\leqslant u_n} p^{(n)}_{F,\sigma}(X_{1:n} + \lambda - \lambda') \leqslant e^{o(n) + n u_n \sum_{i=1}^{n} |X_i|} p^{(n)}_{F,\sigma}(X_{1:n}) \leqslant e^{o(n)} p^{(n)}_{F,\sigma}(X_{1:n})$$

on the event $\sum_{i=1}^{n} |X_i| \leqslant n w_n$ for any sequence w_n going to infinity slowly enough so that $n^2 u_n w_n = o(1)$. With probability one, this event occurs for n large enough. Then, using the tests constructed in Ghosal *et al.* (1999), condition (iv) of Proposition 3 is satisfied. Condition

(iii) comes from Theorem 2 of Wu & Ghosal (2008), combined with (5.1) and

$$\int_{-\infty}^{\infty} p_0(x) \log \left\{ \int_{-\infty}^{\infty} \tilde{g}_\sigma(x-\xi)\, \mathrm{d}F(\xi) \right\} \mathrm{d}x \geqslant \int_{-\infty}^{\infty} p_0(x) \log \left\{ \int_{-\infty}^{\infty} \phi(x \mid \xi,\, \sigma^2)\, \mathrm{d}F(\xi) \right\} \mathrm{d}x - \frac{u_n}{\sigma} \left[\int_{-\infty}^{\infty} |x| p_0(x)\, \mathrm{d}x + \sup\{|\xi| : \xi \in \mathrm{supp}(F)\} \right].$$

Consistency of the empirical Bayes posterior for the unknown density follows from Proposition 3.

5·3. *Example 5: Regression with g-priors*

We provide here more detailed computations for Example 5. Consider the canonical Gaussian regression model $Y = 1\alpha + X\beta + \varepsilon$, with $\varepsilon \sim N(0,\, \sigma^2 I_n)$, where $Y = (Y_1,\, \ldots,\, Y_n)^{\mathrm{T}}$ is the response vector, X is the $(n \times k)$-fixed design matrix of full rank k and I_n is the n-dimensional identity matrix. With abuse of notation, we also denote by X the design matrix whose columns have been re-centered so that $1^{\mathrm{T}} X = 0^{\mathrm{T}}$. Assume that $n^{-1}(X^{\mathrm{T}}X)$ converges to a positive definite matrix V as $n \to \infty$. A popular prior for $\theta = (\alpha,\, \beta,\, \sigma^2)$, especially in the variable selection literature, see, e.g., Clyde & George (2000), George & Foster (2000), is

$$\pi(\alpha,\, \sigma^2) \propto \sigma^{-2}, \qquad \beta \mid \sigma^2 \sim N(0,\, g\sigma^2 (X^{\mathrm{T}}X)^{-1}), \quad g > 0,$$

which is a modified version of the original Zellner (1986)'s g-prior. Since the choice of g has a crucial impact on the shrinking effect in estimation, data-driven choices of g have been suggested. An empirical Bayes selection of g based on the maximum marginal likelihood gives (see equation (9) in Liang *et al.* (2008)) $\hat{g}_n = \max\{F_n - 1,\, 0\}$, where $F_n = R^2(n-1-k)/\{(1-R^2)k\}$, R^2 being the coefficient of determination. Thus, $\hat{g}_n = 0$ if and only if $F_n \leqslant 1$. Suppose that Y is generated by the model with parameter values $\alpha_0,\, \beta_0,\, \sigma_0^2$. It turns out that

$$\begin{cases} \liminf_{n\to\infty} P(\hat{g}_n = 0) = \liminf_{n\to\infty} P(F_n \leqslant 1) > 0 & (\beta_0 = 0), \\ \lim_{n\to\infty} P(\hat{g}_n > 0) = \lim_{n\to\infty} P(F_n > 1) = 1 & (\beta_0 \neq 0). \end{cases} \tag{5.2}$$

Interestingly, when $\beta_0 = 0$, the probability that $\hat{g}_n$ takes the value zero which is in the boundary does not asymptotically vanish. Conversely, when $\beta_0 \neq 0$, the probability that the empirical Bayes posterior is non-degenerate tends to one. To prove (5.2), let $\hat{\beta}$ be the ordinary least squares estimator and

$$\tilde{F}_n = \frac{(\hat{\beta} - \beta_0)^{\mathrm{T}} (X^{\mathrm{T}}X)(\hat{\beta} - \beta_0)/k}{\mathrm{SSE}/(n-1-k)}.$$

If $\beta_0 = 0$, then $F_n \equiv \tilde{F}_n \to \chi^2_k/k$ almost surely because $\mathrm{SSE}/(n-k-1) \to \sigma_0^2$ almost surely, and $\liminf_{n\to\infty} P(\hat{g}_n = 0) = P(\chi^2_k/k \leqslant 1) > 0$.

If $\beta_0 \neq 0$, from consistency of $\hat{\beta}$,

$$R_n = \frac{n^{-1}\{(\beta_0 - 2\hat{\beta})^{\mathrm{T}} (X^{\mathrm{T}}X)\beta_0\}/k}{\mathrm{SSE}/(n-1-k)} \to -\frac{(\beta_0^{\mathrm{T}} V \beta_0)/k}{\sigma_0^2} < 0$$

almost surely, which implies $1 + nR_n \to -\infty$. Consequently,

$$P(\hat{g}_n > 0) = P(F_n > 1) = P\left[\tilde{F}_n > 1 + \frac{\{(\beta_0 - 2\hat{\beta})^{\mathrm{T}} (X^{\mathrm{T}}X)\beta_0\}/k}{\mathrm{SSE}/(n-1-k)} \right] = P(\tilde{F}_n > 1 + nR_n) \to 1.$$

The consequences of (5.2) on strong merging are analyzed. By direct computations, whatever $\beta_0 \in \mathbb{R}^k$, for each $g > 0$, the Bayesian posterior $\Pi(\cdot \mid g, Y)$ for β is consistent at β_0, i.e., $\Pi(\cdot \mid g, Y)$ converges weakly to a point mass at β_0 with probability one. Let $\Omega_n = \{\hat{g}_n = 0\}$. Clearly, $\Omega_n \subseteq \{\Pi(\cdot \mid \hat{g}_n, Y) = \delta_0\}$. If $\beta_0 = 0$, then, for each $g > 0$, $\liminf_{n\to\infty} P(d_{\mathrm{TV}}(\Pi(\cdot \mid g, Y), \Pi(\cdot \mid \hat{g}_n, Y)) = 1) > 0$, where $d_{\mathrm{TV}}(\cdot, \cdot)$ denotes the total variation distance. Therefore, there exists a set with positive probability wherein strong merging cannot take place. If $\beta_0 \neq 0$, for every $g > 0$, by direct computations, $\|\pi(\cdot \mid g, Y) - \pi(\cdot \mid \hat{g}_n, Y)\|_1 = \int_{\mathbb{R}^k} |\pi(\beta \mid g, Y) - \pi(\beta \mid \hat{g}_n, Y)| \,\mathrm{d}\beta \to 0$ in probability. Strong merging takes place on a set with probability tending to one.

5·4. *Proof of Theorem* 1

We begin by proving (4.1). From (ii), for each $\lambda^* \in \tilde{\Lambda}^*$, $P_0^{(\infty)}$-almost surely, for all large n, $m(X_{1:n} \mid \lambda^*) > 0$ and, by definition of $\hat{\lambda}_n$, $0 < m(X_{1:n} \mid \lambda^*) \leqslant \hat{m}(X_{1:n}) < \infty$, whence

$$\frac{\hat{m}(X_{1:n})}{m(X_{1:n} \mid \lambda^*)} \geqslant 1. \tag{5.3}$$

We prove the reverse inequality. Using (i) of Assumption A2, (i) and (ii), for any $\delta > 0$, there exists $\epsilon > 0$ (depending on δ, θ_0 and $g(\theta_0)$) so that, with probability greater than or equal to $1 - c_2(n\epsilon^2)^{-(1+t)}$, for every $\lambda \in \Lambda$,

$$\begin{aligned}
\frac{m(X_{1:n} \mid \lambda)}{p_{\theta_0}^{(n)}(X_{1:n})} &< e^{-c_1 n\epsilon^2} + \int_{U_\epsilon} R(p_\theta^{(n)})\pi(\theta \mid \lambda) \,\mathrm{d}\nu(\theta) \\
&\leqslant e^{-c_1 n\epsilon^2} + \int_{U_\epsilon} R(p_\theta^{(n)}) g(\theta) \,\mathrm{d}\nu(\theta) \\
&< e^{-c_1 n\epsilon^2} + (1 + \delta/3) \int_{U_\epsilon} R(p_\theta^{(n)}) g(\theta_0) \,\mathrm{d}\nu(\theta) \\
&< e^{-c_1 n\epsilon^2} + (1 + 2\delta/3) \int_{U_\epsilon} R(p_\theta^{(n)})\pi(\theta \mid \lambda^*) \,\mathrm{d}\nu(\theta),
\end{aligned}$$

where the second inequality descends from the definition of g because $\pi(\theta \mid \lambda) \leqslant g(\theta)$ for all $\theta \in U_\epsilon$, the third one from the positivity and continuity of g at θ_0 and the last one from the fact that $g(\theta_0) = \pi(\theta_0 \mid \lambda^*)$, together with the continuity of $\pi(\theta \mid \lambda^*)$ at θ_0. By the first Borel-Cantelli lemma, for any $\delta > 0$, there exists $\epsilon > 0$ so that for all large n, for every $\lambda \in \Lambda$, $m(X_{1:n} \mid \lambda)/p_{\theta_0}^{(n)}(X_{1:n}) < e^{-c_1 n\epsilon^2} + (1 + 2\delta/3) \int_{U_\epsilon} R(p_\theta^{(n)})\pi(\theta \mid \lambda^*) \,\mathrm{d}\nu(\theta)$ for all large n, $P_0^{(\infty)}$-almost surely. The Kullback-Leibler support condition on $\Pi(\cdot \mid \lambda^*)$ implies that, on a set of $P_0^{(\infty)}$-probability one, for any constant $a > 0$,

$$\int_{U_\epsilon} R(p_\theta^{(n)})\pi(\theta \mid \lambda^*) \,\mathrm{d}\nu(\theta) > e^{-an} \tag{5.4}$$

for all large n. Therefore, for any $\delta > 0$, on a set of $P_0^{(\infty)}$-probability one, for every $\lambda \in \Lambda$, $m(X_{1:n} \mid \lambda) \leqslant (1 + \delta) m(X_{1:n} \mid \lambda^*)$ for all large n, which, combined with (5.3), proves (4.1).

We now prove the convergence of $\hat{\lambda}_n$. Recall that, by (i) of Assumption A2, for any $\epsilon > 0$, on a set of $P_0^{(\infty)}$-probability one, for every $\lambda \in \Lambda$, $m(X_{1:n} \mid \lambda)/p_{\theta_0}^{(n)}(X_{1:n}) < e^{-c_1 n\epsilon^2} + \int_{U_\epsilon} R(p_\theta^{(n)})\pi(\theta \mid \lambda) \,\mathrm{d}\nu(\theta)$ for all large n. For $\delta > 0$, define $N_\delta = \{\lambda \in \Lambda : d(\lambda, \Lambda^*) \leqslant \delta\}$. For

any fixed $\delta > 0$, by (iii), there exist $\epsilon_1,\ \eta > 0$ so that, on a set of $P_0^{(\infty)}$-probability one,

$$\sup_{\lambda \in N_\delta^c} \frac{m(X_{1:n} \mid \lambda)}{p_{\theta_0}^{(n)}(X_{1:n})} < e^{-c_1 n \epsilon_1^2} + (1-\eta) \int_{U_{\epsilon_1}} R(p_\theta^{(n)}) g(\theta)\, \mathrm{d}\nu(\theta)$$

for all large n, whence, using (i) and (ii) on the continuity of g and $\pi(\cdot \mid \lambda^*)$, $\lambda^* \in \tilde{\Lambda}^*$, at θ_0,

$$\sup_{\lambda \in N_\delta^c} \frac{m(X_{1:n} \mid \lambda)}{p_{\theta_0}^{(n)}(X_{1:n})} < e^{-c_1 n \epsilon_1^2} + (1-\eta/2) \frac{m(X_{1:n} \mid \lambda^*)}{p_{\theta_0}^{(n)}(X_{1:n})}$$

for all large n. Using (5.4), we finally get that $\sup_{\lambda \in N_\delta^c} m(X_{1:n} \mid \lambda) < (1-\eta/4) m(X_{1:n} \mid \lambda^*)$ for all large n, $P_0^{(\infty)}$-almost surely. The fact that η is fixed implies that, with $P_0^{(\infty)}$-probability one, $\hat{\lambda}_n \in N_\delta$ for n large enough. Since Λ^* is included in the interior of Λ, with $P_0^{(\infty)}$-probability one, $\hat{\lambda}_n$ belongs to the interior of Λ and $\Pi(\cdot \mid \hat{\lambda}_n) \ll \nu$ for all large n. This fact, combined with consistency of the empirical Bayes posterior and of every oracle posterior $\Pi(\cdot \mid \lambda^*,\, X_{1:n})$, and the convergence in (4.1), yields that, $P_0^{(\infty)}$-almost surely, for any $\epsilon > 0$,

$$\begin{aligned}
\|\pi(\cdot \mid \hat{\lambda}_n,\, X_{1:n}) - \pi(\cdot \mid \lambda^*,\, X_{1:n})\|_1 &\leqslant \epsilon + \int_{U_\epsilon} p_\theta^{(n)}(X_{1:n}) \left| \frac{\pi(\theta \mid \hat{\lambda}_n)}{\hat{m}(X_{1:n})} - \frac{\pi(\theta \mid \lambda^*)}{m(X_{1:n} \mid \lambda^*)} \right| \mathrm{d}\nu(\theta) \\
&\leqslant \epsilon + \left| \frac{\hat{m}(X_{1:n})}{m(X_{1:n} \mid \lambda^*)} - 1 \right| \\
&\qquad + \int_{U_\epsilon} \frac{p_\theta^{(n)}(X_{1:n})}{\hat{m}(X_{1:n})} |\pi(\theta \mid \hat{\lambda}_n) - \pi(\theta \mid \lambda^*)|\, \mathrm{d}\nu(\theta) \\
&\leqslant 2\epsilon + \int_{U_\epsilon} \frac{p_\theta^{(n)}(X_{1:n})}{\hat{m}(X_{1:n})} |\pi(\theta \mid \hat{\lambda}_n) - \pi(\theta \mid \lambda^*)|\, \mathrm{d}\nu(\theta)
\end{aligned}$$

for n large enough. We split U_ϵ into $D_\epsilon = \{\theta \in U_\epsilon :\, \pi(\theta \mid \hat{\lambda}_n) \geqslant \pi(\theta \mid \lambda^*)\}$ and $D_\epsilon^c = \{\theta \in U_\epsilon :\, \pi(\theta \mid \hat{\lambda}_n) < \pi(\theta \mid \lambda^*)\}$. Since, for any $\delta > 0$, if ϵ is small enough, $\pi(\theta \mid \hat{\lambda}_n) \leqslant \pi(\theta \mid \lambda^*)(1 + \delta/3)$,

$$\begin{aligned}
\int_{D_\epsilon} p_\theta^{(n)}(X_{1:n}) \{\pi(\theta \mid \hat{\lambda}_n) - \pi(\theta \mid \lambda^*)\}\, \mathrm{d}\nu(\theta) &\leqslant \frac{\delta}{3} \int_{D_\epsilon} p_\theta^{(n)}(X_{1:n}) \pi(\theta \mid \lambda^*)\, \mathrm{d}\nu(\theta) \\
&\leqslant \frac{\delta}{3} \hat{m}(X_{1:n}).
\end{aligned} \tag{5.5}$$

From consistency of the empirical Bayes posterior,

$$\begin{aligned}
\int_{U_\epsilon} p_\theta^{(n)}(X_{1:n}) \pi(\theta \mid \lambda^*)\, \mathrm{d}\nu(\theta) \leqslant \hat{m}(X_{1:n}) < \int_{U_\epsilon} p_\theta^{(n)}(X_{1:n}) \pi(\theta \mid \hat{\lambda}_n)\, \mathrm{d}\nu(\theta) \\
+ (\epsilon + \delta/3) \hat{m}(X_{1:n}),
\end{aligned}$$

whence

$$\begin{aligned}
\int_{D_\epsilon^c} p_\theta^{(n)}(X_{1:n}) \{\pi(\theta \mid \lambda^*) - \pi(\theta \mid \hat{\lambda}_n)\}\, \mathrm{d}\nu(\theta) \leqslant \int_{D_\epsilon} p_\theta^{(n)}(X_{1:n}) \{\pi(\theta \mid \hat{\lambda}_n) - \pi(\theta \mid \lambda^*)\}\, \mathrm{d}\nu(\theta) \\
+ (\epsilon + \delta/3) \hat{m}(X_{1:n})
\end{aligned}$$

and, using (5.5), $\int_{D_\epsilon^c} p_\theta^{(n)}(X_{1:n})\{\pi(\theta \mid \lambda^*) - \pi(\theta \mid \hat{\lambda}_n)\}\,\mathrm{d}\nu(\theta) \leqslant (\epsilon + 2\delta/3)\hat{m}(X_{1:n})$, which implies that, for all large n,

$$\int_{U_\epsilon} \frac{p_\theta^{(n)}(X_{1:n})}{\hat{m}(X_{1:n})} |\pi(\theta \mid \hat{\lambda}_n) - \pi(\theta \mid \lambda^*)|\,\mathrm{d}\nu(\theta) \leqslant (\epsilon + \delta).$$

Thus, (4.3) is proved and the proof is complete.

5·5. *Proof of Theorem* 2

Define, for any $\delta > 0$, the set $\Omega_{n,\delta}$ of $x_{1:n}$ such that $e^{l_n(\hat{\theta}_n) - l_n(\theta_0)} \leqslant 1 + \delta$. From assumption (iv), for every $\delta > 0$, $\liminf_{n\to\infty} P_0^{(n)}(\Omega_{n,\delta}) > 0$. From assumption (ii), $\hat{m}(X_{1:n})/p_{\theta_0}^{(n)}(X_{1:n}) \geqslant 1$. We now study the reverse inequality. Using condition (i) of Assumption A2, for any $\epsilon > 0$, on a set A_n with $P_0^{(n)}$-probability converging to one, $\hat{m}(X_{1:n})/p_{\theta_0}^{(n)}(X_{1:n}) = \int_{U_\epsilon} e^{l_n(\theta) - l_n(\theta_0)}\,\mathrm{d}\Pi(\theta \mid \hat{\lambda}_n) + O(e^{-n\delta})$. Moreover, using condition (iii), for every $\theta \in U_\epsilon$,

$$l_n(\theta) - l_n(\theta_0) = l_n(\hat{\theta}_n) - l_n(\theta_0) + \frac{-n(\theta - \hat{\theta}_n)^{\mathrm{T}} I(\theta_0)(\theta - \hat{\theta}_n)}{2}(1 + o_p(1)),$$

so that, if $M_n = M\{(\log n)/n\}^{1/2}$, with $M > 0$, on a set of $P_0^{(n)}$-probability going to one, for all $H > 0$, $\int_{\|\theta - \hat{\theta}_n\| \geqslant M_n} e^{l_n(\theta) - l_n(\hat{\theta}_n)}\,\mathrm{d}\Pi(\theta \mid \hat{\lambda}_n) = O(n^{-H})$ provided M is large enough. This leads to

$$\frac{\hat{m}(X_{1:n})}{p_{\theta_0}^{(n)}(X_{1:n})} = e^{l_n(\hat{\theta}_n) - l_n(\theta_0)} \int_{U_{M_n}} e^{-n(\theta - \hat{\theta}_n)^{\mathrm{T}} I(\theta_0)(\theta - \hat{\theta}_n)/2}\,\mathrm{d}\Pi(\theta \mid \hat{\lambda}_n) + O(n^{-H}),$$

where $U_{M_n} = \{\theta : \|\theta - \hat{\theta}_n\| < M_n\}$. With abuse of notation, we still denote by A_n the set having $P_0^{(n)}$-probability going to one wherein the above computations are valid, so that, on $A_n \cap \Omega_{n,\delta}$, for n large enough, $\hat{m}(X_{1:n})/p_{\theta_0}^{(n)}(X_{1:n}) \leqslant 1 + 2\delta$. Let $\lambda \in \Lambda$ be such that the prior Lebesgue density $\pi(\cdot \mid \lambda)$ is positive and continuous at θ_0. Under (iii) and condition (i) of Assumption A2, usual Laplace expansion of the marginal distribution of $X_{1:n}$ yields

$$\frac{m(X_{1:n} \mid \lambda)}{p_{\theta_0}^{(n)}(X_{1:n})} = \frac{\pi(\theta_0 \mid \lambda) e^{l_n(\hat{\theta}_n) - l_n(\theta_0)} (2\pi)^{k/2}}{n^{k/2} |I(\theta_0)|^{1/2}} \{1 + o_p(1)\},$$

so that $m(X_{1:n} \mid \lambda)/\hat{m}(X_{1:n}) = o_p(1)$. We now study the total variation distance between the two posteriors. If $\Pi(\cdot \mid \hat{\lambda}_n)$ is degenerate, that is, it is not absolutely continuous with respect to Lebesgue measure, then the total variation distance between the empirical Bayes posterior and the Bayesian posterior corresponding to the prior $\Pi(\cdot \mid \lambda)$ is equal to one. Thus, we only need to consider the case where $\Pi(\cdot \mid \hat{\lambda}_n)$ is absolutely continuous with respect to Lebesgue measure. On a set of $P_0^{(n)}$-probability going to one, which we still denote by A_n, intersected with $\Omega_{n,\delta}$,

for each $\theta \in U_{M_n}$,

$$\begin{aligned}\pi(\theta \mid \hat{\lambda}_n,\, X_{1:n}) - \pi(\theta \mid \lambda,\, X_{1:n}) &= e^{l_n(\theta)-l_n(\hat{\theta}_n)} \\ &\quad \times \left\{ e^{l_n(\hat{\theta}_n)-l_n(\theta_0)} \pi(\theta \mid \hat{\lambda}_n) - \frac{n^{k/2}|I(\theta_0)|^{1/2}}{(2\pi)^{k/2}} + o_p(1) \right\} \\ &= e^{-n(\theta-\hat{\theta}_n)^{\mathrm{T}} I(\theta_0)(\theta-\hat{\theta}_n)/2} \frac{n^{k/2}|I(\theta_0)|^{1/2}}{(2\pi)^{k/2}} (1 + o_p(1)) \\ &\quad \times \left\{ e^{l_n(\hat{\theta}_n)-l_n(\theta_0)} \pi(\theta \mid \hat{\lambda}_n) \frac{(2\pi)^{k/2}}{n^{k/2}|I(\theta_0)|^{1/2}} - 1 \right\}.\end{aligned}$$

Set $u = n^{1/2} I(\theta_0)^{1/2}(\theta - \hat{\theta}_n)$ and define $V_n = \{u : g_n(u) \geqslant 1 - 2\delta\}$, where $g_n(u) = (2\pi)^{k/2} n^{-k/2} |I(\theta_0)|^{-1/2} \pi(\hat{\theta}_n + I(\theta_0)^{-1/2} u n^{-1/2} \mid \hat{\lambda}_n)$. To simplify the notation, we also denote by $V_n = \{\theta = \hat{\theta}_n + I(\theta_0)^{-1/2} u n^{-1/2} : u \in V_n\}$. Then, for every $c > 0$, $\int_{V_n \cap \{\|u\| < cM_n n^{1/2}\}} g_n(u)\, \mathrm{d}u = (2\pi)^{k/2} \int_{V_n \cap \{\|\theta - \hat{\theta}_n\| < cM_n n^{1/2}\}} \pi(\theta \mid \hat{\lambda}_n)\, \mathrm{d}\theta \leqslant (2\pi)^{k/2}$ and, by definition of V_n, $\int_{V_n \cap \{\|u\| < cM_n n^{1/2}\}} g_n(u)\, \mathrm{d}u \geqslant (1 - 2\delta) \int_{V_n \cap \{\|u\| < cM_n n^{1/2}\}} \mathrm{d}u$. Hence

$$\int_{V_n \cap \{\|u\| < cM_n n^{1/2}\}} \mathrm{d}u \leqslant (2\pi)^{k/2} (1 - 2\delta)^{-1}. \tag{5.6}$$

Note that on V_n^c, $\pi(\theta \mid \hat{\lambda}_n)(2\pi)^{k/2} n^{-k/2} |I(\theta_0)|^{-1/2} < 1 - 2\delta$, so that $\pi(\theta \mid \hat{\lambda}_n)(1 + \delta)(2\pi)^{k/2} n^{-k/2} |I(\theta_0)|^{-1/2} - 1 < -\delta$ and we can bound from below the L_1-distance between the two posterior densities. On $A_n \cap \Omega_{n,\delta}$, since $I(\theta_0)$ is positive definite,

$$\begin{aligned}\|\pi(\cdot \mid \hat{\lambda}_n,\, X_{1:n}) - \pi(\cdot \mid \lambda,\, X_{1:n})\|_1 &\geqslant \int_{V_n^c \cap U_{M_n}} |\pi(\theta \mid \hat{\lambda}_n,\, X_{1:n}) - \pi(\theta \mid \lambda,\, X_{1:n})|\, \mathrm{d}\theta \\ &\geqslant \delta \int_{V_n^c \cap U_{M_n}} e^{-n(\theta-\hat{\theta}_n)^{\mathrm{T}} I(\theta_0)(\theta-\hat{\theta}_n)/2} \frac{n^{k/2}|I(\theta_0)|^{1/2}}{(2\pi)^{k/2}}\, \mathrm{d}\theta \\ &\geqslant \delta \int_{V_n^c \cap \{\|u\| < cM(\log n)^{1/2}\}} \phi(u)\, \mathrm{d}u,\end{aligned}$$

for some $c > 0$, where ϕ denotes the density of a standard Gaussian distribution on $\mathbb{R}^k$. By choosing $L > 0$ large enough and using (5.6),

$$\begin{aligned}\int_{V_n^c \cap \{\|u\| < cM(\log n)^{1/2}\}} \phi(u)\, \mathrm{d}u &\geqslant \phi(L) \int_{V_n^c \cap \{\|u\| < L\}} \mathrm{d}u \\ &\geqslant \phi(L) \left\{ \frac{\pi^{k/2} L^k}{\Gamma(k/2+1)} - \int_{V_n \cap \{\|u\| < cM(\log n)^{1/2}\}} \mathrm{d}u \right\} \\ &\geqslant \phi(L) \frac{\pi^{k/2} L^k}{2\Gamma(k/2+1)} > 0,\end{aligned}$$

which completes the proof.

[*Received November* 2012]