Bayes and Empirical Bayes : do they merge?

Journal:	*Biometrika*
Manuscript ID:	Draft
Manuscript Type:	Article
Date Submitted by the Author:	n/a
Complete List of Authors:	Petrone, Sonia; Bocconi University, Dept of Decision Sciences Rousseau, Judith; CREST-ENSAE, ; Université Paris Dauphine, CEREMADE Scricciolo, Catia; Bocconi Universtiy, Decision Sciences
Keywords:	Bayeian statistics, empirical Bayes, asymptotics, merging

Biometrika (2012), **99**, 1, *pp.* 1–21
© 2012 Biometrika Trust
Printed in Great Britain
Advance Access publication on 31 July 2012

Bayes and empirical Bayes: do they merge?

BY S. PETRONE
Bocconi University, Milan, Italy
sonia.petrone@unibocconi.it

J. ROUSSEAU
CREST-ENSAE and Université Paris Dauphine, Paris, France
rousseau@ceremade.dauphine.fr

AND C. SCRICCIOLO
Bocconi University, Milan, Italy
CREST-ENSAE, Paris, France
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 empirical estimate of the prior hyperparameters. Despite not rigorous, the underlying idea is that, for sufficiently large sample size, empirical Bayes leads to similar inferential answers as a proper Bayesian inference. However, precise mathematical results seem missing. In this work, we give more rigorous results in terms of merging of Bayesian and empirical Bayesian 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: Consistency, Bayesian weak merging, Frequentist strong merging, Maximum marginal likelihood estimate, Dirichlet process mixtures, g-priors.

1. INTRODUCTION AND MOTIVATION

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

From a Bayesian point of view, an empirical Bayes approach is not justified, but it is attractive as a computationally simple alternative to a more rigorous but usually analytically more complex hierarchical specification of the prior, of the kind $\theta|\lambda \sim \pi(\theta|\lambda)$ and $\lambda \sim h(\lambda)$. Thus, for a Bayesian statistician, empirical Bayes is of interest for two reasons: on one hand, when it is difficult to honestly fix λ, it is expected that a data-driven choice of λ leads to better inferential results; and, empirical Bayes could be a simple approximation of the hierarchical posterior distribution. This is possibly the reason of the wide use of empirical Bayes in practical applications of Bayesian methods. 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 study whether empirical Bayes procedures have some optimality property (versus a fixed choice of λ). To our knowledge, precise general results about such asymptotic agreement and about general optimality property are missing. The aim of this paper is to provide some results in both these directions. First, we will give conditions for the asymptotic agreement, or merging, of empirical Bayes and Bayesian solutions; however, we will also individuate situations where empirical Bayes and Bayes diverge and thus, from a Bayesian viewpoint, require particular care. Then, we 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 samples, empirical Bayes would give a solution that is close to the oracle Bayes and in this sense exploits information more efficiently than a fixed choice of λ.

Despite not rigorously justified, empirical Bayes is quite often used by practitioners and in the literature, see for instance George and Foster (2000), in the context of variable selection in regression; Clyde and George (2000), for wavelets shrinkage estimation; Liu (1996) and McAuliffe, Blei and Jordan (2006) in Bayesian nonparametric mixture models, and Favaro *et al.* (2009), in Bayesian nonparametric inference for species diversity. Systematic comparison of empirical Bayes and Bayesian procedures appears less explored. A careful comparison of empirical Bayes and Bayesian variable selection criteria in regression is developed by Cui and George (2008). In this context, a surprising result has been recently underlined by Scott and Berger (2010), who prove an asymptotic discrepancy between fully Bayesian and empirical Bayes inferences. Empirical Bayes and hierarchical Bayes procedures for nonparametric inverse problems are studied in a recent work by Knapik *et al.* (2012). In their case, the hyperparameter λ has an interpretation as a model index and a direct relation to the true parameter exists a priori. However, generally the hyperparameter merely characterizes some aspects of the prior so that there exists no notion of a true value of λ; thus, it is not immediately clear what would be a desirable limit of the sequence of $\hat{\lambda}_n$. We will propose a notion of oracle value instead of a true value for λ, in Section 4.

The term empirical Bayes is indeed used with different meanings in the literature. Another common use refers to problems where a prior is introduced, but some frequentist interpretation of it is possible, typically in mixture models where $X_i|\theta_i \sim p_{\theta_i}(x)$ and the θ_i are a sample from a latent distribution $G(\theta|\lambda)$. In these problems, maximum likelihood estimation of λ, i.e. of the latent distribution, is often referred to as empirical Bayes. A Bayesian approach would assign a prior distribution on λ. In these cases, a comparison between Bayes and empirical Bayes reduces to the interesting, but more standard, comparison between maximum likelihood and Bayes procedures, and it is not the object of this work.

The first question we address is whether it is true that the empirical Bayes and the Bayesian posterior distributions will be asymptotically close. In fact, a relevant counterexample has been recently pointed out by Scott and Berger (2010), in the special case of variable selection in regression models. They consider a Bayesian approach where variable selection is based on a vector of inclusion $\gamma \in \{0,1\}^k$ which selects among k potential regressors, and the prior on $\gamma = (\gamma_1, \ldots, \gamma_k)$ assumes that the γ_i are independent Bernoulli with parameter λ, $\pi(\gamma_1, \ldots, \gamma_k|\lambda) \propto \lambda^{k_\gamma}(1-\lambda)^{k-k_\gamma}$, $\lambda \in (0, 1)$, where $k_\gamma = \sum \gamma_i$ is the selected number of covariates. In this framework, George and Foster (2000) have shown that an empirical Bayes procedure that estimates the inclusion probability λ from the data, e.g. by maximum marginal likelihood, may be preferable to a Bayesian procedure that uses a fixed value of the prior hyperparameter λ. Scott and Berger (2010) compare this empirical Bayes approach with a hierarchical Bayes procedure that assigns a prior on λ. Surprisingly, they prove an asymptotic discrepancy between the two procedures. In particular, they show that the empirical Bayes posterior distribution on the set of models that 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, however, in terms of posterior distribution is far from being satisfactory. So we shed light on such phenomena by describing when and why marginal maximum likelihood empirical Bayes procedures will be pathological or, on the contrary, when and why they will have some good oracle property. These results have therefore the practical interest of characterizing, at least in the parametric case, those families of priors to be used with regard to empirical Bayes procedures and those to be avoided, in particular if one is not merely interested in point estimation, but in some more general characteristics of the posterior distribution.

We formalize the asymptotic comparison in terms of merging of Bayes and empirical Bayes procedures. We consider two notions of merging. First, we study Bayesian weak merging in the sense of Diaconis and Freedman (1986).Then, we study frequentist strong merging in the sense of Ghosh and Ramamoorthi (2003), which compares posterior distributions in terms of total variation distance, P_0^∞-almost surely, where $P_0^\infty \equiv P_{\theta_0}^\infty$ denotes the probability law of $(X_i)_{i\geq 1}$ under θ_0, the true value, in the frequentist sense, of the parameter θ. It is worth noting that, when strong merging holds, if Bernstein von-Mises holds in the L_1-sense for the Bayes posterior, then it also holds for the empirical Bayes posterior.

Developing from Diaconis and Freedman (1986), we see (Section 3) that weak merging of Bayes and empirical Bayes posterior distributions holds if and only if the empirical Bayes posterior is weakly consistent, in the frequentist sense, at θ_0, for every θ_0 in the parameter space Θ. Thus, conditions for weak consistency of empirical Bayes posteriors are needed. Besides the Bayesian motivations in terms of merging, consistency is of autonomous interest from a frequentist viewpoint, and we consider it in a general context. Conditions for empirical Bayes consistency are generally stronger than those needed for consistency of Bayes posteriors. We provide sufficient conditions that cover both parametric and nonparametric cases. In fact, an empirical Bayes approach is even more tempting in nonparametric problems, since frequentist properties of Bayes procedures are known to crucially depend on fine details of the prior (Diaconis and Freedman, 1986) and on a careful choice of the prior hyperparameters. We exhibit examples to illustrate empirical Bayes consistency for Dirichlet process mixtures, which is a commonly used nonparametric prior.

Even when consistency and weak merging hold, simple examples show that the empirical Bayes posterior can have unexpected and counterintuitive behaviors. Frequentist strong merging is a way to refine the analysis. Obtaining strong merging of Bayes 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 $\hat\lambda_n$. In this setup, 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 $\sup_\lambda \pi(\theta_0|\lambda)$ is reached for a value λ_0 (here unique for simplicity) in the boundary of Λ and this is such that $\pi(\theta|\lambda_0)$ is degenerate on θ_0, then the empirical Bayes posterior will not merge with any (hierarchical) Bayes posterior distribution. We illustrate this behavior in Bayesian regression with g-priors. Conversely, if $\sup_\lambda \pi(\theta_0|\lambda) < \infty$, which is the case if it is reached for λ_0 in the interior of Λ, then $\hat\lambda_n$ converges to λ_0 and frequentist strong merging holds. The value λ_0 can be understood as the prior oracle as it is the value of the hyperparameters such that the prior mostly favors the true θ_0. Under this respect, the empirical Bayes posterior achieves some kind of optimality. The asymptotic selection of the oracle value λ_0 suggests that empirical Bayes may have better finite sample properties than a Bayesian solution with a fixed choice of λ. Finite sample comparisons are beyond the scope of this work, but we have a discussion at the end of the paper.

The paper is organized as follows. In Section 2, we define the context and the notation. In Section 3, we study Bayesian merging and consistency of empirical Bayes posteriors. Parametric and nonparametric examples illustrate these results. In Section 4, we study frequentist strong merging and obtain, as a by-product, that in regular cases the empirical Bayes procedure leads to an oracle choice of the prior hyperparameter. Some open issues are discussed in Section 5.

2. General context and notation

Let $\mathcal{X}$ and Θ denote the observational space and the parameter space, respectively. In order to cover parametric and nonparametric problems, we allow them to be quite general, only requiring that they are complete and separable metric spaces, equipped with their Borel σ-fields, $\mathcal{B}(\mathcal{X})$ and $\mathcal{B}(\Theta)$. Let $(X_i)_{i\geq 1}$ be a sequence of random elements, with the X_i's taking values in $\mathcal{X}$. Suppose that, given θ, the probability measure of the process $(X_i)_{i\geq 1}$ is P_θ^∞ and for $n \geq 1$ denote by $P_\theta^{(n)}$ the joint probability law of $(X_1, \ldots, X_n)$. We assume that $P_\theta^{(n)}$ is dominated by a common σ-finite measure μ and denote by $p_\theta^{(n)}$ the density of $P_\theta^{(n)}$ w.r.t. μ.

In the sequel, we use the short notation $X_{1:n} = (X_1, \ldots, X_n)$ and $x^\infty = (x_1, x_2 \ldots)$. Let $\{\Pi(\cdot|\lambda) : \lambda \in \Lambda\}$ be a family of prior probability measures on Θ. Given a prior $\Pi(\cdot|\lambda)$, we denote by $\Pi(\cdot|\lambda, X_{1:n})$ the corresponding posterior distribution of θ, given $X_{1:n}$.

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, $\hat\lambda_n$ takes values in the closure $\bar\Lambda$ of Λ. For λ_0 in the boundary $\partial\Lambda$ of Λ, we define $\Pi(\cdot|\lambda_0)$ as the σ-additive weak limit of $\Pi(\cdot|\lambda)$ for $\lambda \to \lambda_0$, when it exists. We use the notation $P_n \Rightarrow P$ to mean that P_n converges weakly to P, for any probabilities P_n, P, and the notation $\|f\|_1$ for the L_1-norm of a function f. We shall say that the empirical Bayes posterior is well defined if $\Pi(\cdot|\hat\lambda_n, X_{1:n})$ is a probability measure for all large n, P_θ^∞-almost surely, for all θ. Then, the empirical Bayes posterior is defined, on all Borel sets B, as

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

It will also be denoted, with shorter notation, by $\Pi_{n,\hat\lambda_n}$. Throughout the paper, the empirical Bayes posterior is always tacitly assumed to be well defined.

Many types of estimators $\hat\lambda_n$ can be considered: in particular, the maximum marginal likelihood, defined as $\hat\lambda_n \in \mathrm{argsup}_\lambda\, m(X_{1:n}|\lambda)$, where $m(X_{1:n}|\lambda) = \int_\Theta p_\theta^{(n)}(X_{1:n})\, \mathrm{d}\Pi(\theta|\lambda)$, is the most popular. Whenever we consider the maximum marginal likelihood estimator, we assume that $\sup_\lambda m(X_{1:n}|\lambda) < \infty$ for all large n, P_θ^∞-almost surely, for all θ, and write $\hat m(X_{1:n}) = m(X_{1:n}|\hat\lambda_n)$. We shall present general results for the empirical Bayes posterior without specifying the type of estimator $\hat\lambda_n$, as well as specific results for the maximum likelihood.

3. Bayesian weak merging and consistency

3·1. *Bayesian merging of Bayes and empirical Bayes inferences*

Merging formalizes the idea that two posteriors or, in a predictive setting, two predictive distributions of all future observations, given the past, will eventually be close. A well-known result by Blackwell and Dubins (1962) establishes that, for P and Q probability laws of a process $(X_i)_{i\geq 1}$, if P and Q are mutually absolutely continuous, then there is strong merging of the predictions of future events, given the increasing information provided by the data $X_{1:n}$. For exchangeable P and Q corresponding to priors Π and q, respectively, P and Q are mutually absolutely continuous if and only if Π and q are such. However, the empirical Bayes approach only gives a sequence of posterior distributions $\Pi_{n,\hat{\lambda}_n}$, without having a properly defined probability law of the process $(X_i)_{i\geq 1}$. Thus, the above result on strong merging does not apply. Furthermore, Blackwell and Dubins' result does not apply when the priors are singular, as it is often the case in nonparametric problems.

Diaconis and Freedman (1986) gave 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 $|\int g\,\mathrm{d}p_n - \int g\,\mathrm{d}q_n| \to 0$ for all continuous and bounded functions g. Diaconis and Freedman showed that two Bayesians with different priors will merge weakly if and only if one Bayesian has weakly consistent posterior, in the frequentist sense, at θ, for every $\theta \in \Theta$. We show that an analogous result holds here: the empirical Bayes merges with any Bayesian if and only if the empirical Bayes posterior is weakly consistent at θ, for every $\theta \in \Theta$.

The results are herein restricted to the case of exchangeable sequences, thus, given θ, the X_i's are independent and identically distributed with common distribution P_θ. Given a prior Π on Θ, we use $\Pi_n(\cdot)$ to denote the posterior distribution $\Pi(\cdot|X_1,\ldots,X_n)$, and P_Π for the exchangeable probability law of the process $(X_n)_{n\geq 1}$ defined through Π. Recall that a posterior distribution Π_n is *weakly consistent* at θ if, for any weak neighborhood W of θ, $\Pi_n(W^c) \to 0$, almost surely-$[P_\theta^\infty]$, for all $\theta \in \Theta$, where P_θ^∞ here denotes the infinite product measure on $\mathcal{X}^\infty$.

Let $\Pi_{n,\hat{\lambda}_n}$ be the empirical Bayes posterior as described in Section 2. The following result is a straightforward consequence of Theorem A.1 in Diaconis and Freedman (1986).

Proposition 1. *Let the map $\theta \mapsto P_\theta$ be one-to-one and Borel. Given a family of priors $\{\Pi(\cdot|\lambda) : \lambda \in \Lambda\}$, the empirical Bayes posterior is consistent at any $\theta \in \Theta$ if and only if, for any prior probability q on Θ, the empirical Bayes posterior and the Bayes posterior q_n weakly merge with P_q-probability* 1.

The proof is immediate since it suffices to note that the proof for the equivalences (i)–(iv) in Theorem A.1 of Diaconis and Freedman (1986), page 18, goes through to the present case: in fact, it is based on the properties of the Bayesian posterior q_n, whereas, for the posterior $\Pi_{n,\hat{\lambda}_n}$, only consistency is required.

Proposition 1 shows that any Bayesian can be sure that her estimate with respect to quadratic loss of any continuous and bounded function g will asymptotically agree with the empirical Bayes estimate, if and only if the empirical Bayes posterior is consistent at any θ. If so, in particular, a Bayesian with hierarchical prior $\Pi_h(\theta) = \int \Pi(\theta|\lambda)h(\lambda)\,\mathrm{d}\,\lambda$ is sure that $|\int g(\theta)\,\mathrm{d}\Pi(\theta|\hat{\lambda}_n,\, X_{1:n}) - \int\int g(\theta)\,\mathrm{d}\Pi(\theta|\lambda, X_{1:n})h(\lambda|X_{1:n})\mathrm{d}\lambda| \to 0$, where $h(\lambda|X_{1:n})$ is the posterior density of λ.

The above results show that even a minimal requirement such as weak merging is not guarantee; in fact, it holds if and only if the empirical Bayes posterior is weakly consistent. Note that consistency refers to the posterior distribution on θ, and cannot be refereed to the sequence of estimators $\hat{\lambda}_n$, since in our context there is generally no notion of true value of λ. Besides its Bayesian motivations in terms of merging, from a frequentist viewpoint consistency is a basic property of autonomous interest. Thus, we study consistency of empirical Bayes in a more general case, for dependent sequences beyond the case of exchangeability, and covering both parametric and nonparametric problems. Clearly, consistency of the empirical Bayes posterior distributions requires more care than for standard Bayesian procedures, since the prior is data-dependent through $\hat\lambda_n$ and one has to control the behavior of the sequence of estimators $\hat\lambda_n$. We give two results, one for procedures where $\hat\lambda_n$ is computed by maximum likelihood and one for the case when $\hat\lambda_n$ is a convenient sequence of estimators.

To be more specific, let (Θ, d) be a semi-metric space. For any $\epsilon > 0$, let $U_\epsilon \equiv U_\epsilon(\theta_0) = \{\theta \in \Theta :\ d(\theta, \theta_0) < \epsilon\}$ denote the open ball centered at θ_0 with radius ϵ. The empirical Bayes posterior is consistent at θ_0 if, for any $\epsilon > 0$, $\Pi(U_\epsilon^c \mid \hat\lambda_n,\, X_{1:n}) \to 0$ almost surely-P_0^∞, where P_0^∞ denotes the probability measure of $(X_i)_{i\geq 1}$ under θ_0.

For $\theta \in \Theta$, let $R(p_\theta^{(n)}) = (p_\theta^{(n)}/p_{\theta_0}^{(n)})(X_{1:n})$ denote the likelihood ratio. We shall use the following assumptions.

(**A1**) There exist constants $c_1,\, c_2 > 0$ such that, for any $\epsilon > 0$,

$$P_0^*\left(\sup_{\theta\in U_\epsilon^c} R(p_\theta^{(n)}) \geq e^{-c_1 n\epsilon^2}\right) \leq c_2 (n\epsilon^2)^{-(1+t)}$$

for some $t > 0$, where P_0^* denotes the outer measure.

(**A2**) For each $\theta_0 \in \Theta$, there exists $\lambda_0 \in \Lambda$ such that, for any $\eta > 0$, $\Pi(B_{\mathrm{KL}}(\theta_0;\, \eta) \mid \lambda_0) > 0$, where, for $\mathrm{KL}_\infty(\theta_0;\, \theta) = -\lim_{n\to\infty} n^{-1} \log R(p_\theta^{(n)})$, the set $B_{\mathrm{KL}}(\theta_0;\, \eta) = \{\theta \in \Theta :\ \mathrm{KL}_\infty(\theta_0;\, \theta) < \eta\}$.

When compared to the assumptions usually considered for posterior consistency, (**A1**) is quite strong; it is, however, a common assumption in the maximum likelihood estimation literature. It is verified in most parametric models, see for instance Schervish (1995), and also in nonparametric models. For instance, Wong and Shen (1995) proved that, for independent and identically distributed observations with density p_θ, if U_ϵ is the Hellinger open ball centered at p_{θ_0} with radius ϵ, then a sufficient condition for (**A1**) to hold true is that there exist constants $c_3,\, c_4 > 0$ such that, for each $\epsilon > 0$,

$$\int_{\epsilon^2/2^8}^{\sqrt{2}\epsilon} \sqrt{H_{[]}(u/c_3,\, \Theta,\, h)}\, \mathrm{d}u \leq c_4 \sqrt{n}\epsilon^2 \qquad \text{for } n \text{ large enough,} \tag{3.1}$$

where the function $H_{[]}(\cdot,\, \Theta,\, h)$ denotes the Hellinger bracketing metric entropy of Θ. In some cases (**A1**) can be weakened into

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

where $\Theta_n \subset \Theta$ is such that $\Pi(\Theta_n^c \mid \hat\lambda_n,\, X_{1:n}) = o_{P_0}(1)$. Note that, contrariwise to fully Bayes approaches, the fact that the prior is data-dependent prevents the use of exponential bounds on $\Pi(\Theta_n^c)$ to control the posterior probability of Θ_n^c. However, in Section 3.2.4 we provide a nonparametric example where we can prove directly that the empirical Bayes posterior probability of Θ_n^c goes to 0.

In this paper, (**A1**) is used to handle the numerator of the ratio defining the posterior probability of any neighborhood U_ϵ in the following way. By the first Borel-Cantelli lemma, (**A1**) implies that $\sup_{\theta\in U_\epsilon^c} R(p_\theta^{(n)}) < e^{-c_1 n\epsilon^2}$ for all large n, P_0^∞-almost surely . Thus, for all large n,

$$\int_{U_\epsilon^c} R(p_\theta^{(n)})\,\mathrm{d}\Pi(\theta|\hat\lambda_n) \le \Pi(U_\epsilon^c|\hat\lambda_n)\sup_{\theta\in U_\epsilon^c} R(p_\theta^{(n)}) \le \sup_{\theta\in U_\epsilon^c} R(p_\theta^{(n)}) < e^{-c_1 n\epsilon^2} \tag{3.2}$$

P_0^∞-almost surely. Note that the bound in (3.2) is valid for any type of estimator $\hat\lambda_n$.

Assumption (**A2**) is the usual Kullback-Leibler prior support condition, herein required to hold true for some value $\lambda_0 \in \Lambda$. It is a mild assumption considered in most results on posterior consistency and has been shown to be satisfied for various models and families of priors. The rather abstract definition of $\mathrm{KL}_\infty(\cdot;\,\cdot)$ is mainly considered to deal with dependent data. In the case of independent and identically distributed sequences, $\mathrm{KL}_\infty(\theta_0;\,\theta)$ is simply the Kullback-Leibler divergence between the densities p_{θ_0} and p_θ (per observation). In the present context, it is used when $\hat\lambda_n$ is the maximum marginal likelihood to bound from below $\hat m(X_{1:n})/p_{\theta_0}^{(n)}(X_{1:n})$. For other types of estimator $\hat\lambda_n$, a variant of (**A2**) is considered, cf. conditions (ii)-(iii) of Proposition 3.

We first study consistency of the empirical Bayes posterior when $\hat\lambda_n$ is the maximum marginal likelihood.

3·2. *Case of the maximum marginal likelihood estimator*

Let $\hat\lambda_n$ be the maximum marginal likelihood as defined in Section 2. We have the following result.

PROPOSITION 2. *Under assumptions* (**A1**) *and* (**A2**)*, the posterior* $\Pi(\cdot|\hat\lambda_n,\,X_{1:n})$*, where* $\hat\lambda_n$ *is the maximum marginal likelihood estimator, is consistent at* θ_0*, i.e. for any* $\epsilon > 0$*,* $\Pi(U_\epsilon^c|\hat\lambda_n,\,X_{1:n}) \to 0$*,* P_0^∞*-almost surely.*

The proof of Proposition 2 is deferred to the Appendix.

Although one could expect that the usual Kullback-Leibler prior support condition, here (**A2**), implies weak consistency of the empirical Bayes posterior, as it happens for Bayesian posteriors, it is, however, not the case and additional assumptions on the behavior of the likelihood ratio and/or on the prior need to be required, as illustrated in the following example. Consider Bahadur (1958)'s example, see also Lehmann and Casella (1998), pages 445–447, and Ghosh and Ramamoorthi (2003), pages 29–31. Let $\Theta = \mathbb{N}^*$. For each $\theta = k$, a density p_θ on $[0,\,1]$ is defined as follows. Let $a_0 = 1$ and define recursively a_k by $\int_{a_k}^{a_{k-1}}[h(x) - C]\,\mathrm{d}x = 1 - C$, where $0 < C < 1$ is a given constant and $h(x) = e^{1/x^2}$. Since $\int_0^1 e^{1/x^2}\,\mathrm{d}x = \infty$, the a_k's are uniquely determined and the sequence $a_k \to 0$ as $k \to \infty$. For $\theta \in \Theta$, define

$$p_\theta(x) = \begin{cases} h(x), \text{ if } & a_\theta < x \le a_{\theta-1}, \\ C, \text{ if } & x \in [0,1]\cap(a_\theta,\,a_{\theta-1}]^c, \\ 0, & \text{otherwise.}\end{cases}$$

Let $X_1,\,\ldots,\,X_n|\theta \sim p_\theta$ independently. The maximum likelihood $\hat\theta_n$ exists and tends to ∞ in probability, regardless of the true value $\theta_0 = k_0$ 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 $\theta \in \Theta$. Consider a family of priors $\{\Pi(\cdot|\lambda) : \lambda \in \Lambda\}$ such that, for each θ, there exists $\lambda \in \bar\Lambda$ for which $\Pi(\cdot|\lambda) = \delta_\theta$. It is always possible to construct such a family of priors.

For $\lambda = (m, \sigma)$, let

$$\Pi(1|\lambda) = \Phi((1/2 - m)/\sigma) - \Phi((-1/2 - m)/\sigma),$$
$$\Pi(\theta|\lambda) = \Phi((\theta - 1/2 - m)/\sigma) - \Phi((\theta - 3/2 - m)/\sigma) + \Phi((-\theta - 3/2 - m)/\sigma) - \Phi((-\theta - 1/2 - m)/\sigma) \qquad \text{for } \theta > 1,$$

where Φ is the cumulative distribution function of a standard Gaussian random variable. By taking $m = \theta - 1$ and letting $\sigma \to 0$, we have as a limit the Dirac mass at θ because $\Pi(1|\lambda) = 0$ and $\Pi(\theta|\lambda) \to 1$. Thus, for each $k_0 \in \mathbb{N}^*$, by taking $m = k_0 - 1$ and letting $\sigma \to 0$, we have as a limit the Dirac mass at k_0. Then, the posterior is the Dirac mass at the maximum likelihood estimator $\hat{\theta}_n$, which is inconsistent. To see this, note that

$$\forall\, \lambda \in \Lambda, \quad m(X_{1:n}|\lambda) \le \prod_{i=1}^{n} p_{\hat{\theta}_n}(X_i) \quad \text{and} \quad \hat{m}(X_{1:n}) = m(X_{1:n}|(\hat{\theta}_n - 1, 0)) = \prod_{i=1}^{n} p_{\hat{\theta}_n}(X_i).$$

Proposition 2 gives a result on consistency, however it can be easily turned into a result on posterior concentration rates by replacing ϵ by $\bar{\epsilon}_n$ in $(\mathbf{A1})$ and changing the Kullback-Leibler neighborhood in $(\mathbf{A2})$ with $S_n = \{\theta : \mathrm{KL}(p_{\theta_0}^{(n)}; p_\theta^{(n)}) \le n\tilde{\epsilon}_n^2,\, V(p_{\theta_0}^{(n)}, p_\theta^{(n)}) \le n\tilde{\epsilon}_n^2\}$, with $V(p, p') = \int p(\log(p/p'))^2(x)\mathrm{d}x$ as in Ghosal and van der Vaart (2007).

3·3. *Case of a convergent $\hat{\lambda}_n$*

In some applications, $\hat{\lambda}_n$ is chosen to be a convenient statistic, such as some moment estimator, so that the prior is centered at a plausible value for the parameter. In such cases, $\hat{\lambda}_n$ has often a known asymptotic behavior, which however does not guarantee that the empirical Bayes posterior has a stable behaviour, too. In the following proposition, we give sufficient conditions for consistency of the empirical Bayes posterior in such situations. Suppose that the parameter is split into $\theta = (\tau, \zeta)$, where $\tau \in \mathrm{T}$ and $\zeta \in \mathrm{Z}$ and, given $\lambda \in \Lambda \subseteq \mathbb{R}^\ell$, $\tau \sim \tilde{\Pi}(\cdot|\lambda)$ while $\zeta \sim \tilde{\Pi}$. In other words, the hyperparameter λ only influences the prior distribution of τ. The overall prior is $\Pi(\cdot|\lambda) = \tilde{\Pi}(\cdot|\lambda) \times \tilde{\Pi}(\cdot)$. Let $\theta_0 = (\tau_0, \zeta_0)$ be the true value of $\theta = (\tau, \zeta)$.

PROPOSITION 3. *Let $\tilde{\Pi}(\cdot|\hat{\lambda}_n)$, $n = 1, 2, \ldots$, and $\tilde{\Pi}(\cdot|\lambda_0)$ be probability measures on $\mathcal{B}(\mathrm{T})$. Assume that $(\mathbf{A1})$ is satisfied and*

(i) $\tilde{\Pi}(\cdot|\hat{\lambda}_n) \Rightarrow \tilde{\Pi}(\cdot|\lambda_0)$ P_0^∞-almost surely,

(ii) for each $\eta > 0$, there exists a set $K_\eta \subseteq B_{\mathrm{KL}}(\theta_0; \eta)$ such that $\Pi(K_\eta|\lambda_0) > 0$,

(iii) defined, for each $x^\infty \in \mathcal{X}^\infty$ and any $\eta > 0$, the set

$$E_{x^\infty}^{(\eta)} = \left\{ (\tau, \zeta) \in K_\eta : \frac{1}{n} \log \frac{p_{(\tau_0, \zeta_0)}^{(n)}}{p_{(\tau_n, \zeta)}^{(n)}}(x_{1:n}) \not\to \mathrm{KL}_\infty((\tau_0, \zeta_0); (\tau, \zeta)) \quad \textit{for some } \tau_n \to \tau \right\},$$

$E_{x^\infty}^{(\eta)} \in \mathcal{B}(\mathrm{T}) \otimes \mathcal{B}(\mathrm{Z})$ and, for P_0^∞-almost every $x^\infty \in \mathcal{X}^\infty$,

$$\Pi(E_{x^\infty}^{(\eta)}|\lambda_0) = 0. \tag{3.3}$$

Then, for any $\epsilon > 0$, $\Pi(U_\epsilon^c|\hat{\lambda}_n, X_{1:n}) \to 0$ a.s. $[P_0^\infty]$.

The proof of Proposition 3 is postponed to the Appendix.

Condition (i) is a natural condition when $\hat{\lambda}_n$ is an explicit estimator (as opposed to the maximum marginal likelihood). Condition (ii) is the usual Kullback-Leibler prior support condition, except for the fact that here it concerns the support of the limiting prior. Condition (iii) is more unusual. If, in the definition of $E^{(\eta)}_{x^\infty}$, the τ_n's were fixed at τ, then (iii) would be a basic ergodic condition on the support of $\Pi(\cdot|\lambda_0)$, so the difficulty comes from obtaining an ergodic theorem uniformly over neighborhoods of τ. In the case of independent and identically distributed observations, the following condition implies (iii). If

$$\forall\,\eta,\,\epsilon>0,\ \ \forall\,\theta\in K_\eta,\ \ \exists\,\delta\equiv\delta(\theta,\,\epsilon)>0:\quad \mathbb{E}_0\left[\sup_{\theta'\in\Theta:\,d(\theta',\,\theta)<\delta}\left|\log\frac{p_\theta}{p_{\theta'}}(X_1)\right|\right]<\frac{\epsilon}{2},$$

then standard strong law of large numbers arguments imply that there exists a set $\mathcal{X}_0^\infty\subseteq\mathcal{X}^\infty$, with $P_0^\infty(\mathcal{X}_0^\infty)=1$, such that, for each $x^\infty\in\mathcal{X}_0^\infty$, condition (3.3) is satisfied.

We now show two examples that illustrate the above consistency results. The examples refer to nonparametric problems, where the asymptotic behavior of the empirical Bayes procedure is more delicate.

Example 1. Dirichlet process mixtures of Gaussians are popular Bayesian nonparametric priors, commonly used for density estimation and in a wide range of problems. A univariate Dirichlet process scale-location mixture of Gaussians assumes that $X_i|G$ are independently distributed according to $p_G(\cdot)=\int\phi(\cdot|\mu,\,\sigma^2)\,\mathrm{d}G(\mu,\,\sigma)$, where $\phi(\cdot|\mu,\,\sigma^2)$ denotes the Gaussian density with parameters μ and σ^2. The mixing distribution G is given a Dirichlet process prior with parameter $\lambda\bar{\alpha}(\cdot)$, $G\sim\mathrm{DP}(\lambda\bar{\alpha})$, where λ is a positive scalar and $\bar{\alpha}$ is a probability measure on $\mathbb{R}\times\mathbb{R}^{+*}$ with $\mathbb{R}^{+*}=\{x\in\mathbb{R}:\ x>0\}$. The choice of the scale parameter λ has a crucial impact on inference and this has suggested to treat it as random, assigning it a hyperprior in a hierarchical Bayes approach, or to fix it by empirical Bayes (Liu (1996); McAuliffe, Blei and Jordan (2006)), which has computational advantages. In particular, Liu (1996) considers the marginal maximum likelihood estimator of λ for Dirichlet process mixtures of Binomial distributions, but his argument remains valid for more general kernels (Petrone and Raftery, 1997). Liu shows that the marginal maximum likelihood $\hat{\lambda}_n$ is the solution of

$$\sum_{j=1}^{n}\frac{\lambda}{\lambda+j-1}=\mathrm{E}[K_n|\lambda,\,X_{1:n}],\tag{3.4}$$

where $\mathrm{E}[K_n|\lambda,\,X_{1:n}]$ is the expected number of occupied clusters under the conditional posterior distribution, given λ. Note that, even if the model is parametrized in the mixing distribution G, Dirichlet Process mixtures of Gaussians are usually thought as priors on the space of densities p on $\mathcal{X}$. Let $U_\epsilon=\{p:h(p,\,p_0)<\epsilon\}$ where h is the Hellinger metric and p_0 the true density. If we assume that $\bar{\alpha}$ has support $A\times[\underline{\sigma},\,\bar{\sigma}]$, with A a compact interval of $\mathbb{R}$, $0<\underline{\sigma}<\bar{\sigma}<\infty$ and $\Theta=\{G:\ \mathrm{supp}(G)\subseteq A\times[\underline{\sigma},\,\bar{\sigma}]\}$, then, from Theorem 3.2 of Ghosal and van der Vaart (2001), page 1244, $\{p_G:\ G\in\Theta\}$ has bracketing Hellinger metric entropy satisfying condition (3.1), so that assumption $(\mathbf{A1})$ is fulfilled. Moreover, if the true density is a mixture of Gaussian distributions, $p_0=p_{G_0}$, with support$(G_0)\subseteq A\times[\underline{\sigma},\,\bar{\sigma}]$, then also condition $(\mathbf{A2})$ is satisfied. The existence of a solution of (3.4) implies that the empirical Bayes posterior for the unknown density is well defined, and, using Proposition 2, we get consistency.

Example 2. Let us now consider a Dirichlet Process location mixture of Gaussians: $X_i|(F,\,\sigma)$ independently distributed according to $p_{F,\sigma}(\cdot)=\int\phi(\cdot|\mu,\,\sigma^2)\,\mathrm{d}F(\mu)$, with $F\sim\mathrm{DP}(\alpha_{\mathbb{R}}\,\mathrm{N}(\lambda,\,\tau^2))$, where $\alpha_{\mathbb{R}}$ is a positive constant and $\mathrm{N}(\lambda,\,\tau^2)$ denotes the Gaussian distribution with mean λ and variance τ^2, and σ having prior distribution H with support $[\underline{\sigma},\,\bar{\sigma}]$, $0<$

$\underline{\sigma} < \bar{\sigma} < \infty$. We consider empirical Bayes selection of λ and a natural candidate is $\hat{\lambda}_n = \bar{X}_n$. The prior on F is then a $\mathrm{DP}(\alpha_{\mathbb{R}}\,\mathrm{N}(\bar{X}_n,\, \tau^2))$.

We prove consistency of the empirical Bayes posterior for the unknown density of the data with respect to the Hellinger or the L_1-distance. Let the true density be a mixture $p_{F_0,\,\sigma_0}$, with $\sigma_0 \in [\underline{\sigma},\, \bar{\sigma}]$ and F_0 satisfying $F_0([-a, a]^c) \lesssim e^{-c_0 a^2}$ for all large a and a constant $c_0 > 0$, and let $m_0 = \mathrm{E}_0[X_1]$ be the mean of X_1 under $p_{F_0,\,\sigma_0}$. For fixed $\epsilon > 0$, choose $0 < \delta < \epsilon^2$ small enough and $a_n = n^q$, with $1/4 < q < 1$. Consider the sieve set $\Theta_n = \{(F,\, \sigma) :\, F([-a_n,\, a_n]) > 1 - \delta,\;\; \sigma \in [\underline{\sigma},\, \bar{\sigma}]\}$. From Theorem 6 in Ghosal and van der Vaart (2007b), combined with the proof of Theorem 7 in Ghosal and van der Vaart (2007b), if $\tilde{\Theta}_n = \{(F,\, \sigma) :\, F([-a_n,\, a_n]) = 1,\;\; \sigma \in [\underline{\sigma},\, \bar{\sigma}]\}$, for all $\epsilon^2/2^8 < u < \sqrt{2}\epsilon$, $H_{[]}(u/c_3,\, \Theta_n,\, h) \lesssim a_n(\log a_n + \log(1/\epsilon))^2$. Thus, for n large enough, $\int_{\epsilon/2^8}^{\sqrt{2}\epsilon}(H_{[]}(u/c_3,\, \Theta_n,\, h))^{1/2}\,\mathrm{d}u \lesssim \epsilon\sqrt{a_n}(\log a_n) < \epsilon^2\sqrt{n}$, since $a_n = n^q$ with $1/4 < q < 1$. By (3.2), P_0^∞-almost surely, $\int_{H_\epsilon^c \cap \Theta_n} R(p_{F,\,\sigma}^{(n)})\,\mathrm{d}\Pi(F,\, \sigma|\hat{\lambda}_n) < e^{-c_1 n\epsilon^2}$ for all large n.

We now show that $\mathrm{E}_0[\int_{\Theta_n^c} R(p_{F,\,\sigma}^{(n)})\,\mathrm{d}\Pi(F,\, \sigma|\hat{\lambda}_n)] = o(1)$. Since $\bar{X}_n \xrightarrow{\text{a.s.}} m_0$, we have $\mathrm{N}(\bar{X}_n,\, \tau^2) \overset{\text{a.s.}}{\Longrightarrow} \mathrm{N}(m_0,\, \tau^2)$. By Theorem 3.2.6 in Ghosh and Ramamoorthi (2003), pages 105–106, $\mathrm{DP}(\alpha_{\mathbb{R}}\,\mathrm{N}(\bar{X}_n,\, \tau^2)) \overset{\text{a.s.}}{\Longrightarrow} \mathrm{DP}(\alpha_{\mathbb{R}}\,\mathrm{N}(m_0,\, \tau^2))$ and condition (i) of Proposition 3 is fulfilled with $\lambda_0 = m_0$. Denote by $\Pi(\cdot|\lambda_0)$ the overall limiting prior $\mathrm{DP}(\alpha_{\mathbb{R}}\,\mathrm{N}(m_0,\, \tau^2)) \times H$. When $F \sim \mathrm{DP}(\alpha_{\mathbb{R}}\,\mathrm{N}(\bar{X}_n,\, \tau^2))$, using the stick-breaking representation, we have $p_{F,\,\sigma}(\cdot) \overset{\text{a.s.}}{=} \sum_{j=1}^\infty p_j\phi(\cdot|\xi_j,\, \sigma^2)$, with $\xi_j \sim \mathrm{N}(\bar{X}_n,\, \tau^2)$ and independent. As $\phi(\cdot|\xi_j,\, \sigma^2) = \phi(\cdot|(\bar{X}_n - m_0) + \xi_j',\, \sigma^2)$, with $\xi_j' \sim \mathrm{N}(m_0,\, \tau^2)$, we have $p_{F,\,\sigma}(\cdot) = p_{F',\,\sigma}(\cdot - (\bar{X}_n - m_0))$, with $F' \sim \mathrm{DP}(\alpha_{\mathbb{R}}\,\mathrm{N}(m_0,\, \tau^2))$. Let A_n be the set wherein the inequality $|\bar{X}_n - m_0| \le L/\sqrt{n}$ holds for some constant $L > 0$. Note that $P_{\theta_0}^{(n)}(A_n^c)$ can be made as small as needed by choosing L large enough. Using the above representation of the Dirichlet prior,

$$p_{F',\,\sigma}^{(n)}(X_{1:n} - (\bar{X}_n - m_0)) \le \prod_{i=1}^n \int \phi(X_i|\xi,\, \sigma^2) e^{\frac{L|X_i - \xi|}{\sqrt{n}\sigma^2}}\,\mathrm{d}F'(\xi) = c_{n,\,\sigma}^n \prod_{i=1}^n \int g_\sigma(X_i - \xi)\,\mathrm{d}F'(\xi),$$

where g_σ is the probability density proportional to $\phi(y|0,\, \sigma^2)e^{L|y|/(\sqrt{n}\sigma^2)}$ and

$$c_{n,\,\sigma} = \int \phi(y|0,\, \sigma^2)e^{L|y|/(\sqrt{n}\sigma^2)}\,\mathrm{d}y \le e^{L^2/(2n\sigma^2)}\left(1 + \frac{2L}{\sigma\sqrt{n}}\right),$$

which implies that

$$\mathrm{E}_0\left[\mathbf{I}_{A_n}(X_{1:n})\int_{\Theta_n^c} R(p_{F,\,\sigma}^{(n)})\,\mathrm{d}\Pi(F,\, \sigma|\hat{\lambda}_n)\right] \lesssim \left(1 + \frac{2L}{\underline{\sigma}\sqrt{n}}\right)^n \Pi(\Theta_n^c|\lambda_0) \lesssim e^{-c_1 a_n^2 + c_2\sqrt{n}} \le e^{-c_1 a_n^2/2},$$

for n large enough, by definition of a_n.

Next, we bound from below the denominator of the ratio defining the empirical Bayes posterior probability of the set H_ϵ^c. Using similar computations to those above, on A_n,

$$\int_\Theta R(p_{F,\,\sigma}^{(n)})\,\mathrm{d}\Pi(F,\, \sigma|\hat{\lambda}_n) \gtrsim e^{-\frac{L^2}{2\underline{\sigma}^2}}\, c_{n,\,\sigma}^n \int_\Theta R(\tilde{g}_{F',\,\sigma}^{(n)})\,\mathrm{d}\Pi(F',\, \sigma|\lambda_0),$$

with $\tilde{g}_{F',\,\sigma}(\cdot) = c_{n,\,\sigma}^{-1}\int \phi(\cdot|\xi,\, \sigma^2)e^{-L\frac{|\cdot - \xi|}{\sqrt{n}\sigma^2}}\,\mathrm{d}F'(\xi)$, where, with abuse of notation, we still denote by $c_{n,\,\sigma}$ the normalizing constant. Using computations similar to those in the proof of (5.21) in Ghosal and van der Vaart (2001), we obtain that, for any $\eta > 0$, $\int_\Theta R(\tilde{g}_{F',\,\sigma}^{(n)})\,\mathrm{d}\Pi(F',\, \sigma|\lambda_0) \ge$

$e^{-n\eta}$ for n large enough. Consistency of the empirical Bayes posterior for the unknown density follows.

We conclude this section with a parametric example which, while very simple, is illuminating in showing that, even when consistency and weak merging hold, the empirical Bayes posterior distribution may exhibit very different behaviour and in fact can diverge from any Bayesian posterior and underestimate the uncertainty on θ.

Example 3. Consider $X_i \mid \theta \sim \mathrm{N}(\theta, \sigma^2)$ independently, with σ^2 known; this model satisfies condition $(\mathbf{A1})$. Let $\theta \sim \mathrm{N}(m, \tau^2)$.

Case 1. If τ^2 is fixed, and $m = \lambda$ is estimated by maximum marginal likelihood, then $\hat\lambda_n = \bar X_n$ and the resulting empirical Bayes posterior distribution is $\mathrm{N}(\bar X_n,\, (1/\tau^2 + n/\sigma^2)^{-1})$, a completely regular density. This sequence of posteriors can be seen to be consistent by direct computations, thus it merges weakly with any Bayesian posterior distribution.

Case 2. Let us now consider empirical Bayes inference when the prior variance $\lambda = \tau^2$ is estimated by maximum marginal likelihood, while m is fixed. Then (see e.g. Lehmann and Casella (1998), page 263) $\sigma^2 + n\hat\tau_n^2 = \max\{\sigma^2,\, n(\bar X_n - m)^2\}$, so that $\hat\tau_n^2 = \sigma^2 n^{-1} \max\left\{n(\bar X_n - m)^2/\sigma^2 - 1,\, 0\right\}$. The resulting posterior $\Pi(\cdot|\hat\tau_n^2,\, X_{1:n})$ is Gaussian with mean $m_n = (\sigma^2/n)/(\hat\tau_n + \sigma^2/n)m + \hat\tau_n^2/(\hat\tau_n + \sigma^2/n)\bar x_n$ and variance $(1/\hat\tau_n + n/\sigma^2)^{-1}$. A hierarchical Bayes approach would assign a prior on τ^2 such as $1/\tau^2 \sim \mathrm{Gamma}(a, b)$. This leads to a Student's-t prior distribution for θ, with flatter tails, that may give better frequentist properties, see, for example, Berger and Robert (1990), Berger and Strawderman (1996). However, the Student's-t prior is no longer conjugate and the empirical Bayes posterior is simpler to compute. Yet, in this example, the empirical Bayes posterior is only partially regular, in the sense that $\hat\tau_n^2$ can be equal to zero so that $\Pi(\cdot|\hat\tau_n^2,\, X_{1:n})$ can be degenerate at m. Simple computations show that the probability that $\hat\tau_n = 0$ goes to zero if the true $\theta_0 \neq m$, but it remains strictly positive if $\theta_0 = m$. This suggests that if $\theta_0 \neq m$, the hierarchical and the empirical Bayes posterior densities can be asymptotically close; however, if $\theta_0 = m$, 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 could object that, despite unsatisfactory from a Bayesian viewpoint, the empirical Bayes posterior would be degenerate on the true value of θ. However, the case where $\lambda = (m, \tau^2)$ shows that empirical Bayes may dramatically underestimate the posterior uncertainty. In this case, the maximum marginal likelihood 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 any location-scale family of priors. Indeed, if the model $p_\theta^{(n)}$ admits a maximum likelihood $\hat\theta_n$ and $\pi(\cdot|\lambda)$ is of the form $\sigma^{-1}g((\cdot - \mu)/\sigma)$, with $\lambda = (\mu,\, \sigma)$, for some unimodal density g which is maximum at 0, then $\hat\lambda_n = (\hat\theta_n,\, 0)$ and the empirical Bayes posterior is the point mass at $\hat\theta_n$. This shows that such families of priors should not be used in combination with maximum marginal likelihood empirical Bayes procedures.

4. Frequentist strong merging and asymptotic behavior of $\hat\lambda_n$

As illustrated in the simple Gaussian example above, stronger forms of merging are needed to capture and explain possible divergent behavior. In this section we study frequentist strong merging (Ghosh and Ramamoorthi, 2003) of empirical Bayes and Bayesian posterior distributions. We recall that two sequences Π_n and q_n of probability measures on Θ are said to merge strongly if their total variation distance converges to zero P_0^∞-almost surely.

We limit here our attention to the parametric case, thus $\Theta \subseteq \mathbb{R}^k$ for finite k, and we suppose that the prior probability measure $\Pi(\theta|\lambda)$ has density $\pi(\theta|\lambda)$ with respect to Lebesgue measure for all $\lambda \in \Lambda \subseteq \mathbb{R}^\ell$. Before formally stating a general result which describes the asymptotic behaviour of empirical Bayes posteriors, we present an informal argument to explain the heuristics behind it. Under usual regularity conditions on the model, the marginal likelihood, given λ, can be approximated as

$$m(X_{1:n}|\lambda) = \pi(\theta_0|\lambda)\frac{p^{(n)}_{\hat{\theta}_n}(X_{1:n})}{n^{k/2}} \times O_{P_{\theta_0}}(1)$$

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

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

An interesting phenomenon occurs: assuming the above argument is correct, the maximum marginal likelihood estimate is asymptotically maximizing the value $\pi(\theta_0|\lambda)$ of the prior density at true value θ_0 of the parameter. In other words, it selects the most interesting value of the hyperparameter λ in the prior. We call the set of values of λ maximizing $\pi(\theta_0|\lambda)$ the prior oracle set of hyperparameters and denote it by Λ_0. In terms of (strong) merging, however, Λ_0 may correspond to unpleasant values, typically if the supremum is reached for values of λ on the boundary of Λ and these correspond to a prior Dirac mass at θ_0; for continuous θ, this may happen if $\sup_{\lambda\in\Lambda} \pi(\theta_0|\lambda) = \infty$. Then, the empirical Bayes posterior is degenerate. This is what happens in case 2 of Example 3 or, more generally, when $\pi(\cdot|\lambda)$ is a location-scale family and λ contains the scale parameter. Obviously, in such cases, we cannot interchange the limit and the maximization. We now present these ideas more rigorously.

The map $g : \theta \mapsto \sup_{\lambda\in\Lambda} \pi(\theta|\lambda)$ 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 Sections 4·1 and 4·2 below, if $\theta_0 \in \Theta_0$, then the empirical Bayes posterior is regular; thus we refer to the case $\theta_0 \in \Theta_0$ as the non-degenerate case. Instead, if $\theta_0 \in \Theta_0^c$, referred as the degenerate case, the empirical Bayes posterior may be degenerate and fail to merge strongly with any regular Bayes posterior.

4·1. *Non-degenerate case*

In the non-degenerate case, we give sufficient conditions for the EB posterior $\Pi(\cdot|\hat{\lambda}_n, X_{1:n})$, where $\hat{\lambda}_n$ is the maximum marginal likelihood, to merge strongly with any posterior $\Pi(\cdot|\lambda, X_{1:n})$, $\lambda \in \Lambda$. In fact, the ultimate goal is to establish strong merging with hierarchical Bayes, as empirical Bayes is commonly used as an approximation of a hierarchical Bayes posterior. We make the comparison with a Bayesian posterior corresponding to a fixed choice of λ only for technical reasons, but note that, if both the prior $\pi(\cdot|\lambda)$ and the hierarchical prior $\int \pi(\theta|\lambda)h(\lambda)\mathrm{d}\lambda$ are positive and continuous at θ_0, and the corresponding posteriors are consistent, then, by Theorem 1.3.1 of Ghosh and Ramamoorthi (2003), pages 18–20, the latter merge strongly. Thus, strong merging of $\Pi(\cdot|\hat{\lambda}_n, X_{1:n})$ and $\Pi(\theta|\lambda, X_{1:n})$ implies strong merging of the EB posterior with any hierarchical Bayes posterior satisfying the above conditions. Let $\theta_0 \in \Theta_0$, and define the set $\Lambda_0 = \{\lambda_0 \in \Lambda : \pi(\theta_0|\lambda_0) = g(\theta_0)\}$.

THEOREM 1. *Suppose that $\theta_0 \in \Theta_0$. Assume that $(\mathbf{A1})$ is satisfied and*

(i) *the map $g : \theta \mapsto \sup_{\lambda\in\Lambda} \pi(\theta|\lambda)$ is positive and continuous at θ_0;*

(ii) *there exists a non-empty subset $\tilde{\Lambda}$ of Λ_0 such that, for any $\lambda_0 \in \tilde{\Lambda}$, $\Pi(U_\epsilon \cap B_{\mathrm{KL}}(\theta_0;\, \eta)|\lambda_0) > 0$ for all $\epsilon,\, \eta > 0$.*

Then, for each $\lambda_0 \in \tilde{\Lambda}_0$,

$$\frac{\hat{m}(X_{1:n})}{m(X_{1:n}|\lambda_0)} \to 1 \qquad \text{a.s. } [P_0^\infty]. \tag{4.1}$$

If, in addition to (i) and (ii), the following assumption is satisfied

(iii) *$\tilde{\Lambda}_0 = \Lambda_0$ is included in the interior of Λ and, for any $\delta > 0$, there exist $\epsilon,\, \eta > 0$ so that*

$$\sup_{\theta \in U_\epsilon} \sup_{\lambda \in \Lambda:\, d(\lambda,\, \Lambda_0) > \delta} \frac{\pi(\theta|\lambda)}{g(\theta)} \leq 1 - \eta,$$

where $d(\lambda,\, \Lambda_0) = \inf_{\lambda_0 \in \Lambda_0} d(\lambda,\, \lambda_0)$,

then

$$d(\hat{\lambda}_n,\, \Lambda_0) \to 0 \qquad \textit{and} \qquad \|\pi(\cdot|\hat{\lambda}_n,\, X_{1:n}) - \pi(\cdot|\lambda_0,\, X_{1:n})\|_1 \to 0 \qquad \text{a.s. } [P_0^\infty]. \tag{4.2}$$

The proof of Theorem 1 is presented in the Appendix.

Equation (4.2) shows that, if $\theta_0 \in \Theta_0$, the maximum marginal likelihood $\hat{\lambda}_n$ converges to the oracle sets of hyperameters Λ_0, almost surely-P_0^∞; thus, asymptotically it gives the best selection of the prior hyperparameter. Furthermore, strong merging holds. Roughly speaking, this means that, in the limit, the differences between the empirical Bayes and the Bayesian posterior densities tend to disappear; yet, the asymptotic oracle selection of λ is still of interest, suggesting more efficient finite sample properties of empirical Bayes with respect to a fixed choice of λ.

4·2. *Degenerate case and extension to the model choice framework*

Example 3 in Section 3 suggests that strong merging may fail when $g(\theta_0) = \infty$. We generalize this finding and show that such pathological behaviours are not so much related to strange behaviours of the sampling model $p_\theta^{(n)}$, but rather to the choice of the family of priors $\{\Pi(\cdot|\lambda),\, \lambda \in \Lambda\}$.

THEOREM 2. *Suppose that $\theta_0 \in \Theta_0^c$. Assume that $(\mathbf{A1})$ is satisfied and*

(i) *there exists $\lambda_0 \in \partial\Lambda_0$ such that $\Pi(\cdot|\lambda_0) = \delta_{\theta_0}$,*
(ii) *with $P_{\theta_0}^{(n)}$-probability going to 1, $\hat{m}(X_{1:n}) \geq p_{\theta_0}^{(n)}(X_{1:n})$,*
(iii) *the model admits a local asymptotic normality expansion in the following form: for each $\epsilon > 0$, there exists a set, with $P_{\theta_0}^{(n)}$-probability going to 1, wherein, uniformly in $\theta \in U_\epsilon$,*

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

$\hat{\theta}_n$ denoting the maximum likelihood estimator and $l_n(\theta) = \log(p_\theta^{(n)}(X_{1:n}))$.
(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 cannot merge strongly with any Bayes posterior $\Pi(\cdot|\lambda,\, X_{1:n})$, with $\lambda \in \Lambda$ such that the prior density $\pi(\cdot|\lambda)$ is positive and continuous at θ_0.

Thus, under regularity assumptions, if $\theta_0 \in \Theta_0^c$ and there exists $\lambda_0 \in \partial\Lambda$ such that $\Pi(\cdot|\lambda_0) = \delta_{\theta_0}$, the empirical Bayes posterior cannot merge strongly with $\Pi(\cdot|\lambda,\, X_{1:n})$, neither, by Theorem 1.3.1 of Ghosh and Ramamoorthi (2003), with any consistent posterior $\Pi(\cdot|X_{1:n})$ associated to a prior Π which is positive and continuous at θ_0. This includes in particular any smooth hierarchical prior.

Interestingly, Scott and Berger (2010) also encounter an analogous phenomenon in their comparison between fully Bayes and empirical Bayes approaches for variable selection in regression models. We believe this is due to the same reasons as described above, although it does not completely fit our setup because we have restricted ourselves to priors that are absolutely continuous with respect to Lebesgue measure. However, this is not a crucial difference. We describe in an informal way the link between our explanation above and their findings. First, we briefly recall their setup. They consider a regression model $Y_i = x_i^T\beta + \epsilon_i$, with $\epsilon_i \sim \mathrm{N}(0,\, \phi^{-1})$ and independent, where x_i is the $k \times 1$ vector of possible regressors; their aim is to select the best set of covariates among the k candidates. Variable selection is based on an inclusion vector $\gamma = (\gamma_1, \dots, \gamma_k) \in \{0,\, 1\}^k$, where $\gamma_j = 1$ means that jth covariate has to be included. The prior on $\theta = (\beta,\, \phi,\, \gamma)$ is defined as $\pi(\theta|\lambda) = \pi(\beta,\, \phi|\gamma)\, \pi(\gamma|\lambda)$, where $\pi(\beta,\, \phi|\gamma)$ 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, $\pi(\gamma|\lambda) = \lambda^{k_\gamma}\, (1-\lambda)^{k-k_\gamma}$. A crucial issue is how to fix the probability of inclusion λ. An empirical Bayes selection of λ based on the maximum marginal likelihood estimator considers

$$\hat\lambda_n = \operatorname{argsup}\, m(Y_{1:n}|\lambda) = \operatorname{argsup} \sum_\gamma m(Y_{1:n}|\gamma)\lambda^{k_\gamma}(1-\lambda)^{k-k_\gamma}.$$

Here $m(Y_{1:n}|\gamma)$ acts as the likelihood. Each model is regular so that, under $P_{\theta_0}^{(n)}$, with $\theta_0 = (\beta_0,\, \phi_0,\, \gamma_0)$, β_0 denoting the true k-dimensional vector of regression coefficient, with some elements possibly equal to zero, which also gives the indicators γ_0 associated to the true model, and writing $\boldsymbol{\beta}_{0,\gamma}$ the restriction of the $\boldsymbol{\beta}_0$ to the coefficients present in model γ:

$$\frac{m(Y_{1:n}|\gamma)}{p_{\theta_0}^{(n)}} \approx \frac{c_\gamma \pi(\phi_0,\, \beta_{0,\gamma}|\gamma)\, e^{l_n(\hat\theta_\gamma)-l_n(\theta_0)}}{n^{(k_\gamma+1)/2}}, \qquad c_\gamma = O_{P_{\theta_0}}(1)$$

Hence the marginal likelihood $m(Y_{1:n}|\gamma)$ concentrates asymptotically at $\gamma = \gamma_0$, in the sense that $m(Y_{1:n}|\gamma_0)/m(Y_{1:n}|\gamma)$ goes to infinity under P_{θ_0} for any $\gamma \neq \gamma_0$ and

$$m(Y_{1:n}|\lambda) \approx m(Y_{1:n}|\gamma_0)\pi(\gamma_0|\lambda) \tag{4.3}$$

as n goes to infinity. We can thus apply our notion of oracle value for λ, which is $\lambda_0 = \operatorname{argsup} \pi(\gamma_0|\lambda) = k_{\gamma_0}/k$. If $\gamma_0 = (0, \dots, 0)$, then $\lambda_0 = 0$, and if $\gamma_0 = (1, \dots, 1)$, $\lambda_0 = 1$, which correspond to degenerate distributions on γ. Note that in the case of model selection, the discrete nature of problem does not prevent the merging of the empirical Bayes posterior with a posterior associated to a fixed λ or with a hierarchical prior.

This is not merely specific of the linear regression example, and, in a general model choice framework with competing models M_j, $j = 1, \dots, J$, where $\theta = (\beta_j,\, M_j)$ is decomposed into a model indicator M_j and the parameter β_j within the model M_j, with prior of the kind $P(M_j|\lambda)\pi_j(\beta_j|M_j)$, the behaviour of marginal maximum likelihood estimator $\hat\lambda_n$ is driven by the asymptotic behaviour of $m(Y_{1:n}|M_j)$. In many model selections procedures $m(Y_{1:n}|M^0)/m(Y_{1:n}|M_j)$ converges to infinity for all $M_j \neq M^0$ under $P_{\theta_0}^{(n)}$, with M^0 denoting the true model so that (4.3) remains valid and $\hat{\lambda}_n$ asymptotically maximizes $P(M^0|\lambda)$. Depending on the form of $P(M^0|\lambda)$, the empirical Bayes prior distribution can be degenerate or not.

4·3. *Example: Regression with g-priors*

As a final example, we consider the canonical Gaussian regression model $Y = 1\alpha + X\beta + \varepsilon$, with $\varepsilon \sim \mathrm{N}(0, \sigma^2 I)$, where $Y = (Y_1, \ldots, Y_n)^T$ is the response vector, X is the $(n \times k)$ fixed design matrix of full rank and I denotes the n-dimensional identity matrix. With abuse of notation, we denote by X also the design matrix whose columns have been re-centered so that $1^T X = 0^T$, in which case β can be estimated separately from α using OLS estimators. We assume that the matrix $n^{-1}(X^T X)$ is positive definite and 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 for instance Clyde and George (2000), George and Foster (2000), Liang *et al.* (2008), is

$$\pi(\alpha, \sigma^2) \propto \sigma^{-2}, \qquad \beta|\sigma^2 \sim \mathrm{N}(0, g\sigma^2 (X^T X)^{-1}), \quad g > 0, \tag{4.4}$$

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 estimate gives (see Liang *et al.* (2008), equation (9) page 413), $\hat{g}_n = \max\{F_n - 1, 0\}$, $F_n = R^2(n-1-k)/[k(1-R^2)]$, R^2 being the ordinary coefficient of determination. Thus, $\hat{g}_n = 0$ if $F_n \leq 1$. Suppose that Y is generated by the model with parameter values $\alpha_0, \beta_0, \sigma_0^2$. It turns out that

$$\begin{cases} \underline{\lim}_{n\to\infty} \mathrm{P}(\hat{g}_n = 0) = \underline{\lim}_{n\to\infty} \mathrm{P}(F_n \leq 1) \geq \gamma > 0, & \text{if } \beta_0 = 0, \\ \lim_{n\to\infty} \mathrm{P}(\hat{g}_n > 0) = \lim_{n\to\infty} \mathrm{P}(F_n > 1) = 1, & \text{if } \beta_0 \neq 0. \end{cases} \tag{4.5}$$

Interestingly, when $\beta_0 = 0$, the probability that the $\hat{g}_n$ takes the value zero 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 1, as $n \to \infty$. To prove (4.5), let $\hat{\beta}$ be the ordinary least squares estimators and

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

If $\beta_0 = 0$, then $F_n \equiv \tilde{F}_n \xrightarrow{\text{a.s.}} \chi_k^2/k$, because $\mathrm{SSE}/(n-k-1) \xrightarrow{\text{a.s.}} \sigma_0^2$, and $\underline{\lim}_{n\to\infty} \mathrm{P}(\hat{g}_n = 0) \geq \mathrm{P}(\chi_k^2/k \leq 1) =: \gamma > 0$.

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

$$R_n = \frac{n^{-1}[(\beta_0 - 2\hat{\beta})^T (X^T X)\beta_0]/k}{\mathrm{SSE}/(n-1-k)} \xrightarrow{\text{a.s.}} -\frac{(\beta_0^T V \beta_0)/k}{\sigma_0^2} < 0,$$

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

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

The consequences of (4.5) on strong merging are clear. By direct computations, whatever $\beta_0 \in \mathbb{R}^k$, for each $g > 0$, the Bayesian posterior $\Pi(\cdot|g, Y)$ is consistent: $\Pi(\cdot|g, Y) \overset{\text{a.s.}}{\Longrightarrow} \delta_{\beta_0}$. Let $\Omega_n = \{\hat{g}_n = 0\}$. Clearly, $\Omega_n \subseteq \{\Pi(\cdot|\hat{g}_n, Y) = \delta_0\}$. Then, if $\beta_0 = 0$, for each $g > 0$, $\underline{\lim}_{n\to\infty} \mathrm{P}(d_{\mathrm{TV}}(\Pi(\cdot|g, Y), \Pi(\cdot|\hat{g}_n, Y)) = 1) > 0$, where d_{TV} denotes the total variation distance. Therefore, there is a set of positive probability wherein strong merging cannot take place. If

$\beta_0 \neq 0$, for each $g > 0$, by direct computations, $\mathrm{P}(\|\pi(\cdot|g,\, Y) - \pi(\cdot|\hat{g}_n,\, Y)\|_1 \to 0) \to 1$. Thus, in this case strong merging takes place on a set with probability tending to 1.

5. Final remarks

In this paper, we formalized some common knowledge about the asymptotic equivalence of Bayes and empirical Bayes methods. Our aim was not to encourage the use of empirical Bayes; of course, when honest prior information is available, the Bayesian approach is the natural way of incorporating it in the analysis. Yet empirical Bayes is commonly used when a data-driven choice of the prior hyperparameters is desirable, as a computationally simpler alternative to a Bayesian hierarchical specification of the prior. We gave a more rigorous justification of this common practice. Further finite sample comparison of empirical and hierarchical Bayes is a natural development. Our results about the empirical Bayes asymptotic oracle selection of the prior hyperparameters suggest more efficient use of the information than a Bayesian solution based on a fixed choice of λ, when honest information for a subjective selection of λ is not available. Similar asymptotic oracle selection of λ can be envisaged for hierarchical Bayes procedures. These issues are beyond the scope of this work, but our results underline some key aspects that we believe provide the ground for further general finite sample comparisons.

Acknowledgements. This work originated from a question by Persi Diaconis. We are grateful to him and Jim Berger for stimulating discussions. S.P. has been partially supported by the Italian Ministry of University and Research, grant 2008MK3AFZ and by the ANR grant Bandhits ANR-2010-BLAN-0113-01

6. Appendix

6·1. *Proof of Proposition* 2

Fix $\epsilon > 0$. The posterior probability $\Pi(U_\epsilon^c|\hat{\lambda}_n,\, X_{1:n})$ can be written as

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

By definition of $\hat{m}(X_{1:n})$, with P_0^∞-probability 1, $D_n \geq m(X_{1:n}|\, \lambda_0)/p_{\theta_0}^{(n)}(X_{1:n}) \equiv D_n(\lambda_0)$ for all large n, where λ_0 is as required in $(\mathbf{A2})$. Thus, $\Pi(U_\epsilon^c|\hat{\lambda}_n,\, X_{1:n}) \leq N_n/D_n(\lambda_0)$. Under $(\mathbf{A1})$, by (3.2), $N_n < e^{-c_1 n\epsilon^2}$ for all large n, P_0^∞-almost surely. Reasoning as in Lemma 10 of Barron (1988), page 23, for any $\eta > 0$, $D_n(\lambda_0) > e^{-n\eta}$ for all large n, P_0^∞-almost surely. Choosing $0 < \eta < c_1\epsilon^2$, for $\delta = (c_1\epsilon^2 - \eta) > 0$, we have $\Pi(U_\epsilon^c|\hat{\lambda}_n,\, X_{1:n}) = N_n/D_n \leq N_n/D_n(\lambda_0) < e^{-n\delta}$ for all large n, P_0^∞-almost surely. The assertion follows.

6·2. *Proof of Proposition* 3

Fix $\epsilon > 0$. Set $N_n = \int_{U_\epsilon^c} R(p_\theta^{(n)})\, \mathrm{d}\Pi(\theta|\hat{\lambda}_n)$, under $(\mathbf{A1})$, $N_n < e^{-c_1 n\epsilon^2}$ for all large n, P_0^∞-almost surely. Let $D_n = \int_\Theta R(p_\theta^{(n)})\, \mathrm{d}\Pi(\theta|\hat{\lambda}_n)$. In order to bound from below D_n, it is convenient to refer to the probability space, say $(\Omega,\, \mathcal{F},\, \mathrm{P})$, wherein the X_i's are defined. Let $\mu_n^{(\omega)}(\cdot) = \tilde{\Pi}(\cdot|\hat{\lambda}_n(\omega))$, $n = 1,\, 2,\, \ldots,$ and $\mu_0(\cdot) = \tilde{\Pi}(\cdot|\lambda_0)$. Let $\Omega_0 = \{\omega \in \Omega : \mu_n^{(\omega)} \Rightarrow \mu_0\}$. By assumption (i), $\mathrm{P}(\Omega_0) = 1$. For any $\omega \in \Omega_0$, by Skorohod's theorem (cf. Theorem 1.8 in Ethier and Kurtz (1986), pages 102–103), there exists a probability space $(\Omega',\, \mathcal{F}',\, \rho)$ on which

T-valued random elements $Y_n^{(\omega)}$, $n = 1, 2, \ldots$, and Y_0 are defined such that $Y_n^{(\omega)} \sim \mu_n^{(\omega)}$, $n = 1, 2, \ldots$, $Y_0 \sim \mu_0$ and $d(Y_n^{(\omega)}(\omega'), Y_0(\omega')) \to 0$ for ρ-almost every $\omega' \in \Omega'$. Let $\Omega_1 = \{\omega \in \Omega :$ (3.3) holds true$\}$. Clearly, $\mathrm{P}(\Omega_0 \cap \Omega_1) = 1$. Fix $\omega \in (\Omega_0 \cap \Omega_1)$. For any $\eta > 0$, let

$$S_{\eta/2}^{(\omega)} = \left\{ (\tau, \zeta) \in K_{\eta/2} : \lim_{n\to\infty} \frac{1}{n} \log \frac{p^{(n)}_{(\tau_0,\zeta_0)}}{p^{(n)}_{(\tau_n,\zeta)}}(X_{1:n}(\omega)) = \mathrm{KL}_\infty((\tau_0, \zeta_0); (\tau, \zeta)) \quad \text{for all } \tau_n \to \tau \right\}.$$

By assumptions (ii)-(iii), $\Pi(S_{\eta/2}^{(\omega)} | \lambda_0) > 0$. Defined $D_{\eta/2}^{(\omega)} = \{(\omega', \zeta) : (Y_0(\omega'), \zeta) \in S_{\eta/2}^{(\omega)}\}$,

$$\int_{\mathrm{Z}} \int_{\Omega'} \mathbf{I}_{D_{\eta/2}^{(\omega)}}(\omega', \zeta)\, \mathrm{d}\rho(\omega')\, \mathrm{d}\tilde{\Pi}(\zeta) = \Pi(S_{\eta/2}^{(\omega)} | \lambda_0) > 0. \tag{6.1}$$

By Fubini's theorem, a change of measure and Fatou's lemma,

$$\begin{aligned}
\varliminf_{n\to\infty} e^{n\eta} D_n &\geq \int_{\mathrm{Z}} \int_{\Omega'} \varliminf_{n\to\infty} \exp\left\{ n \left[\eta - \frac{1}{n} \log \frac{p^{(n)}_{(\tau_0,\zeta_0)}}{p^{(n)}_{(Y_n^{(\omega)}(\omega'),\zeta)}}(X_{1:n}(\omega)) \right] \right\} \mathrm{d}\rho(\omega')\, \mathrm{d}\tilde{\Pi}(\zeta) \\
&\geq \int_{\mathrm{Z}} \int_{\Omega'} \mathbf{I}_{D_{\eta/2}^{(\omega)}}(\omega', \zeta) \\
&\quad \times \varliminf_{n\to\infty} \exp\left\{ n \left[\eta - \frac{1}{n} \log \frac{p^{(n)}_{(\tau_0,\zeta_0)}}{p^{(n)}_{(Y_n^{(\omega)}(\omega'),\zeta)}}(X_{1:n}(\omega)) \right] \right\} \mathrm{d}\rho(\omega')\, \mathrm{d}\tilde{\Pi}(\zeta) = \infty,
\end{aligned}$$

because the integrand is equal to ∞ over a set of positive probability, see (6.1). Thus, $D_n > e^{-n\eta}$ for all large n, P_0^∞-almost surely. Choosing $0 < \eta < c_1\epsilon^2$, for $\delta = (c_1\epsilon^2 - \eta) > 0$, we have $\Pi(U_\epsilon^c | \hat{\lambda}_n, X_{1:n}) = N_n/D_n < e^{-n\delta}$ for all large n, P_0^∞-almost surely. The assertion follows.

6·3. *Proof of Theorem* 1

We begin by proving (4.1). From (ii), for each $\lambda_0 \in \tilde{\Lambda}_0$, P_0^∞-almost surely, $m(X_{1:n}|\lambda_0) > 0$ for all large n. By definition of $\hat{\lambda}_n$, $0 < m(X_{1:n}|\lambda_0) \leq \hat{m}(X_{1:n}) < \infty$ for all large n, whence for all large n

$$\frac{\hat{m}(X_{1:n})}{m(X_{1:n}|\lambda_0)} \geq 1, \tag{6.2}$$

P_0^∞-almost surely. We prove the reverse inequality. Using (**A1**), (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 $1 - c_2(n\epsilon^2)^{-(1+t)}$, $\forall\, \lambda \in \Lambda$

$$\begin{aligned}
\frac{m(X_{1:n}|\lambda)}{p^{(n)}_{\theta_0}(X_{1:n})} &< e^{-c_1 n\epsilon^2} + \int_{U_\epsilon} R(p^{(n)}_\theta) \pi(\theta|\lambda)\, \mathrm{d}\nu(\theta) \leq e^{-c_1 n\epsilon^2} + \int_{U_\epsilon} R(p^{(n)}_\theta) g(\theta)\, \mathrm{d}\nu(\theta) \\
&< e^{-c_1 n\epsilon^2} + (1 + \delta/3) \int_{U_\epsilon} R(p^{(n)}_\theta) g(\theta_0)\, \mathrm{d}\nu(\theta) \\
&< e^{-c_1 n\epsilon^2} + (1 + 2\delta/3) \int_{U_\epsilon} R(p^{(n)}_\theta) \pi(\theta|\lambda_0)\, \mathrm{d}\nu(\theta),
\end{aligned}$$

where the second inequality descends from the definition of g, because $\pi(\theta|\lambda) \leq g(\theta)$ 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 $g(\theta_0) = \pi(\theta_0|\lambda_0)$, together with the continuity of $\pi(\theta|\lambda_0)$ at θ_0. By the first Borel-Cantelli lemma, for any $\delta > 0$, there exists $\epsilon > 0$ so that for all large n, $\forall\,\lambda \in \Lambda$, $m(X_{1:n}|\lambda)/p_{\theta_0}^{(n)}(X_{1:n}) < e^{-c_1 n\epsilon^2} + (1+2\delta/3)\int_{U_\epsilon} R(p_\theta^{(n)})\pi(\theta|\lambda_0)\,\mathrm{d}\nu(\theta)$, P_0^∞-almost surely. The Kullback-Leibler condition on $\Pi(\cdot|\lambda_0)$ implies that, on a set of P_0^∞-probability 1, for all large n

$$\forall\, a > 0, \qquad \int_{U_\epsilon} R(p_\theta^{(n)})\pi(\theta|\lambda_0)\,\mathrm{d}\nu(\theta) > e^{-an}. \tag{6.3}$$

Therefore, for any $\delta > 0$, on a set of P_0^∞-probability 1, for each $\lambda \in \Lambda$, $m(X_{1:n}|\lambda) \leq (1+\delta)m(X_{1:n}|\lambda_0)$ for all large n, which, combined with (6.2), proves (4.1).

We now prove the convergence of $\hat{\lambda}_n$. Recall that, by assumption $(\mathbf{A1})$, for any $\epsilon > 0$, on a set of P_0^∞-probability 1, $\forall\,\lambda \in \Lambda$, for all large n, $m(X_{1:n}|\lambda)/p_{\theta_0}^{(n)}(X_{1:n}) < e^{-c_1 n\epsilon^2} + \int_{U_\epsilon} R(p_\theta^{(n)})\pi(\theta|\lambda)\,\mathrm{d}\nu(\theta)$. For $\delta > 0$, define $N_\delta = \{\lambda \in \Lambda : \, d(\lambda,\, \Lambda_0) \leq \delta\}$. For any fixed $\delta > 0$, by assumption (iii), there exist $\epsilon_1,\, \eta > 0$ so that, on a set of P_0^∞-probability 1,

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

whence, using (i) and (ii) on the continuity of g and $\pi(\cdot|\lambda_0)$, $\lambda_0 \in \tilde{\Lambda}_0$, at θ_0,

$$\sup_{\lambda \in N_\delta^c} \frac{m(X_{1:n}|\lambda)}{p_{\theta_0}^{(n)}(X_{1:n})} < e^{-c_1 n\epsilon_1^2} + (1-\eta/2)\frac{m(X_{1:n}|\lambda_0)}{p_{\theta_0}^{(n)}(X_{1:n})}.$$

Using (6.3), we finally get that $\sup_{\lambda \in N_\delta^c} m(X_{1:n}|\lambda) < (1-\eta/4)m(X_{1:n}|\lambda_0)$ for all large n, P_0^∞-almost surely. The fact that η is fixed implies that, for n large enough, $\hat{\lambda}_n \in N_\delta$, a.s. $[P_0^\infty]$. Since Λ_0 is included in the interior of Λ, with P_0^∞-probability 1, $\hat{\lambda}_n$ belongs to the interior of Λ and $\Pi(\cdot|\hat{\lambda}_n) \ll \nu$ for all large n. This fact, combined with consistency of both the empirical Bayes posterior and $\Pi(\cdot|\lambda_0,\, X_{1:n})$, and the convergence in (4.1), yields that, P_0^∞-almost surely, for any $\epsilon > 0$,

$$\begin{aligned}
\|\pi(\cdot|\hat{\lambda}_n,\, X_{1:n}) - \pi(\cdot|\lambda_0,\, X_{1:n})\|_1 &\leq \epsilon + \int_{U_\epsilon} p_\theta^{(n)}(X_{1:n})\left|\frac{\pi(\theta|\hat{\lambda}_n)}{\hat{m}(X_{1:n})} - \frac{\pi(\theta|\lambda_0)}{m(X_{1:n}|\lambda_0)}\right|\mathrm{d}\nu(\theta) \\
&\leq \epsilon + \left|\frac{\hat{m}(X_{1:n})}{m(X_{1:n}|\lambda_0)} - 1\right| \\
&\quad + \int_{U_\epsilon} \frac{p_\theta^{(n)}(X_{1:n})}{\hat{m}(X_{1:n})}|\pi(\theta|\hat{\lambda}_n) - \pi(\theta|\lambda_0)|\,\mathrm{d}\nu(\theta) \\
&\leq 2\epsilon + \int_{U_\epsilon} \frac{p_\theta^{(n)}(X_{1:n})}{\hat{m}(X_{1:n})}|\pi(\theta|\hat{\lambda}_n) - \pi(\theta|\lambda_0)|\,\mathrm{d}\nu(\theta)
\end{aligned}$$

for n large enough. We split U_ϵ into $D_\epsilon = \{\theta \in U_\epsilon : \, \pi(\theta|\hat{\lambda}_n) \geq \pi(\theta|\lambda_0)\}$ and $D_\epsilon^c = \{\theta \in U_\epsilon : \pi(\theta|\hat{\lambda}_n) < \pi(\theta|\lambda_0)\}$. Since, for any $\delta > 0$, if ϵ is small enough, $\pi(\theta|\hat{\lambda}_n) \leq \pi(\theta|\lambda_0)(1+\delta/3)$,

$$\int_{D_\epsilon} p_\theta^{(n)}(X_{1:n})[\pi(\theta|\hat{\lambda}_n) - \pi(\theta|\lambda_0)]\,\mathrm{d}\nu(\theta) \leq \frac{\delta}{3}\int_{D_\epsilon} p_\theta^{(n)}(X_{1:n})\pi(\theta|\lambda_0)\,\mathrm{d}\nu(\theta) \leq \frac{\delta}{3}\hat{m}(X_{1:n}). \tag{6.4}$$

From consistency of the empirical Bayes posterior,

$$\int_{U_\epsilon} p_\theta^{(n)}(X_{1:n})\pi(\theta|\lambda_0)\,\mathrm{d}\nu(\theta) \le \hat m(X_{1:n}) < \int_{U_\epsilon} p_\theta^{(n)}(X_{1:n})\pi(\theta|\hat\lambda_n)\,\mathrm{d}\nu(\theta) + (\epsilon+\delta/3)\hat m(X_{1:n}),$$

whence

$$\int_{D_\epsilon^c} p_\theta^{(n)}(X_{1:n})[\pi(\theta|\lambda_0)-\pi(\theta|\hat\lambda_n)]\,\mathrm{d}\nu(\theta) \le \int_{D_\epsilon} p_\theta^{(n)}(X_{1:n})[\pi(\theta|\hat\lambda_n)-\pi(\theta|\lambda_0)]\,\mathrm{d}\nu(\theta) + (\epsilon+\delta/3)\hat m(X_{1:n})$$

and, using (6.4), $\int_{D_\epsilon^c} p_\theta^{(n)}(X_{1:n})[\pi(\theta|\lambda_0)-\pi(\theta|\hat\lambda_n)]\,\mathrm{d}\nu(\theta) \le (\epsilon+2\delta/3)\hat m(X_{1:n})$, which implies that for all large n

$$\int_{U_\epsilon} \frac{p_\theta^{(n)}(X_{1:n})}{\hat m(X_{1:n})}|\pi(\theta|\hat\lambda_n)-\pi(\theta|\lambda_0)|\,\mathrm{d}\nu(\theta) \le (\epsilon+\delta).$$

Thus, (4.2) is proved and the proof is complete.

6·4. *Proof of Theorem* 2

Define, for any $\delta>0$, the set $\Omega_{n,\delta}$ of $x_{1:n}$'s such that $e^{l_n(\hat\theta_n)-l_n(\theta_0)} \le 1+\delta$. From assumption (iv), for every $\delta>0$, $\underline{\lim}_{n\to\infty} P_{\theta_0}^{(n)}(\Omega_{n,\delta})>0$. From assumption (ii), $\hat m(X_{1:n})/p_{\theta_0}^{(n)}(X_{1:n}) \ge 1$. We now study the reverse inequality. Using $(\mathbf{A1})$, for any $\epsilon>0$, on a set A_n with $P_{\theta_0}^{(n)}$-probability going to 1, $\hat m(X_{1:n})/p_{\theta_0}^{(n)}(X_{1:n}) = \int_{U_\epsilon} e^{l_n(\theta)-l_n(\theta_0)}\mathrm{d}\Pi(\theta|\hat\lambda_n) + O(e^{-n\delta})$. Moreover, using the LAN condition (iii), for every $\theta\in U_\epsilon$,

$$l_n(\theta)-l_n(\theta_0) = l_n(\hat\theta_n)-l_n(\theta_0) + \frac{-n(\theta-\hat\theta_n)'I(\theta_0)(\theta-\hat\theta_n)}{2}(1+o_p(1)),$$

so that, if $M_n = M\sqrt{(\log n)/n}$, with $M>0$, on a set of $P_{\theta_0}^{(n)}$-probability going to 1, for all $H>0$,

$$\int_{\|\theta-\hat\theta_n\|>M_n} e^{l_n(\theta)-l_n(\hat\theta_n)}\,\mathrm{d}\Pi(\theta|\hat\lambda_n) = O(n^{-H})$$

provided M is large enough. This leads to

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

where $U_{M_n} = \{\theta : \|\theta-\hat\theta_n\| \le M_n\}$. With abuse of notation, we still denote by A_n the set having $P_{\theta_0}^{(n)}$-probability going to 1 wherein the above computations are valid, so that, on $A_n\cap\Omega_{n,\delta}$, n large enough. $\hat m(X_{1:n})/p_{\theta_0}^{(n)}(X_{1:n}) \le 1+2\delta$ Let $\lambda\in\Lambda$ be such that the prior density $\pi(\cdot|\lambda)$ is positive and continuous at θ_0. Under assumptions (iii) and $(\mathbf{A1})$, usual Laplace expansion of the marginal distribution of $X_{1:n}$ yields

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

so that $m(X_{1:n}|\lambda)/\hat m(X_{1:n}) = o_p(1)$. We now study the L_1-distance between the two posteriors. If $\Pi(\cdot|\hat\lambda_n)$ is degenerate, that is it is not absolutely continuous w.r.t. Lebesgue measure, which plays here the role of ν), then the L_1-distance between the empirical Bayes posterior and the posterior corresponding to $\Pi(\cdot|\lambda)$ is 1. Thus, we only need to consider the case where $\Pi(\cdot|\hat{\lambda}_n)$ is absolutely continuous w.r.t. Lebesgue measure. On a set of $P_{\theta_0}^{(n)}$-probability going to 1, which we still denote by A_n, intersected with $\Omega_{n,\delta}$, for each $\theta \in U_{M_n}$,

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

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

$$\int_{V_n \cap \{\|u\| \leq cM_n\sqrt{n}\}} \mathrm{d}u \leq (2\pi)^{-k/2}(1-2\delta)^{-1}. \tag{6.5}$$

Note that, on V_n^c, $\pi(\theta|\hat{\lambda}_n)(2\pi)^{k/2}/(n^{k/2}|I(\theta_0)|^{1/2}) < 1 - 2\delta$, so that

$$\pi(\theta|\hat{\lambda}_n)(1+\delta)\frac{(2\pi)^{k/2}}{n^{k/2}|I(\theta_0)|^{1/2}} - 1 < -\delta$$

and we can bound from below the L_1-distance between the two posteriors: on $A_n \cap \Omega_{n,\delta}$,

$$\begin{aligned}
\int_\Theta |\pi(\theta|\hat{\lambda}_n,\, X_{1:n}) - \pi(\theta|\lambda,\, X_{1:n})|\,\mathrm{d}\theta &\geq \int_{V_n^c \cap U_{M_n}} |\pi(\theta|\hat{\lambda}_n,\, X_{1:n}) - \pi(\theta|\lambda,\, X_{1:n})|\,\mathrm{d}\theta \\
&\geq \delta \int_{V_n^c \cap U_{M_n}} e^{-n(\theta-\hat{\theta}_n)'I(\theta_0)(\theta-\hat{\theta}_n)/2}\frac{n^{k/2}|I(\theta_0)|^{1/2}}{(2\pi)^{k/2}}\,\mathrm{d}\theta \\
&\geq \delta \int_{V_n^c \cap \{\|u\| \leq cM\sqrt{\log n}\}} \phi(u)\,\mathrm{d}u,
\end{aligned}$$

for some $c > 0$, since $I(\theta_0)$ is positive definite and where $\phi(\cdot)$ is the density of a standard Gaussian distribution on $\mathbb{R}^k$. By choosing $L > 0$ large enough and using (6.5),

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

which completes the proof.

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.

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

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*, 2nd ed. New York: Springer-Verlag.

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.

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

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.

Ethier, S. N. & Kurtz, T. G. (1986). *Markov Processes: Characterization and Convergence*. John Wiley & Sons, Inc.

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.

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

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

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. (2007a). Convergence rates of posterior distributions for noniid observations. *Ann. Statist.* **35**, 192–223.

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

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.

Petrone, S. & Raftery, A. E. (1997). A note on the Dirichlet process prior in Bayesian nonparametric inference with partial exchangeability. *J. Statist. Plann. Inference* **36**, 69–83.

Schervish, M. J. (1995). *Theory of Statistics*. New York: Springer-Verlag.

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

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

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.

[*Received November* 2012]