METRON (2014) 72:201–215
DOI 10.1007/s40300-014-0044-1

Empirical Bayes methods in classical and Bayesian inference

Sonia Petrone · Stefano Rizzelli · Judith Rousseau · Catia Scricciolo

Received: 27 April 2014 / Accepted: 5 May 2014 / Published online: 3 June 2014

Abstract Empirical Bayes methods are often thought of as a bridge between classical and Bayesian inference. In fact, in the literature the term empirical Bayes is used in quite diverse contexts and with different motivations. In this article, we provide a brief overview of empirical Bayes methods highlighting their scopes and meanings in different problems. We focus on recent results about merging of Bayes and empirical Bayes posterior distributions that regard popular, but otherwise debatable, empirical Bayes procedures as computationally convenient approximations of Bayesian solutions.

Keywords Bayesian weak merging · Compound experiments · Frequentist strong merging · Hyper-parameter oracle value · Latent distribution · Maximum marginal likelihood estimation · Shrinkage estimation

1 Introduction

Empirical Bayes methods are popularly employed by researchers and practitioners and are attractive in appearing to bridge frequentist and Bayesian approaches to inference. In fact, a frequentist statistician would find just a formal Bayesian flavor in empirical Bayes methods, while a Bayesian statistician would say that there is nobody less Bayesian than an empirical Bayesian (Lindley, in [6]). Further confusing, in the literature the term empirical Bayes is used in quite diverse contexts, with different motivations. Classical empirical Bayes methods arose in the context of compound experiments, where a latent distribution driving experiment-specific parameters formally acts as a prior on each one such parameter and is estimated from the data, usually by maximum likelihood. The term empirical Bayes is also used in the context of purely Bayesian inference when hyper-parameters of a subjective prior distribu-

S. Petrone (✉) · S. Rizzelli · C. Scricciolo
Bocconi University, Milan, Italy
e-mail: sonia.petrone@unibocconi.it

J. Rousseau
CREST-ENSAE and CEREMADE, Université Paris Dauphine, Paris, France

tion are selected through the data. Empirical Bayes estimates are also popularly employed to deal with nuisance parameters. All these situations are different and require specific analysis.

In this article, we give a brief overview of classical and recent results on empirical Bayes methods, discussing their use in these different contexts. Section 2 recalls classical empirical Bayes methods in compound sampling problems and mixture models. Although arising as a way to by-pass the need of specifying a prior distribution in computing optimal Bayesian solutions [31], here the approach is purely frequentist. The empirical Bayes solution is basically shrinkage estimation, where the introduction of a latent distribution may facilitate interpretation and modeling thus helping to design efficient shrinkage.

In a broader sense, the term empirical Bayes is used to denote a data-driven selection of prior hyper-parameters in Bayesian inference. We discuss this case in Sect. 3. Here, the prior distribution can only have an interpretation in terms of subjective probability on an unknown, but fixed, parameter. Although not rigorous, it is a common practice to try to overcome difficulties in specifying the prior distribution by plugging in some estimates of the prior hyper-parameters. From a Bayesian viewpoint, in such cases one should rather assign a hyper-prior distribution, which however makes computations more involved. The empirical Bayes selection thus appears as a convenient way out that is expected to give similar inferential results as the hierarchical Bayesian solution for large sample sizes and better results for finite samples than a "wrong choice" of the prior hyper-parameters. Although commonly trusted, these facts are not rigorously proved. Recent results [30] address the presumed asymptotic equivalence of Bayesian and empirical Bayes solutions in terms of merging. Roughly speaking, they show that, in regular parametric problems, the empirical Bayes and the Bayesian posterior distributions generally tend to merge, that is, to be asymptotically close, but also that possible divergent behavior may arise. Thus, the use of empirical Bayes prior selection requires much care.

Section 4 discusses another popular use of empirical Bayes methods in problems with nuisance parameters. We extend, in particular, the results of [30] on weak merging of empirical Bayes procedures to nuisance parameter problems, which we illustrate with partial linear regression models.

The result about merging recalled in Sect. 3 only gives first-order asymptotic comparison between empirical Bayes and any Bayesian posterior distributions. A higher-order comparison would be needed to distinguish among them. We conclude the article with some hints on the finite-sample behavior of the empirical Bayes posterior distribution in a simple but insightful example (Sect. 5). The results suggest that, when merging holds, the empirical Bayes posterior distribution can indeed be a computationally convenient approximation of an efficient, in a sense to be specified, Bayesian solution.

2 Classical empirical Bayes

The introduction of the empirical Bayes method is traditionally associated with Robbins' article [31] on compound sampling problems. Compound sampling models arise in a variety of situations including multi-site clinical trials, estimation of disease rates in small geographical areas, longitudinal studies. In this setting, n values $\theta_1, \ldots, \theta_n$ are drawn at random from a latent distribution G. Then, conditionally on $\theta_1, \ldots, \theta_n$, observable random variables $X_1, \ldots, X_n$ are drawn independently from probability distributions $p(\cdot \mid \theta_1), \ldots, p(\cdot \mid \theta_n)$, respectively. The framework can be thus described:

$$X_i \mid \theta_i \overset{\text{indep}}{\sim} p(\cdot \mid \theta_i)$$
$$\theta_i \mid G \overset{\text{iid}}{\sim} G(\cdot), \qquad i = 1, \ldots, n,$$

where the index i refers to the ith experiment. Interest lies in estimating an experiment specific parameter θ_i when all the n observations $X_1, \ldots, X_n$ are available. For the generic ith experiment, one has $X_i \mid \theta_i \sim p(\cdot \mid \theta_i)$ and $\theta_i \sim G$; thus, the latent distribution G formally plays the role of a prior distribution on θ_i in a Bayesian flavor. Were G known, inference on θ_i would be carried out through the Bayes' rule, computing the posterior distribution of θ_i given X_i, $\mathrm{d}G(\theta_i \mid X_i) \propto p(X_i \mid \theta_i)\,\mathrm{d}G(\theta_i)$, and θ_i could be estimated by the Bayes' estimator with respect to squared error loss, i.e., the posterior mean $\mathbb{E}_G[\theta_i \mid X_i] = \int \theta \mathrm{d}G(\theta \mid X_i)$. In fact, in general G is unknown and the Bayes' estimator $\mathbb{E}_G[\theta_i \mid X_i]$ is not computable. One can however use an estimate of the "prior distribution" G based on the available observations $X_1, \ldots, X_n$, which is what originated the term "empirical Bayes". Were $\theta_1, \ldots, \theta_n$ observable, their common distribution G could be pointwise consistently estimated by the empirical cumulative distribution function (cdf) $\widehat{G}_n(\theta) = \sum_{i=1}^n 1_{(-\infty,\,\theta]}(\theta_i)$. As the θ_i are not observable, the empirical Bayes approach suggests estimating G from the data $X_1, \ldots, X_n$ exploiting the fact that

$$X_i \mid G \overset{\text{iid}}{\sim} f_G(\cdot) = \int p(\cdot \mid \theta)\,\mathrm{d}G(\theta), \qquad i = 1, \ldots, n.$$

We still denote by $\widehat{G}_n$ any estimator for G based on $X_1, \ldots, X_n$. As in [31], consider $i = n$, that is, estimating θ_n. The unit-specific unknown θ_n can be estimated by the empirical Bayes version $\mathbb{E}_{\widehat{G}_n}[\theta \mid X_n]$ of the posterior mean. Empirical Bayes methods considered in [31] have been named *nonparametric* empirical Bayes, because G is assumed to be completely unknown, to distinguish them from *parametric* empirical Bayes methods later developed by Efron and Morris [11–15], where G is assumed to be known up to a finite-dimensional parameter. If G is completely unknown, then the cdf $F_G(x) = \int_{-\infty}^x f_G(u)\,\mathrm{d}u$, $x \in \mathbb{R}$, can be estimated from the empirical cdf $\widehat{F}_n(x) = \sum_{i=1}^n 1_{(-\infty,\,x]}(X_i)$ which, for every fixed x, tends to $F_G(x)$, as $n \to \infty$, whichever the mixing distribution G. Thus, depending on the kernel density $p(\cdot \mid \theta)$ and the class $\mathcal{G}$ to which G belongs, the estimator $\widehat{G}_n$ entailed by $\widehat{F}_n$ approximates G for large n and the corresponding empirical Bayes' estimator $\mathbb{E}_{\widehat{G}_n}[\theta \mid X_n]$ for θ_n approximates the posterior mean $\mathbb{E}_G[\theta \mid X_n]$. To illustrate this, we consider the following example due to Robbins [31], which deals with the Poisson case. Here, whatever the unknown distribution G, the posterior mean can be written as the ratio of the probability mass function $f_G(\cdot)$ evaluated at different points. These terms can be estimated by the corresponding values of the empirical mass function.

Example 1 Let $X_i \mid \theta_i \sim \text{Poisson}(\theta_i)$ independently, with $\theta_i \mid G \overset{\text{iid}}{\sim} G$, $i = 1, \ldots, n$, where G is a cdf on $\mathbb{R}^+$. In this case, $\mathbb{E}_G[\theta \mid X = x] = (x+1) f_G(x+1)/f_G(x), x = 0,\ 1, \ldots$, which can be estimated by $\varphi_n(x) = (x+1) \sum_{i=1}^n 1_{\{x+1\}}(X_i) / \sum_{i=1}^n 1_{\{x\}}(X_i)$. Then, whatever the unknown distribution G, for any fixed x, $\varphi_n(x) \to \mathbb{E}_G[\theta \mid X = x]$ as $n \to \infty$, with probability 1. This naturally suggests using $\varphi_n(X_n)$ as an estimator for θ_n. Robbins [31] extended this technique to the cases where the X_i has geometric, binomial or Laplace distribution.

As discussed by Morris [29], parametric empirical Bayes procedures are needed to deal with those cases where n is too small to well approximate the Bayesian solution, but still a substantial improvement over standard methods can be made as for the James–Stein's estimator. When the mixing distribution is assumed to have a specific parametric form $G(\cdot \mid \psi)$, it is common practice to estimate the unknown parameter ψ from the data by maximum

likelihood, computing $\widehat{\psi}_n \equiv \widehat{\psi}(X_1, \ldots, X_n)$ as $\widehat{\psi}_n = \text{argmax}_\psi \prod_{i=1}^n \int p(X_i \mid \theta)\, \text{d}G(\theta \mid \psi)$. Inference on θ_n is then carried out using $G(\cdot \mid \widehat{\psi}_n)$ to compute $\mathbb{E}_{G(\cdot \mid \widehat{\psi}_n)}[\theta \mid X_n]$. Empirical Bayes estimation of θ_n has the advantage of doing asymptotically as well as the Bayes' estimator without knowing the "prior" distribution. However, the Bayesian approach and the empirical Bayes approach are only seemingly related: there is, indeed, a clearcut difference between them. In the empirical Bayes approach to compound problems, although G formally acts as a prior distribution on a single parameter, its introduction is motivated in a frequentist sense, as the common distribution of the random sample $(\theta_1, \ldots, \theta_n)$; indeed, estimation of G is carried out by frequentist methods. In the Bayesian approach, a prior distribution can be assigned to a *fixed* unknown parameter, being interpreted as a formalization of subjective information. In the context of multiple independent experiments, a Bayesian statistician would rather assume probabilistic dependence across experiments by regarding the θ_i as exchangeable and assigning a prior probability law to G (in the nonparametric case) or to ψ (in the parametric case); see, e.g., [1,4,8].

Rather than in comparison with Bayesian inference, the advantage of empirical Bayes methods can be appreciated in comparison with classical maximum likelihood estimators. The empirical Bayes estimate of θ_i makes efficient use of the available information because all data are used when estimating G. In other terms, empirical Bayes techniques involve learning from the experience of others or, using Tukey's evocative expression, "borrowing strength". To illustrate this crucial aspect, we consider the following classical example.

Example 2 Let $(X_1, \ldots, X_p)' \sim \mathcal{N}_p(\theta, \sigma^2 I_p)$ be a p-variate Gaussian distribution, where $\theta = (\theta_1, \ldots, \theta_p)$ and I_p is the p-dimensional identity matrix. The X_i can be the mean of a random sample $X_{i,j}$, $j = 1, \ldots, n$, within the ith experiment. Suppose σ^2 is known. Let $\theta_i \mid \psi \overset{\text{iid}}{\sim} \mathcal{N}(0, \psi)$, with unknown variance ψ. Then, the maximum marginal likelihood estimator for ψ is $\widehat{\psi}_p = \max\{0, s^2 - \sigma^2\}$, where $s^2 = \sum_{i=1}^p x_i^2/p$. The empirical Bayes' estimator for θ_i is $\mathbb{E}_{\mathcal{N}(0, \widehat{\psi}_p)}[\theta \mid X_i] = [1-(p-2)\sigma^2/\sum_{i=1}^p x_i^2]X_i$, which coincides with the James–Stein's estimator [23,33], that dominates the maximum likelihood estimator $\hat{\theta}_i = X_i$ for $p \geq 3$ with respect to the overall quadratic loss $\sum_{i=1}^p (\theta_i - \hat{\theta}_i)^2$.

As remarked by Morris [29], James–Stein's estimator is minimax-optimal for the sum of the individual squared error loss functions only in the equal variances case. Optimality is lost, for example, if global loss functions that weight differently the individual squared losses are used. Other forms of shrinkage, possibly suggested by the empirical Bayes approach, are then necessary.

We conclude this section with a historical note. Although the introduction of the empirical Bayes method is traditionally associated with Robbins [31]'s article, the idea was partially anticipated by, among others, Gini [21] who, as pointed out by Forcina [18], pioneerly provided empirical Bayes solutions for estimating the parameter of a binomial distribution, and by Fisher et al. [17] who applied the parametric empirical Bayes technique to the so-called species sampling problem assuming a Gamma "prior" distribution, see also Good [22]. Since then, the field has witnessed a tremendous growth both in terms of theoretical developments as well as in diversity of applications, see, e.g., the monographs [27] and [10].

3 Empirical Bayes selection of prior hyper-parameters

In a broader sense, the term empirical Bayes is commonly associated with general techniques that make use of a data-driven choice of the prior distribution in Bayesian inference. Here,

the basic setting is inference on an exchangeable sequence (X_i). Exchangeability is intended in a subjective sense: the data are physically independent, but probabilistic dependence is expressed among them, as past observations give information on future values and such incomplete information is described probabilistically through the conditional distribution of $X_{n+1}, X_{n+2}, \ldots$, given $X_1 = x_1, \ldots, X_n = x_n$. Exchangeability is the basic dependence assumption, which is equivalent to assuming a statistical model $p(\cdot \mid \theta)$ such that the X_i are conditionally independent and identically distributed (iid) given θ and expressing a prior distribution on θ, by de Finetti's representation theorem for exchangeable sequences. Thus, the statistical model and the prior are together a way of expressing the probability law of the observable sequence (X_i) and in such way they should be chosen. In fact, choosing a honest subjective prior in Bayesian inference can be a difficult task. A way of formulating such uncertainty is to assign the prior on θ hierarchically, assuming $\theta \mid \lambda \sim \Pi(\cdot \mid \lambda)$, a parametric distribution depending on hyper-parameters λ, and $\lambda \sim H(\lambda)$. However, this often complicates computations, so that it is a common practice to plug in some estimate $\hat{\lambda}_n$ of the prior hyper-parameters as a shortcut. The resulting data-dependent prior $\Pi(\cdot \mid \hat{\lambda}_n)$, combined with the likelihood, results into a pseudo-posterior distribution $\Pi(\cdot \mid \hat{\lambda}_n, X_1, \ldots, X_n)$ that is commonly referred to as empirical Bayes. Many types of estimators for λ are considered, the most popular being the maximum marginal likelihood estimator, defined as

$$\hat{\lambda}_n \in \operatorname*{argmax}_{\lambda \in \bar{\Lambda}} \int \prod_{i=1}^{n} p(X_i \mid \theta)\, \Pi(\mathrm{d}\theta \mid \lambda),$$

where $\bar{\Lambda}$ is the closure of Λ.

Such empirical Bayes approach is appealing in offering the possibility of making Bayesian inference by-passing a complete specification of the prior and it is largely used in practical applications and in the literature: see, e.g., [7,19,25,32] in the context of variable selection in regression, [5] for wavelet shrinkage estimation, [26] and [28] in Bayesian nonparametric mixture models, [16] in Bayesian nonparametric inference for species diversity, [2,3] and [34] in Bayesian nonparametric procedures for curve estimation.

Although popular, this mixed approach is not rigorous from a Bayesian point of view. Its interest mainly lies in being a computationally simpler alternative to a more rigorous, but usually analytically more complex, hierarchical specification of the prior: one expects that, when the sample size is large, the empirical Bayes posterior distribution will be close to some Bayesian posterior distribution. Moreover, for finite samples, a data-driven empirical Bayes selection of the prior hyper-parameters is expected to give better inferential results than a "wrong choice" of λ. These commonly believed facts do not seem to be rigorously proved in the literature. A recent work by Petrone et al. [30] addresses the supposed asymptotic equivalence between empirical Bayes and Bayesian posterior distributions in terms of merging.

Two notions of merging are considered: Bayesian weak merging in the sense of [9], and frequentist strong merging in the sense of [20]. Bayesian weak merging compares posterior distributions in terms of weak convergence, with respect to (wrt) the exchangeable probability law of (X_i). Roughly speaking, we have weak merging of the empirical Bayes and Bayesian posterior distributions if any Bayesian statistician is sure that her/his posterior distribution and the empirical Bayes posterior distribution will eventually be close, in the sense of weak convergence. This is a minimal requirement, but it is not guaranteed. From results in [9], it can be proved that weak merging holds if and only if the empirical Bayes posterior distribution is consistent in the frequentist sense at the true value θ_0, whatever θ_0. Consistency at θ_0 means

that the sequence of empirical Bayes posterior distributions weakly converges to a point mass at θ_0, almost surely wrt $P_{\theta_0}^{\infty}$, where $P_{\theta_0}^{\infty}$ denotes the probability law of (X_i) such that the X_i are iid according to P_{θ_0}.

Sufficient conditions for consistency of empirical Bayes posterior distributions are provided in [30], Section 3. In general, consistency of Bayesian posterior distributions does not imply consistency of empirical Bayes posteriors. For the latter, one has to control the asymptotic behavior of the estimator $\hat{\lambda}_n$, too. If $\hat{\lambda}_n$ is the maximum marginal likelihood estimator, its properties can be exploited to show that the empirical Bayes posterior distribution is consistent at θ_0 under essentially the same conditions which ensure consistency of Bayesian posterior distributions. For more general estimators, conditions become more cumbersome. When $\hat{\lambda}_n$ is a convergent sequence, sufficient conditions are given in Proposition 3 of [30], based on a change of the prior probability measure such that the dependence on the data is transferred from the prior to the likelihood.

Even when consistency and weak merging hold, the empirical Bayes posterior distribution may underestimate the uncertainty on θ and diverge from any Bayesian posterior, relatively to a stronger metric than the one of weak convergence. This behavior is illustrated in the following example.

Example 3 Let $X_i \mid \theta \sim \mathcal{N}(\theta, \sigma^2)$ independently, with σ^2 known, and $\theta \sim \mathcal{N}(\mu, \tau^2)$. Consider empirical Bayes inference where the prior variance $\lambda = \tau^2$ is estimated by the maximum marginal likelihood estimator, the prior mean μ being fixed. Then, see, e.g., [24], p. 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 empirical Bayes posterior distribution $\Pi(\cdot \mid \hat{\tau}_n^2,\ X_1, \ldots, X_n)$ 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 $(1/\hat{\tau}_n^2 + n/\sigma^2)^{-1}$. Since $\hat{\tau}_n^2$ can be equal to zero with positive probability, the empirical Bayes posterior can be degenerate at μ. The probability of the event $\hat{\tau}_n = 0$ converges to zero when $\theta_0 \neq \mu$, but 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 relatively to some distance; however, if $\theta_0 = \mu$, there is a positive probability that the empirical Bayes and the Bayesian posterior distributions are singular. The possible degeneracy of the empirical Bayes posterior distribution is pathological in the sense that the uncertainty on the parameter is a posteriori underestimated.

Such behaviour is not restricted to the Gaussian distribution and applies more generally to location-scale families of priors. If the model admits a maximum likelihood estimator $\hat{\theta}_n$ and the prior density is of the form $\tau^{-1}g((\cdot - \mu)/\tau)$, with $\lambda = (\mu,\ \tau)$ for some unimodal density g that attains the maximum at zero, then $\hat{\lambda}_n = (\hat{\theta}_n,\ 0)$ and the empirical Bayes posterior is a point mass at $\hat{\theta}_n$. These families of priors should not be jointly used with maximum marginal likelihood empirical Bayes procedures.

A way to refine the analysis to better understand the impact of a data-dependent prior on the posterior distribution is to study frequentist strong merging in the sense of [20]. Two sequences of posterior distributions are said to merge strongly if their total variation distance converges to zero almost surely wrt $P_{\theta_0}^{\infty}$.

Strong merging of Bayesian posterior distributions in nonparametric contexts is often impossible since pairs of priors are typically singular. Petrone et al. [30] study the problem for regular parametric models, comparing Bayesian posterior distributions and empirical Bayes posterior distributions based on the maximum marginal likelihood estimator of λ. Informally, their results show that strong merging may hold for some true values θ_0, but may fail for others. That is, for values of θ_0 in an appropriate set, say Θ_0, the empirical Bayes

posterior distribution strongly merges with any Bayesian posterior distribution corresponding to a prior distribution q which is continuous and bounded at θ_0,

$$\|\Pi(\cdot \mid \hat{\lambda}_n,\ X_1, \ldots, X_n) - q(\cdot \mid X_1, \ldots, X_n)\|_{\mathrm{TV}} \to 0 \tag{1}$$

almost surely wrt $P_{\theta_0}^{\infty}$, where $\|\cdot\|_{\mathrm{TV}}$ denotes the total variation distance. However, for $\theta_0 \notin \Theta_0$, strong merging fails: the empirical Bayes posterior can indeed be singular wrt any smooth Bayesian posterior distribution.

More precisely, suppose that the prior distribution has density $\pi(\cdot)$ with respect to some dominating measure and includes θ_0 in its Kullback–Leibler support. Furthermore, suppose that the parameter space is totally bounded. Assume that conditions hold that guarantee consistency for the empirical Bayes and the Bayesian posterior distributions. Under such assumptions, and some additional requirements that are satisfied by regular parametric models, it can be shown ([30], Theorem 1) that the maximum marginal likelihood estimator $\hat{\lambda}_n$ converges to a value λ^* (here assumed to be unique for brevity) such that $\pi(\theta_0 \mid \lambda^*) \geq \pi(\theta_0 \mid \lambda)$ for every λ in the hyper-parameter space Λ. Such value can be interpreted as the "oracle value" of the hyper-parameter, that is the value of the hyper-parameter for which the prior mostly favors the true value θ_0. Furthermore, it is proved that if θ_0 is such that $\pi(\theta_0 \mid \lambda^*) < \infty$, then strong merging holds, namely

$$\|\Pi(\cdot \mid \hat{\lambda}_n,\ X_1, \ldots, X_n) - \Pi(\cdot \mid \lambda^*,\ X_1, \ldots, X_n)\|_{\mathrm{TV}} \to 0 \tag{2}$$

almost surely wrt $P_{\theta_0}^{\infty}$. Since $\|\Pi(\cdot \mid \lambda^*,\ X_1, \ldots, X_n) - q(\cdot \mid X_1, \ldots, X_n)\|_{\mathrm{TV}}$ goes to zero $P_{\theta_0}^{\infty}$-almost surely for any prior q that is continuous and bounded at θ_0 ([20], Theorem 1.3.1), by the triangular inequality, one has (1). However, if θ_0 is such that $\pi(\theta_0 \mid \lambda^*) = \infty$, then strong merging fails. This is the case if, for such θ_0, λ^* is in the boundary of Λ and the prior distribution is degenerate at θ_0 for $\lambda \to \lambda^*$. In this case, the empirical Bayes posterior distribution is degenerate too, thus it is singular wrt any smooth Bayesian posterior.

Result (1), which holds only in the non-degenerate case, ensures that the empirical Bayes posterior distribution will be close in total variation to the Bayesian posterior, whatever the prior distribution. But this result only provides a first-order asymptotic comparison that does not distinguish among Bayesian solutions. In fact, from (2), one could expect that the empirical Bayes approach can actually give a closer approximation of an efficient, in the sense of using the prior distribution that mostly favors the true value of θ_0, Bayesian solution. Higher-order asymptotic results are beyond the scope of this note, but we will return to this issue in Sect. 5, providing a simple, but we believe insightful, example.

4 Empirical Bayes selection of nuisance parameters

Another relevant context of application of empirical Bayes methods concerns Bayesian analysis in semi-parametric models, where estimation of nuisance parameters is preliminarily considered to carry out inference on the component of interest. The framework can be thus described: observations $X_1, \ldots, X_n$ are drawn independently from a distribution with density $p_{\psi,\lambda}(\cdot)$,

$$X_i \mid (\psi,\ \lambda) \overset{\text{iid}}{\sim} p_{\psi,\lambda}(\cdot), \qquad i = 1, \ldots, n,$$

where $\psi \in \Psi \subseteq \mathbb{R}^k$ is the parameter of interest and $\lambda \in \Lambda \subseteq \mathbb{R}^{\ell}$ a nuisance parameter. Bayesian inference with nuisance parameters does not conceptually present particular difficulties: a prior distribution is assigned to the overall parameter $(\psi,\ \lambda)$,

$$\Pi(\mathrm{d}\psi,\ \mathrm{d}\lambda) = \Pi(\mathrm{d}\psi \mid \lambda)\Pi(\mathrm{d}\lambda),$$

and inference on ψ is carried out marginalizing the joint posterior distribution $\Pi(\mathrm{d}\psi,\ \mathrm{d}\lambda \mid X_1, \ldots, X_n)$. However, this can be computationally cumbersome. A common approach is thus to plug in some estimator $\hat{\lambda}_n$ of λ and use a data-dependent prior

$$\Pi(\mathrm{d}\psi \mid \hat{\lambda}_n)\ \delta_{\hat{\lambda}_n}(\mathrm{d}\lambda). \tag{3}$$

We highlight the difference between the present context and the one described in Sect. 3: there, $\hat{\lambda}_n$ is used to estimate a hyper-parameter λ, a parameter of the prior only, whereas here $\hat{\lambda}_n$ is used to estimate λ which is, in the first place, a component of the overall parameter of the model, it is part of the model.

The results developed in [30] can be extended to prove the asymptotic equivalence, in terms of weak merging, between the empirical Bayes posterior and any Bayesian posterior for ψ, provided $\hat{\lambda}_n$ is a sequence of *consistent* estimators for the true value λ_0 corresponding to the density p_{ψ_0,λ_0} generating the observations. It is known from Proposition 1 in [30] that a necessary and sufficient condition for weak merging is that the empirical Bayes posterior for ψ is consistent at ψ_0, namely, $\Pi(\cdot \mid \hat{\lambda}_n,\ X_1, \ldots, X_n)$ weakly converges to a point mass at ψ_0, $\Pi(\cdot \mid \hat{\lambda}_n,\ X_1, \ldots, X_n) \Rightarrow \delta_{\psi_0}$ along almost all sample paths when sampling from the infinite product measure $P^{\infty}_{\psi_0,\lambda_0}$. To illustrate how this assertion can be shown, we present an example on partially linear regression.

Example 4 Suppose we observe a random sample from the distribution of $X = (Y,\ V,\ W)$, in which, for some unobservable error e independent of $(V,\ W)$, the relationship among the components is described by

$$Y = \psi V + \eta_\lambda(W) + e.$$

The independent variable Y is a regression on $(V,\ W)$ that is linear in V with slope ψ, but may depend on W in a nonlinear way through $\eta_\lambda(W)$ which represents an additive contamination of the linear structure of Y. We assume that V and W take values in $[0,\ 1]$ and that, for $\lambda \in \Lambda \subseteq \mathbb{R}^\ell$, the function $w \mapsto \eta_\lambda(w)$ is known up to λ. If the error e is assumed to be normal, $e \sim \mathcal{N}(0,\ \sigma_0^2)$ with known variance σ_0^2, then the density of X is given by

$$p_{\psi,\lambda}(x) = \phi_{\sigma_0}(y - \psi v - \eta_\lambda(w))\, p_{V,W}(v,\ w), \qquad x = (y,\ v,\ w) \in \mathbb{R} \times [0,\ 1]^2,$$

where $\phi_{\sigma_0}(\cdot) = \sigma_0^{-1}\phi(\cdot/\sigma_0)$, with ϕ the standard Gaussian density, and $p_{V,W}$ the joint density of $(V,\ W)$. Consider an empirical Bayes approach that estimates λ by any sequence $\hat{\lambda}_n$ of consistent estimators for λ_0 and use the empirical Bayes posterior $\Pi(\cdot \mid \hat{\lambda}_n,\ X_1, \ldots, X_n)$ corresponding to a prior of the form in (3) to carry out inference on ψ. The empirical Bayes posterior $\Pi(\cdot \mid \hat{\lambda}_n,\ X_1, \ldots, X_n)$ weakly merges with the posterior for ψ corresponding to any genuine prior on $(\psi,\ \lambda)$ if only $\Pi(\cdot \mid \hat{\lambda}_n,\ X_1, \ldots, X_n)$ is consistent at ψ_0. We show that, for every $\delta > 0$, the empirical Bayes posterior probability $\Pi(|\psi - \psi_0| > \delta \mid \hat{\lambda}_n,\ X_1, \ldots, X_n) \to 0$ in $P^n_{\psi_0,\lambda_0}$-probability. Let $m_{\psi,\lambda}(v,\ w) = \psi v + \eta_\lambda(w)$. Assume there exists a constant $B > 0$ such that $\sup_{\psi,\lambda} \|m_{\psi,\lambda}\|_\infty \leq B$. Since the Hellinger distance $h(p_{\psi,\lambda_0},\ p_{\psi_0,\lambda_0}) \geq (\mathbb{E}[V^2])^{1/2} e^{-B^2/4\sigma_0^2} |\psi - \psi_0|/2\sigma_0$, the inclusion $\{\psi : |\psi - \psi_0| > \delta\} \subseteq \{\psi : h(p_{\psi,\lambda_0},\ p_{\psi_0,\lambda_0}) > M\delta\}$ holds for a suitable positive constant M. To prove the claim, it is therefore enough to study the asymptotic behavior of $\Pi(h(p_{\psi,\lambda_0},\ p_{\psi_0,\lambda_0}) > M\delta \mid \hat{\lambda}_n,\ X_1, \ldots, X_n)$ which, if the prior for ψ, given λ, belongs to a location family of ν-densities generated by $\pi_0(\cdot)$, i.e., $\pi(\cdot \mid \lambda) = \pi_0(\cdot - \lambda)$, is equal to

$$\begin{aligned}
&\Pi(h(p_{\psi,\lambda_0},\ p_{\psi_0,\lambda_0}) > M\delta \mid \hat{\lambda}_n,\ X_1,\ldots,X_n)\\
&\quad = \frac{\int_{h(p_{\psi,\lambda_0},\ p_{\psi_0,\lambda_0})>M\delta} \prod_{i=1}^n \phi_{\sigma_0}(Y_i - m_{\psi,\hat{\lambda}_n}(V_i,\ W_i))\pi_0(\psi - \hat{\lambda}_n)\nu(\mathrm{d}\psi)}{\int \prod_{i=1}^n \phi_{\sigma_0}(Y_i - m_{\psi_0,\lambda_0}(V_i,\ W_i))\pi_0(\psi - \hat{\lambda}_n)\nu(\mathrm{d}\psi)}\\
&\quad = \frac{N(X_1,\ldots,X_n)}{D(X_1,\ldots,X_n)}.
\end{aligned}$$

Assume there exists a continuous function $g : [0,\ 1] \to \mathbb{R}$ and $\alpha > 0$ such that, for any $\lambda,\ \lambda_0 \in \Lambda$, the difference $|\eta_\lambda(w) - \eta_{\lambda_0}(w)| \le |g(w)|\|\lambda - \lambda_0\|^\alpha \le \|g\|_\infty \|\lambda - \lambda_0\|^\alpha$ for every $w \in [0,\ 1]$. Then, on the event $\Omega_n = (-a_n \le \min_i Y_i \le \max_i Y_i \le a_n,\ \ \|\hat{\lambda}_n - \lambda_0\| \le u_n)$, which, for sequences $u_n \downarrow 0$ and $a_n = O((\log n)^\kappa)$, $\kappa > 0$, has probability $P^n_{\psi_0,\lambda_0}(\Omega_n) = 1 + o(1)$, we have

$$\begin{aligned}
N(X_1,\ldots,X_n)/D(X_1,\ldots,X_n) \le \exp\{2n(u_n + \|g\|_\infty u_n^\alpha)(a_n + B) + nu_n^2\}\\
\times \Pi(h(p_{\psi,\lambda_0},\ p_{\psi_0,\lambda_0}) > M\delta \mid \lambda_0,\ X_1,\ldots,X_n).
\end{aligned}$$

If the Bayesian posterior $\Pi(\cdot \mid \lambda_0,\ X_1,\ldots,X_n)$ is Hellinger consistent at P_{ψ_0,λ_0} and the convergence is exponentially fast, then also the empirical Bayes posterior $\Pi(\cdot \mid \hat{\lambda}_n,\ X_1,\ldots,X_n)$ is consistent at P_{ψ_0,λ_0} and the claim that $\Pi(|\psi - \psi_0| > \delta \mid \hat{\lambda}_n,\ X_1,\ldots,X_n) \to 0$ follows.

5 Higher-order comparisons and finite-sample properties

We return to the discussion at the end of Sect. 3, providing a simple example. Although limited to the Gaussian case, this gives some hints about finer comparisons between Bayesian and empirical Bayes posterior distributions. The evidence in this example could be extended to more general contexts, such as Bayesian inference and variable selection in linear regression with g-priors.

As discussed, an empirical Bayes choice of the prior hyper-parameters in Bayesian inference is not rigorous, but can be of interest as an approximation of a computationally more involved hierarchical Bayesian posterior distribution. In fact, the results recalled in Sect. 3 show that, even for regular parametric models, the empirical Bayes posterior distribution can be singular wrt any smooth Bayesian posterior, depending on the form of the prior distribution and on the nature of the prior hyper-parameters. Thus, care is needed when using empirical Bayes methods as an approximation of Bayesian solutions. On a positive side, these results show that, in non-degenerate cases, the empirical Bayes posterior distribution does merge strongly with any smooth Bayesian posterior distribution. However, this first-order asymptotic comparison does not distinguish among Bayesian posterior distributions arising from different priors. The aim here is to grasp some evidence for finer comparisons. We explore the following two issues.

Asymptotically, in regular parametric models, any smooth Bayesian posterior distribution is approximated by a Gaussian distribution centered at the maximum likelihood estimate $\hat{\theta}_n$, by the Bernstein–von Mises theorem. Strong merging of Bayesian and empirical Bayes posterior distributions implies that, when a Bernstein–von Mises behavior holds for the Bayesian posterior distribution, it also holds for the empirical Bayes posterior; which is a particularly interesting implication in nonparametric problems. In fact, one would expect that the empirical Bayes posterior distribution can provide a closer approximation to a hierarchical Bayesian posterior than the Bernstein–von Mises Gaussian distribution.

Less informally, based on the results of Sect. 3, one would conjecture that the hierarchical posterior distribution concentrates around the oracle value λ^* of the prior hyperparameters for increasing sample sizes and, since $\hat{\lambda}_n \to \lambda^*$, the empirical Bayes posterior distribution $\Pi(\theta \mid \hat{\lambda}_n, X_1, \ldots, X_n)$ and the hierarchical Bayesian posterior distribution $\Pi_h(\theta \mid X_1, \ldots, X_n)$ can be close even for moderate sample sizes. The following example suggests that, although this is the case asymptotically by the results in Sect. 3, the posterior distribution $h(\lambda \mid X_1, \ldots, X_n)$ slowly incorporates the sample information so that, for finite samples, the empirical Bayes posterior distribution $\Pi(\theta \mid \hat{\lambda}_n, X_1, \ldots, X_n)$ is a close approximation of $\Pi_h(\theta \mid X_1, \ldots, X_n)$ only if the prior distribution on λ is enough concentrated around the oracle value λ^*. In other words, the example suggests that, in the non-degenerate case, the empirical Bayes posterior distribution is a high-order approximation of the posterior distribution of a "well informed" Bayesian researcher whose prior highly favors the true value of θ.

Example 5 Consider the simple example of the Gaussian conjugate model introduced in Sect. 3 with now a hierarchical specification of the prior. Let $X_i \mid \theta \sim \mathcal{N}(\theta, \sigma^2)$ independently, with σ^2 known. Let $\theta \mid \lambda \sim \mathcal{N}(0, \lambda)$ and $1/\lambda \sim \mathcal{G}(\alpha, \beta)$, a Gamma distribution where $\beta > 0$ is the scale parameter. Then, $\mathbb{E}(\lambda) = \beta/(\alpha - 1)$ and $\mathbb{V}(\lambda) = \beta^2/[(\alpha - 1)^2(\alpha - 2)]$. The prior of θ obtained by integrating out λ is a Student's-t with zero mean, 2α degrees of freedom and scaling factor β/α. The prior variance of θ equals the prior guess on the hyperparameter λ, $\mathbb{V}(\theta) = \mathbb{E}(\lambda)$. Although this is a simple model, computations of the posterior distribution of θ become analytically complicated. The conditional distribution of θ, given λ and the data, is

$$\theta \mid (\lambda, x_1, \ldots, x_n) \sim \mathcal{N}\left((n\lambda + \sigma^2)^{-1} n\lambda \bar{x}_n, (n\lambda + \sigma^2)^{-1}\sigma^2\lambda\right)$$

and the posterior distribution of θ is obtained by integrating λ out wrt its posterior distribution $h(\lambda \mid x_1, \ldots, x_n)$. This integration step is not analytically manageable and approximation by Markov chain Monte Carlo (MCMC) is usually employed.

The empirical Bayes selection of λ is an attractive, computationally simpler, shortcut. Estimation of λ via maximum marginal likelihood gives $\hat{\lambda}_n = \max\{0, \bar{x}_n^2 - \sigma^2/n\}$. Thus, the maximum marginal likelihood estimator $\hat{\lambda}_n$ may take value zero in the boundary of $\Lambda = (0, \infty)$ with positive probability. If $\hat{\lambda}_n = 0$, then the empirical Bayes prior distribution of θ is a point mass at the prior guess and the resulting posterior distribution is degenerate. As seen in Sect. 3, if the true value $\theta_0 = \mathbb{E}[\theta] = 0$, the probability of degeneracy remains positive even when $n \to \infty$, thus determining an asymptotic divergence between the empirical Bayes posterior distribution and the hierarchical Bayesian posterior distribution. If $\theta_0 \neq 0$, such probability goes to zero and strong merging holds. Interest is in investigating higher-order approximations in this case.

We first focus on point estimation with quadratic loss. The Bayes' estimate is the posterior expectation

$$\mathbb{E}[\theta \mid x_1, \ldots, x_n] = \frac{\int (1 + \theta^2/2\beta)^{-(2\alpha+1)/2} \exp\left\{-n\left[-\frac{1}{n}\log\theta + \frac{1}{2\sigma^2}(\theta - \bar{x}_n)^2\right]\right\} \mathrm{d}\theta}{\int (1 + \theta^2/2\beta)^{-(2\alpha+1)/2} \exp\left\{-\frac{n}{2\sigma^2}(\theta - \bar{x}_n)^2\right\} \mathrm{d}\theta}, \tag{4}$$

Table 1 Comparing empirical Bayes and Laplace point estimates as approximations of the hierarchical Bayes' point estimates

	$\mathbb{E}[\lambda] = 1/3$	$\mathbb{E}[\lambda] = 1$	$\mathbb{E}[\lambda] = 3$	$\mathbb{E}[\lambda] = 4$	$\mathbb{E}[\lambda] = 10$
(a) $n = 20$; $\bar{x}_n = 1.835$					
$\hat{\mathbb{E}}^{\mathrm{LC}}[\theta \mid x_1, \ldots, x_n]$ (Laplace appr.)	1.797	1.769	1.749	1.745	1.738
$\mathbb{E}[\theta \mid x_1, \ldots, x_n]$ (hierarc. Bayes, by MCMC) (standard dev.)	1.683 (0.0029)	1.750 (0.0031)	1.801 (0.0033)	1.805 (0.0034)	1.821 (0.0033)
$\mathbb{E}[\lambda \mid x_1, \ldots, x_n]$ (Bayes, by MCMC) (standard dev.)	0.074 (0.0051)	1.301 (0.0113)	3.018 (0.0261)	3.902 (0.0332)	8.915 (0.0721)
$\hat{\lambda}_n$ (maximum marginal lik.)	3.320	3.320	3.320	3.320	3.320
$\mathbb{E}[\theta \mid \hat{\lambda}_n, x_1, \ldots, x_n]$ (empirical Bayes)	1.809	1.809	1.809	1.809	1.809
(b) $n = 50$; $\bar{x}_n = 2.009$					
$\hat{\mathbb{E}}^{\mathrm{LC}}[\theta \mid x_{1:n}]$ (Laplace appr.)	1.994	1.982	1.972	1.970	1.967
$\mathbb{E}[\theta \mid x_1, \ldots, x_n]$ (hierarc. Bayes, by MCMC) (standard dev.)	1.951 (0.0018)	1.974 (0.0019)	1.993 (0.0022)	1.999 (0.0022)	2.005 (0.0024)
$\mathbb{E}[\lambda \mid x_1, \ldots, x_n]$ (Bayes, by MCMC) (standard dev.)	0.833 (0.0074)	1.403 (0.0123)	3.117 (0.0251)	4.012 (0.0342)	9.103 (0.0911)
$\hat{\lambda}_n$ (maximum marginal lik.)	4.016	4.016	4.016	4.016	4.016
$\mathbb{E}[\theta \mid \hat{\lambda}_n . x_1, \ldots, x_n]$ (empirical Bayes)	1.999	1.999	1.999	1.999	1.999

Simulated data with $\theta_0 = 2$ and $\sigma^2 = 1$

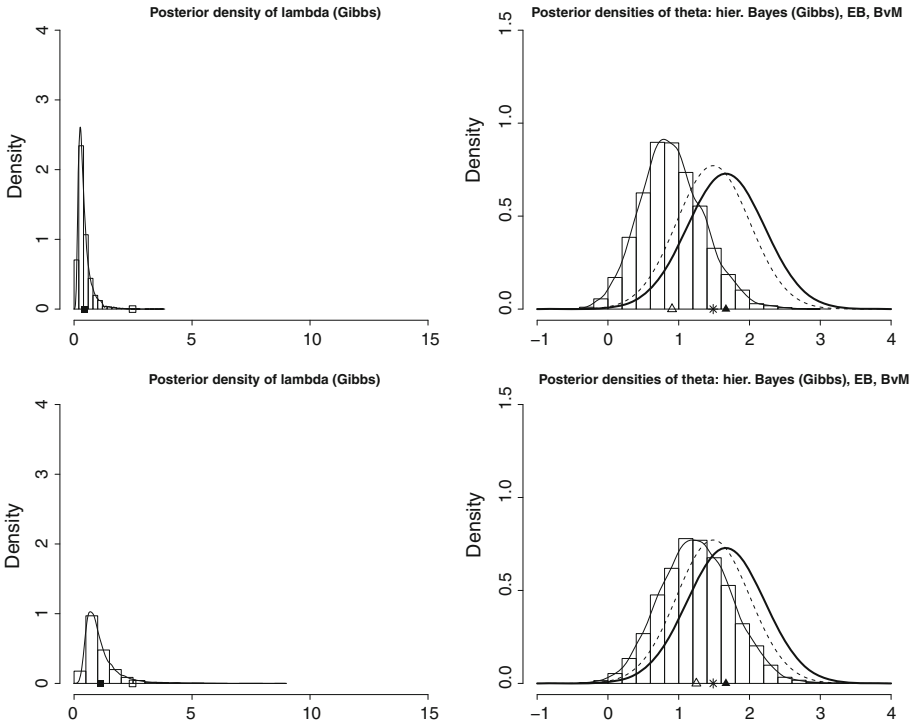

Fig. 1 Comparing empirical Bayes and hierarchical Bayesian posterior densities. Simulated data from a Gaussian distribution $\mathcal{N}(2, 6)$; $n = 20$; $\bar{x}_n = 1.667$. $\mathbb{E}[\lambda] = 1/3$ (first row) and $\mathbb{E}[\lambda] = 1$ (second row). First column: MCMC estimate of the posterior density of λ; the *full square* denotes $E(\lambda \mid x_1, \ldots, x_n)$ and the *empty square* denotes the marginal maximum likelihood estimate $\hat{\lambda}_n$. Second column: hierarchical Bayesian posterior density of θ (MCMC estimate; *solid curve*), empirical Bayes posterior density of θ (*dashed curve*) and limit Gaussian density $d\mathcal{N}(\bar{x}_n, \sigma^2/n)$ (*bold solid curve*). The *empty triangle* denotes $\mathbb{E}[\theta \mid x_1, \ldots, x_n]$; the *star* denotes $\mathbb{E}[\theta \mid \hat{\lambda}_n, x_1, \ldots, x_n]$; the *full triangle* denotes the sample mean, $\bar{x}_n$

for which a closed form expression is not available. Its empirical Bayes approximation is obtained by plugging $\hat{\lambda}_n$ into the expression of $\mathbb{E}[\theta \mid \lambda, x_1, \ldots, x_n]$:

$$\mathbb{E}[\theta \mid \hat{\lambda}_n, x_1, \ldots, x_n] = \frac{n\hat{\lambda}_n}{n\hat{\lambda}_n + \sigma^2}\bar{x}_n = \left(1 - \frac{\sigma^2}{n\hat{\lambda}_n + \sigma^2}\right)\bar{x}_n. \tag{5}$$

We may expect that

$$\begin{aligned}\mathbb{E}[\theta \mid x_1, \ldots, x_n] &= \int \mathbb{E}[\theta \mid \lambda, x_1, \ldots, x_n] h(\lambda \mid x_1, \ldots, x_n)\, \mathrm{d}\lambda \\ &= \mathbb{E}[\theta \mid \hat{\lambda}_n, x_1, \ldots, x_n] + O(n^{-k}),\end{aligned}$$

since, as n increases, $\hat{\lambda}_n$ tends to the oracle value λ^* and $h(\lambda \mid x_1, \ldots, x_n)$ could collapse to a point mass at λ^*. It is interesting to investigate on the order of the error term $O(n^{-k})$. To grasp some evidence, we compare the empirical Bayes point estimate with the Laplace approximation developed by [24], p. 270:

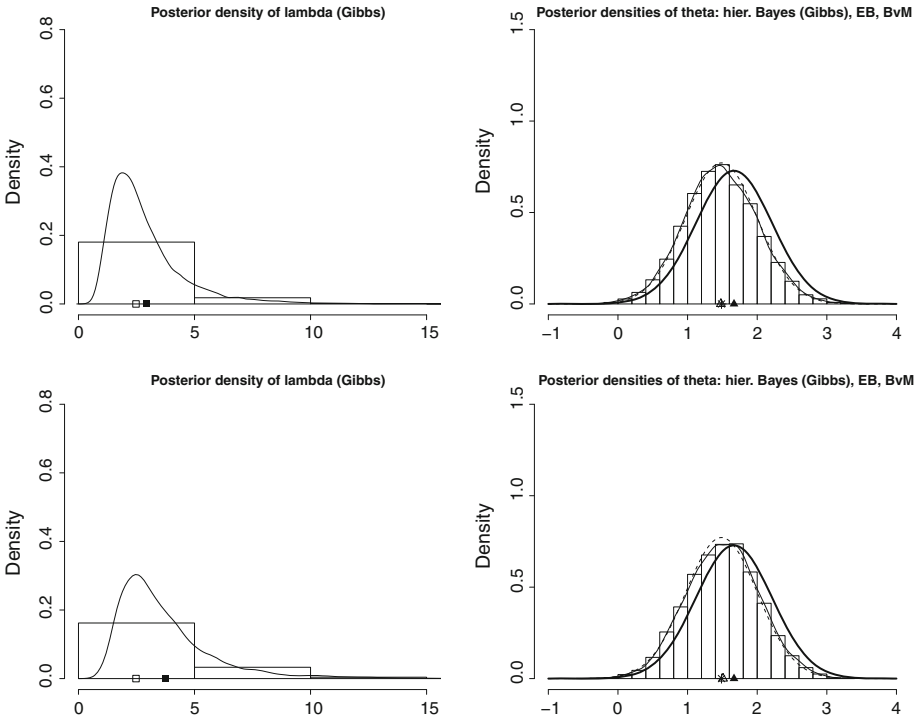

Fig. 2 Comparing empirical Bayes and hierarchical Bayesian posterior densities. Simulated data from a Gaussian distribution $\mathcal{N}(2, 6)$; $n = 20$; $\bar{x}_n = 1.667$. $\mathbb{E}[\lambda] = 3$ (first row) and $\mathbb{E}[\lambda] = 4$ (second row). Legenda as for Fig. 1

$$\hat{\mathbb{E}}^{\mathrm{LC}}[\theta \mid x_1, \ldots, x_n] = \left(1 - \frac{(2\alpha + 1)/2\alpha}{\left(1 + \bar{x}_n^2/2\beta\right)} \frac{\sigma^2}{n}\right) \bar{x}_n \tag{6}$$

that is a special case of the Laplace approximation with error term $O(n^{-3/2})$.

Table 1 compares $\mathbb{E}[\theta \mid \hat{\lambda}_n, x_1, \ldots, x_n]$ and $\hat{\mathbb{E}}^{\mathrm{LC}}[\theta \mid x_1, \ldots, x_n]$ as approximations of the hierarchical Bayes' point estimate $\mathbb{E}[\theta \mid x_1, \ldots, x_n]$ in a simulation study where $\theta_0 = 2$ and $\sigma^2 = 1$. Along the columns, the value of α is fixed at 4, while β varies, thus resulting into different prior guesses $\mathbb{E}[\lambda]$. Since $\mathbb{E}[\lambda] = \mathbb{V}(\theta)$, increasing values of β correspond to smaller precision of the hierarchical prior. When $\beta = 12$, the prior guess equals the oracle value, i.e., $\mathbb{E}[\lambda] = \lambda^* = 4$. In this case, the empirical Bayes' point estimate provides a clearly better approximation of $\mathbb{E}[\theta \mid x_1, \ldots, x_n]$ than $\hat{\mathbb{E}}^{\mathrm{LC}}[\theta \mid x_1, \ldots, x_n]$. For example, Table 1b shows how $\mathbb{E}[\theta \mid x_1, \ldots, x_n]$ and $\mathbb{E}[\theta \mid \hat{\lambda}_n, x_1, \ldots, x_n]$ coincide up to the thousandths digit for $n = 50$ and $\mathbb{E}[\lambda] = 4$. This suggests a higher-order form of merging between the empirical Bayes posterior distribution and the hierarchical posterior distribution of a "more informed" Bayesian statistician, i.e., the one who assigns a hyper-prior such that $\mathbb{E}[\lambda] = \lambda^*$. In order to shade light on this point, we now consider density approximation.

We first want to check whether the empirical Bayes posterior distribution provides a better approximation of the hierarchical Bayesian posterior distribution than the Bernstein–von Mises Gaussian approximating distribution, $\mathcal{N}(\bar{x}_n, \sigma^2/n)$. This comparison has been investigated in several simulation studies, each one giving similar indications. We report the results for simulated data from a Gaussian distribution with mean $\theta_0 = 2$ and variance

$\sigma^2 = 6$ (Figs. 1, 2). The hierarchical Bayesian posterior densities are computed by Gibbs sampling. The first column in the plots shows the posterior density $h(\lambda \mid x_1, \ldots, x_n)$ of λ. This appears to slowly concentrate towards the oracle value $\lambda^* = 4$. The second column shows the MCMC approximation of the hierarchical Bayesian posterior density of θ, together with the empirical Bayes posterior density (dashed curve) and the limit Gaussian density $\mathcal{N}(\bar{x}_n, \sigma^2/n)$ (bold curve). What emerges is that for a prior guess of λ close to the oracle value, the empirical Bayes posterior density provides a better approximation of the hierarchical Bayesian posterior density already for the small sample size $n = 20$. This seems to confirm the previously formulated conjecture: for finite sample sizes, empirical Bayes provides a good approximation of the hierarchical Bayesian procedure adopted by the more informed statistician and strong merging may hold up to a higher-order approximation.

References

1. Antoniak, C.E.: Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. Ann. Stat. **2**, 1152–1174 (1974)
2. Belitser, E., Enikeeva, F.: Empirical Bayesian test of the smoothness. Math. Methods Stat. **17**, 1–18 (2008)
3. Belitser, E., Levit, B.: On the empirical Bayes approach to adaptive filtering. Math. Methods Stat. **12**, 131–154 (2003)
4. Berry, D.A., Christensen, R.: Empirical Bayes estimation of a binomial parameter via mixtures of Dirichlet processes. Ann. Stat. **7**, 558–568 (1979)
5. Clyde, M.A., George, E.I.: Flexible empirical Bayes estimation for wavelets. J. R. Stat. Soc. Ser. B **62**, 681–698 (2000)
6. Copas, J.B.: Compound decisions and empirical Bayes (with discussion). J. R. Stat. Soc. Ser. B **31**, 397–425 (1969)
7. Cui, W., George, E.I.: Empirical Bayes vs. fully Bayes variable selection. J. Stat. Plann. Inference **138**, 888–900 (2008)
8. Deely, J.J., Lindley, D.V.: Bayes empirical Bayes. J. Am. Stat. Assoc. **76**, 833–841 (1981)
9. Diaconis, P., Freedman, D.: On the consistency of Bayes estimates. Ann. Stat. **14**, 1–26 (1986)
10. Efron, B.: Large-scale inference. Empirical Bayes methods for estimation, testing, and prediction. Cambridge University Press, Cambridge (2010)
11. Efron, B., Morris, C.: Limiting the risk of Bayes and empirical Bayes estimators. II. The empirical Bayes case. J. Am. Stat. Assoc. **67**, 130–139 (1972a)
12. Efron, B., Morris, C.: Empirical Bayes on vector observations: an extension of Stein's method. Biometrika **59**, 335–347 (1972b)
13. Efron, B., Morris, C.: Stein's estimation rule and its competitors-an empirical Bayes approach. J. Am. Stat. Assoc. **68**, 117–130 (1973a)
14. Efron, B., Morris, C.: Combining possibly related estimation problems. (With discussion by Lindley, D.V., Copas, J.B., Dickey, J.M., Dawid, A.P., Smith, A.F.M., Birnbaum, A., Bartlett, M.S., Wilkinson, G.N., Nelder, J.A., Stein, C., Leonard, T., Barnard, G.A., Plackett, R.L.). J. R. Stat. Soc. Ser. B **35**, 379–421 (1973b)
15. Efron, B., Morris, C.N.: Data analysis using Stein's estimator and its generalizations. J. Am. Stat. Assoc. **70**, 311–319 (1973c)
16. Favaro, S., Lijoi, A., Mena, R.H., Prünster, I.: Bayesian nonparametric inference for species variety with a two parameter Poisson–Dirichlet process prior. J. R. Stat. Soc. Ser. B **71**, 993–1008 (2009)
17. Fisher, R.A., Corbet, A.S., Williams, C.B.: The relation between the number of species and the number of individuals in a random sample of an animal population. J. Anim. Ecol. **12**, 42–58 (1943)
18. Forcina, A.: Gini's contributions to the theory of inference. Int. Stat. Rev. **50**, 65–70 (1982)
19. George, E.I., Foster, D.P.: Calibration and empirical Bayes variable selection. Biometrika **87**, 731–747 (2000)
20. Ghosh, J.K., Ramamoorthi, R.V.: Bayesian nonparametrics. Springer, New York (2003)
21. Gini, C.: Considerazioni sulla probabilità a posteriori e applicazioni al rapporto dei sessi nelle nascite umane. Studi Economico-Giuridici. Università di Cagliari. III. Reprinted in Metron, vol. 15, pp. 133–172 (1911)

22. Good, I.J.: Breakthroughs in statistics: foundations and basic theory. In: Johnson, N.L., Kotz, S. (eds.) Introduction to Robbins (1992) An empirical Bayes approach to statistics, pp. 379–387. Springer, Berlin (1995)
23. James, W., Stein, C.: Estimation with quadratic loss. In: Proceedings of Fourth Berkeley Symposium on Mathematics Statistics and Probability, vol. 1, pp. 361–379. University of California Press, California (1961)
24. Lehmann, E.L., Casella, G.: Theory of point estimation, 2nd edn. Springer, New York (1998)
25. Liang, F., Paulo, R., Molina, G., Clyde, M.A., Berger, J.O.: Mixtures of g-priors for Bayesian variable selection. J. Am. Stat. Assoc. **103**, 410–423 (2008)
26. Liu, J.S.: Nonparametric hierarchical Bayes via sequential imputation. Ann. Stat. **24**, 911–930 (1996)
27. Maritz, J.S., Lwin, T.: Empirical Bayes methods, 2nd edn. Chapman and Hall, London (1989)
28. McAuliffe, J.D., Blei, D.M., Jordan, M.I.: Nonparametric empirical Bayes for the Dirichlet process mixture model. Stat. Comput. **16**, 5–14 (2006)
29. Morris, C.N.: Parametric empirical Bayes inference: theory and applications. J. Am. Stat. Assoc. **78**, 47–55 (1983)
30. Petrone, S., Rousseau, J., Scricciolo, C.: Bayes and empirical Bayes: do they merge? Biometrika **101**(2), 285–302 (2014)
31. Robbins, H.: An empirical Bayes approach to statistics. In: Proceedings of Third Berkeley Symposium on Mathematics, Statistics and Probability, vol. 1, pp. 157–163. University of California Press, California (1956)
32. Scott, J.G., Berger, J.O.: Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. Ann. Stat. **38**, 2587–2619 (2010)
33. Stein, C.: Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In: Proceedings of Third Berkeley Symposium on Mathematics, Statistics and Probability, vol. 1, pp. 197–206. University of California Press, California (1956)
34. Szabó, B.T., van der Vaart, A.W., van Zanten, J.H.: Empirical Bayes scaling of Gaussian priors in the white noise model. Electron. J. Stat. **7**, 991–1018 (2013)