On Convergence Rates of Empirical Bayes Procedures

Sulla Velocità di Convergenza di Procedure Bayesiane Empiriche

Sophie Donnet, Vincent Rivoirard, Judith Rousseau and Catia Scricciolo

Abstract Empirical Bayes procedures are commonly used based on the supposed asymptotic equivalence with fully Bayesian procedures, which, however, has not so far received full theoretical support in terms of uncertainty quantification. In this note, we provide some results on contraction rates of empirical Bayes posterior distributions which are illustrated in nonparametric curve estimation using Dirichlet process mixture models.

Abstract *Le procedure bayesiane empiriche sono comunemente utilizzate sulla base di una presunta equivalenza asintotica con quelle propriamente bayesiane, la quale, tuttavia, non ha finora ricevuto piena giustificazione teorica in termini di quantificazione dell'incertezza. In questa nota si forniscono alcuni risultati sulla velocità di convergenza della distribuzione finale bayesiana empirica che sono illustrati nel caso della stima non parametrica di curve mediante modelli miscuglio di processi di Dirichlet.*

Key words: Dirichlet process mixtures, empirical Bayes selection of prior hyperparameters, nonparametric curve estimation

---

Sophie Donnet
CEREMADE, Université Paris Dauphine, Place du Maréchal de Lattre de Tassigny, 75775 Paris Cedex 16, France, e-mail: donnet@ceremade.dauphine.fr

Vincent Rivoirard
CEREMADE, Université Paris Dauphine, Place du Maréchal de Lattre de Tassigny, 75775 Paris Cedex 16, France, e-mail: rivoirard@ceremade.dauphine.fr

Judith Rousseau
CREST-ENSAE, 3, Avenue P. Larousse, 92240 Malakoff, France, e-mail: rousseau@ceremade.dauphine.fr

Catia Scricciolo
Department of Decision Sciences, Bocconi University, Via Röntgen 1, I-20136 Milano, Italy, e-mail: catia.scricciolo@unibocconi.it

1 Introduction

In a Bayesian approach to inference, the prior distribution should be chosen independently of the data. However, not always enough information is *a priori* available to honestly elicit the prior distribution by either fixing the values of its hyper-parameters or specifying a prior for them in a hierarchical approach and it is a common practice to just replace them with reasonable estimates. A data-driven selection of the prior hyper-parameters can be appealing as a computationally simpler alternative to an analytically more complex hierarchical specification of the prior. The prior is then data-dependent and this mixed approach, which arises as a way to bypass a complete specification of the prior distribution, appears as a bridge between Bayesian and frequentist approaches to inference, thus being commonly referred to as empirical Bayes. Indeed, the expression "empirical Bayes" has been originally used by [20] in compound sampling problems to denote estimation of a latent distribution generating the unknown parameters of n distributions from each one of which an observation is then drawn at random. A data-driven choice of the prior hyper-parameters is only one of the possible contexts in which the term empirical Bayes is used. Another relevant context of application of empirical Bayes methods is the analysis of semi-parametric models, where preliminary estimation of nuisance parameters is considered to then carry out inference on the component of interest. It has been illustrated in Example 4 of [18] that some of the techniques employed to work out the case of empirical Bayes selection of prior hyper-parameters can also be fruitfully used to deal with the case of a data-driven choice of the nuisance parameters.

Empirical Bayes methods are widely used based on the idea that, for large sample sizes, there is a substantial agreement between empirical Bayes and fully Bayesian procedures, in the sense that they lead to the same inferential conclusions. However, empirical Bayes procedures have so far been mainly adopted to explore possible data-dependent prior distributions with the aim of showing that empirical Bayes inference can give satisfactory results in applications, see, *e.g.*, [11] in the context of variable selection in regression, [3] for wavelet shrinkage estimation, [16], [17] in Bayesian nonparametric mixture models and [9] in Bayesian nonparametric inference for species diversity. A careful comparison of empirical Bayes and Bayesian variable selection criteria in regression is developed by [5] and [22]. However, there seems to be a substantial lack of results establishing the presumed asymptotic equivalence between empirical Bayes and fully Bayesian procedures. The recent article of [19] addresses this issue in terms of weak merging in the sense of [6], which, roughly speaking, means that 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. From results in [6], it follows that Bayesian weak merging holds if and only if the empirical Bayes posterior is consistent in the frequentist sense, see Proposition 1 in [19]. However, there are no general results concerning the asymptotic behaviour of empirical Bayes posterior distributions like rates of convergence. Studying the asymptotic behaviour of posterior distributions corresponding to data-dependent priors may allow for a better understanding of the

impact of the prior measure on the posterior distribution and the entailed inferential procedures. There have been so far hardly general results in the literature on the asymptotic behaviour of empirical Bayes procedures and most of the existing ones only concern specific families of models and priors, see, for example, [2], [25] on Gaussian priors in the white noise model and [26] for finite mixture models.

In this note, we present some general results that describe the large sample behavior of empirical Bayes posterior distributions and quantify the uncertainty in terms of contraction rates. The illustration concerns empirical Bayes estimation in the context of linear regression with unknown error distribution modeled as a Dirichlet process (DP) mixture of Gaussian distributions. Other relevant examples concerning empirical Bayes estimation of monotone non-increasing intensity functions of Aalen point processes and empirical Bayes adaptive density estimation with DP mixtures of Gaussian densities are studied in [7]. The main message is that a large class of empirical Bayes procedures have optimal frequentist asymptotic properties as fully Bayesian procedures. That such an optimal recovery is possible using empirical Bayes methods is theoretically important and encouraging to the use of these procedures in various applied settings such as the analysis of DNA microarrays, see [8]; it is also promising for the use of dynamic empirical Bayes models in finance and insurance, see, for instance, [10] and [4].

2 Main Results

In this section, we blend ideas from [19] and [7] to give a heuristic explanation of the requirements and ingredients involved in the approach followed to tackle the problem of assessing contraction rates for empirical Bayes posterior distributions which attempts at giving conditions resembling those considered in the literature for fully Bayesian posteriors, see, *e.g.*, [13].

Consider a statistical model $(P_\theta^{(n)}, \theta \in \Theta)$ over a sample space $\mathscr{X}^{(n)}$, together with a family of prior distributions $\Pi(\cdot \mid \gamma)$, $\gamma \in \Gamma \subseteq \mathbb{R}^\ell$, on the parameter space Θ equipped with a semi-metric $d(\cdot,\cdot)$. A Bayesian statistician would either set γ to a specific value or integrate it out using a probability distribution in a hierarchical specification of the prior for θ. Both approaches would lead to a prior distribution for θ that does not depend on the data. However, it is often the case that not enough information is *a priori* available to either fix a value for γ or elicit a prior distribution for it so that γ can be more easily estimated from the data using an estimator $\hat{\gamma}_n$. In our approach, a crucial assumption concerns the large sample behavior of $\hat{\gamma}_n$: we assume that, with probability going to 1, $\hat{\gamma}_n$ belongs to a subset K_n of Γ,

$$P_{\theta_0}^{(n)}(\hat{\gamma}_n \in K_n^c) = o(1), \tag{1}$$

where $P_{\theta_0}^{(n)}$ denotes the probability measure generating the data. Clearly, requirement (1) is satisfied if $\hat{\gamma}_n$ converges to some value γ^*: one can then simply take K_n to

be a fixed neighborhood of γ^*. Existence of a sequence of sets K_n wherein, with high probability, the estimator $\hat{\gamma}_n$ is going to take values, allows to get a uniform control, over K_n or some suitably chosen subset thereof, of the likelihood ratio when the dependence on the data through $\hat{\gamma}_n$ is transferred from the prior to the likelihood. In fact, the main difficulty when studying frequentist asymptotic properties of posterior distributions corresponding to data-dependent priors arises from the dependence of the prior on the data. The key idea is then to get control of the impact of $\hat{\gamma}_n$ on the prior by constructing a transformation $\psi_{\gamma,\gamma'} : \Theta \to \Theta$ such that, for all $\gamma, \gamma' \in \Gamma$, if $\theta \sim \Pi(\cdot \mid \gamma)$ then $\psi_{\gamma,\gamma'}(\theta) \sim \Pi(\cdot \mid \gamma')$. This mapping allows to make a change of the prior measure so that the dependence on the data through $\hat{\gamma}_n$ is conveniently transferred from the prior to the likelihood, in the spirit of what is done in Proposition 3 and, more concretely, in Example 3 of [19] which deals with nonparametric mixture models for which the mapping $\psi_{\gamma,\gamma'}$ remains naturally identified when the prior involves a Dirichlet process on the mixing distribution.

Another important requirement is the one according to which the prior mass should be sufficiently spread out over the parameter space Θ so that, wherever the true value θ_0 of θ lies, there is enough mass in suitable neighborhoods of $P_{\theta_0}^{(n)}$ as defined by the Kullback-Leibler divergence. To formalize this requirement, for $\varepsilon_n \downarrow 0$, define a Kullback-Leibler type neighborhood of $P_{\theta_0}^{(n)}$ as $B_{n,k}(\theta_0) = \{\theta : \mathrm{KL}(\theta_0; \theta) \leq n\varepsilon_n^2,\ \mathrm{V}_k(\theta_0; \theta) \leq n^{k/2}\varepsilon_n^k\}$, $k \geq 2$, where $\mathrm{V}_k(\theta_0; \theta)$ is the centered absolute moment of order k of $\log(p_{\theta_0}^{(n)}/p_\theta^{(n)})$ under $P_{\theta_0}^{(n)}$. Then, the refined Kullback-Leibler prior support condition is that, for fixed $k \geq 2$, $\inf_{\gamma \in K_n} \Pi(B_{n,k}(\theta_0) \mid \gamma) \gtrsim e^{-2n\varepsilon_n^2}$ for sufficiently large n, where "$\gtrsim$" denotes inequality valid up to a constant multiple which is universal or inessential for our purposes. A formal statement providing sufficient conditions for contraction rates of empirical Bayes posterior distributions can be obtained as a refinement of Proposition 3 in [19]. We now present an example on empirical Bayes posterior contraction rates in the context of linear regression with unknown error distribution that is modeled as a DP mixture of Gaussian densities.

Example. ***Linear regression with unknown error distribution***
Suppose we observe a random sample $X^{(n)} = (X_1, \ldots, X_n)$ of independent response random variables each one satisfying the relationship $X_i = \alpha + \beta z_i + \varepsilon_i$, $i = 1, \ldots, n$, where the z_i are deterministic covariates taking values in a compact set and the errors ε_i are independent and identically distributed according to a distribution with Lebesgue density p that is modeled as a DP mixture of Gaussian densities:

$$p(\cdot) = p_{F,\sigma}(\cdot) := \int \phi_\sigma(\cdot - \theta)\, \mathrm{d}F(\theta)$$

$$F \sim \mathrm{DP}(\alpha_{\mathbb{R}}\mathrm{N}(m, s^2)) \text{ independent of } \sigma \sim \mathrm{IG}(\nu_1, \nu_2), \quad \nu_1, \nu_2 > 0,$$

where $\phi_\sigma(\cdot) = \sigma^{-1}\phi(\cdot/\sigma)$, with ϕ the density of a standard Gaussian distribution and $\mathrm{N}(m, s^2)$ the Gaussian base distribution with mean m and variance s^2. Let $\gamma = (m, s^2)$. The priors for α and β are assumed to be compactly supported and to possess positive Lebesgue densities. All parameters are assumed to be *a priori*

independent. Let $\Pi(\cdot \mid \gamma)$ denote the joint prior distribution of (p, α, β). Assume that the true value (p_0, α_0, β_0) of (p, α, β) is a point in the prior support. We consider rates of contraction for the empirical Bayes posterior distribution corresponding to a DP mixture of Gaussians as the prior on the unknown error density, when a data-driven choice of the hyper-parameter γ of the base measure of the DP is performed. Posterior consistency has been studied for this model by [1], while [13] have assessed rates of convergence for fully Bayesian posterior distributions in general and, specifically, for the case where the prior on the mixing distribution is a DP with compactly supported base measure and the true error density is also a Gaussian mixture, finding the nearly parametric rate $n^{-\frac{1}{2}}(\log n)$. The key point is that, when considering rates of convergence for the parameter (p, α, β), the Euclidean part (α, β) does not really play a role: the driving component is only the density p, so that working out rates for this model essentially reduces to working out rates for DP mixtures of Gaussian densities in the problem of density estimation. Posterior contraction rates for density estimation with DP mixtures of Gaussian densities have been first investigated by [12] and [14]. Following an idea of [21], [15] have proven that finite location mixtures of Gaussian densities lead to full rate-adaptation over the whole collection of locally Hölder log-densities on the real line. This result has been extended to a multivariate set-up and to other families of super-smooth kernels by [24] and [23]. Here, we are concerned with full rate-adaptation when the true error density is ordinary δ-smooth, $\delta > 0$, using a DP mixture of Gaussians with a data-driven choice of $\gamma = (m, s^2)$ like $\hat{\gamma}_n = (\bar{X}_n, S_n^2)$, with $\bar{X}_n = \sum_{i=1}^n X_i/n$ and $S_n^2 = \sum_{i=1}^n (X_i - \bar{X}_n)^2/n$. Set $\mathrm{E}_0[X_1] = m_0$ and $\sigma_0^2 = \mathrm{Var}_0[X_1]$, since $\hat{\gamma}_n$ is a consistent estimator for (m_0, σ_0^2), the set $K_n = [m_0 - \log n/\sqrt{n}, m_0 + \log n/\sqrt{n}] \times [\sigma_0^2 - \log n/\sqrt{n}, \sigma_0^2 + \log n/\sqrt{n}]$ satisfies condition (1). Having identified the transformation $\psi_{\gamma,\gamma'}(p, \alpha, \beta)$ which remains essentially determined by the stick-breaking construction of the DP, the density $p^{(n)}_{\psi_{\gamma,\gamma'}(p,\alpha,\beta)}$ is, uniformly over all γ, γ', bounded above and below by suitable densities. This allows to then proceed using standard tools. In particular, letting $\varepsilon_n = n^{-\delta/(2\delta+1)}(\log n)^a$ for some $a \geq 0$, in order to lower bound the prior probability of $B_{n,k}(p_0, \alpha_0, \beta_0)$, the fact that $\mathrm{KL}((p_0, \alpha_0, \beta_0); (p, \alpha, \beta)) \lesssim \|p^{1/2} - p_0^{1/2}\|_2^2 + |\alpha - \alpha_0|^2 + |\beta - \beta_0|^2$ is used. The key idea is that, for a δ-smooth density p_0, there exists a finite mixing distribution F^*, with N_σ points included in $[-a_\sigma, a_\sigma]$, satisfying $\mathrm{KL}(p_0, p_{F^*,\sigma}) = O(\sigma^{2\delta})$ and, for $k \geq 2$, $\mathrm{V}_k(p_0, p_{F^*,\sigma}) = O(\sigma^{k\delta})$, with $a_\sigma = |\log \sigma|^{\rho_1}$ and $N_\sigma = O(\sigma^{-1}|\log \sigma|^{\rho_2})$, see, *e.g.*, Lemma 4 in [15]. Consequently, for α and β such that $\max\{|\alpha - \alpha_0|, |\beta - \beta_0|\} \leq \sigma$, we have $\mathrm{KL}((p_0, \alpha_0, \beta_0); (p_{F^*,\sigma}, \alpha, \beta)) = O(\sigma^{2\delta})$. Once a finite mixture is constructed, it is shown that there exists a set of finite Gaussian mixtures close to it and contained in a Kullback-Leibler type ball which, at worst, is charged enough prior mass. We thus have the following result, where the distance d_n defining the ball $U_{M_n\varepsilon_n}$ centered at (p_0, α_0, β_0) is $d_n^2(P^{(n)}_{p,\alpha,\beta}, P^{(n)}_{p_0,\alpha_0,\beta_0}) = n^{-1}\sum_{i=1}^n \|p(\cdot - \alpha - \beta z_i)^{1/2} - p_0(\cdot - \alpha_0 - \beta_0 z_i)^{1/2}\|_2^2$. The joint posterior distribution of (p, α, β) concentrates around (p_0, α_0, β_0) at rate ε_n relative to d_n, in symbols, $\mathrm{E}^{(n)}_{p_0,\alpha_0,\beta_0}[\Pi(U^c_{M_n\varepsilon_n}|\hat{\gamma}_n, X^{(n)})] = o(1)$.

References

1. Amewou-Atisso, M., Ghosal, S., Ghosh, J.K., Ramamoorthi, R.V.: Posterior consistency for semiparametric regression problems. Bernoulli **9**, 291–312 (2003)
2. Belitser, E., Levit, B.: On the empirical Bayes approach to adaptive filtering. Math. Methods Statist. **12**, 131–154 (2003)
3. Clyde, M.A., George, E.I.: Flexible empirical Bayes estimation for wavelets. J. R. Statist. Soc. Ser. B **62**, 681–698 (2000)
4. Cover, T.M., Gluss, D.H.: Empirical Bayes stock market portfolios. Adv. Appl. Math. **7**, 170–181 (1986)
5. Cui, W., George, E.I.: Empirical Bayes vs. fully Bayes variable selection. J. Statist. Plann. Inference **138**, 888–900 (2008)
6. Diaconis, P., Freedman, D.: On the consistency of Bayes estimates. Ann. Statist. **14**, 1–26 (1986)
7. Donnet, S., Rivoirard, V., Rousseau, J., Scricciolo, C.: Posterior concentration rates for empirical Bayes procedures with applications to Dirichlet Process mixtures. *Working Paper* (2014)
8. Efron, B., Tibshirani, R., Storey, J.D., Tusher, V.: Empirical Bayes analysis of a microarray experiment. J. Amer. Statist. Assoc. **96**, 151–160 (2001)
9. 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. Statist. Soc. Ser. B **71**, 993–1008 (2009)
10. Frost, P.A., Savarino, J.E.: An empirical Bayes approach to efficient portfolio selection. J. Financ. Quant. Anal. **21**, 293–305 (1986)
11. George, E.I., Foster, D.P.: Calibration and empirical Bayes variable selection. Biometrika **87**, 731–747 (2000)
12. Ghosal, S., van der Vaart, A.: Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities. Ann. Statist. **29**, 1233–1263 (2001)
13. Ghosal, S., van der Vaart, A.: Convergence rates of posterior distributions for noniid observations. Ann. Statist. **35**, 192–223 (2007)
14. Ghosal, S., van der Vaart, A.: Posterior convergence rates of Dirichlet mixtures at smooth densities. Ann. Statist. **35**, 697–723 (2007)
15. Kruijer, W., Rousseau, J., van der Vaart, A.: Adaptive Bayesian density estimation with location-scale mixtures. Electron. J. Stat. **4**, 1225–1257 (2010)
16. Liu, J.S.: Nonparametric hierarchical Bayes via sequential imputation. Ann. Statist. **24**, 911–930 (1996)
17. McAuliffe, J.D., Blei, D.M., Jordan, M.I.: Nonparametric empirical Bayes for the Dirichlet process mixture model. Stat. Comput. **16**, 5–14 (2006)
18. Petrone, S., Rizzelli, S., Rousseau, J., Scricciolo, C.: Empirical Bayes methods in classical and Bayesian inference. METRON *To appear* (2014)
19. Petrone, S., Rousseau, J., Scricciolo, C.: Bayes and empirical Bayes: do they merge? Biometrika *To appear* (2014)
20. Robbins, H.: An empirical Bayes approach to statistics. Proc. Third Berkeley Symp. on Math. Statist. and Prob., Vol. **1** (Univ. of Calif. Press, 1956), 157–163 (1956)
21. Rousseau, J.: Rates of convergence for the posterior distributions of mixtures of Betas and adaptive nonparametric estimation of the density. Ann. Statist. **38**, 146–180 (2010)
22. Scott, J.G., Berger, J.O.: Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. Ann. Statist. **38**, 2587–2619 (2010)
23. Scricciolo, C.: Adaptive Bayesian density estimation in L^p-metrics with Pitman-Yor or normalized inverse-Gaussian process kernel mixtures. Bayesian Analysis *To appear* (2014)
24. Shen, W., Tokdar, S., Ghosal, S.: Adaptive Bayesian multivariate density estimation with Dirichlet mixtures. Biometrika **100**, 623–640 (2013)
25. 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. Statist. **7**, 991–1018 (2013)
26. Wasserman L.: Asymptotic inference for mixture models using data-dependent priors. J. R. Statist. Soc. B. **62**, 159–180 (2000)