Electronic Journal of Statistics
Vol. 13 (2019) 3049–3081
ISSN: 1935-7524
https://doi.org/10.1214/19-EJS1600

Data-driven priors and their posterior concentration rates

Ryan Martin and Stephen G. Walker

Department of Statistics
North Carolina State University
2311 Stinson Dr.
Raleigh, NC 27695, USA
e-mail: rgmarti3@ncsu.edu

Department of Mathematics
University of Texas at Austin
2515 Speedway Stop C1200
Austin, TX 78712, USA
e-mail: s.g.walker@math.utexas.edu

Abstract: In high-dimensional problems, choosing a prior distribution such that the corresponding posterior has desirable practical and theoretical properties can be challenging. This begs the question: can the data be used to help choose a prior? In this paper, we develop a general strategy for constructing a data-driven or *empirical prior* and sufficient conditions for the corresponding posterior distribution to achieve a certain concentration rate. The idea is that the prior should put sufficient mass on parameter values for which the likelihood is large. An interesting byproduct of this data-driven centering is that the asymptotic properties of the posterior are less sensitive to the prior shape which, in turn, allows users to work with priors of computationally convenient forms while maintaining the desired rates. General results on both adaptive and non-adaptive rates based on empirical priors are presented, along with illustrations in density estimation, nonparametric regression, and high-dimensional normal models.

AMS 2000 subject classifications: Primary 62C12, 62E20; secondary 62G07, 62G08.
Keywords and phrases: Adaptation, data-dependent prior, density estimation, empirical Bayes, nonparametric regression.

Received June 2018.

1. Introduction

The Bayesian framework is ideally suited for updating prior beliefs. However, applications often do not come equipped with genuine prior beliefs, so the data analyst must make a choice. For low-dimensional problems, the posterior is relatively insensitive to the choice of prior, at least asymptotically, so default non-informative priors can be used. For modern high-dimensional problems, on the other hand, the prior matters, and the present way of thinking is to choose a prior such that the corresponding posterior distribution has certain desirable properties. For example, in sparse high-dimensional normal linear models, conjugate normal priors are attractive due to their computational simplicity. However,

it was shown in Castillo and van der Vaart (2012, Theorem 2.8) that, for priors with thin normal tails, the posterior has certain suboptimal asymptotic properties, so these are out and more sophisticated priors like the horseshoe (Carvalho, Polson and Scott, 2010) and its variants (e.g., Armagan, Dunson and Lee, 2013; Bhattacharya et al., 2015; Bhadra et al., 2017) are now in. The point is that, at least in high-dimensional problems, the interpretation of prior distributions has changed—their role is simply to facilitate efficient posterior inference and, therefore, only priors whose corresponding posterior has good properties are used. So if an *empirical* or *data-dependent* prior had some practical or theoretical benefit, then there would be no reason not to use it. This begs the two-part question: are there any benefits to the use of an empirical prior and, if so, how to construct one for which these benefits are realized?

The idea of letting the prior depend on data is not new. Classical empirical Bayes, as described in Berger (1985, Ch. 4.5), Carlin and Louis (1996), and more recently in Efron (2010), leaves certain prior hyperparameters unspecified and then uses the data to construct plug-in estimates of these parameters, usually via marginal maximum likelihood. That is, if θ is the parameter of interest, then a class $\{Q_\gamma : \gamma \in \Gamma\}$ of prior distributions for θ is considered, and rather than introducing another prior for γ, one simply gets an estimator, $\hat{\gamma}$, based on data, and uses the plug-in prior $Q_{\hat{\gamma}}$. The primary motivation for such a strategy is to let the data help carry some of the data analyst's prior specification burden. This, in turn, can provide some computational benefits, since the posterior for γ does not need to be evaluated. These computational savings are usually minimal in the high-dimensional settings we have in mind here, since γ is usually of very low dimension compared to the interest parameter θ. Posterior distribution properties for these classical empirical Bayes strategies have been investigated recently in, e.g., Szabó, van der Vaart and van Zanten (2013) and van der Pas, Szabó and van der Vaart (2017a,b) for a high-dimensional Gaussian model, and more generally in Petrone, Rousseau and Scricciolo (2014), Rousseau and Szabo (2017), and Donnet et al. (2018). These results confirm a natural conjecture that the use of the data-dependent prior $Q_{\hat{\gamma}}$ is asymptotically equivalent to the use of data-independent prior $Q_{\gamma^\star}$, where $\gamma^\star$ an appropriately defined "best" value. But they do not reveal any theoretical benefit to the use of a data-dependent prior, it only says the performance is no worse than it would be with a special data-independent prior $Q_{\gamma^\star}$. What is missing from the classical approach is a direct use of the information the data contains about θ itself; it only uses information indirectly through a marginal likelihood that is of little relevance to the actual problem.

Fortunately, there are other strategies for constructing empirical priors. Martin and Walker (2014) and Martin, Mess and Walker (2017) recently employed a new type of empirical Bayes procedure, in two structured high-dimensional Gaussian linear models; related approaches to these problems can be found in Belitser (2017), Belitser and Nurushev (2017), Belitser and Ghosal (2019), and Arias-Castro and Lounici (2014). Their main idea was to suitably center the prior for θ around a good estimator, and they were able to establish various optimal posterior concentration rate and structure learning results. An important

practical consequence of their approach is that the computationally convenient conjugate normal priors, shown to be suboptimal in the classical Bayesian setting, do actually meet the conditions for optimality in this new empirical Bayes context. The practical and theoretical benefits in these cases have been refined and extended in Martin and Shen (2017), Martin and Ning (2019), and Martin and Tang (2019); see, also, Lee, Lee and Lin (2017). However, their empirical prior construction and the asymptotic properties rely heavily on the Gaussian linear model structure, so whether there is a general framework underlying these developments remains an open question. Our main contribution here is to give an affirmative answer to this question, by presenting a general empirical prior construction and establishing general posterior concentration rate results.

To set the scene, let X^n be the data, indexed by $n \geq 1$, not necessarily independent and identically distributed (iid) or even independent, with joint distribution P_θ^n with density p_θ^n indexed by a parameter θ in Θ, possibly high- or infinite-dimensional. For a sequence of prior distributions, Π_n, on Θ, the posterior distribution, Π^n, for θ is defined, according to Bayes's formula, as

$$\Pi^n(A) = \frac{\int_A L_n(\theta)\,\Pi_n(d\theta)}{\int_{\Theta_n} L_n(\theta)\,\Pi_n(d\theta)}, \quad A \subseteq \Theta_n,$$

where $L_n(\theta) = p_\theta^n(X^n)$ is the likelihood function. A relevant property of the posterior Π^n is its concentration rate relative to the Hellinger distance on the set of joint densities $\{p_\theta^n : \theta \in \Theta\}$. Recall that the Hellinger distance between two densities, say, f and g, with dominating measure μ, is given by $H^2(f,g) = \frac{1}{2}\int (f^{1/2} - g^{1/2})^2\,d\mu$. If ε_n is a sequence with $\varepsilon_n \to 0$ no faster than $n^{-1/2}$, then we say that the posterior distribution has (Hellinger) concentration rate (at least) ε_n at $\theta^\star$ if $\mathsf{E}_{\theta^\star}^n\{\Pi^n(A_{M\varepsilon_n})\} \to 0$ as $n \to \infty$, where

$$A_{M\varepsilon_n} = \bigl\{\theta : H^2(p_{\theta^\star}^n, p_\theta^n) > 1 - e^{-M^2 n\varepsilon_n^2}\bigr\}$$

and $M > 0$ is a sufficiently large constant. Here $\mathsf{E}_{\theta^\star}^n$ denotes expectation with respect to the joint distribution $\mathsf{P}_{\theta^\star}^n$. For a deterministic or data-independent sequence of priors, Π_n, this property has been investigated in Ghosal, Ghosh and van der Vaart (2000) and Walker, Lijoi and Prünster (2007) for the iid case and by Ghosal and van der Vaart (2007a) in the non-iid case. Here we investigate this property for certain data-dependent priors.

To motivate our specific empirical prior construction, recall an essential part of the posterior concentration rate proofs for standard Bayesian posteriors. If ε_n is the desired rate, then it is typical to consider a "neighborhood" of the true $\theta^\star$ of the form

$$\bigl\{\theta : K(p_{\theta^\star}^n, p_\theta^n) \leq n\varepsilon_n^2,\, V(p_{\theta^\star}^n, p_\theta^n) \leq n\varepsilon_n^2\bigr\}, \tag{1}$$

where K is the Kullback–Leibler divergence and V is the corresponding second moment,

$$K(f,g) = \int \log(f/g)\, f\,d\mu \quad \text{and} \quad V(f,g) = \int \log^2(f/g)\, f\,d\mu.$$

A crucial step in proving that the posterior attains the ε_n rate is to demonstrate that the prior allocates a sufficient amount of mass to the set in (1). If the prior could be suitably centered at $\theta^\star$, then this prior concentration would be trivial. The difficulty, of course, is that $\theta^\star$ is unknown, so care is needed to construct a prior satisfying this prior concentration property simultaneously for a sufficiently wide range of $\theta^\star$. In fact, this placement of prior mass can be problematic and is one reason why examples like monotone density estimation are challenging; see Salomond (2014).

Our proposed alternative is motivated by considering an "empirical version" of the neighborhood in (1), namely,

$$\left\{\theta : \int \log \frac{p_{\hat{\theta}_n}(x)}{p_\theta(x)} \, \mathbb{P}_n(dx) \leq n\,\varepsilon_n^2\right\},$$

where $\hat{\theta}_n$ is a suitable estimator and $\mathbb{P}_n$ is the empirical distribution function. We do not need a term corresponding to the second moment, V, as in (1). This is equivalent to

$$\mathcal{L}_n = \{\theta : L_n(\theta) \geq e^{-n\varepsilon_n^2} L_n(\hat{\theta}_n)\}.$$

This is effectively a neighborhood of $\hat{\theta}_n$, which is known, unlike the $\theta^\star$ in (1), so it is straightforward to construct a prior to assign a sufficient amount of mass to $\mathcal{L}_n$. The consequence is that a prior satisfying this mass condition would depend on the data, since it must be suitably centered at $\hat{\theta}_n$. But aside from the data-dependent centering and some care in its spread (see Remark 3), the specific shape of the empirical prior distribution satisfying this property is not particularly important. Therefore, the conditions can be checked with relatively simple—often conjugate—priors, which greatly simplifies posterior computations. Moreover, the method in general is quite versatile, providing simple solutions with optimal concentration rates in challenging problems like monotone (Martin, 2018) and heavy-tailed density estimation (Section 4.3), and other shape-constrained problems (Martin and Shen, 2017), while giving improved rates in a classical nonparametric regression problem (Section 4.5).

The discussion above focused on cases where the target rate ε_n was known, which can be unrealistic in high-dimensional problems. For example, in a nonparametric regression problem, the optimal rate will depend on the smoothness of the true mean function. If this smoothness is known, then it is possible to tune the prior so that the attainable and targeted rates agree. However, if the smoothness is unknown, as is often the case, the prior cannot make direct use of this information, so one needs to make the prior more flexible so that it can adapt to the unknown rate. Adaptive posterior concentration rate results have received considerable attention in the recent literature, see van der Vaart and van Zanten (2009), Kruijer, Rousseau and van der Vaart (2010), Arbel, Gayraud and Rousseau (2013), Scricciolo (2015), and Shen and Ghosal (2015). The common feature in all this work is that the prior should be a mixture over an appropriate model complexity index. The empirical prior approach described above can readily handle this modification, and we provide general sufficient conditions for adaptive empirical Bayes posterior concentration.

The remainder of this paper is organized as follows. In Section 2, we introduce the notion of an empirical prior and present the conditions needed for the corresponding posterior distribution to concentrate at the true parameter value at a particular rate. This discussion is split into two parts, depending on whether the complexity is known or unknown. Section 3 presents the proofs of the two main theorems, and a take-away point is that the arguments are quite straightforward, suggesting that the particular empirical prior construction is indeed very natural. Several examples are presented in Section 4, starting from a relatively simple parametric problem and ending with a challenging adaptive nonparametric density estimation problem. We conclude, in Section 5, with a brief discussion. Details for the examples are in the Appendix.

2. Empirical priors and posterior concentration

2.1. Known complexity

For our first case, suppose the complexity of $\theta^\star$, e.g., the smoothness of the true density or regression function, is known. Then we know the target rate, ε_n, and we can make use of this information to design an appropriate sieve on which to construct an empirical prior. For this case, below we present a set of sufficient conditions that imply the posterior corresponding to our empirical prior has Hellinger concentration rate ε_n. Applications of this result will be given in Section 4.

Our prior construction here and in the next subsection relies on a sieve, Θ_n, an increasing sequence of finite-dimensional subsets of the parameter space Θ. Let $\hat{\theta}_n = \arg\max_{\theta\in\Theta_n} L_n(\theta)$ be a sieve maximum likelihood estimator (MLE). As is always the case, what distinguishes a sieve from some other subset of the parameter space is its approximation properties. Condition S1 below states specifically what will be required.

Condition S1. Given ε_n, there exists a deterministic sequence $\theta^\dagger = \theta_n^\dagger$ in Θ_n such that

$$\max\bigl\{K(p_{\theta^\star}^n, p_{\theta^\dagger}^n), V(p_{\theta^\star}^n, p_{\theta^\dagger}^n)\bigr\} \leq n\varepsilon_n^2, \quad \text{all large } n.$$

Remark 1. The sequence $\theta^\dagger = \theta_n^\dagger$ in Condition S1 can be interpreted as "pseudo-true" parameter values in the sense that $n^{-1}K(p_{\theta^\star}^n, p_{\theta^\dagger}^n) \to 0$. In the case that Θ_n eventually contains $\theta^\star$, then we can trivially take $\theta^\dagger = \theta^\star$. However, in examples like that in Section 4.5, the model does not include the true distribution, so identifying $\theta^\dagger$ is more challenging. Fortunately, appropriate sieves are already known in many of the key examples.

Remark 2. Define the likelihood ratio, $R_n(\theta) = L_n(\theta)/L_n(\theta^\star)$. An important consequence of Condition S1 is a bound on $R_n(\hat{\theta}_n)$ at the sieve MLE, which will be used in what follows. That is, there exists a constant $c > 1$ such that

$$R_n(\hat{\theta}_n) \geq e^{-cn\varepsilon_n^2} \quad \text{with } \mathsf{P}_{\theta^\star}^n\text{-probability converging to 1.} \tag{2}$$

Indeed, for $\theta^\dagger$ in Condition S1, by definition of $\hat{\theta}_n$, we trivially have $R_n(\hat{\theta}_n) \geq R_n(\theta^\dagger)$, and for the iid case it follows from Lemma 8.1 in Ghosal, Ghosh and van der Vaart (2000)—with their "Π" a point mass at $\theta^\dagger$—that $R_n(\theta^\dagger) \geq e^{-cn\varepsilon_n^2}$ with $\mathsf{P}_{\theta^\star}^n$-probability converging to 1. The general case is handled in Lemma 10 of Ghosal and van der Vaart (2007b).

The sieve Θ_n will also serve as the support of our yet-to-be-defined empirical prior Π_n. Since it is finite-dimensional, we will assume that it is equipped with a *data-independent* measure ν_n, e.g., Lebesgue measure, and Π_n will have a density π_n with respect to ν_n. The reason the measure must be data-independent is that it rules out the case of a degenerate prior supported at $\hat{\theta}_n$, a situation we are not interested in investigating.

The next two conditions—LP1 and GP1—concern the prior supported on Θ_n. The first, a local prior condition, formally describes how the empirical prior Π_n should concentrate on that empirical version of the Kullback–Leibler neighborhood (1) eluded to in Section 1, namely,

$$\mathcal{L}_n = \big\{\theta \in \Theta_n : L_n(\theta) \geq e^{-dn\varepsilon_n^2} L_n(\hat{\theta}_n)\big\}, \quad \text{some } d > 0. \tag{3}$$

On one hand, requiring that a sufficient amount of mass be assigned to $\mathcal{L}_n$ is similar to the standard local prior support conditions in Ghosal, Ghosh and van der Vaart (2000), Shen and Wasserman (2001), and Walker, Lijoi and Prünster (2007), inspired by the developments in Barron (1988). On the other hand, the neighborhood's dependence on the data is our chief novelty and the main driver of our empirical prior construction.

Condition LP1. Given ε_n, there exists $C > 0$ such that the prior Π_n satisfies

$$\mathsf{P}_{\theta^\star}^n\{\Pi_n(\mathcal{L}_n) < e^{-Cn\varepsilon_n^2}\} \to 0, \quad n \to \infty, \tag{4}$$

where $\mathcal{L}_n$ is as in (3), depending implicitly on ε_n.

Remark 3. LP1 often requires the spread of Π_n to be decreasing with n. For example, in a scalar normal mean problem, to satisfy LP1 with $\varepsilon_n = n^{-1/2}$ requires, say, a normal empirical prior, centered at the sample mean, with variance $v_n = vn^{-1}$ for some $v > 0$. Of course, LP1 is a sufficient but not necessary condition, so it is possible, at least in simple cases like this, to get the desired posterior concentration rate with other priors, e.g., with constant v_n. We are conditioned to believe that a tight prior is undesirable because it might be overly informative, but this rationale is based on the prior center being fixed. In the present case, the prior gets its "non-informativeness" from the data-driven center. And from this perspective, relatively tight prior concentration is actually quite reasonable, since one cannot expect a real benefit from the prior centering without putting a substantial amount of prior mass there. And based on results presented elsewhere (see Section 5), the relatively tight empirical prior *does not* negatively affect the frequentist validity of the posterior uncertainty quantification. Finally, when n is fixed, the empirical prior spread involves constants, e.g., v in v_n above, that can be chosen by the data analyst, so there is no flexibility lost in practice.

The second prior condition is global and effectively controls the tails of the empirical prior density π_n, i.e., how heavy can the tails be and still achieve the desired rate. This is an empirical prior version of the more familiar prior tail condition (Ghosal, Ghosh and van der Vaart, 2000) or the prior summability condition (Walker, Lijoi and Prünster, 2007) in the classical Bayesian nonparametric setting.

Condition GP1. Given ε_n, there exists constants $K > 0$ and $p > 1$, such that the density function π_n of the empirical prior Π_n satisfies

$$\int_{\Theta_n} \left[\mathsf{E}_{\theta^\star}^n \{\pi_n(\theta)^p\}\right]^{1/p} \nu_n(d\theta) \lesssim e^{Kn\varepsilon_n^2},$$

where "$\lesssim$" means less than or equal to up to a multiplicative constant.

Condition GP1 points to π_n not having too heavy tails, but in a distributional sense, taking into account its dependence on the data. While it might be unfamiliar, our examples in Section 4 show that it can be verified for commonly used priors, such as a normal distribution, centered at the MLE $\hat{\theta}$, with suitable variance, and any $p > 1$.

With the empirical prior Π_n on Θ_n, having density π_n with respect to ν_n, the posterior distribution is defined as

$$\Pi^n(A) = \frac{\int_A L_n(\theta)\, \pi_n(\theta)\, \nu_n(d\theta)}{\int_{\Theta_n} L_n(\theta)\, \pi_n(\theta)\, \nu_n(d\theta)}, \quad A \subseteq \Theta_n. \tag{5}$$

Then the following theorem considered the asymptotic behavior of the random variable $\Pi^n(A_{M\varepsilon_n})$, where $A_{M\varepsilon_n}$ is the Hellinger neighborhood described in Section 1. While this Hellinger neighborhood is relatively specific, the result entails rates with respect to other metrics in the examples of Section 4. And, for example, in the usual iid case, if the posterior mass assigned to $A_{M\varepsilon_n}$ vanishes, then so does that of

$$\{\theta : H(p_{\theta^\star}, p_\theta) > M\varepsilon_n\},$$

in which case ε_n is the usual Hellinger rate.

Theorem 1. *Let ε_n be such that $\varepsilon_n \to 0$ and $n\varepsilon_n^2 \to \infty$. If, for this ε_n, Condition S1 holds and the empirical prior satisfies LP1 and GP1, then there exists a constant $M > 0$ such that $\mathsf{E}_{\theta^\star}^n \{\Pi^n(A_{M\varepsilon_n})\} \to 0$ as $n \to \infty$. If $\varepsilon_n = O(n^{-1/2})$, then the same conclusion holds but with the constant M replaced by an arbitrary sequence $M_n \to \infty$.*

Proof. See Section 3. □

2.2. Unknown complexity

As discussed above, the attainable concentration rates depend on certain complexity features of the unknown $\theta^\star$, e.g., smoothness of a regression function. If that feature is known, as in Section 2.1, then so is the desired rate, ε_n, and that

information can be used to construct a suitable sieve on which to define a prior, empirical or otherwise. When that feature is unknown, the standard practice (e.g., Ghosal and van der Vaart, 2017, Chap. 10) is to work with a prior that mixes over models of different complexity levels, and leads to a posterior that adapts to the "right" complexity for the unknown $\theta^\star$. Here we adopt that same mixture strategy, but with an empirical twist.

Start with a representation of θ as a pair (S, θ_S), where S is some model index, taking values in some finite set $\mathcal{S}_n$, and θ_S is the corresponding model parameter, taking values in $\Theta_{n,S}$. This suggests a sieve

$$\Theta_n = \bigcup_{S \in \mathcal{S}_n} \Theta_{n,S}.$$

The particular form of this decomposition can vary across applications. One that is common is to represent a log-density or regression function in terms of a basis expansion and let $\theta = (\theta_1, \theta_2, \ldots)$ denote the coefficients. Then S could correspond to a finite set of indices that are "turned on,"

$$\Theta_{n,S} = \{\theta = (\theta_1, \theta_2, \ldots) : \theta_j = 0, \, j \notin S\},$$

and $\mathcal{S}_n$ a collection of subsets of $\{1, 2, \ldots, \}$ whose cardinality is bounded by some specified T_n. This version of S is used in Sections 4.4 and 4.5 for a sparse normal means model and nonparametric regression, respectively. In mixture models, on the other hand, S would be an integer the represents the number of mixture components. An important feature of S or of $\Theta_{n,S}$ is its dimension, which we will denote by $|S|$; in our examples, each $\Theta_{n,S}$ will be finite-dimensional and $|S|$ is literally its dimension, but this could also apply to infinite-dimensional $\Theta_{n,S}$ with $|S|$ a suitable entropy of $\Theta_{n,S}$. The key point is that $|S|$ measures the complexity of model S, both intuitively and in the technical sense that a more complex S, one with larger $|S|$, will have a slower associated rate.

Here, compared to Section 2.1, we do not know the complexity of the true $\theta^\star$ or, more specifically, we do not know which $\Theta_{n,S}$, if any, contains $\theta^\star$. If there happens to exist a true model $S^\star$, so that $\theta^\star \in \Theta_{n,S^\star}$, then the rate we would hope to achieve is $\varepsilon_n = \varepsilon_{n,S^\star}$, in which case we say that the posterior concentration rate is *adaptive*. More generally, if there exists a "best" model $S^\dagger$—see Condition S2 below—then adaptation entails that the posterior concentrates at the associated *oracle rate* $\varepsilon_n = \varepsilon_{n,S^\dagger}$.

The driving assumption behind recent developments in high-dimensional inference is that the truth is not too complex, and we can incorporate such a belief into our prior for S. Towards this, start with a marginal prior w_n for S, supported on $\mathcal{S}_n$, and a conditional prior $\Pi_{n,S}$ for θ_S, given S, supported on $\Theta_{n,S}$. Since $\Theta_{n,S}$ is finite-dimensional, there is some non-data-dependent measure, $\nu_{n,S}$, such as Lebesgue measure, with respect to which $\Pi_{n,S}$ has a density, $\pi_{n,S}$. Then the prior distribution Π_n on Θ_n is a mixture

$$\Pi_n(A) = \sum_{S \in \mathcal{S}_n} w_n(S) \, \Pi_{n,S}(A \cap \Theta_{n,S}), \quad A \subseteq \Theta_n, \tag{6}$$

where $\Pi_{n,S}(B) = \int_B \pi_{n,S}(\theta)\,\nu_{n,S}(d\theta)$. In practice, we often have prior information in the form of a "low-complexity assumption," i.e., small $w_n(S)$ for complex S, but we can be non-informative about θ_S, as before, by letting data control its prior center.

As before, various conditions are needed in order to prove that the posterior concentrates at a certain rate. Again, these come in the form of a condition on the sieve and local and global conditions on the prior. In this case, the complexity is unknown and we seek a more general adaptive concentration result so, naturally, the conditions here are more complicated than in Section 2.1.

Condition S2. Given ε_n, there exists $S^\dagger = S_n^\dagger$ in $\mathcal{S}_n$, with $|S^\dagger| \leq n\varepsilon_n^2$, and an associated $\theta^\dagger = \theta_n^\dagger$ in $\Theta_{n,S^\dagger}$ such that

$$\max\bigl\{K(p_{\theta^\star}^n, p_{\theta^\dagger}^n), V(p_{\theta^\star}^n, p_{\theta^\dagger}^n)\bigr\} \leq n\varepsilon_n^2, \quad \text{all large } n.$$

Recall that the complexity of the model, as measured by $|S|$, and the quality of approximation are at odds with one another, i.e., a simple model with small $|S|$ will tend to have large Kullback–Leibler approximation error and vice versa. The smallest ε_n for which Condition S2 holds will be called the *oracle rate*. In some examples, it is known that $\theta^\star$ belongs to $\Theta_{n,S^\star}$ for some $S^\star$, in which case we can take $S^\dagger = S^\star$ and $\theta^\dagger = \theta^\star$ so that Condition S2 is trivial and the corresponding oracle rate is simply the rate $\varepsilon_{n,S^\star}$ associated with the true parameter space. In cases where $\theta^\star$ does not belong to any sieve, approximation-theoretic results are needed to check Condition S2. Examples of both types are presented in Section 4. Regardless, $S^\dagger$ acts like the "pseudo-true" model, $\theta^\dagger$ a deterministic sequence of "pseudo-true" parameters, and $\varepsilon_n = \varepsilon_{n,S^\dagger}$ is the oracle rate; see Remark 1. Moreover, like in Remark 2, Condition S2 implies a bound on the likelihood ratio, i.e.,

$$\mathsf{P}_{\theta^\star}^n\{R_n(\hat{\theta}_{n,S^\dagger}) < e^{-cn\varepsilon_n^2}\} \to 0, \tag{7}$$

where $\hat{\theta}_{n,S}$ denotes the sieve MLE over $\Theta_{n,S}$, $S \in \mathcal{S}_n$.

Next, similar to what we did in Section 2.1, let us define the sets

$$\mathcal{L}_{n,S} = \bigl\{\theta \in \Theta_{n,S} : L_n(\theta) \geq e^{-d|S|} L_n(\hat{\theta}_{n,S})\bigr\}, \quad S \in \mathcal{S}_n, \quad d > 0,$$

which are just neighborhoods of $\hat{\theta}_{n,S}$ in $\Theta_{n,S}$. Then we have the following versions of the local and global prior conditions, suitable for the adaptive case, which dictate how the prior $\Pi_{n,S}$ allocates mass to $\mathcal{L}_{n,S}$ and $\Theta_{n,S} \cap \mathcal{L}_{n,S}^c$, respectively.

Condition LP2. Given ε_n and the pseudo-true model $S^\dagger$ from Condition S2, there exist constants $A > 0$ and $C > 0$ such that, as $n \to \infty$,

$$\mathsf{P}_{\theta^\star}^n\{\Pi_{n,S^\dagger}(\mathcal{L}_{n,S^\dagger}) > e^{-Cn\varepsilon_n^2}\} \to 0, \quad \text{and} \quad w_n(S^\dagger) \gtrsim e^{-An\varepsilon_n^2}.$$

Condition GP2. Given ε_n, there exists constants $K \geq 0$ and $p > 1$ such that

$$\sum_{S \in \mathcal{S}_n} w_n(S) \int_{\Theta_{n,S}} \bigl[\mathsf{E}_{\theta^\star}^n\{\pi_{n,S}(\theta)^p\}\bigr]^{1/p} \nu_{n,S}(d\theta) \lesssim e^{Kn\varepsilon_n^2}, \quad \text{all large } n. \tag{8}$$

In certain examples, such as those in Sections 4.4–4.5, it can be shown that the integral in Condition GP2 above is bounded by $e^{\kappa|S|}$ for some constant κ. Then the condition is satisfied with $K = 0$ if the prior w_n for S is such that the marginal prior for $|S|$ has exponential tails (e.g., Arbel, Gayraud and Rousseau, 2013; Shen and Ghosal, 2015).

For adaptive concentration rates, some extra regularization is needed in addition to the prior centering. This additional regularization amounts to a second way in which the prior depends on the data, so we refer to these as *double empirical priors*, and below we will consider two types of regularization.

- *Type 1 Regularization.* For an $\alpha \in (0,1)$ to be specified, if Π_n is the empirical prior above, then we set the double empirical prior as

$$\widetilde{\Pi}_n(d\theta) \propto \frac{\Pi_n(d\theta)}{L_n(\theta)^{1-\alpha}}. \tag{9}$$

  Dividing by a portion of the likelihood penalizes those parameters that "track the data too closely" (Walker and Hjort, 2001), hence regularization. A range of acceptable α values is identified below. In fact, α can often be arbitrarily close to 1, so this is indeed a very minor adjustment.
- *Type 2 Regularization.* Even though the regularization step in the above construction is very mild, some readers might be uncomfortable with what can be viewed as even a minor adjustment to the likelihood. An alternative approach is to place the additional regularization on the prior w_n for S. That is, if w_n is as above, then for an $\alpha \in (0,1)$ to be specified, define

$$\widetilde{w}_n(S) \propto \frac{w_n(S)}{L_n(\hat{\theta}_{n,S})^{1-\alpha}}, \quad S \in \mathcal{S}_n. \tag{10}$$

  This has the effect of putting an even smaller weight on those models that fit the data "too well" in the sense that their maximum likelihood is large, hence regularization. But, as above, often any $\alpha < 1$ is allowed, so this extra regularization is quite mild. This amounts to a double empirical prior of the form

$$\widetilde{\Pi}_n(A) = \sum_{S \in \mathcal{S}_n} \widetilde{w}_n(S)\, \Pi_{n,S}(A \cap \Theta_{n,S}), \quad A \subseteq \Theta_n. \tag{11}$$

In either case, for a suitable (and implicit) α to be defined below, the posterior distribution based on the double empirical prior can be expressed as

$$\Pi^n(A) = \frac{\int_A R_n(\theta)\, \widetilde{\Pi}_n(d\theta)}{\int_{\Theta_n} R_n(\theta)\, \widetilde{\Pi}_n(d\theta)}, \quad A \subseteq \Theta_n. \tag{12}$$

Theorem 2. *Let ε_n be such that $\varepsilon_n \to 0$ and $n\varepsilon_n^2 \to \infty$, and assume that Conditions S2, LP2, and GP2 hold for this ε_n. For the constant $p > 1$ in Condition GP2, take any*

$$\alpha \in (0, 1 - p^{-1}).$$

Then there exists $M > 0$ such that Π^n in (12), whether it be based on Type I or Type II regularization, satisfies $\mathsf{E}_{\theta^\star}^n\{\Pi^n(A_{M\varepsilon_n})\} \to 0$ as $n \to \infty$.

Proof. See Section 3. □

We automatically have the Condition S2 holds for any ε_n larger than the oracle rate, and since Condition LP2 depends specifically on the pseudo-true model $S^\dagger$, it can typically be shown that it too holds for the oracle rate. So as long as Condition GP2 also holds for the oracle rate, we get the advertised adaptation property. Otherwise, the rate is the larger of the oracle rate in Conditions S2 and LP2 and that which satisfies Condition GP2. Moreover, if the integral in (8) is exponential in the dimension $|S|$, then Condition GP2 can be well-controlled with weights $w_n(S)$ that are exponentially small in $|S|$. Finally, note that Condition GP2 can often be verified for any $p > 1$; see the examples in Section 4 and the results in Martin, Mess and Walker (2017), Martin and Shen (2017), etc. In such cases, any $\alpha < 1$ is allowed in either Type I or Type II regularization.

3. Proofs

3.1. Proof of Theorem 1

Start by expressing the posterior Π^n in (5) as

$$\Pi^n(A) = \frac{N_n(A)}{D_n} = \frac{\int_A R_n(\theta)\,\Pi_n(d\theta)}{\int_{\Theta_n} R_n(\theta)\,\Pi_n(d\theta)}, \quad A \subseteq \Theta_n. \tag{13}$$

The dependence of the prior on data requires some modification of the usual arguments for establishing concentration properties of Π^n. In particular, in Lemma 1, the lower bound on the denominator D_n in (13) is obtained quite simply thanks to the data-dependent prior, formalizing the motivation for this empirical Bayes approach described in Section 1, while Lemma 2 applies Hölder's inequality to get an upper bound on the numerator $N_n(A_{M\varepsilon_n})$.

Lemma 1. $D_n \geq e^{-dn\varepsilon_n^2}\, R_n(\hat\theta_n)\,\Pi_n(\mathcal{L}_n)$.

Proof. The denominator D_n can be trivially lower-bounded as follows:

$$D_n \geq \int_{\mathcal{L}_n} R_n(\theta)\,\pi_n(\theta)\,\nu_n(d\theta) = R_n(\hat\theta_n) \int_{\mathcal{L}_n} \frac{L_n(\theta)}{L_n(\hat\theta_n)}\,\pi_n(\theta)\,\nu_n(d\theta).$$

Now use the definition of $\mathcal{L}_n$ to complete the proof. □

Lemma 2. *Assume Condition GP1 holds for ε_n with constants (K, p), and let $q > 1$ be the Hölder conjugate of p. Then*

$$\mathsf{E}_{\theta^\star}^n\left\{\frac{N_n(A_{M\varepsilon_n})}{R_n(\hat\theta_n)^{1-\frac{1}{2q}}}\right\} \lesssim e^{-Gn\varepsilon_n^2},$$

where $G = M^2 q^{-1} - K$.

Proof. Start with the following simple bound:

$$N_n(A_{M\varepsilon_n}) = \int_{A_{M\varepsilon_n}} R_n(\theta)\pi_n(\theta)\,\nu_n(d\theta)$$
$$\leq R_n(\hat{\theta}_n)^{1-\frac{1}{2q}} \int_{A_{M\varepsilon_n}} R_n(\theta)^{\frac{1}{2q}}\pi_n(\theta)\,\nu_n(d\theta).$$

Dividing both sides by $R_n(\hat{\theta}_n)^{1-\frac{1}{2q}}$, and taking expectations, moving this expectation inside the integral, and applying Hölder's inequality, gives

$$\mathsf{E}^n_{\theta^\star}\Big\{\frac{N_n(A_{M\varepsilon_n})}{R_n(\hat{\theta}_n)^{1-\frac{1}{2q}}}\Big\} \leq \int_{A_{M\varepsilon_n}} \big[\mathsf{E}^n_{\theta^\star}\{R_n(\theta)^{\frac{1}{2}}\}\big]^{\frac{1}{q}} \big[\mathsf{E}^n_{\theta^\star}\{\pi_n(\theta)^p\}\big]^{\frac{1}{p}}\,\nu_n(d\theta).$$

A standard argument (e.g., Walker and Hjort, 2001) shows that the first expectation on the right hand side above equals $1 - H^2(p^n_{\theta^\star}, p^n_\theta)$ and, therefore, is upper bounded by $e^{-M^2 n\varepsilon_n^2}$, uniformly in $\theta \in A_{M\varepsilon_n}$. Under Condition GP1, the integral of the second expectation is $\lesssim e^{Kn\varepsilon_n^2}$. Combining these two bounds proves the claim. □

Proof of Theorem 1. To start, set

$$a_n = e^{-cn\varepsilon_n^2} \quad \text{and} \quad b_n = c_0\, e^{-(C+d)n\varepsilon_n^2}\, R_n(\hat{\theta}_n),$$

where the constants (C, c, d) are as in Condition LP1, Remark 2, and Equation (3), respectively, and c_0 is another sufficiently small constant. Also, abbreviate $N_n = N_n(A_{M\varepsilon_n})$ and $R_n = R_n(\hat{\theta}_n)$. If $1(\cdot)$ denotes the indicator function, then

$$\Pi^n(A_{M\varepsilon_n}) = \frac{N_n}{D_n} 1(R_n \geq a_n \text{ and } D_n \geq b_n) + \frac{N_n}{D_n} 1(R_n < a_n \text{ or } D_n < b_n)$$
$$\leq \frac{R_n^{1-\frac{1}{2q}}}{b_n} \frac{N_n}{R_n^{1-\frac{1}{2q}}} 1(R_n \geq a_n) + 1(R_n < a_n) + 1(D_n < b_n)$$
$$\leq \frac{e^{(C+d)n\varepsilon_n^2}}{a_n^{\frac{1}{2q}}} \frac{N_n}{R_n^{1-\frac{1}{2q}}} + 1(R_n < a_n) + 1(D_n < b_n)$$
$$= e^{(C+\frac{c}{2q}+d)n\varepsilon_n^2} \frac{N_n}{R_n^{1-\frac{1}{2q}}} + 1(R_n < a_n) + 1(D_n < b_n).$$

Taking expectation and applying Lemma 2, we get

$$\mathsf{E}^n_{\theta^\star}\{\Pi^n(A_{M\varepsilon_n})\} \lesssim e^{(C+\frac{c}{2q}+d)n\varepsilon_n^2} e^{-Gn\varepsilon_n^2} + \mathsf{P}^n_{\theta^\star}(R_n < a_n) + \mathsf{P}^n_{\theta^\star}(D_n < b_n). \quad (14)$$

The second and third terms are $o(1)$ by Remark 2 and Lemma 1, respectively. If we take $G > C + \frac{c}{2q} + d$ or, equivalently, $M^2 > q(K + C + \frac{c}{2q} + d)$, then the first term is $o(1)$ as well, completing the proof of the first claim.

For the second claim, when $n\varepsilon_n^2$ is bounded, the conclusion (14) still holds, and the latter two terms are still $o(1)$. The first term in the upper bound is decreasing in G or, equivalently, in M, so the upper bound vanishes for any $M_n \to \infty$. □

3.2. Proof of Theorem 2

The proof approach here is similar to that of Theorem 1 above, with a few differences. We will start with the posterior defined by the double empirical prior with Type 1 regularization described in Section 2.2. For that version of the prior, the the posterior probability $\Pi^n(A_{M\varepsilon_n})$ is a ratio $N_n(A_{M\varepsilon_n})/D_n$, where

$$N_n(A_{M\varepsilon_n}) = \sum_{S\in\mathcal{S}_n} w_n(S) \int_{A_{M\varepsilon_n}\cap\Theta_{n,S}} R_n(\theta)^\alpha \, \pi_{n,S}(\theta) \, \nu_{n,S}(d\theta)$$

and

$$D_n = \sum_{S\in\mathcal{S}_n} w_n(S) \int_{\Theta_{n,S}} R_n(\theta)^\alpha \, \pi_{n,S}(\theta) \, \nu_{n,S}(d\theta).$$

After proving Theorem 2 for this case, we will describe the adjustments needed to get the same result with the Type 2 regularized double empirical prior. Throughout, we will assume Conditions S2, LP2, and GP2 hold with ε_n.

Lemma 3. $D_n \geq e^{-d|S^\dagger|} \, w_n(S^\dagger) \, R_n(\hat{\theta}_{n,S^\dagger})^\alpha \, \Pi_{n,S^\dagger}(\mathcal{L}_{n,S^\dagger})$.

Proof. Almost identical to the proof of Lemma 1. □

Lemma 4. *Let $K \geq 0$ and $p > 1$ be the constants in Condition GP2, let $q > 1$ be the Hölder conjugate of p, and take α in $(0, 1 - p^{-1})$. Then*

$$\mathsf{E}^n_{\theta^\star}\{N_n(A_{M\varepsilon_n})\} \lesssim e^{-Gn\varepsilon_n^2},$$

where $G = M^2k - K$ and k depends only on α and q.

Proof. Abbreviate $N_n(A_{M\varepsilon_n})$ by N_n. Taking expectation of N_n, moving expectation inside the integral, and applying Hölder's inequality, we get

$$\mathsf{E}^n_{\theta^\star}(N_n) \leq \sum_{S\in\mathcal{S}_n} w_n(S) \int_{A_{M\varepsilon_n}\cap\Theta_{n,S}} \left[\mathsf{E}^n_{\theta^\star}\{R_n(\theta)^{\alpha q}\}\right]^{\frac{1}{q}} \left[\mathsf{E}^n_{\theta^\star}\{\pi_{n,S}(\theta)^p\}\right]^{\frac{1}{p}} \nu_{n,S}(d\theta).$$

Consider the first expectation on the right-hand side, $\mathsf{E}^n_{\theta^\star}\{R_n(\theta)^{\alpha q}\}$. Let $r = \alpha q$, which is in $(0, 1)$ by the choice of α. Then the expected likelihood ratio can be written as $\int f^r g^{1-r} \, d\mu$, where f and g are joint densities corresponding to θ and $\theta^\star$, respectively. This latter integral is related to the Rényi divergence of order r which, in turn, is related to the Hellinger distance. Indeed, by Theorem 16 in van Erven and Harremoës (2014), it is easy to see that

$$\mathsf{E}^n_{\theta^\star}\{R_n(\theta)^{\alpha q}\} \leq \{1 - H^2(p^n_{\theta^\star}, p^n_\theta)\}^{k'},$$

where k' only depends on αq. Therefore, by definition of $A_{M\varepsilon_n}$, the right-hand side is upper bounded by $e^{-M^2k'n\varepsilon_n^2}$, uniformly in $\theta \in A_{M\varepsilon_n} \cap \Theta_{n,S}$ and in S, so

$$\mathsf{E}^n_{\theta^\star}(N_n) \leq e^{-(M^2k'/q)n\varepsilon_n^2} \sum_{S\in\mathcal{S}_n} w_n(S) \int_{A_{M\varepsilon_n}\cap\Theta_{n,S}} \left[\mathsf{E}^n_{\theta^\star}\{\pi_{n,S}(\theta)^p\}\right]^{\frac{1}{p}} \nu_{n,S}(d\theta).$$

Under Condition GP2, the summation on the right-hand side above is bounded by a constant times $e^{Kn\varepsilon_n^2}$ and the claim now follows with $k = k'q^{-1}$. □

Proof of Theorem 2. By Lemma 3 and Condition LP2,

$$D_n \geq e^{-d|S^\dagger|} e^{-An\varepsilon_n^2} R_n(\hat{\theta}_{n,S_n^\star})^\alpha e^{-Cn\varepsilon_n^2}.$$

And by (7) we have $R_n(\hat{\theta}_{n,S^\dagger}) \geq e^{-cn\varepsilon_n^2}$ for some $c > 1$, with $\mathsf{P}_{\theta^\star}^n$-probability converging to 1. Since $|S^\dagger| \leq n\varepsilon_n^2$, this lower bound for the denominator can be combined with the upper bound in the numerator from Lemma 4 using an argument very similar to that in the proof of Theorem 1, to get

$$\mathsf{E}_{\theta^\star}^n\{\Pi^n(A_{M\varepsilon_n})\} \leq e^{-\{M^2k-(K+A+C+c\alpha+d)\}n\varepsilon_n^2} + o(1).$$

So, for M sufficiently large, the upper bound vanishes, proving the claim. □

It turns out that the proof for the Type 2 regularized double empirical prior follows along almost the same lines. The key is that we do not need to be concerned about the normalizing constant in the definition of $\widetilde{w}_n$ in (10) because it appears in both the numerator and denominator of the posterior probability. Similarly, we can replace $L_n(\hat{\theta}_{n,S})$ in (10) by $R_n(\hat{\theta}_{n,S})$ so, for the proof, we are free to assume that

$$\widetilde{w}_n(S) = \frac{w_n(S)}{R_n(\hat{\theta}_{n,S})^{1-\alpha}}, \quad S \in \mathcal{S}_n.$$

With this, the bound on the denominator, D_n, of the posterior probability from Lemma 3 is unchanged. For the numerator, $N_n(A_{M\varepsilon_n})$, note the following trivial inequality:

$$R_n(\theta) = R_n(\theta)^{1-\alpha} R_n(\theta)^\alpha \leq R_n(\hat{\theta}_{n,S})^{1-\alpha} R_n(\theta)^\alpha.$$

Consequently,

$$\widetilde{w}_n(S) \int_{A_{M\varepsilon_n} \cap \Theta_{n,S}} R_n(\theta)\, \pi_{n,S}(\theta)\, \nu_{n,S} \leq w_n(S) \int_{A_{M\varepsilon_n} \cap \Theta_{n,S}} R_n(\theta)^\alpha\, \pi_{n,S}(\theta)\, \nu_{n,S},$$

and the right-hand side is exactly what we bounded in the proof of Lemma 4. So we can put together the bounds on the numerator and denominator exactly like we did above to obtain the ε_n posterior convergence rate for the Type 2 regularized version of the double empirical prior.

4. Examples

4.1. Fixed finite-dimensional parameter estimation

Suppose that the parameter space Θ is a fixed subset of $\mathbb{R}^d$, for a fixed $d < \infty$. Under the usual regularity conditions, the log-likelihood $\ell_n = \log L_n$ is twice

continuously differentiable, its derivative $\dot{\ell}_n$ satisfies $\dot{\ell}_n(\hat{\theta}_n) = 0$ at the (unique) global MLE $\hat{\theta}_n$, and the following expansion holds:

$$\ell_n(\theta) - \ell_n(\hat{\theta}_n) = -\tfrac{1}{2}(\theta - \hat{\theta}_n)^\top \hat{\Sigma}_n(\theta - \hat{\theta}_n) + o(n\|\theta - \hat{\theta}_n\|^2), \tag{15}$$

where $\hat{\Sigma}_n = -\ddot{\ell}_n(\hat{\theta}_n)$. Then the set $\mathcal{L}_n$ can be expressed as

$$\mathcal{L}_n = \{\theta : (\theta - \hat{\theta}_n)^\top \Sigma_n(\theta - \hat{\theta}_n) < an\varepsilon_n^2\}.$$

For rate $\varepsilon_n = n^{-1/2}$, this suggests an empirical prior of the form:

$$\Pi_n = \mathsf{N}_d\big(\hat{\theta}_n, n^{-1}\Psi^{-1}\big), \tag{16}$$

for some fixed positive definite matrix Ψ in order to ensure S1. The proposition below states that this empirical prior yields a posterior that concentrates at the parametric rate $\varepsilon_n = n^{-1/2}$. Note that we do not need any additional fine-tuning, like in Theorem 2.4 of Ghosal, Ghosh and van der Vaart (2000), to get optimal rates in the finite-dimensional case.

Proposition 1. *Assume that each component θ_j in the d-dimensional parameter θ are on $(-\infty, \infty)$, and that the regularity conditions necessary to establish the quadratic approximation (15) hold. Then Conditions LP1 and GP1 hold for the empirical prior (16) with $\varepsilon_n = n^{-1/2}$. Therefore, the posterior, with $\alpha = 1$, concentrates at the rate $\varepsilon_n = n^{-1/2}$ relative to any metric on Θ.*

Proof. See the Appendix. □

4.2. Density estimation via histograms

Consider estimation of a density function, p, supported on the compact interval $[0, 1]$, based on iid samples $X_1, \ldots, X_n$. A simple approach to develop a Bayesian model for this problem is a random histogram prior (e.g., Scricciolo, 2007, 2015). That is, we consider a partition of the interval $[0, 1]$ into S bins of equal length, i.e., $[0, 1] = \bigcup_{s=1}^{S} E_s$, where $E_s = [\frac{s-1}{S}, \frac{s}{S})$, $s = 1, \ldots, S$. For a given S, write the model

$$p_\theta(x) = \sum_{s=1}^{S} \theta_s \, \mathsf{Unif}(x \mid E_s), \quad x \in [0, 1],$$

consisting of mixtures of uniforms, i.e., piecewise constant densities, where the parameter θ is a vector in the S-dimensional probability simplex,

$$\Delta(S) = \{(\theta_1, \ldots, \theta_S) : \theta_s \geq 0, \textstyle\sum_{s=1}^{S} \theta_s = 1\}.$$

That is, p_θ is effectively a histogram with S bins, all of the same width, S^{-1}, and the height of the s^{th} bar is $S^{-1}\theta_s$, $s = 1, \ldots, S$. Here, assuming the regularity of the true density is known, we construct an empirical prior for the vector parameter θ such that, under conditions on the true density, the corresponding posterior on the space of densities has Hellinger concentration rate within a

logarithmic factor of the minimax rate. More sophisticated models for density estimation will be presented in Sections 4.3 and 4.6.

Let $S = S_n$ be the number of bins, specified below. This defines a sieve $\Theta_n = \Delta(S_n)$ and, under the proposed histogram model, the data can be treated as multinomial, so the (sieve) MLE is $\hat{\theta}_n = (\hat{\theta}_{n,1}, \ldots, \hat{\theta}_{n,S})$, where $\hat{\theta}_{n,s}$ is just the proportion of observations in the s^{th} bin, $s = 1, \ldots, S$. Here we propose a Dirichlet prior Π_n for θ, namely,

$$\theta \sim \Pi_n = \mathsf{Dir}_S(\hat{\alpha}), \quad \hat{\alpha}_s = 1 + c\,\hat{\theta}_{n,s}, \quad s = 1, \ldots, S,$$

which is centered on the sieve MLE in the sense that the mode of the empirical prior density is $\hat{\theta}_n$; the factor $c = c_n$ will be specified below. Finally, this empirical prior for θ determines an empirical prior for the density via the mapping $\theta \mapsto p_\theta$.

Proposition 2. *Suppose that the true density, $p^\star$, is uniformly bounded away from 0 and is Hölder continuous with smoothness parameter β, where $\beta \in (0, 1]$ is assumed to be known. Set $\varepsilon_n = n^{-\kappa} \log^{\kappa} n$, where $\kappa = \beta/(2\beta + 1)$. For the empirical prior Π_n described above, if $S = S_n = n\varepsilon_n^2 (\log n)^{-1}$ and $c = c_n = n\varepsilon_n^{-2}$, then there exists $M > 0$ such that the corresponding posterior Π^n, with $\alpha = 1$, satisfies*

$$\mathsf{E}_{p^\star}^n \left[\Pi^n(\{\theta : H(p^\star, p_\theta) > M\varepsilon_n\}) \right] \to 0.$$

Proof. See the Appendix. □

4.3. Mixture density estimation

Let $X_1, \ldots, X_n$ be iid samples from a density p_θ of the form

$$p_\theta(x) = \int k(x \mid \mu)\, \theta(d\mu), \tag{17}$$

where $k(x \mid \mu)$ is a known kernel and the mixing distribution θ is unknown. Here we focus on the normal mixture case, where $k(x \mid \mu) = \mathsf{N}(x \mid \mu, \sigma^2)$, where σ is known, but see Remark 4. The full parameter space Θ, which contains the true mixing distribution $\theta^\star$, is the set of all probability measures on the μ-space, but we consider here a finite mixture model of the form

$$\theta = (\omega, \mu) \mapsto p_\theta(\cdot) = \sum_{s=1}^{S} \omega_s\, k(\cdot \mid \mu_s), \tag{18}$$

for an integer S, a vector $\omega = (\omega_1, \ldots, \omega_S)$ in the simplex $\Delta(S)$, and a set of distinct support points $\mu = (\mu_1, \ldots, \mu_S)$. For fixed S, let $\hat{\theta} = (\hat{\omega}, \hat{\mu})$ be the MLE for the mixture weights and locations, respectively, where the optimization is restricted so that $|\hat{\mu}_s| \leq B$, where $B = B_n$ is to be determined. We propose to "center" an empirical prior on the S-specific MLE as follows:

- ω and μ are independent;

- the vector ω is $\mathsf{Dir}_S(\hat{\alpha})$ like in Section 4.2, where $\hat{\alpha}_s = 1 + c\,\hat{\omega}_s$, $s = 1, \ldots, S$;
- the components $(\mu_1, \ldots, \mu_S)$ of μ are independent, with

$$\mu_s \sim \mathsf{Unif}(\hat{\mu}_s - \delta_n, \hat{\mu}_s + \delta_n), \quad s = 1, \ldots, S,$$

  where δ_n is a sequence of positive constants to be determined.

To summarize, we have an empirical prior Π_n for $\theta = (\omega, \mu)$, supported on the sieve $\Theta_n = \Delta(S) \times \mathbb{R}^S$, where $S = S_n$ will be specified, with density function

$$\pi_n(\theta) = \mathsf{Dir}_S(\omega \mid \hat{\alpha}) \times \prod_{s=1}^{S} \mathsf{Unif}(\mu_s \mid \hat{\mu}_s - \delta_n, \hat{\mu}_s + \delta_n).$$

This determines an empirical prior for the density through the mapping (18).

Proposition 3. *Suppose that the true mixing distribution $\theta^\star$ in (17) has compact support. Set $\varepsilon_n = (\log n)^{1/2} n^{-1/2}$. If $S_n \propto n\varepsilon_n^2 (\log n)^{-1} = \log n$, $B_n \propto \log^{1/2}(\varepsilon_n^{-1})$, $c_n = n\varepsilon_n^{-2} = n^2/(\log n)^2$, and $\delta_n \propto \varepsilon_n$, then there exists $M > 0$ such that the posterior Π^n, with $\alpha = 1$, corresponding to the empirical prior described above satisfies*

$$\mathsf{E}^n_{\theta^\star}[\Pi^n(\{\theta \in \Theta_n : H(p_{\theta^\star}, p_\theta) > M\varepsilon_n\})] \to 0.$$

Proof. See the Appendix. □

Remark 4. The proof of Proposition 3 is not especially sensitive to the choice of kernel. More specifically, the local prior support condition, LP1, can be verified for kernels other than normal, the key condition being Equation (24) in the Appendix. For example, that condition can be verified for the Cauchy kernel

$$k(x \mid \mu) = \frac{1}{\sigma\pi}\Big\{1 + \frac{(x-\mu)^2}{\sigma^2}\Big\}^{-1},$$

where σ is a fixed scale parameter. Therefore, using the same empirical prior formulation as for the normal case, the same argument in the proof of Proposition 3 shows that the Cauchy mixture posterior achieves the rate $\varepsilon_n = (\log n) n^{-1/2}$ when the true density $p^\star = p_{\theta^\star}$ is a finite Cauchy mixture. That the rate achieved is nearly parametric is not surprising—the finite Cauchy mixture is effectively finite-dimensional—but, to our knowledge, the Bayesian literature does not say anything about rates for heavy-tailed density estimation while it fits quite easily into our setup. Of course, the challenge is in going from a finite to infinite Cauchy mixture and, if suitable bounds on the error in approximating the latter by the former were available, then our analysis would immediately give a rate for the more general case.

4.4. Estimation of a sparse normal mean vector

Consider inference on the mean vector $\theta = (\theta_1, \ldots, \theta_n)^\top$ of a normal distribution, $\mathsf{N}_n(\theta, I_n)$, based on a single sample $X = (X_1, \ldots, X_n)$. That is,

$X_i \sim \mathsf{N}(\theta_i, 1)$, for $i = 1, \ldots, n$, independent. The mean vector is assumed to be sparse in the sense that most of the components, θ_i, are zero, but the locations and values of the non-zero components are unknown. This problem was considered by Martin and Walker (2014) and they show that a version of the double empirical Bayes posterior contracts at the optimal minimax rate. Here we propose an arguably simpler empirical prior and demonstrate the same asymptotic optimality of the posterior based on the general results in Section 2.2.

Write the mean vector θ as a pair (S, θ_S), where $S \subseteq \{1, 2, \ldots, n\}$ identifies the non-zero entries of θ, and θ_S is the $|S|$-vector of non-zero values. Assume that the true mean vector $\theta^\star$ has $|S_n^\star| = s_n^\star$ such that $s_n^\star = o(n)$. The sieves $\Theta_{n,S}$ are subsets of $\mathbb{R}^n$ that constrain the components of the vectors corresponding to indices in S^c to be zero; no constraint on the non-zero components is imposed. Note that we can trivially restrict to subsets S of cardinality no more than $T_n = n$. Furthermore, Condition S2 is trivially satisfied because $\theta^\star$ belongs to the sieve $S_n^\star$ by definition, so we can take $\theta^\dagger = \theta^\star$.

For this model, the Hellinger distance for joint densities satisfies

$$H^2(p_{\theta^\star}^n, p_\theta^n) = 1 - e^{-\frac{1}{8}\|\theta - \theta^\star\|^2},$$

where $\|\cdot\|$ is the usual ℓ_2-norm on $\mathbb{R}^n$. In this sparse setting, as demonstrated by Donoho et al. (1992), the ℓ_2-minimax rate of convergence is $s_n^\star \log(n/s_n^\star)$; we set this rate equal to $n\varepsilon_n^2$, so that $\varepsilon_n^2 = (s_n^\star/n)\log(n/s_n^\star)$. Therefore, if we can construct a prior such that Conditions LP2 and GP2 hold for this ε_n, then it will follow from Theorem 2 that the corresponding empirical Bayes posterior concentrates at the optimal minimax rate.

Let the prior distribution w_n for S be given by

$$w_n(S) \propto \binom{n}{|S|}^{-1} e^{-g(|S|)|S|}, \quad S \subseteq \{1, 2, \ldots, n\},$$

where $g(s)$ is a non-decreasing slowly varying function as $s \to \infty$, which includes the case where $g(s) \equiv B$ for a sufficiently large constant B; see the proof of the proposition. For the conditional prior for θ_S, given S, we let

$$\theta_S \mid S \sim \mathsf{N}_{|S|}(\hat{\theta}_{n,S}, \gamma^{-1} I_{|S|}), \quad \text{for any } \gamma \in (0,1),$$

where the sieve MLE is $\hat{\theta}_{n,S} = X_S = (X_i : i \in S)$.

Proposition 4. *Suppose the normal mean vector $\theta^\star$ is $s_n^\star$-sparse in the sense that only $s_n^\star = o(n)$ of the entries in $\theta^\star$ are non-zero. For the empirical prior described above, there exists a constant $M > 0$ such that the corresponding posterior distribution Π^n, using Type I or Type II regularization, with any $\alpha < 1$, satisfies*

$$\mathsf{E}_{\theta^\star}^n \left[\Pi^n(\{\theta : \|\theta - \theta^\star\|^2 > M s_n^\star \log(n/s_n^\star)\}) \right] \to 0.$$

Proof. See the Appendix. □

Note that the prior being employed in this empirical Bayes formulation is conjugate, leading to some computational savings compared to the non-conjugate priors shown to be optimal in Castillo and van der Vaart (2012) under a classical Bayesian formulation; see Martin, Mess and Walker (2017) and Martin (2017) for more on computational benefits, and Martin and Ning (2019) for results on coverage of credible sets based on this empirical Bayes model. A similar approach to the one described above is considered in Martin and Shen (2017) to get minimax optimal posterior concentration rates and fast computation for the case where θ is known to be piecewise constant.

4.5. Regression function estimation

Consider a nonparametric regression model

$$Y_i = f(t_i) + \sigma z_i, \quad i = 1, \ldots, n,$$

where $z_1, \ldots, z_n$ are iid $\mathsf{N}(0,1)$, $t_1, \ldots, t_n$ are equi-spaced design points in $[0,1]$, i.e., $t_i = i/n$, and f is an unknown function. Following Arbel, Gayraud and Rousseau (2013), we consider a Fourier basis expansion for $f = f_\theta$, so that $f(t) = \sum_{j=1}^\infty \theta_j \phi_j(t)$, where $\theta = (\theta_1, \theta_2, \ldots)$ and $(\phi_1, \phi_2, \ldots)$ are the basis coefficients and functions, respectively. They give conditions such that their Bayesian posterior distribution for f, induced by a prior on the basis coefficients θ, concentrates at the true $f^\star$ at the minimax rate corresponding to the unknown smoothness of $f^\star$. Here we derive a similar result, with a better rate, for the posterior derived from an empirical prior.

Following the calculations in Section 4.4, the Hellinger distance between the joint distribution of $(Y_1, \ldots, Y_n)$ for two different regression functions, f and g, satisfies

$$H^2(p_f^n, p_g^n) = 1 - e^{-\frac{n}{8\sigma^2}\|f-g\|_n^2},$$

where $\|f\|_n^2 = n^{-1}\sum_{i=1}^n f(t_i)^2$ is the squared L_2-norm corresponding to the empirical distribution of the covariate t. So, if the conditions of Theorem 2 are satisfied, then we get a posterior concentration rate relative to the metric $\|\cdot\|_n$.

Suppose that the true regression function $f^\star$ is in a Sobolev space of index $\beta > \frac{1}{2}$. That is, there is an infinite coefficient vector $\theta^\star$ such that $f^\star = f_{\theta^\star}$ and $\sum_{j=1}^\infty \theta_j^{\star 2} j^{2\beta} \lesssim 1$. This implies that the coefficients $\theta_j^\star$ for large j are of relatively small magnitude and suggests a particular formulation of the model and empirical prior. As before, we rewrite the infinite vector θ as (S, θ_S), but this time S is just an integer in $\{1, 2, \ldots, n\}$, and $\theta_S = (\theta_1, \ldots, \theta_S, 0, 0, \ldots)$ is an infinite vector with only the first S terms non-zero. That is, we will restrict our prior to be supported on vectors whose tails vanish in this sense. For the prior w_n for the integer S, we take

$$w_n(s) \propto e^{-g(s)s}, \quad s = 1, \ldots, n,$$

where $g(s)$, is a non-decreasing slowly varying function, which includes the case of $g(s) \equiv B$ for B sufficiently large; see the proof of the proposition. Next, for the

conditional prior for θ_S, given S, note first that the sieve MLE is a least-squares estimator

$$\hat{\theta}_{n,S} = (\Phi_S^\top \Phi_S)^{-1} \Phi_S^\top Y,$$

where Φ_S is the $n \times |S|$ matrix determined by the basis functions at the observed covariates, i.e., $\Phi_S = (\phi_j(t_i))_{ij}$, $i = 1, \ldots, n$ and $j = 1, \ldots, |S|$. As in Martin, Mess and Walker (2017), this suggests a conditional prior of the form

$$\theta_S \mid S \sim \mathsf{N}_{|S|}\big(\hat{\theta}_{n,S}, \gamma^{-1}(\Phi_S^\top \Phi_S)^{-1}\big), \quad \text{for any } \gamma \in (0,1).$$

This empirical prior for $\theta \equiv (S, \theta_S)$ induces a corresponding empirical prior for f through the mapping $\theta \mapsto f_\theta$.

Proposition 5. *Suppose that the true regression function $f^\star$ is in a Sobolev space of index $\beta > \frac{1}{2}$. For the empirical prior described above, there exists a constant $M > 0$ such that the corresponding posterior distribution Π^n, using Type I or Type II regularization, with any $\alpha < 1$, satisfies*

$$\mathsf{E}_{f^\star}^n \big[\Pi^n(\{\theta : \|f_\theta - f^\star\|_n > M n^{-\beta/(2\beta+1)}\})\big] \to 0.$$

Proof. See the Appendix. □

Note that the rate obtained in Proposition 5 is *exactly* the optimal minimax rate, i.e., there are no extra logarithmic factors. This, like in Section 4.4, is a consequence of $f^\star$ eventually being in the specified sieve; these extra log factors are a result of having to approximate the true parameter by an element in the sieve. A similar result, without the additional logarithmic terms, is given in Gao and Zhou (2016).

4.6. Nonparametric density estimation

Consider the problem of estimating a density p supported on the real line. Like in Section 4.3, we propose a normal mixture model and demonstrate the asymptotic concentration properties of the posterior based on an empirical prior, but with the added feature that the rate is adaptive to the unknown smoothness of the true density function. Specifically, as in Kruijer, Rousseau and van der Vaart (2010), we assume that data $X_1, \ldots, X_n$ are iid from a true density $p^\star$, where $p^\star$ satisfies the conditions C1–C4 in their paper; in particular, we assume that $\log p^\star$ is Hölder with smoothness parameter β. They propose a fully Bayesian model—one that does not depend on the unknown β—and demonstrate that the posterior concentration rate, relative to the Hellinger distance, is $\varepsilon_n = (\log n)^t n^{-\beta/(2\beta+1)}$ for suitable constant $t > 0$, which is within a logarithmic factor of the optimal rate.

Here we extend the approach presented in Section 4.3 to achieve adaptation by incorporating a prior for the number of mixture components, S, as well as the S-specific kernel variance σ_S^2 as opposed to fixing their values. For the prior w_n for S, we let

$$w_n(S) \propto e^{-D(\log S)^r S}, \quad S = 1, \ldots, n,$$

where $r > 1$ and $D > 0$ are specified constants. Given S, we consider a mixture model with S components of the form

$$p_{S,\theta_S}(\cdot) = \sum_{s=1}^{S} \omega_{s,S}\, \mathsf{N}(\cdot \mid \mu_{s,S}, \lambda_S^{-1}),$$

where $\theta_S = (\omega_S, \mu_S, \lambda_S)$, $\omega_S = (\omega_{1,S}, \ldots, \omega_{S,S})$ is a probability vector in $\Delta(S)$, $\mu_S = (\mu_{1,S}, \ldots, \mu_{S,S})$ is a S-vector of mixture locations, and λ_S is a precision (inverse variance) that is the same in all the kernels for a given S. We can fit this model to data using, say, the EM algorithm, and produce a given S sieve MLE: $\hat{\omega}_S = (\hat{\omega}_{1,S}, \ldots, \hat{\omega}_{S,S})$, $\hat{\mu}_S = (\hat{\mu}_1, \ldots, \hat{\mu}_S)$, and $\hat{\lambda}_S$. Following our approach in Section 4.3, consider an empirical prior for ω_S obtained by taking

$$\omega_S \mid S \sim \mathsf{Dir}_S(\hat{\alpha}_S)$$

where $\hat{\alpha}_{s,S} = 1 + c\hat{\omega}_{s,S}$ and $c = c_S$ is to be determined. The prior for μ_S follows the same approach as in Section 4.3, i.e.,

$$\mu_{S,s} \sim \mathsf{Unif}(\hat{\mu}_{S,s} - \delta, \hat{\mu}_{S,s} + \delta), \quad s = 1, \ldots, S, \quad \text{independent},$$

where $\delta = \delta_S$ is to be determined. The prior for λ_S is also uniform,

$$\lambda_S \sim \mathsf{Unif}(\hat{\lambda}_S(1 - \psi), \hat{\lambda}_S(1 + \psi)),$$

where $\psi = \psi_S$ is to be determined. Also, as with $\hat{\mu}_S$ being restricted to the interval $(-B, +B)$, we restrict the $\hat{\lambda}_S$ to lie in (B_l, B_u), to be determined. Then we get a prior on the density function through the mapping $(S, \theta_S) \mapsto p_{S,\theta_S}$. For this choice of empirical prior, the following proposition shows that the corresponding posterior distribution concentrates around a suitable true density $p^\star$ at the optimal rate, up to a logarithmic factor, exactly as in Kruijer, Rousseau and van der Vaart (2010).

Proposition 6. *Suppose that the true density $p^\star$ satisfies Conditions C1–C4 in Kruijer, Rousseau and van der Vaart (2010), in particular, $\log p^\star$ is Hölder continuous with smoothness parameter β. For the empirical prior described above, if $B = (\log n)^2$, $B_l = n^{-1}$, $B_u = n^{b-2}$, and, for each S, $c = c_s = n^2 S^{-1}$, $\delta = \delta_S = S^{1/2} n^{-(b+3/2)}$, and $\psi = \psi_S = S n^{-1}$, for a sufficiently large $b > 2$, then there exists constants $M > 0$ and $t > 0$ such that the corresponding posterior distribution Π^n, using Type I or Type II regularization, with any $\alpha < 1$, satisfies*

$$\mathsf{E}_{p^\star}^n \left[\Pi^n(\{\theta : H(p^\star, p_\theta) > M(\log n)^t n^{-\beta/(2\beta+1)}\}) \right] \to 0.$$

Proof. See the Appendix. □

5. Conclusion

This paper considers the construction of an empirical or data-dependent prior such that, when combined with the likelihood via Bayes's formula, gives a posterior distribution with desirable asymptotic concentration properties. The details

vary a bit depending on whether the complexity of the true $\theta^\star$ is known to the user or not (Sections 2.1–2.2), but the basic idea is to first choose a suitable sieve and then center the prior for the sieve parameters on the sieve MLE. This makes establishing the necessary local prior support condition and lower-bounding the posterior denominator straightforward, which is a major obstacle in the standard Bayesian nonparametric setting. Having the data involved in the prior complicates the usual argument to upper-bound the posterior numerator, but compared to the usual global prior conditions involving entropy, here we only need to suitably control the spread of the empirical prior. The end result is a data-dependent measure that achieves a certain—often optimal—concentration rate, adaptively, if necessary.

The approach presented here is quite versatile, so there are many potential applications beyond those examples studied here. A more general question to be considered in a follow-up work, one that has attracted a lot of attention in the Bayesian nonparametric community recently, concerns the coverage probability of credible regions derived from our empirical Bayes posterior distribution. Having suitable concentration rates is an important first step, but coverage properties will require new insights. The theoretical results presented in Martin and Ning (2019) for the sparse normal means problem and in Martin and Tang (2019) for regression, along the numerical results in Martin (2018) for the monotone density estimation, are promising, but more work is needed.

Acknowledgments

The authors are grateful to the associate editor and anonymous referees for their detailed comments and suggestions. This work is partially supported by the U. S. National Science Foundation, grants DMS–1506879 and DMS–1737933.

Appendix A: Details for the examples

A.1. Proof of Proposition 1

For Condition LP1, under the proposed normal prior, we have

$$\Pi_n(\mathcal{L}_n) = \int_{n\,(\theta-\hat{\theta}_n)^\top \Psi(\theta-\hat{\theta}_n)<a} \mathsf{N}\big(\theta \mid \hat{\theta}_n, n^{-1}\Psi^{-1}\big)\,d\theta.$$

Making a change of variable, $z = n^{1/2}\Psi^{1/2}(\theta - \hat{\theta}_n)$, the integral above can be rewritten as

$$\Pi_n(\mathcal{L}_n) = \int_{\|z\|^2<a} \frac{1}{(2\pi)^{d/2}} e^{-\frac{1}{2}\|z\|^2}\,dz,$$

and, therefore, $\Pi_n(\mathcal{L}_n)$ is lower-bounded by a constant not depending on n so $\Pi_n(\mathcal{L}_n)$ is bounded away from zero; hence Condition LP1 holds with $\varepsilon_n = n^{-1/2}$. For Condition GP1, write the prior as $\theta \sim \mathsf{N}_d(\hat{\theta}_n, n^{-1}\Psi^{-1})$ and the

asymptotic distribution of the MLE as $\hat{\theta} \sim \mathsf{N}_d(\theta^\star, n^{-1}\Sigma^{\star -1})$, where $\Sigma^\star$ is the Fisher information matrix evaluated at $\theta^\star$. Then we have,

$$\pi_n(\theta)^p \propto |pn\Psi|^{-1/2}|n\Psi|^{p/2}\mathsf{N}_d(\theta \mid \hat{\theta}_n, (pn\Psi)^{-1}).$$

Thus

$$\mathsf{E}_{\theta^\star}\{\pi_n(\theta)^p\} \propto |pn\Psi|^{-1/2}|n\Psi|^{p/2}\,\mathsf{N}_d\big(\theta \mid \theta^\star, (pn\Psi)^{-1} + n^{-1}\Sigma^{\star -1}\big)$$

and so

$$\int \big[\mathsf{E}_{\theta^\star}\{\pi_n(\theta)^p\}\big]^{\frac{1}{p}}\, d\theta \propto |I_d + p\Psi\Sigma^{\star -1}|^{\frac{1}{2}-\frac{1}{2p}}.$$

As long as Ψ is non-singular, the right-hand side above is not dependent on n and is finite, which implies we can take $\varepsilon_n = n^{-1/2}$. It follows from Theorem 1 that the Hellinger rate is $\varepsilon_n = n^{-1/2}$ and, since all metrics on the finite-dimensional Θ are equivalent, the same rate obtains for any other metric.

We should highlight the result that the integral involved in checking Condition GP1 is at most exponential in the dimension of the parameter space:

$$\int \big[\mathsf{E}_{\theta^\star}\{\pi_n(\theta)^p\}\big]^{\frac{1}{p}}\, d\theta \leq e^{\kappa d}, \quad \kappa > 0. \tag{19}$$

This result will be useful in the proof of some of the other propositions.

A.2. Proof of Proposition 2

We start by verifying Condition LP1. Note that, for those models in the support of the prior, the data are multinomial, so the likelihood function is

$$L_n(\theta) = \theta_1^{n_1} \cdots \theta_S^{n_S},$$

where $(n_1, \ldots, n_S)$ are the bin counts, i.e., $n_s = |\{i : X_i \in E_s\}|$, $s = 1, \ldots, S$. Taking expectation with respect to $\theta \sim \mathsf{Dir}_S(\hat{\alpha})$ gives

$$\begin{aligned}
\mathsf{E}(\theta_1^{n_1} \cdots \theta_S^{n_S}) &= \frac{\Gamma(c+S)}{\Gamma(c+S+n)} \prod_{s=1}^{S} \frac{\Gamma(n_s + 1 + c\hat{\theta}_s)}{\Gamma(1 + c\hat{\theta}_s)} \\
&= \frac{\Gamma(c+S)}{\Gamma(c+S+n)} \prod_{s=1}^{S} \prod_{k=1}^{n_s} (k + c\hat{\theta}_s) \\
&\geq \frac{\Gamma(c+S)}{\Gamma(c+S+n)} \prod_{s=1}^{S} (1 + c\hat{\theta}_s)^{n_s} \\
&\geq \frac{\Gamma(c+S)\, c^n}{\Gamma(c+S+n)} \prod_{s=1}^{S} \hat{\theta}_s^{n_s}.
\end{aligned}$$

Therefore,

$$\mathsf{E}\{L_n(\theta)\} \geq \frac{\Gamma(c+S)\, c^n}{\Gamma(c+S+n)} L_n(\hat{\theta}). \tag{20}$$

Next, a simple "reverse Markov inequality" says, for any random variable $Y \in (0,1)$,

$$\mathsf{P}(Y > a) \geq \frac{\mathsf{E}(Y) - a}{1-a}, \quad a \in (0,1). \tag{21}$$

Recall that $\mathcal{L}_n = \{\theta \in \Theta_n : L_n(\theta) > e^{-dn\varepsilon_n^2} L_n(\hat{\theta})\}$ as in (3), so we can apply (21) to get

$$\Pi_n(\mathcal{L}_n) \geq \frac{\mathsf{E}\{L_n(\theta)\}/L_n(\hat{\theta}) - e^{-dn\varepsilon_n^2}}{1 - e^{-dn\varepsilon_n^2}}.$$

Then it follows from (20) that

$$\Pi_n(\mathcal{L}_n) \geq \frac{\Gamma(c+S)\, c^n}{\Gamma(c+S+n)} - e^{-dn\varepsilon_n^2}$$

and, therefore, Condition LP1 is satisfied, with $C > d$, if

$$\frac{\Gamma(c+S+n)}{\Gamma(c+S)c^n} \leq e^{dn\varepsilon_n^2}. \tag{22}$$

Towards this, we have

$$\frac{\Gamma(c+S+n)}{\Gamma(c+S)\, c^n} = \prod_{j=1}^{n} \Big(1 + \frac{S+j}{c}\Big) \leq \Big(1 + \frac{S+n+1}{c}\Big)^n.$$

So, if $c = n\varepsilon_n^{-2}$ as in the proposition statement, then the right-hand side above is upper-bounded by $e^{n\varepsilon_n^2(1+S/n)}$. Since $S \leq n$, (22) holds for, say, $d > 2$, hence, Condition LP1.

Towards Condition GP1, note that the Dirichlet component for θ satisfies

$$\mathsf{Dir}_S(\theta \mid \hat{\alpha}) \leq \mathsf{Dir}_S(\hat{\theta} \mid \hat{\alpha}) \approx (c+S)^{c+S+1/2} \prod_{s: n_s > 0} \frac{1}{(1 + c\hat{\theta}_s)^{c\hat{\theta}_s + 1/2}}\, \hat{\theta}_s^{c\hat{\theta}_s},$$

where the "$\approx$" is by Stirling's formula, valid for all $n_s > 0$ due to the value of c. This has a uniform upper bound:

$$\mathsf{Dir}_S(\theta \mid \hat{\alpha}) \leq \frac{(c+S)^{c+S+1/2}}{c^c}, \quad \forall\, \theta \in \Delta(S).$$

Then Condition GP1 holds if we can bound the product of this and $\Gamma(S)^{-1}$, the volume of $\Delta(S)$, by $e^{Kn\varepsilon_n^2}$ for a constant $K > 0$. Using Stirling's formula again, and the fact that $c/S \to \infty$, we have

$$\frac{(c+S)^{c+S+1/2}}{c^{c+S/2}\,\Gamma(S)} = \frac{S^{S+1/2}}{c^{S/2}\,\Gamma(S)} \Big(1 + \frac{S}{c}\Big)^c \Big(1 + \frac{c}{S}\Big)^{S+1/2} \leq e^{K' S \log(1+c/S)}, \quad K' > 0.$$

We need $S\log(1+c/S) \leq n\varepsilon_n^2$. Since $c/S \ll n^2$, the logarithmic term is $\lesssim \log n$. But we assumed that $S \leq n\varepsilon_n^2 (\log n)^{-1}$, so the product is $\lesssim n\varepsilon_n^2$, proving Condition GP1.

It remains to check Condition S1. A natural candidate for the pseudo-true parameter $\theta^\dagger$ in Condition S1 is one that sets θ_s equal to the probability assigned by the true density $p^\star$ to E_s. Indeed, set

$$\theta_s^\dagger = \int_{E_s} p^\star(x)\,dx, \quad s = 1, \ldots, S.$$

It is known (e.g., Scricciolo, 2015, p. 93) that, if $p^\star$ is β-Hölder, with $\beta \in (0, 1]$, then the sup-norm approximation error of $p_{\theta^\dagger}$ is

$$\|p^\star - p_{\theta^\dagger}\|_\infty \lesssim S^{-\beta}.$$

Since $p^\star$ is uniformly bounded away from 0, it follows from Lemma 8 in Ghosal and van der Vaart (2007a) that $\max\{K(p^\star, p_{\theta^\dagger}), V(p^\star, p_{\theta^\dagger})\} \lesssim H^2(p^\star, p_{\theta^\dagger})$ which, in turn, is upper-bounded by $S^{-2\beta}$ by the above display. Therefore, we need $S = S_n$ to satisfy $S^{-\beta} \leq \varepsilon_n$, and this is achieved by choosing $S = n\varepsilon_n^2(\log n)^{-1}$ as in the proposition. This establishes Condition S1, completing the proof.

A.3. Proof of Proposition 3

We start by verifying Condition LP1. Towards this, we first note that, for mixtures in the support of the prior, the likelihood function is

$$L_n(\theta) = \prod_{i=1}^n \sum_{s=1}^S \omega_s\, k(X_i \mid \mu_s), \quad \theta = (\omega, \mu),$$

which can be rewritten as

$$L_n(\theta) = \sum_{(n_1, \ldots, n_S)} \omega_1^{n_1} \cdots \omega_S^{n_S} \sum_{(s_1, \ldots, s_n)} \prod_{s=1}^S \prod_{i: s_i = s} k(X_i \mid \mu_s), \tag{23}$$

where the first sum is over all S-tuples of non-negative integers $(n_1, \ldots, n_S)$ that sum to n, the second sum is over all n-tuples of integers $1, \ldots, S$ with $(n_1, \ldots, n_S)$ as the corresponding frequency table, and $k(x \mid \mu) = \mathsf{N}(x \mid \mu, \sigma^2)$ for known σ^2. We also take the convention that, if $n_s = 0$, then the product $\prod_{i: s_i = s}$ is identically 1. Next, since the prior has ω and μ independent, we only need to bound

$$\mathsf{E}(\omega_1^{n_1} \cdots \omega_S^{n_S}) \quad \text{and} \quad \mathsf{E}\Big\{\prod_{s=1}^S \prod_{i: s_i = s} k(X_i \mid \mu_s)\Big\}$$

for a generic $(n_1, \ldots, n_S)$. The first expectation is with respect to the prior for ω and can be handled exactly like in the proof of Proposition 2. For the second expectation, which is with respect to the prior for the μ, since the prior has the components of μ independent, we have

$$\mathsf{E}\Big\{\prod_{s=1}^S \prod_{i: s_i = s} k(X_i \mid \mu_s)\Big\} = \prod_{s=1}^S \mathsf{E}\Big\{\prod_{i: s_i = s} k(X_i \mid \mu_s)\Big\},$$

so we can work with a generic s. Writing out the product of kernels, we get

$$\mathsf{E}\Big\{ \prod_{i:s_i=s} k(X_i \mid \mu_s) \Big\} = \Big(\frac{1}{2\pi\sigma^2}\Big)^{n_s/2} e^{-\frac{1}{2\sigma^2}\sum_{i:s_i=s}(X_i-\bar{X})^2} \mathsf{E}\big\{ e^{-\frac{n_s}{2\sigma^2}(\mu_s-\bar{X})^2} \big\}.$$

By Jensen's inequality, i.e., $\mathsf{E}(e^Z) \geq e^{\mathsf{E}(Z)}$, the expectation on the right-hand side is lower bounded by

$$e^{-\frac{n_s}{2\sigma^2}\mathsf{E}(\mu_s-\bar{X})^2} = e^{-\frac{n_s}{2\sigma^2}\{v_n+(\hat{\mu}_s-\bar{X})^2\}},$$

where $v_n = \delta_n^2/3$ is the variance of $\mu_s \sim \mathsf{Unif}(\hat{\mu}_s - \delta_n, \hat{\mu}_s + \delta_n)$. This implies

$$\mathsf{E}\Big\{ \prod_{s=1}^{S} \prod_{i:s_i=s} k(X_i \mid \mu_s) \Big\} \geq e^{-\frac{nv_n}{2\sigma^2}} \prod_{s=1}^{S} \prod_{i:s_i=s} k(X_i \mid \hat{\mu}_s). \tag{24}$$

Putting the two expectations back together, from (23) we have that

$$\mathsf{E}\{L_n(\theta)\} \geq \frac{\Gamma(c+S)\, c^n}{\Gamma(c+S+n)} e^{-\frac{nv_n}{2\sigma^2}} L_n(\hat{\theta}) \tag{25}$$

where now the expectation is with respect to both priors. Recall that $\mathcal{L}_n = \{\theta \in \Theta_n : L_n(\theta) > e^{-dn\varepsilon_n^2} L_n(\hat{\theta})\}$ as in (3), and define $\mathcal{L}_n' = \{\theta \in \mathcal{L}_n : L_n(\theta) \leq L_n(\hat{\theta}_n)\}$. Since, $\mathcal{L}_n \supseteq \mathcal{L}_n'$ and, for $\theta \in \mathcal{L}_n'$, we have $L_n(\theta)/L_n(\hat{\theta}_n) \leq 1$, we can apply the reverse Markov inequality (21) again to get

$$\Pi_n(\mathcal{L}_n) \geq \frac{\mathsf{E}\{L_n(\theta)\}/L_n(\hat{\theta}) - e^{-dn\varepsilon_n^2}}{1 - e^{-dn\varepsilon_n^2}}.$$

Then it follows from (25) that

$$\Pi_n(\mathcal{L}_n) \geq \frac{\Gamma(c+S)\, c^n}{\Gamma(c+S+n)} e^{-\frac{nv_n}{2\sigma^2}} - e^{-dn\varepsilon_n^2}$$

and, therefore, Condition LP1 is satisfied if

$$\frac{nv_n}{2\sigma^2} \leq bn\varepsilon_n^2 \quad \text{and} \quad \frac{\Gamma(c+S+n)}{\Gamma(c+S)c^n} \leq e^{an\varepsilon_n^2}, \tag{26}$$

where $a + b < d$. The first condition is easy to arrange; it requires that

$$v_n \leq 2b\sigma^2\varepsilon_n^2 \iff \delta_n \leq (6b\sigma^2)^{1/2}\varepsilon_n,$$

which holds by assumption on δ_n. The second condition holds with $a = 2$ by the argument in the proof of Proposition 2. Therefore, Condition LP1 holds.

Towards Condition GP1, putting together the bound on the Dirichlet density function in the proof of Proposition 2 and the following bound on the uniform densities,

$$\prod_{s=1}^{S} \mathsf{Unif}(\mu_s \mid \hat{\mu}_s - \delta_n, \hat{\mu}_s + \delta_n) \leq \Big(\frac{1}{2\delta_n}\Big)^S \prod_{s=1}^{S} I_{[-B_n-\delta_n, B_n+\delta_n]}(\mu_s),$$

we have that, for any $p > 1$,

$$\int_{\Theta_n} \left[\mathsf{E}_{\theta^\star}\{\pi_n(\theta)^p\}\right]^{1/p} d\theta \le \frac{(c+S)^{c+S+1/2}}{c^c\,\Gamma(S)} \cdot \Big(\frac{1}{2\delta_n}\Big)^S \{2(B_n+\delta_n)\}^S.$$

Then Condition GP1 holds if we can make both terms in this product to be like $e^{Kn\varepsilon_n^2}$ for a constant $K > 0$. The first term in the product, coming from the Dirichlet part, is handled just like in the proof of Proposition 2 and, for the second factor, we have

$$\Big(\frac{1}{2\delta_n}\Big)^S \{2(B_n+\delta_n)\}^S \le e^{S\log(1+\frac{B_n}{\delta_n})}.$$

Since $\delta_n \propto \varepsilon_n$ and $B_n \propto \log^{1/2}(\varepsilon_n^{-1})$, we have $B_n/\delta_n \propto n^{1/2}$, so the exponent above is $\lesssim S\log n \lesssim n\varepsilon_n^2$. This takes care of the second factor, proving Condition GP1.

Finally, we refer to Section 4 in Ghosal and van der Vaart (2001) where they show that there exists a finite mixture, characterized by $\theta^\dagger$, with S components and locations in $[-B_n, B_n]$, such that $\max\{K(p_{\theta^\star}, p_{\theta^\dagger}),\, V(p_{\theta^\star}, p_{\theta^\dagger})\} \le \varepsilon^2$. This $\theta^\dagger$ satisfies our Condition S1, so the proposition follows from Theorem 1.

In the context of Remark 4, when the normal kernel is replaced by a Cauchy kernel, we need to verify (24) in order to meet LP1. To this end, let us start with

$$\mathsf{E}\exp\Big[-\log \prod_{i:s_i=s} \{1+\sigma^{-2}(X_i-\mu_s)^2\}\Big]$$

where the expectation is with respect to the prior for the μ_s and the σ is assumed known. This log of this expectation is easily seen to be lower-bounded by

$$-\sum_{i:s_i=s} \log[1+\sigma^{-2}\mathsf{E}(X_i-\mu_s)^2] = -\sum_{i:s_i=s} \log[1+\sigma^{-2}(X_i-\hat\mu_s)^2 + \sigma^{-2}v_n].$$

Exponentiating the right-hand term, we get

$$\Big\{\prod_{s_i=s} \frac{1}{1+\sigma^{-2}(X_i-\hat\mu_s)^2}\Big\} \frac{1}{\prod_{i:s_i=s}\Big(\sigma^2 + \frac{v_n}{1+(X_i-\hat\mu_s)^2}\Big)}$$

and the second term here is lower-bounded by $\exp(-n_s\, v_n/\sigma^2)$. Therefore, Condition LP1 holds with the same ε_n as in the normal case.

Condition GP1 in this case does not depend on the form of the kernel, whether it be normal or Cauchy. And S1 is satisfied if we assume the true density $p^\star = p_{\theta^\star}$ is a finite mixture of densities, for example, the Cauchy. This proves the claim in Remark 4, namely, that the empirical Bayes posterior, based on a Cauchy kernel, concentrates at the rate $\varepsilon_n = (\log n)n^{-1/2}$ when the true density is a finite Cauchy mixture.

A.4. Proof of Proposition 4

The proportionality constant depends on n (and g) but it is bounded away from zero and infinity as $n \to \infty$ so can be ignored in our analysis. Here we can check the second part of Condition LP2. Indeed, for the true model $S_n^\star$ of size $s_n^\star$, using the inequality $\binom{n}{s} \leq (en/s)^s$, we have

$$w_n(S_n^\star) \propto \binom{n}{s_n^\star}^{-1} e^{-Bs_n^\star} \geq e^{-[B+1+\log(n/s^\star)]s_n^\star}$$

and, since $n\varepsilon_n^2 = s_n^\star \log(n/s_n^\star)$, the second condition in Condition LP2 holds for all large n with $A > 1$. Next, for Condition GP2, note that the prior w_n given above corresponds to a hierarchical prior for S that starts with a truncated geometric prior for $|S|$ and then a uniform prior for S, given $|S|$. Then it follows directly that Condition GP2 on the marginal prior for $|S|$ is satisfied.

For Condition LP2, we first write the likelihood ratio for a generic $\theta \in \Theta_S$:

$$\frac{L_n(\theta)}{L_n(\hat{\theta}_{n,S})} = e^{-\frac{1}{2}\|\theta_S - \hat{\theta}_{n,s}\|^2}.$$

Therefore, $\mathcal{L}_{n,S} = \{\theta \in \Theta_S : \frac{1}{2}\|\theta - \hat{\theta}_{n,S}\|^2 < |S|\}$. This is just a ball in $\mathbb{R}^{|S|}$ so we can bound the Gaussian measure assigned to it. Indeed,

$$\begin{aligned}
\Pi_n(\mathcal{L}_{n,S}) &= \int_{\|z\|^2 < 2|S|} (2\pi)^{-d/2} \gamma^{d/2} e^{-\frac{\gamma}{2}\|z\|^2} \, dz \\
&> (2\pi)^{-|S|/2} \gamma^{|S|/2} e^{-\gamma|S|} \frac{\pi^{|S|/2}}{\Gamma(\frac{|S|}{2}+1)} (2|S|)^{|S|/2} \\
&= \gamma^{|S|/2} e^{-\gamma|S|} \frac{1}{\Gamma(\frac{|S|}{2}+1)} |S|^{|S|/2}.
\end{aligned}$$

Stirling's formula gives an approximation of the lower bound:

$$e^{-\gamma|S|} \gamma^{|S|/2} 2^{|S|/2} e^{|S|/2} \Big(\frac{|S|/2}{2\pi}\Big)^{1/2}.$$

For moderate to large $|S|$, the above display is $\gtrsim \exp\big\{\big(1 - 2\gamma + \log\gamma + \log 2\big)\frac{|S|}{2}\big\}$ and, therefore, plugging in $S_n^\star$ for the generic S above, we see that Condition LP2 holds if $1 - 2\gamma + \log\gamma + \log 2 < 0$. For Condition GP2, the calculation is similar to that in the finite-dimensional case handled in Proposition 1. Indeed, the last part of the proof showed that, for a d-dimensional normal mean model with covariance matrix Σ^{-1} and a normal empirical prior of with mean $\hat{\theta}_n$ and covariance matrix proportional to Σ^{-1}, then the integral specified in the second part of Condition GP2 is exponential in the dimension d. In the present case, we have that

$$\int_{\Theta_S} \big[\mathsf{E}_{\theta^\star}\{\pi_{n,S}(\theta)^p\}\big]^{\frac{1}{p}} \, d\theta = e^{\kappa|S|}$$

for some $\kappa > 0$ and then, clearly, Condition GP2 holds with $K = \kappa$. If we take B in the prior w_n for S to be larger than this K, then the conditions of Theorem 2 are met with $\varepsilon_n^2 = (s_n^\star/n)\log(n/s_n^\star)$.

A.5. Proof of Proposition 5

By the choice of marginal prior for S and the normal form of the conditional prior for θ_S, given S, Conditions LP2 and GP2 follow almost exactly like in the proof of Proposition 4. Indeed, the second part of Condition GP2 holds with K the same as was derived above. Therefore, we have only to check Condition S2. Let p_θ denote the density corresponding to regression function $f = f_\theta$. If $\theta^\star$ is the coefficient vector in the basis expansion of $f^\star$, then it is easy to check that

$$K(p^n_{\theta^\star}, p^n_{\theta^\star_S}) = \frac{n}{2\sigma^2}\|\theta^\star - \theta^\star_S\|^2 = \frac{n}{2\sigma^2}\sum_{j>|S|}\theta^{\star 2}_j.$$

If $f^\star$ is smooth in the sense that it belongs to a Sobolev space indexed by $\beta > \frac{1}{2}$, i.e., the basis coefficient vector $\theta^\star$ satisfies $\sum_{j=1}^\infty \theta^{\star 2}_j j^{2\beta} \lesssim 1$, then it follows that

$$K(p^n_{\theta^\star}, p^n_{\theta^\star_S}) \lesssim n|S|^{-2\beta}.$$

So, if we take $\varepsilon_n = n^{-\beta/(2\beta+1)}$ and $|S^\star_n| = \lfloor n\varepsilon_n^2 \rfloor = \lfloor n^{1/(2\beta+1)} \rfloor$, then a candidate $\theta^\dagger$ in Condition S2 is $\theta^\dagger = \theta^\star_S$. That the desired bound on the Kullback–Leibler second moment V also holds for this $\theta^\dagger$ follows similarly, as in Arbel, Gayraud and Rousseau (2013, p. 558). This establishes Condition S2 so the conclusion of the proposition follows from Theorem 2.

A.6. Proof of Proposition 6

Write $\varepsilon_n = (\log n)^t n^{-\beta/(2\beta+1)}$ for a constant $t > 0$ to be determined. For Condition S2, we appeal to Lemma 4 in Kruijer, Rousseau and van der Vaart (2010) which states that there exists a finite normal mixture, $p^\dagger$, having $S^\star_n$ components, with

$$S^\star_n \lesssim n^{1/(2\beta+1)}(\log n)^{k-t} = n\varepsilon_n^2(\log n)^{k-3t},$$

such that $\max\{K(p^\star, p^\dagger), V(p^\star, p^\dagger)\} \leq \varepsilon_n^2$, where $k = 2/\tau_2$ and τ_2 is related to the tails of $p^\star$ in their Condition C3. So, if t is sufficiently large, then our Condition S2 holds.

For Condition GP2, we first note that, by a straightforward modification of the argument given in the proof of Proposition 3, we have

$$\int_{\Delta(S)\times\mathbb{R}^S\times\mathbb{R}_+} \bigl[\mathsf{E}_{p^\star}\{\pi_{n,S}(\theta)^p\}\bigr]^{1/p}\,d\theta \leq e^{bS\log n}\Bigl(1+\frac{B}{\delta}\Bigr)^S \frac{B_u(1+\psi) - B_l(1-\psi)}{2\psi B_l},$$

for some $b > 0$. The logarithmic term appears in the first product because, as in the proof of Proposition 3, the exponent can be bounded by a constant times $S\log(1 + c/S) \lesssim S\log n$ since $c/S = n^2/S^2 < n^2$. To get the upper bound in the above display to be exponential in S, we can take

$$\delta \gtrsim \frac{B}{n^b} \quad \text{and} \quad \psi \gtrsim \frac{B_u - B_l}{B_l}\frac{1}{e^{bS\log n} - (B_l + B_u)/(2B_l)}.$$

With these choices, it follows that the right-hand side in the previous display is upper bounded by $e^{3b \log n}$, independent of S. Therefore, trivially, the summation in (8) is also upper bounded by $e^{3b \log n}$. Since $\log n \leq n\varepsilon_n^2$, we have that Condition GP2 holds.

Condition LP2 has two parts to it. For the first part, which concerns the prior concentration on $\mathcal{L}_n$, we can follow the argument in the proof of Proposition 3. In particular, with the additional prior on λ, the version of (25) is

$$\mathsf{E} L_n(\theta_S) \geq \frac{\Gamma(c+S)\, c^n}{\Gamma(c+S+n)} e^{-\frac{1}{6} n \delta^2 \hat{\lambda}}\, e^{-nz\psi} L_n(\hat{\theta}_S)$$

for some $z \in (0,1)$. This is based on the result that if $\lambda \sim \mathsf{Unif}(\hat{\lambda}(1-\psi), \hat{\lambda}(1+\psi))$ then $\mathsf{E}\lambda = \hat{\lambda}$ and $\mathsf{E} \log \lambda > \log \hat{\lambda} - z\psi$ for some $z \in (0,1)$. With $c = n^2 S^{-1}$ as proposed, the argument in the proof of Proposition 2 shows that the first term on the right-hand side of the above display is lower-bounded by e^{-CS} for some $C > 0$. To make other other terms lower-bounded by something of the order $e^{-C'S}$, we need δ and ψ to satisfy

$$\delta^2 \lesssim \frac{1}{B_u^2} \frac{S}{n} \quad \text{and} \quad \psi \lesssim \frac{S}{n}.$$

Given these constraints and those coming from checking Condition GP2 above, we require

$$\frac{B}{n^b} \lesssim \frac{1}{B_u} \Big(\frac{S}{n}\Big)^{1/2} \quad \text{and} \quad n^{bS} - \frac{1}{2}\Big(1 + \frac{B_u}{B_l}\Big) \lesssim \frac{n\, B_u}{B_l}.$$

From Lemma 4 in Kruijer, Rousseau and van der Vaart (2010), we can deduce that the absolute value of the locations for $p^\dagger$ are smaller than a constant times $\log \varepsilon_n^{-\beta}$. Hence, we can take $B = (\log n)^2$. Also, we need $B_l \lesssim \varepsilon_n^\beta$ which is met by taking $B_l = n^{-1}$. To meet our constraints, we can take $B_u = n^{b-2}$, so we need $b \geq 2$. These conditions on $(B, B_l, B_u, \delta, \psi)$ are met by the choices stated in the proposition. For the second part of Condition LP2, which concerns the concentration of w_n around $S_n^\star$, we have

$$w_n(S_n^\star) \geq e^{-D(\log S_n^\star)^r S_n^\star} \gtrsim e^{-Dn\varepsilon_n^2 (\log n)^{k+r-3t}}.$$

So, just like in Kruijer, Rousseau and van der Vaart (2010), as long as $3t > k+r$, we get $w_n(S_n^\star) \geq e^{-Dn\varepsilon_n^2}$ as required in Condition LP2.

References

Arbel, J., Gayraud, G. and Rousseau, J. (2013). Bayesian optimal adaptive estimation using a sieve prior. *Scand. J. Stat.* **40** 549–570. MR3091697

Arias-Castro, E. and Lounici, K. (2014). Estimation and variable selection with exponential weights. *Electron. J. Stat.* **8** 328–354. MR3195119

Armagan, A., Dunson, D. B. and Lee, J. (2013). Generalized double Pareto shrinkage. *Statist. Sinica* **23** 119–143. MR3076161

Barron, A. (1988). The exponential convergence of posterior probabilities with implications for Bayes estimators of density functions, Technical Report No. 7, Department of Statistics, University of Illinois, Champaign, IL.

Belitser, E. (2017). On coverage and local radial rates of credible sets. *Ann. Statist.* **45** 1124–1151. MR3662450

Belitser, E. and Ghosal, S. (2019). Empirical Bayes oracle uncertainty quantification. *Ann. Statist.*, to appear, http://www4.stat.ncsu.edu/~ghoshal/papers/oracle_regression.pdf.

Belitser, E. and Nurushev, N. (2017). Needles and straw in a haystack: robust confidence for possibly sparse sequences. Unpublished manuscript, arXiv:1511.01803.

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*, Second ed. Springer-Verlag, New York. MR0804611

Bhadra, A., Datta, J., Polson, N. G. and Willard, B. (2017). The horseshoe+ estimator of ultra-sparse signals. *Bayesian Anal.* **12** 1105–1131. MR3724980

Bhattacharya, A., Pati, D., Pillai, N. S. and Dunson, D. B. (2015). Dirichlet-Laplace priors for optimal shrinkage. *J. Amer. Statist. Assoc.* **110** 1479–1490. MR3449048

Carlin, B. P. and Louis, T. A. (1996). *Bayes and Empirical Bayes Methods for Data Analysis. Monographs on Statistics and Applied Probability* **69**. Chapman & Hall, London. MR1427749

Carvalho, C. M., Polson, N. G. and Scott, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika* **97** 465–480. MR2650751

Castillo, I. and van der Vaart, A. (2012). Needles and straw in a haystack: posterior concentration for possibly sparse sequences. *Ann. Statist.* **40** 2069–2101. MR3059077

Donnet, S., Rivoirard, V., Rousseau, J. and Scricciolo, C. (2018). Posterior concentration rates for empirical Bayes procedures with applications to Dirichlet process mixtures. *Bernoulli* **24** 231–256. MR3706755

Donoho, D. L., Johnstone, I. M., Hoch, J. C. and Stern, A. S. (1992). Maximum entropy and the nearly black object. *J. Roy. Statist. Soc. Ser. B* **54** 41–81. With discussion and a reply by the authors. MR1157714

Efron, B. (2010). *Large-Scale Inference. Institute of Mathematical Statistics Monographs* **1**. Cambridge University Press, Cambridge. MR2724758

Gao, C. and Zhou, H. H. (2016). Rate exact Bayesian adaptation with modified block priors. *Ann. Statist.* **44** 318–345. MR3449770

Ghosal, S., Ghosh, J. K. and van der Vaart, A. W. (2000). Convergence rates of posterior distributions. *Ann. Statist.* **28** 500–531. MR1790007

Ghosal, S. and van der Vaart, A. W. (2001). Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities. *Ann. Statist.* **29** 1233–1263. MR1873329

Ghosal, S. and van der Vaart, A. W. (2007a). Posterior convergence rates of Dirichlet mixtures at smooth densities. *Ann. Statist.* **35** 697–723. MR2336864

Ghosal, S. and van der Vaart, A. (2007b). Convergence rates of posterior distributions for non-i.i.d. observations. *Ann. Statist.* **35** 192–223. MR2332274

Ghosal, S. and van der Vaart, A. (2017). *Fundamentals of Nonparametric Bayesian Inference. Cambridge Series in Statistical and Probabilistic Mathematics* **44**. Cambridge University Press, Cambridge. MR3587782

Kruijer, W., Rousseau, J. and van der Vaart, A. (2010). Adaptive Bayesian density estimation with location-scale mixtures. *Electron. J. Stat.* **4** 1225–1257. MR2735885

Lee, K., Lee, J. and Lin, L. (2017). Minimax posterior convergence rates and model selection consistency in high-dimensional DAG models based on sparse Cholesky factors. *Ann. Statist.*, to appear, `arXiv:1811.06198`.

Martin, R. (2017). Invited comment on the article by van der Pas, Szabó, and van der Vaart. *Bayesian Anal.* **12** 1254–1258. MR3724985

Martin, R. (2018). Empirical priors and posterior concentration rates for a monotone density. *Sankhya A*, to appear, `arXiv:1706.08567`.

Martin, R., Mess, R. and Walker, S. G. (2017). Empirical Bayes posterior concentration in sparse high-dimensional linear models. *Bernoulli* **23** 1822–1847. MR3624879

Martin, R. and Ning, B. (2019). Empirical priors and coverage of posterior credible sets in a sparse normal mean model. `arXiv:1812.02150`.

Martin, R. and Shen, W. (2017). Asymptotically optimal empirical Bayes inference in a piecewise constant sequence model. `arXiv:1712.03848`.

Martin, R. and Tang, Y. (2019). Empirical priors for prediction in sparse high-dimensional linear regression. `arXiv:1903.00961`.

Martin, R. and Walker, S. G. (2014). Asymptotically minimax empirical Bayes estimation of a sparse normal mean vector. *Electron. J. Stat.* **8** 2188–2206. MR3273623

Petrone, S., Rousseau, J. and Scricciolo, C. (2014). Bayes and empirical Bayes: do they merge? *Biometrika* **101** 285–302. MR3215348

Rousseau, J. and Szabo, B. (2017). Asymptotic behaviour of the empirical Bayes posteriors associated to maximum marginal likelihood estimator. *Ann. Statist.* **45** 833–865. MR3650402

Salomond, J.-B. (2014). Concentration rate and consistency of the posterior distribution for selected priors under monotonicity constraints. *Electron. J. Stat.* **8** 1380–1404. MR3263126

Scricciolo, C. (2007). On rates of convergence for Bayesian density estimation. *Scand. J. Statist.* **34** 626–642. MR2368802

Scricciolo, C. (2015). Bayesian adaptation. *J. Statist. Plann. Inference* **166** 87–101. MR3390136

Shen, W. and Ghosal, S. (2015). Adaptive Bayesian procedures using random series priors. *Scand. J. Stat.* **42** 1194–1213. MR3426318

Shen, X. and Wasserman, L. (2001). Rates of convergence of posterior distributions. *Ann. Statist.* **29** 687–714. MR1865337

Szabó, B. T., van der Vaart, A. W. and van Zanten, J. H. (2013). Empirical Bayes scaling of Gaussian priors in the white noise model. *Electron.*

J. Stat. **7** 991–1018. MR3044507

VAN DER PAS, S., SZABÓ, B. and VAN DER VAART, A. (2017a). Uncertainty quantification for the horseshoe (with discussion). *Bayesian Anal.* **12** 1221–1274. With a rejoinder by the authors. MR3724985

VAN DER PAS, S., SZABÓ, B. and VAN DER VAART, A. (2017b). Adaptive posterior contraction rates for the horseshoe. *Electron. J. Stat.* **11** 3196–3225. MR3705450

VAN DER VAART, A. W. and VAN ZANTEN, J. H. (2009). Adaptive Bayesian estimation using a Gaussian random field with inverse gamma bandwidth. *Ann. Statist.* **37** 2655–2675. MR2541442

VAN ERVEN, T. and HARREMOËS, P. (2014). Rényi divergence and Kullback-Leibler divergence. *IEEE Trans. Inform. Theory* **60** 3797–3820. MR3225930

WALKER, S. and HJORT, N. L. (2001). On Bayesian consistency. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **63** 811–821. MR1872068

WALKER, S. G., LIJOI, A. and PRÜNSTER, I. (2007). On rates of convergence for posterior distributions in infinite-dimensional models. *Ann. Statist.* **35** 738–746. MR2336866