Received: 2 September 2019 | Revised: 12 May 2020 | Accepted: 13 July 2020

DOI: 10.1111/sjos.12485

ORIGINAL ARTICLE

Scandinavian Journal of Statistics

Objective priors in the empirical Bayes framework

Ilja Klebanov[1] | **Alexander Sikorski**[1] | **Christof Schütte**[1,2] | **Susanna Röblitz**[3]

[1]Zuse Institute Berlin

[2]Department of Mathematics and Computer Science, Free University Berlin

[3]Department of Informatics University of Bergen

Correspondence
Ilja Klebanov, Zuse Institute Berlin, Takustr. 7, 14195 Berlin, Germany.
Email: klebanov@zib.de

Funding information
Federal Ministry for Education and Research (BMBF), Grant/Award Number: Project BovSys (FKZ031A311); Matheon supported by Einstein Foundation Berlin, Grant/Award Number: ECMath Project CH13

Abstract

When dealing with Bayesian inference the choice of the prior often remains a debatable question. Empirical Bayes methods offer a data-driven solution to this problem by estimating the prior itself from an ensemble of data. In the nonparametric case, the maximum likelihood estimate is known to overfit the data, an issue that is commonly tackled by regularization. However, the majority of regularizations are ad hoc choices which lack invariance under reparametrization of the model and result in inconsistent estimates for equivalent models. We introduce a nonparametric, transformation-invariant estimator for the prior distribution. Being defined in terms of the missing information similar to the reference prior, it can be seen as an extension of the latter to the data-driven setting. This implies a natural interpretation as a trade-off between choosing the least informative prior and incorporating the information provided by the data, a symbiosis between the objective and empirical Bayes methodologies.

KEYWORDS

nonparametric inference, information, invariance, Jeffreys prior, maximum penalized likelihood estimator, reference prior

1 | INTRODUCTION

Inferring a parameter $\theta \in \Theta$ from a measurement $x \in \mathcal{X}$ using Bayes' rule requires prior knowledge about θ, which is not given in many applications. This has led to a lot of controversy in the statistical community and to harsh criticism concerning the objectivity of the Bayesian approach. In the past decades, Bayesian statisticians have given (at least) two solutions to the dilemma of missing prior information:

(A) *(noninformative) objective priors:* Objective Bayesian analysis and reference priors in particular (Berger & Bernardo, 1992; Berger, Bernardo, & Sun, 2009; Bernardo, 1979) apply mostly information theoretic ideas to construct priors that are invariant under reparametrization and can be argued to be noninformative.

(B) *empirical Bayes methods:* If independent measurements $x_m \in \mathcal{X}$, $m = 1, \dots, M$, are given for a large number M of "individuals" with individual parametrizations $\theta_m \in \Theta$, which is the case in many statistical studies, empirical Bayes methods (Carlin & Louis, 1996; Casella, 1985; Efron, 2010; Maritz & Lwin, 1989; Robbins, 1956) use this knowledge to construct an *informative* prior as a first step and then apply it for the Bayesian inference of the individual parametrizations θ_m (or any future parametrization θ_* with measurement x_*) in a second step. A typical application is the retrieval of patient-specific parametrizations in large clinical studies, see Neal and Kerckhoffs (2009). However, as discussed below, many of these methods fail to be consistent under reparametrization.

The aim of this article is to extend the construction of reference priors to the empirical Bayes framework in order to derive transformation invariant and informative priors from such "cohort data". We will perform this construction along the lines of the definition of reference priors in Berger et al. (2009) and use a similar notation. Likewise, we concentrate mainly on problems with one continuous parameter ($\Theta \subseteq \mathbb{R}^d$, $d = 1$), possible generalizations to the multiparameter case $d > 1$ are addressed shortly in Section 7.

This article is organized as follows. After introducing the notation in Section 2, we discuss empirical Bayes methods and the inconsistency of maximum penalized likelihood estimation (MPLE) under reparametrization, which is the main motivation for our work, in Section 3. Section 4 provides a solution to this issue by following the same ideas as in the construction of reference priors, resulting in a prior estimate we term the *empirical reference prior*. A rigorous definition and analysis of empirical reference priors is given. Section 5 provides two algorithms which show how the empirical reference prior can be implemented in practice (under the assumption of asymptotic normality)—one based on optimization and the other on a fixed point iteration. In Section 6 we then apply our methodology to a synthetic data set, illustrating invariance of the empirical reference prior under reparametrizations, as well as to the famous baseball data set of Efron and Morris (1975), where we compare our approach with the James–Stein estimator. In Section 7 we discuss possible generalizations of our approach to the multiparameter case $d > 1$, followed by a short conclusion in Section 8.

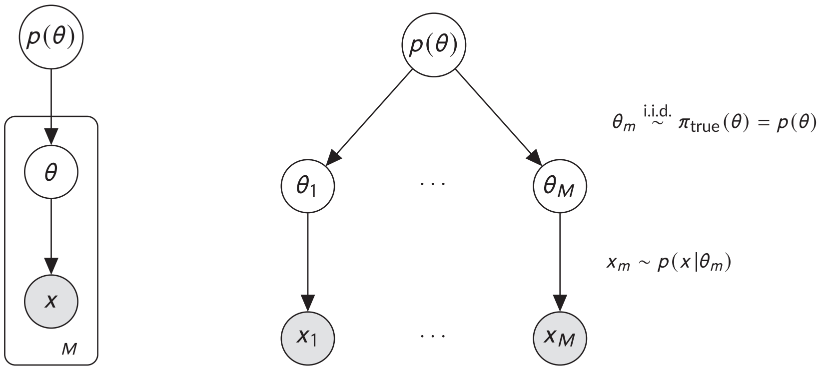

FIGURE 1 Graphical model and schematic representation of the underlying probabilistic model. Such a setup will be referred to as empirical Bayes framework

2 | SETUP AND NOTATION

We will work in the empirical Bayes framework described above and visualized in Figure 1. Specifically, denoting the likelihood model by

$$\mathcal{M} = \{p(x|\theta),\ x \in \mathcal{X},\ \theta \in \Theta\}, \quad \mathcal{X} \subseteq \mathbb{R}^n, \quad \Theta \subseteq \mathbb{R}^d, \quad d = 1,$$

the data $X = (x_1, \dots, x_M)$ are generated by a two-stage process, where the "individual" parametrizations $\theta_1, \dots, \theta_M$ are independent draws from the unknown prior distribution $p(\theta)$ and each data point $x_m \in \mathcal{X}$ is drawn independently from $p(x|\theta_m)$, $m = 1, \dots, M$. Note that $\mathcal{M}$ denotes the model corresponding to the complete observation *vector* $x \in \mathcal{X} \subseteq \mathbb{R}^n$ and that the data $X = (x_1, \dots, x_M)$ consists of M observation vectors $x_1, \dots, x_M \in \mathcal{X}$. This convention is necessary because our theory requires the introduction of (artificial) independent replications of the entire experiment, denoted by the model

$$\mathcal{M}^k = \left\{ p(\vec{x}|\theta) = \prod_{i=1}^{k} p(x^{(i)}|\theta),\ \vec{x} = (x^{(1)}, \dots, x^{(k)}) \in \mathcal{X}^k,\ \theta \in \Theta \right\}.$$

We adopt the standard abuse of notation, denoting all density functions by the letter p and letting the argument indicate which random variable it belongs to, for example, $p(x)$ is the marginal density of x, while $p(\theta|x)$ denotes the posterior density of θ given x. In addition, $\pi(\theta)$ will denote any other possible (proposed, guessed, or estimated) prior on Θ from some class $\mathcal{P}$ of admissible priors, $p(x|\pi) := \int p(x|\theta)\pi(\theta)\mathrm{d}\theta$ the corresponding prior predictive distribution and $\pi(\theta|x) \propto \pi(\theta)p(x|\theta)$ the corresponding posterior. Furthermore, $\pi_{\text{true}}(\theta) := p(\theta)$ denotes the "true" data-generating prior.

The prior may be viewed as a hyperparameter $\pi \in \mathcal{P}$, in which case we assume conditional independence of π and x given θ. The marginal likelihood of π given the entire data $X = (x_1, \dots, x_M)$ is given by

$$L(\pi) = L(\pi \mid \mathcal{M}, X) = \prod_{m=1}^{M} p(x_m|\pi), \quad p(x|\pi) = \int p(x|\theta)\ \pi(\theta)\ \mathrm{d}\theta. \tag{1}$$

We will also assume both the parametric and the hyperparametric model to be identifiable, see van der Vaart (1998, section 5.5), that is,

$$p(x|\theta_1) = p(x|\theta_2) \quad \Leftrightarrow \quad \theta_1 = \theta_2, \quad \theta_1, \theta_2 \in \Theta, \tag{2}$$

$$p(x|\pi_1) = p(x|\pi_2) \quad \Leftrightarrow \quad \pi_1 = \pi_2, \quad \pi_1, \pi_2 \in \mathcal{P}, \tag{3}$$

since otherwise there would be no chance to recover the true distribution $\pi_{\text{true}}(\theta) = p(\theta)$ from no matter how many measurements.

3 | INCONSISTENCY OF EMPIRICAL BAYES METHODS

Roughly speaking, empirical Bayes methods perform statistical inference in two steps—first, they estimate the prior from the entire data $X = (x_1, \dots, x_M)$, and second, they apply Bayes' rule with that prior for each x_m separately[1]. We concentrate on the first step, which is often performed by maximizing the marginal likelihood $L(\pi)$, for example, by means of the EM algorithm introduced by Dempster, Laird, and Rubin (1977). This procedure can be viewed as an interplay of frequentist and Bayesian statistics: The prior is chosen by maximum likelihood estimation (MLE), the actual individual parametrizations are then inferred using Bayes' rule. However, in the nonparametric case (NPMLE), that is, if no finite-dimensional parametric form of the prior is assumed, it can be proven that the marginal likelihood $L(\pi)$ is maximized by a discrete distribution π_{MLE}

$$\pi_{\text{MLE}} = \arg\max_\pi \ \log L(\pi) = \sum_{\nu=1}^{N} w_\nu \delta_{\check{\theta}_\nu}, \quad N \le M, \ w_\nu > 0, \quad \sum_{\nu=1}^{N} w_\nu = 1, \tag{4}$$

with at most M nodes $\check{\theta}_\nu$, see Laird (1978, theorems 2–5) or Lindsay (1995, theorem 21). This typical issue of *overfitting the data* is often dealt with by subtracting a *roughness penalty* (or *regularization term*) $\Phi(\pi) = \Phi(\pi \mid \mathcal{M})$ from the marginal log-likelihood function $\log L(\pi)$, such that smooth or noninformative priors are favored, resulting in the so-called maximum penalized likelihood estimate (MPLE):

$$\pi_{\text{MPLE}} = \arg\max_\pi \ \log L(\pi) - \gamma\Phi(\pi). \tag{5}$$

The constant $\gamma > 0$ balances the trade-off between goodness of fit and smoothness or noninformativity of the prior. This approach can be viewed from the Bayesian perspective as choosing a hyperprior $p(\pi) \propto e^{-\gamma\Phi(\pi)}$ for the hyperparameter π and performing a maximum a posteriori (MAP) estimation for π:

$$\pi_{\text{MAP}} = \arg\max_\pi \ L(\pi)\, p(\pi) = \arg\max_\pi \ \log L(\pi) - \gamma\Phi(\pi) = \pi_{\text{MPLE}} \ . \tag{6}$$

Desirable properties of the roughness penalty function $\Phi(\pi) = \Phi(\pi \mid \mathcal{M})$ are:

[1]For theoretical purposes one should rather think of applying Bayes' rule to some future measurement x_* to infer its parametrization θ_* in order to avoid reusing the data. In practice and for a large number M of measurements this distinction is of little relevance, since the influence of one data point x_m on the prior estimation can usually be neglected.

(a) noninformativity: without any extra information about the parameter or the prior, we want to keep our assumptions to a minimum (in the sense of objective Bayes methods);
(b) invariance under transformations of the parameter space Θ (reparametrizations);
(c) invariance under transformations of the measurement space $\mathcal{X}$;
(d) convexity: Since $\log L(\pi)$ is concave in the NPMLE case (Lindsay, 1995, section 5.1.3), a convex penalty function $\Phi(\pi)$ would guarantee a concave optimization problem (5);
(e) natural and intuitive justification.

The penalty functions currently used are mostly ad hoc and rather brute force solutions that confine amplitudes, for example, ridge regression (McLachlan & Krishnan, 2008, section 1.6), or derivatives (Good & Gaskins, 1971; Silverman, 1982) of the prior, which are neither invariant under reparametrizations nor have a natural derivation. A more contemporary alternative is to use Dirichlet process hyperpriors $p(\pi)$, which have similar limitations. In order to incorporate the notion of noninformativity, Good (1963) suggested to use the entropy as a roughness penalty,

$$\Phi_{H_\theta}(\pi) = -\, H_\theta(\pi) = \int_{\mathcal{X}} \pi(\theta) \log \pi(\theta)\ \mathrm{d}\theta, \tag{7}$$

which is a natural approach from an information-theoretic viewpoint, since high entropy corresponds to high uncertainty and thereby noninformativity of the prior. However, Φ_{H_θ} is not invariant under reparametrizations, making it, as Good puts it, "somewhat arbitrary" (Good, 1963, p. 912):

> It could be objected that, especially for a continuous distribution, entropy is somewhat arbitrary, since it is variant under a transformation of the independent variable.

Following Shannon's derivation of the entropy H_θ, we will explain why the concepts of mutual information and missing information, both of which are invariant under transformations, are far more natural quantities to use in our setup. Prior to that, let us make the notion of invariance more precise.

3.1 | Invariance under transformations

Invariance under transformations guarantees consistency of the resulting probability density estimate and follows the following principle: If two statisticians use equivalent models to explain equivalent data their results must be consistent, or, as Shore and Johnson (1980) put it,

> […] reasonable methods of inductive inference should lead to consistent results when there are different ways of taking the same information into account (for example, in different coordinate systems).

Let us start with the definition:

Definition 1. Let $X = (x_1, \dots, x_M)$ be data generated by the hierarchical model described in Section 2 with $\mathcal{M} = \{p(x|\theta),\ x \in \mathcal{X},\ \theta \in \Theta \subseteq \mathbb{R}^d\}$. Furthermore, let $F = F[\mathcal{M}, X]$ be a function operating on probability densities in $\mathbb{R}^d$.

(I) We call F *invariant under transformations of the parameter* θ or *invariant under reparametrization*, if $F[\mathcal{M}, X](\pi) = F[\mathcal{M}_\varphi, X](\varphi_* \pi)$ for any diffeomorphism $\varphi : \Theta \to \tilde{\Theta},\ \theta \mapsto \tilde{\theta}$, where the transformed model $\mathcal{M}_\varphi$ is defined by

$$p(x|\tilde{\theta}) = p\left(x|\theta = \varphi^{-1}(\tilde{\theta})\right). \tag{8}$$

(II) We call F *invariant under transformations of the measurement* x, if $F[\mathcal{M}, X] = F[\mathcal{M}_\psi, X_\psi]$ for any diffeomorphism $\psi : \mathcal{X} \to \tilde{\mathcal{X}},\ x \mapsto \tilde{x}$, where the transformed model $\mathcal{M}_\psi$ and data $X_\psi = (\tilde{x}_1, \ldots, \tilde{x}_M)$ are defined by

$$p(\tilde{x}|\theta) = \psi_* p(x|\theta)|_{\tilde{x}} = |\det\,(\psi^{-1})'(\tilde{x})| \cdot p\left(x = \psi^{-1}(\tilde{x}) \mid \theta\right), \quad \tilde{x}_m = \psi(x_m). \tag{9}$$

Here and in the following, φ_* and ψ_* denote the pushforward of some measure (or density) under φ and ψ, respectively.

If we wish to use the function F for the estimation of the prior π (as in Equation (5) for $F = \log L - \gamma\Phi$), invariance of F causes the diagrams in Figure 2 to commute. Note that, if we restrict the considerations to some class $\mathcal{P}$ of admissible priors, then this class needs to be transformed correspondingly, $\mathcal{P}_\varphi := \{\varphi_* \pi,\ \pi \in \mathcal{P}\}$, such that the considered class of priors is consistent under reparametrization.

One class of functions that fulfills these invariance properties is introduced in the following theorem, where we also show that the marginal likelihood $L(\pi)$ is transformation invariant up to a constant.

Theorem 1. *Let $X = (x_1, \ldots, x_M)$ be data stemming from a model $\mathcal{M} = \{p(x|\theta),\ x \in \mathcal{X},\ \theta \in \Theta\}$ and*

$$F_g(\pi) = F_g[\mathcal{M}](\pi) = \int_\Theta \int_{\mathcal{X}} \pi(\theta)\ p(x|\theta)\ g\left(\frac{p(x|\theta)}{p(x|\pi)}\right) \mathrm{d}x\ \mathrm{d}\theta$$

for some measurable function $g : \mathbb{R} \to \mathbb{R}$, such that the above integral is well defined. Then F_g is invariant under transformations of θ and x. Furthermore, the marginal likelihood $L(\pi) = L(\pi \mid \mathcal{M}, X)$ defined by (1) is invariant under transformations of θ and x up to a multiplicative constant.

Proof. This is a straightforward application of the change of variables formula. ∎

$$\begin{tikzcd}
(\mathcal{M}, \mathcal{P}, X) \arrow[r, shift left, mapsto, "\varphi"] \arrow[d, mapsto, "\text{estimation of } \pi"'] & (\mathcal{M}_\varphi, \mathcal{P}_\varphi, X) \arrow[l, shift left, mapsto, "\varphi^{-1}"] \arrow[d, mapsto, "\text{estimation of } \pi"] \\
\pi_{\text{est}}(\theta) \arrow[r, shift left, mapsto, "\varphi_*"] & \pi_{\text{est}}^{\varphi}(\tilde{\theta}) \arrow[l, shift left, mapsto, "\varphi_*^{-1}"]
\end{tikzcd}
\qquad
\begin{tikzcd}
(\mathcal{M}, \mathcal{P}, X) \arrow[r, shift left, mapsto, "\psi"] \arrow[d, mapsto, "\text{estimation of } \pi"'] & (\mathcal{M}_\psi, \mathcal{P}, X_\psi) \arrow[l, shift left, mapsto, "\psi^{-1}"] \arrow[d, mapsto, "\text{estimation of } \pi"] \\
\pi_{\text{est}}(\theta) \arrow[r, shift left, mapsto, "\psi_*"] & \pi_{\text{est}}^{\psi}(\theta) \arrow[l, shift left, mapsto, "\psi_*^{-1}"]
\end{tikzcd}$$

FIGURE 2 Commutative diagrams illustrating the consistency of the density estimate $\pi_{\text{est}}(\theta)$. If the estimation is performed in a transformed parameter space $\tilde{\Theta} = \varphi(\Theta)$ or measurement space $\tilde{\mathcal{X}} = \psi(\mathcal{X})$ (the model $\mathcal{M}$, the class $\mathcal{P}$ of admissible priors and the data $X = (x_1, \ldots, x_M)$ are transformed accordingly), the results should be consistent. Note that, in the empirical Bayes framework, the data X enter in the estimation of the prior, hence the commutativity of the right diagram is not trivial

4 | OBJECTIVE BAYESIAN APPROACH TO EMPIRICAL BAYES METHODS

The lack of invariance of common empirical Bayes methods described above will now be tackled by an approach similar to the construction of reference priors and performed along the lines of Berger et al. (2009). The two key ingredients for defining reference priors are permissibility, which yields a rigorous justification for dealing with improper priors, and the maximizing missing information (MMI) property, which is derived from information theoretic considerations and can be argued to guarantee the least informative prior.

Permissibility allows for the (positive and continuous) prior to be improper as long as it yields a proper posterior for each measurement (which is the object Bayesian statisticians are actually interested in) and these posteriors can be approximated by using proper priors arising from restricting the prior to compact subspaces $\Theta_i \subseteq \Theta$.

In the empirical Bayes framework, where the aim is to approximate the true (proper) prior $\pi_{\text{true}}(\theta) = p(\theta)$, improper priors are less of an issue and we will limit ourselves to proper priors. Furthermore, it is completely unclear how to deal with improper priors in this framework, since the "restriction property" of reference priors is neither achievable nor desirable, see the remark and example below. For this reason and since the concept of permissibility has been elaborated extensively in Berger et al. (2009), we will just state its definition.

Definition 2. A strictly positive continuous function $\pi(\theta)$ is a *permissible prior* for model $\mathcal{M}$, if

(I) for each $x \in \mathcal{X}$, $\pi(\theta|x)$ is proper, that is, $\int_{\Theta} p(x|\theta)\ \pi(\theta)\ \mathrm{d}\theta < \infty$,

(II) for any increasing sequence of compact subsets $\Theta_i \subseteq \Theta$, $i \in \mathrm{N}$, with $\cup_i \Theta_i = \Theta$, the corresponding posterior sequence $\pi_i(\theta|x)$ is expected logarithmically convergent to $\pi(\theta|x)$,

$$\lim_{i\to\infty} D_{\mathrm{KL}}(\pi_i(\cdot|x)\|\pi(\cdot|x)) = 0, \qquad \pi_i = \frac{\pi\ \mathbb{1}_{\Theta_i}}{\int_{\Theta_i} \pi(\theta)\ \mathrm{d}\theta}.$$

Here, $\mathbb{1}_{\Theta_i}$ is the indicator function of the subset $\Theta_i \subseteq \Theta$ and $D_{\mathrm{KL}}(\cdot\|\cdot)$ denotes the Kullback–Leibler divergence defined by

$$D_{\mathrm{KL}}(p\|q) := \int_{\Theta} p(\theta)\ \log\left(\frac{p(\theta)}{q(\theta)}\right)\ \mathrm{d}\theta.$$

While permissibility is rather a technicality for dealing with improper priors, the MMI property should be seen as the defining property of reference priors and will now be discussed in more detail.

As motivated in the introduction, penalizing by means of entropy provides a natural approach to incorporate the idea of noninformativity about the parameter into the inference process. However, if we follow Shannon's derivation of the entropy H_θ, we see that it is not the appropriate notion in our setup. Shannon (1948) derived the entropy H_θ from the insight that the proper way to quantify the information gain, when an event with probability $p \in [0, 1]$ actually occurs, is $-\log(p) \in [0, \infty]$. He then defined the entropy as the expected information gain. However, the continuous analogue to this notion, the differential entropy H_θ given by (7), faces several complications:

- The information gain $-\log(\pi(\theta))$ from observing the value θ, as well as the entropy H_θ itself can become negative, which is difficult to interpret.
- H_θ is variant under transformations of θ, leading to an inconsistent notion of information.
- The information gain $-\log(\pi(\theta))$ relies on a *direct* and *exact* (error-free) measurement of θ, which is not plausible in the continuous case.

The last point becomes even more relevant in the empirical Bayes framework, where θ is not (and usually cannot be) measured directly, but is inferred from the measurement x of another quantity. The appropriate notion for the information gain in this setup is the Kullback–Leibler divergence $D_{\mathrm{KL}}(\pi(\cdot\,|x)\,\|\pi)$ between posterior and prior (Kullback & Leibler, 1951). Its expected value (from one observation of model $\mathcal{M}$), the so-called *expected information* or *mutual information* of θ and x, is always nonnegative and invariant under transformations.

Definition 3. The *expected information* gained from one observation of model $\mathcal{M}$ on a parameter θ with prior $\pi(\theta)$ is given by

$$\mathcal{I}[\pi \mid \mathcal{M}] = \int_{\mathcal{X}} p(x|\pi)\; D_{\mathrm{KL}}(\pi(\cdot|x)\|\pi)\;\mathrm{d}x = \int_{\Theta}\int_{\mathcal{X}} \pi(\theta)\; p(x|\theta)\;\log\left(\frac{p(x|\theta)}{p(x|\pi)}\right)\;\mathrm{d}x\;\mathrm{d}\theta.$$

The expected information has very appealing properties for a penalty term:

Theorem 2. *The expected information $\mathcal{I}[\pi \mid \mathcal{M}]$ is concave in π and invariant under transformations of θ and x.*

Proof. Concavity is proven in Cover and Thomas (2006, theorem 2.7.4) while invariance follows directly from Theorem 1. ■

As argued in Bernardo (1979), the quantity $\mathcal{I}[\pi \mid \mathcal{M}^k]$, the expected information on θ gained from k independent observations of $\mathcal{M}$, describes the *missing information* on θ as k goes to infinity:

> By performing infinite replications of $\mathcal{M}$ one would get to know precisely the value of θ. Thus, $\mathcal{I}[\pi \mid \mathcal{M}^\infty]$ measures the amount of missing information about θ when the prior is $\pi(\theta)$. It seems natural to define "vague initial knowledge" about θ as that described by the density $\pi(\theta)$ which maximizes the missing information in the class $\mathcal{P}$.[2]

Following this idea, maximizing the missing information (MMI) results in the least informative prior, making $\Phi(\pi) = -\mathcal{I}[\pi \mid \mathcal{M}^\infty]$ an appealing penalty term in (5). It is now tempting to define empirical reference priors by

$$\pi_* = \arg\max_\pi \;\left(\log L(\pi) + \gamma\; \mathcal{I}[\pi \mid \mathcal{M}^\infty]\right). \tag{10}$$

[2]The notation in this quotation has been adapted to ours. Note that the first statement makes use of the identifiability (2) of θ.

However, since $\mathcal{I}[\pi \mid \mathcal{M}^k]$ typically diverges for $k \to \infty$, the following detour around the optimization formulation (10) appears necessary (as we will see in Section 4.3, some simplifications are possible under certain regularity conditions):

Definition 4. Let $\mathcal{M} = \{p(x|\theta),\ x \in \mathcal{X},\ \theta \in \Theta\}$ be a model, $\mathcal{P}$ be a class of prior functions π with $\int p(x|\theta)\ \pi(\theta)\ \mathrm{d}\theta < \infty$ and $X = (x_1, \dots, x_M)$ be the data consisting of M independent samples from $p(x)$. The function $\pi \in \mathcal{P}$ is said to have

(I) the MMI = MMI($\mathcal{M}, \mathcal{P}$) property for model $\mathcal{M}$ given $\mathcal{P}$ if, for any compact set $\Theta_0 \subseteq \Theta$ and any $\tilde{\pi} \in \mathcal{P}$,

$$\lim_{k\to\infty} (\mathcal{I}[\pi_0 \mid \mathcal{M}^k] - \mathcal{I}[\tilde{\pi}_0 \mid \mathcal{M}^k]) \geq 0, \tag{11}$$

where π_0 and $\tilde{\pi}_0$ denote the (renormalized) restrictions of π and $\tilde{\pi}$ to Θ_0;

(II) the MMIL(X) = MMIL($\mathcal{M}, \mathcal{P}, X, \gamma$) property for model $\mathcal{M}$ given $X, \mathcal{P}$ and $\gamma > 0$ if π is a proper probability density and if, for any proper probability density $\tilde{\pi} \in \mathcal{P}$,

$$\lim_{k\to\infty} \left((\log L(\pi) + \gamma \mathcal{I}[\pi \mid \mathcal{M}^k]) - (\log L(\tilde{\pi}) + \gamma \mathcal{I}[\tilde{\pi} \mid \mathcal{M}^k])\right) \geq 0. \tag{12}$$

Here, MMIL stands for "maximizing (a trade-off between) missing information and log-likelihood". Both definitions are only useful if all the values of the expected information in (11) and (12) are finite. This can be guaranteed by restricting ourselves to some convenient class $\mathcal{P}$ of admissible priors. Typically, one requires strict positivity and continuity of the priors as well as the existence of proper posteriors, see Berger et al. (2009, section 3.3), but different choices of $\mathcal{P}$ are also possible.

4.1 | The formal definition of empirical reference priors

Similar to Berger et al. (2009), and in accordance with their definition of reference priors, we now define empirical reference priors, which constitute the main contribution of this article:

Definition 5. A function $\pi_{\mathrm{ref}}(\theta) = \pi_{\mathrm{ref}}(\theta \mid \mathcal{M}, \mathcal{P})$ is a *reference prior* for model $\mathcal{M}$ given prior class $\mathcal{P}$, if it is permissible and has the MMI property. A probability density $\pi_{\mathrm{erp}}(\theta) = \pi_{\mathrm{erp}}(\theta \mid \mathcal{M}, \mathcal{P}, X, \gamma)$ is an *empirical reference prior* for model $\mathcal{M}$ given prior class $\mathcal{P}$, data $X = (x_1, \dots, x_M)$ and smoothing parameter $\gamma > 0$, if it has the MMIL(X) property.

Let us now formulate and prove the key properties of empirical reference priors.

Theorem 3 (Invariance of the empirical reference prior). *The empirical reference prior is invariant under transformations of θ and x in the following sense:*

$$\pi_{\mathrm{erp}}(\theta \mid \mathcal{M}_\varphi, \mathcal{P}_\varphi, X, \gamma) = \varphi_* \pi_{\mathrm{erp}}(\theta \mid \mathcal{M}, \mathcal{P}, X, \gamma),$$
$$\pi_{\mathrm{erp}}(\theta \mid \mathcal{M}_\psi, \mathcal{P}, X_\psi, \gamma) = \pi_{\mathrm{erp}}(\theta \mid \mathcal{M}, \mathcal{P}, X, \gamma),$$

where we adopted the notation from the definition of invariance and $\mathcal{P}_\varphi := \{\varphi_* \pi,\ \pi \in \mathcal{P}\}$.

Proof. Since $\log L(\pi)$ is invariant under transformations of θ and x up to an additive constant by Theorem 1, this is a direct consequence of Theorems 1 and 2. ∎

Theorem 4 (Compatibility with sufficient statistics). *If the model $\mathcal{M} = \{p(x|\theta),\ x \in \mathcal{X},\ \theta \in \Theta\}$ has a sufficient statistic $t = t(x) \in \mathcal{T}$, then*

$$\pi_{\text{erp}}(\theta \mid \mathcal{M}, \mathcal{P}, X, \gamma) = \pi_{\text{erp}}(\theta \mid \mathcal{M}_t, \mathcal{P}, T, \gamma),$$

where $T = t(X) \in \mathcal{T}^M$ and $\mathcal{M}_t = \{p(t|\theta),\ t \in \mathcal{T},\ \theta \in \Theta_0\}$ is the corresponding model in terms of t.

Proof. Since t is a function of x and a sufficient statistic for θ, we obtain

$$p(x|\theta) = p(x|t(x))\ p(t(x)|\theta).$$

This implies that the marginal log-likelihoods $\log L(\pi)$ and $\log L_t(\pi)$ agree up to an additive constant, where $L_t(\pi) = \prod_{m=1}^{M} p(t_m|\pi),\ t_m := t(x_m)$, denotes the marginal likelihood in terms of t:

$$\log L(\pi) = \sum_{m=1}^{M} \log \int p(x_m|\theta)\ \pi(\theta)\ \mathrm{d}\theta = \sum_{m=1}^{M} \log p(x_m|t_m) + L_t(\pi).$$

Since the expected information is also invariant under such transformations, $\mathcal{I}[\pi \mid \mathcal{M}^k] = \mathcal{I}[\pi \mid \mathcal{M}_t^k]$, see Berger et al. (2009, theorem 5), this proves the claim. ∎

Remark 1. Reference priors have the appealing property that their restrictions to any compact subset Θ_0 coincide with the reference priors on Θ_0, see Berger et al. (2009, section 5):

$$\pi_{\text{ref}}(\theta \mid \mathcal{M}, \mathcal{P})|_{\Theta_0} = \pi_{\text{ref}}(\theta \mid \mathcal{M}_0, \mathcal{P}_0), \quad \mathcal{M}_0 = \{p(x|\theta),\ x \in \mathcal{X},\ \theta \in \Theta_0\}, \quad \mathcal{P}_0 = \left\{\pi|_{\Theta_0},\ \pi \in \mathcal{P}\right\}.$$

However, unlike for objective priors in the absence of data, this property is not desirable in the empirical Bayes framework, as explained in the example below, and usually will not be fulfilled by empirical reference priors. Therefore, a definition of MMIL(X) using restrictions of possibly improper priors (as in the definition of MMI) is not meaningful and we are forced to limit ourselves to proper priors. This limitation is not too restrictive since the aim of empirical Bayes methods is to approximate the true prior $\pi_{\text{true}}(\theta)$ and improper priors do not play a major role. For compact parameter spaces Θ (and in all other cases for which the reference prior turns out to be proper), empirical reference priors provide a meaningful generalization of reference priors, which then correspond to the case $M = 0$, the absence of data X.

Example 1. Let the true data-generating prior be the uniform prior $\pi_{\text{true}} \equiv \frac{1}{2}$ on $\Theta = [0, 2]$, that is, $\theta \sim \text{Unif}([0, 2])$, and $\mathcal{M}$ be the location model given by $x|\theta \sim \mathcal{N}(\theta, 0.5^2)$. For a "large" data set X consisting of $M = 100$ measurements, the empirical reference prior $\pi_{\text{erp}}(\theta \mid \mathcal{M}, \mathcal{P}, X)$ can be expected to provide a good approximation of $\pi_{\text{true}}(\theta)$, hence its restriction to $\Theta_0 = [0, 1]$ will be approximately uniform, see Figure 3. However, the empirical reference prior $\pi_{\text{erp}}(\theta \mid \mathcal{M}_0, \mathcal{P}_0, X)$ on Θ_0 has to put much more weight on values close to 1 in order to explain the many measurements x_m which are larger than 1.

Hence, unlike for reference priors, the equality $\pi_{\text{erp}}(\theta \mid \mathcal{M}, \mathcal{P}, X)|_{\Theta_0} = \pi_{\text{erp}}(\theta \mid \mathcal{M}_0, \mathcal{P}_0, X)$ is neither fulfilled nor desirable. Of course, in practice, the parameter space Θ should agree with (or at least include) the domain of the true prior in order to be consistent with the data generating distribution.

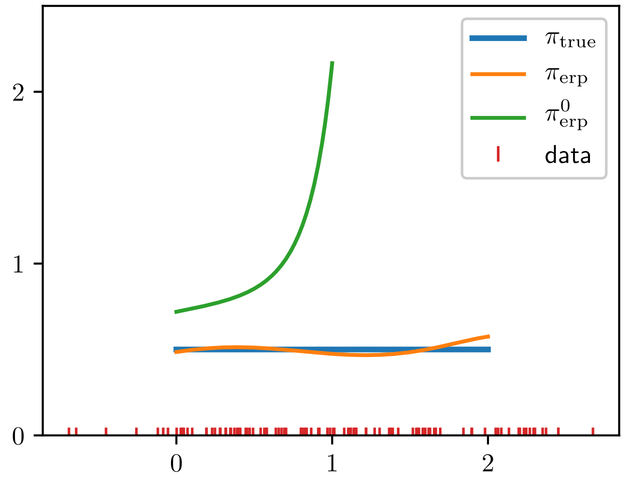

FIGURE 3 Inconsistency of the empirical reference prior under restrictions of the parameter space: $\pi_{\text{erp}} := \pi_{\text{erp}}(\theta \mid \mathcal{M}, \mathcal{P}, X)|_{\Theta_0}$ restricted to $\Theta_0 = [0, 1]$ does not agree with $\pi^0_{\text{erp}} := \pi_{\text{erp}}(\theta \mid \mathcal{M}_0, \mathcal{P}_0, X)$. Instead, π^0_{erp} puts a lot of weight close to the right boundary in order to explain all the data points larger than 1. Unlike in the case of reference priors, this disagreement is reasonable in the presence of data. For simplicity, we chose $\gamma = 1$ in this example [Color figure can be viewed at wileyonlinelibrary.com]

4.2 | Choice of the smoothing parameter γ

So far, it is completely unclear how the smoothing parameter $\gamma > 0$ should be chosen. In fact, even entirely different ways of performing the trade-off between $L(\pi)$ and $\mathcal{I}[\pi \mid \mathcal{M}^\infty]$ are possible. All theoretical results remain unchanged if we replace

$$\log L(\pi) + \gamma \mathcal{I}[\pi \mid \mathcal{M}^k] \quad \text{by} \quad \Psi(\log L(\pi), \mathcal{I}[\pi \mid \mathcal{M}^k])$$

in (12), where $\Psi : \mathbb{R}^2 \to \mathbb{R}$ can be any concave function that is monotonically increasing in both arguments. We will stick to the former formulation and choose γ via likelihood cross-validation (Silverman, 1986, equation (3.43)),

$$\gamma_* = \arg\max_\gamma \sum_{m=1}^{M} \log p\left(x_m \mid \pi_{\text{erp}}(\theta \mid \mathcal{M}, \mathcal{P}, X_{-m}, \gamma)\right), \tag{13}$$

where $X_{-m} := \{x_{m'} \in X \mid m' \neq m\}$ denotes the data set X with the mth point x_m left out.

Proposition 1. *The smoothing parameter (13) is invariant under transformations of θ and x, as long as the class $\mathcal{P}$ of admissible priors is transformed accordingly, $\mathcal{P}_\varphi = \{\varphi_*\pi,\ \pi \in \mathcal{P}\}$.*

Proof. From (8), (9), and Theorem 3 we obtain

$$p\left(x \mid \pi_{\text{erp}}(\tilde{\theta} \mid \mathcal{M}_\varphi, \mathcal{P}_\varphi, X, \gamma)\right) = p\left(x \mid \pi_{\text{erp}}(\theta \mid \mathcal{M}, \mathcal{P}, X, \gamma)\right),$$
$$p\left(\tilde{x} \mid \pi_{\text{erp}}(\theta \mid \mathcal{M}_\psi, \mathcal{P}, X_\psi, \gamma)\right) = C(\tilde{x})\, p\left(x = \psi^{-1}(\tilde{x}) \mid \pi_{\text{erp}}(\theta \mid \mathcal{M}, \mathcal{P}, X, \gamma)\right),$$

where we adopted the notation from the definition of invariance and $C(x) = |\det\, (\psi^{-1})'(\tilde{x})| > 0$ does not depend on γ. This proves the claim. ∎

4.3 | Empirical reference priors under asymptotic normality

As proven in Clarke and Barron (1994), the reference prior coincides with the Jeffreys prior (Jeffreys, 1946, 1961),

$$\pi^J(\theta) \propto J(\theta) := \sqrt{|\det i(\theta)|}, \quad i(\theta) := \int_{\mathcal{X}} p(x|\theta)\, \left(\nabla_\theta \log p(x|\theta)\right) \left(\nabla_\theta \log p(x|\theta)\right)^\intercal \, dx, \tag{14}$$

under certain regularity conditions, which basically ensure asymptotic posterior normality. Let us recall the basic results which we state for any dimension $d \in \mathbb{N}$ of the parameter space Θ.

Condition 1. The likelihood $p(x|\theta)$ is twice continuously differentiable in θ for almost every $x \in \mathcal{X}$. There exists $\epsilon > 0$ and for every θ there exists $\delta > 0$ such that for all $j, k = 1, \dots, d$ the functions

$$\mathbb{E}\left[\left|\frac{\partial}{\partial \theta_j} \log p(x|\theta)\right|^{2+\epsilon}\right] \quad \text{and} \quad \mathbb{E}\left[\sup_{\{\theta : \|\theta-\theta'\|<\delta\}} \left|\frac{\partial^2}{\partial \theta'_j \partial \theta'_k} \log p(x|\theta')\right|^2\right]$$

are finite and continuous in θ. The Fisher information matrix $i(\theta)$ defined by (14) is positive definite for each $\theta \in \Theta$ and equals $\tilde{i}(\theta) = -\mathbb{E}[D_\theta^2 \log p(x|\theta)]$, where D_θ^2 denotes the Hessian matrix with respect to θ. The model $\mathcal{M}$ is identifiable as defined by (2), the parameter space Θ is compact and $\mathcal{P}$ consists of all positive and continuous probability density functions on Θ (this implies that $\mathcal{P}$ is a convex set).

Proposition 2. *Under Condition 1, there exist positive constants $C_1, C_2 > 0$ such that, as $k \to \infty$,*

$$\mathcal{I}[\pi \mid \mathcal{M}^k] = C_1 \log(C_2 k) - D_{\mathrm{KL}}(\pi \| J) + o(1), \quad J(\theta) = \sqrt{|\det i(\theta)|}. \tag{15}$$

Proof. See Clarke and Barron (1994). ∎

Since the first term on the right-hand side of (15) does not depend on the prior $\pi(\theta)$ and the second one is independent of k, the reference prior coincides with Jeffreys prior π_J and the definition of empirical reference priors recovers the form of the MPLE (5):

Corollary 1. *Under Condition 1, Jeffreys prior π_J is the unique reference prior (up to scaling).*

Corollary 2. *Under Condition 1, the empirical reference prior is given by the following MPLE (5),*

$$\pi_{\mathrm{erp}} = \arg\max_{\pi \in \mathcal{P}} \ \log L(\pi) - \gamma \Phi_{\mathcal{I}}(\pi), \quad \Phi_{\mathcal{I}}(\pi) := D_{\mathrm{KL}}(\pi \| J), \tag{16}$$

where the penalty term $\Phi_{\mathcal{I}}$ will be referred to as the missing information penalty.

Theorems 1 and 2 imply desirable properties of the missing information penalty $\Phi_{\mathcal{I}}$ and, in particular, the existence of a unique empirical reference prior that is invariant under transformations of θ and x.

Corollary 3. *Under Condition 1, the optimization problem (16) is strictly concave in π and invariant under transformations of θ and x. Hence, since $\mathcal{P}$ is convex, it has a unique solution π_{erp}.*

Proof. For the concavity of $\log L(\pi)$ see Lindsay (1995, section 5.1.3), while the strict convexity of $\Phi_{\mathcal{I}}$ follows from van Erven and Harremoës (2014, theorem 11). Invariance is a direct consequence of Theorems 1 and 2. ∎

In order to realize the optimization formulation (16), we now compute analytically the gradient of the functional to be maximized. In addition, we characterize π_{erp} as the unique fixed point of a certain function $F^* : \mathcal{P} \to \mathcal{P}$, which motivates the fixed point iteration in Algorithm 2.

Theorem 5. *Let Condition 1 hold. Then the gradient of the functional in (16),*

$$\Psi(\pi) = \log L(\pi) - \gamma \Phi_{\mathcal{I}}(\pi) = \sum_{m=1}^{M} \log \int_{\Theta} p(x_m|\theta)\, \pi(\theta)\, \mathrm{d}\theta - \gamma D_{\mathrm{KL}}(\pi \| J), \tag{17}$$

with respect to $\langle \cdot, \cdot \rangle_{L^2(\Theta)}$ *is given by*

$$(\nabla_\pi \Psi(\pi))(\theta) = v^*(\theta) - \int v^*(\theta')\, \mathrm{d}\theta', \quad v^*(\theta) = \sum_{m=1}^{M} \frac{p(x_m|\theta)}{\int p(x_m|\theta')\, \pi(\theta')\, \mathrm{d}\theta'} - \gamma \log \frac{\pi(\theta)}{J(\theta)}. \tag{18}$$

It follows that the empirical reference prior π_{erp} *is the unique fixed point of* $F^* : \mathcal{P} \to \mathcal{P}$,

$$F^*(\pi) := \frac{F(\pi)}{\int (F(\pi))(\theta)\, \mathrm{d}\theta}, \quad F(\pi) := J\, \exp\left(\gamma^{-1} \sum_{m=1}^{M} \frac{p(x_m|\cdot)}{\int_\Theta p(x_m|\theta')\, \pi(\theta')\, \mathrm{d}\theta'} \right). \tag{19}$$

Proof. We consider perturbations of $\Psi(\pi)$ in arbitrary directions $v \in C(\Theta)$ with $\int v = 0$:

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}\varepsilon} \Psi(\pi + \varepsilon v) &= \sum_{m=1}^{M} \frac{\int p(x_m|\theta)\, v(\theta)\, \mathrm{d}\theta}{\int p(x_m|\theta')\, \pi(\theta')\, \mathrm{d}\theta'} - \gamma \int v(\theta)\, \log \frac{\pi(\theta)}{J(\theta)}\, \mathrm{d}\theta - \gamma \underbrace{\int v(\theta)\, \mathrm{d}\theta}_{=0} \\ &= \left\langle v\, ,\, \sum_{m=1}^{M} \frac{p(x_m|\theta)}{\int p(x_m|\theta')\, \pi(\theta')\, \mathrm{d}\theta'} - \gamma \log \frac{\pi(\theta)}{J(\theta)} \right\rangle_{L^2(\Theta)}, \end{aligned}$$

which proves (18). The above inner product is zero for any $v \in C(\Theta)$ with $\int v = 0$ if and only if its second argument is constant in θ. Hence,

$$\pi_{\mathrm{erp}}(\theta) \propto J(\theta)\, \exp\left(\gamma^{-1} \sum_{m=1}^{M} \frac{p(x_m|\theta)}{\int_\Theta p(x_m|\theta')\, \pi(\theta')\, \mathrm{d}\theta'} \right),$$

which proves the second claim. Note that $F^* : \mathcal{P} \to \mathcal{P}$, since we assumed the parameter space Θ to be compact, and that the uniqueness of the fixed point follows from the strict convexity of Ψ (Corollary 3). ■

5 | PRACTICAL REALIZATION OF THE EMPIRICAL REFERENCE PRIOR

In this section we demonstrate how the empirical reference prior can be computed in the asymptotically normal case (i.e., under Condition 1). We suggest two algorithms, one of which is a straightforward optimization of the strictly concave functional in (16) and the other being a fixed point iteration, the convergence of which, however, is not yet fully understood.

For simplicity, we treat only the case where $\Theta = [a, b]$ is an interval and is discretized by the equidistant grid G_K defined below. Integrals over Θ will be approximated by the midpoint rule

$$Q(f) := \frac{b-a}{K} \sum_{k=1}^{K} f(\theta_k), \quad G_K = \left\{ \theta_k := \frac{2k-1}{2K} | k = 1, \ldots, K \right\}.$$

Furthermore, we assume that the Jeffreys prior $J(\theta)$ can be evaluated pointwise in each θ_k, either because its analytic form is available or by a numerical approximation of the integral in (14). Finally, we consider the computation of $\pi_{\text{erp}}(\theta) = \pi_{\text{erp}}(\theta \mid \mathcal{M}, \mathcal{P}, X, \gamma)$ only for a given smoothing parameter γ. The choice of γ requires another optimization as discussed in Section 4.2, this time in γ, where the algorithms below have to be executed many times for the computation of $\pi_{\text{erp}}(\theta \mid \mathcal{M}, \mathcal{P}, X_{-m}, \gamma)$ with varying data sets X_{-m}, see the likelihood cross-validation formula (13). While this "brute force" solution seems computationally challenging, the overall procedure could be performed in all our examples within just a few seconds (mainly due to the strict concavity of the optimization problem (16)). A more elegant solution which combines the two optimization problems is imaginable, but goes beyond the scope of this article. In the two algorithms below, we identify each prior π with its discretized version $\hat{\pi} = (\pi(\theta_1), \dots, \pi(\theta_K))$ (and similar for J, π_t, etc.).

Algorithm 1.

(I) Formulate the discretized functional

$$\hat{\Psi}(\pi) = \sum_{m=1}^{M} \log Q\left(p(x_m|\cdot)\ \pi\right) - \gamma\ Q\left(\pi\ (\log \pi - \log J)\right)$$

and its gradient in accordance with (17) and (18).

(II) Optimize $\Psi(\pi)$ as a function of π by an optimization algorithm of your choice (we use the method of moving asymptotes [MMA] of Svanberg (2001/2002) *from the NLopt package provided by Johnson* (n.d.)). *Renormalize $\hat{\pi}$ in each iteration step.*

Motivated by the fixed point characterization of π_{erp} in Theorem 5, it is tempting to implement the fixed point iteration $\pi_t = F^*(\pi_{t-1})$, $t \in \mathbb{N}$. However, our empirical studies showed that the resulting sequence $(\pi_t)_{t \in \mathrm{N}}$ often fails to converge, indicating that F^* is, in general, not a contraction. This suggests to decelerate the iteration by choosing $\pi_t = (1 - \tau_t)\ \pi_{t-1} + \tau_t\ F^*(\pi_{t-1})$ with step sizes $\tau_t \in [0, 1]$. While this gave satisfactory results in our examples, it is hard to characterize the step sizes for which the new sequence converges.

Algorithm 2. *Choose a threshold* $\varepsilon > 0$.

1. *Choose an arbitrary positive probability density π_0, for example, $\pi_0(\theta) = (b - a)^{-1}$.*
2. *For $t \in \mathbb{N}$, iterate:*
   (a) compute $\tilde{\pi}_{\text{temp}}(\theta) = J(\theta)\ \exp\left(\gamma^{-1} \sum_{m=1}^{M} \frac{p(x_m|\theta)}{Q(p(x_m|\cdot)\ \pi_{t-1})}\right)$;
   (b) normalize: $\pi_{\text{temp}} = \tilde{\pi}_{\text{temp}} / Q(\tilde{\pi}_{\text{temp}})$;
   (c) choose a step size $\tau_t \in [0, 1]$ *and* $\pi_t = (1 - \tau_t)\ \pi_{t-1} + \tau_t\ \pi_{\text{temp}}$;
   until $Q(|\pi_{\text{temp}} - \pi_{t^*-1}|) < \epsilon$ *for some* $t^* \in \mathbb{N}$.
3. *Set* $\pi_{\text{erp}} = \pi_{t^*}$.

6 | NUMERICAL COMPUTATIONS

In the following, we apply the empirical reference prior approach to a synthetic data set as well as to the famous real-life data set of Efron and Morris (1975). Our code implements Algorithm 1 and is available in the form of a Julia package at https://github.com/axsk/ObjectiveEmpiricalBayes.jl.

6.1 | Example with synthetic data

We illustrate the invariance under reparametrization using the location model

$$x|\theta \sim \mathcal{N}(\theta, \sigma^2), \quad \theta \sim \pi_{\text{true}}, \tag{20}$$

with $\sigma = 0.3$ and π_{true} being an equal mixture of $\mathcal{N}(1, 0.5^2)$ and $\mathcal{N}(3, 0.5^2)$, truncated to the interval $\Theta = [0, 4]$. We compare the performance of MPLE (5) using Tikhonov regularization $\Phi = \|\cdot\|_{L^2}^2$, also known as ridge regression, and the empirical reference prior (16), applied in both the untransformed space $\Theta = [0, 4]$ and the space $\tilde{\Theta}$ transformed by $\varphi : \theta \mapsto \tilde{\theta} = \exp(\theta)$. According to (8), the transformed model takes the form

$$x|\tilde{\theta} \sim \mathcal{N}(\log \tilde{\theta}, \sigma^2), \quad \tilde{\theta} \sim \varphi_* \pi_{\text{true}}. \tag{21}$$

Both optimization problems (5) and (16) are solved by Algorithm 1 using the method of moving asymptotes (MMA) of Svanberg (2001/2002) from the NLopt package provided by Johnson (n.d.), after discretizing the parameter space with 200 equidistant grid points. All priors are estimated using the same data consisting of $M = 100$ synthetic measurements. In both cases, the Jeffreys prior can be computed analytically. We obtain a constant Jeffreys prior for the likelihood model in (20) and $J(\tilde{\theta}) \propto |\tilde{\theta}|^{-1}$ for the one in (21).

As expected from the theory in Section 4, we observe how the lack of invariance of conventional penalty terms is resolved by the missing information penalty $\Phi_{\mathcal{I}}$, without losing the effect of regularization, see Figures 4 and 5. In addition, the empirical reference prior is strictly positive on the whole interval, making $\Phi_{\mathcal{I}}$ preferable to Tikhonov regularization from an objective Bayesian viewpoint: Excluding certain parameter values completely from a finite number of measurements appears unreasonable.

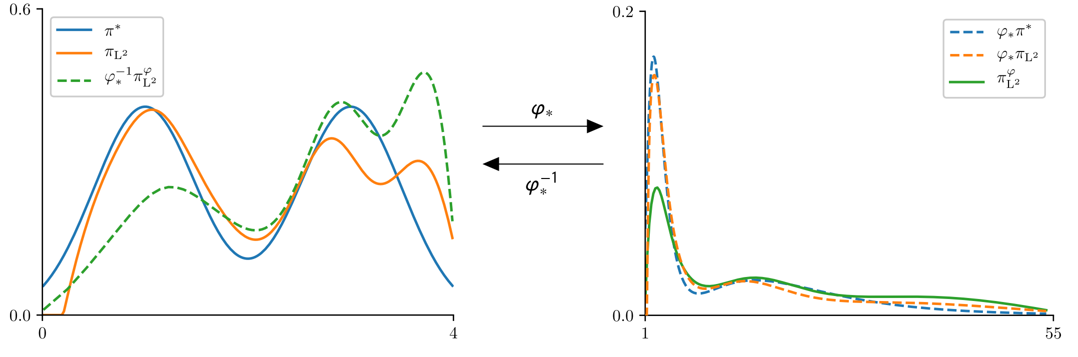

FIGURE 4 Conventional penalty terms, here the Tikhonov-regularization, are variant under transformations $\varphi : \Theta \to \tilde{\Theta}$, resulting in inconsistent prior estimates (see also Figure 2). If the estimation is performed in a transformed space $\tilde{\Theta}$ it gives a different estimate than the pushforward of the estimate in Θ, $\pi_{\text{est}}^{\varphi} \neq \varphi_* \pi_{\text{est}}$. Dashed lines indicate transformed densities. Here, π_{L^2} denotes the MPLE using Tikhonov regularization as penalty. In order to demonstrate the lack of invariance of the density estimate, we chose the same smoothing parameter γ in both spaces [Color figure can be viewed at wileyonlinelibrary.com]

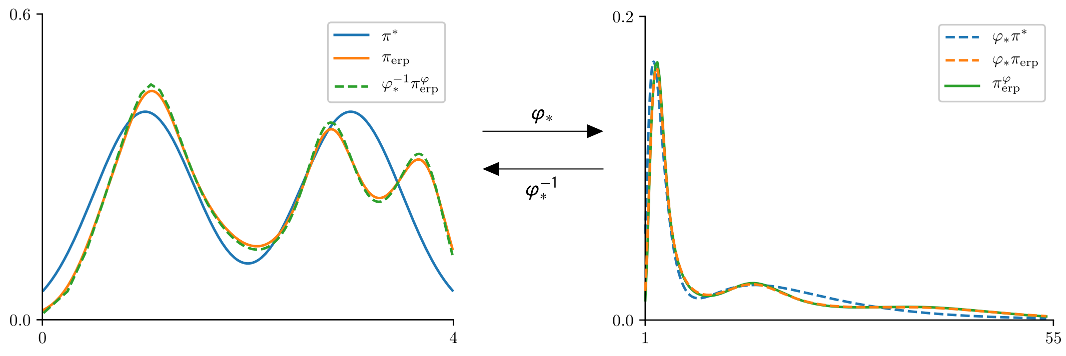

FIGURE 5 Due to the transformation invariance of the missing information penalty $\Phi_{\mathcal{I}}$, the empirical reference prior estimates are consistent under reparametrization (up to a negligible numerical error), $\pi_{\mathrm{erp}}^{\varphi} = \varphi_*\pi_{\mathrm{erp}}$ [Color figure can be viewed at wileyonlinelibrary.com]

6.2 | Example with real-life data

To illustrate our approach on real-life data, let us consider the historical batting averages (or baseball) example of Efron and Morris (1975), which is one of the most famous small data sets in statistics. The batting averages x_m (number of successful hits divided by the number of tries) for $M = 18$ major league baseball players early in the 1970 season—to be precise, the first $N = 45$ at bats of each—are used to estimate their true success rates $\theta_m \in \Theta = [0, 1]$, $m = 1, \dots, M$, which are taken as the averages over the remaining season (around 370 at bats for each player). This example was used by Efron and Morris (1975) to illustrate the strength of the James–Stein (JS) estimator $\hat{\theta}_{\mathrm{JS}}$, a particular parametric empirical Bayes method, compared with the MLE $\hat{\theta}_{\mathrm{MLE}} = x$ in terms of the (empirical) mean squared error

$$\mathrm{MSE} := \frac{1}{M}\sum_{m=1}^{M}(\hat{\theta}_m - \theta_m)^2.$$

Since the JS estimator assumes both the prior and the likelihood model to be normal, the natural binomial likelihood model $\mathcal{M}^{\mathrm{Bin}}$ given by

$$(Nx_m)|\theta_m \overset{\text{independent}}{\sim} \mathrm{Bin}(\cdot|\theta_m, N), \quad \theta_m \overset{\text{i.i.d.}}{\sim} \pi_{\mathrm{true}}, \quad m = 1, \dots, M, \tag{22}$$

is replaced by its normal approximation $\mathcal{M}^{\mathcal{N}}$,

$$x_m|\theta_m \overset{\text{independent}}{\sim} \mathcal{N}(\theta_m, \sigma^2), \quad \theta_m \overset{\text{i.i.d.}}{\sim} \pi_{\mathrm{true}}, \quad m = 1, \dots, M, \tag{23}$$

where $\sigma^2 := \bar{x}(1 - \bar{x})/N$ is chosen to approximate the variance estimate of the binomial distribution and $\bar{x} := M^{-1}\sum_{m=1}^{M} x_m$. Note that in the original paper by Efron and Morris (1975) an additional arcsin transformation is applied to each of the measurements x_m in order to stabilize their variance. For simplicity and following the presentation in Efron (2010, section 1.2), we omit this technical detail. We will apply our empirical reference prior approach to both, the original formulation (23) using $\mathcal{M}^{\mathcal{N}}$ as well as the more meaningful likelihood model $\mathcal{M}^{\mathrm{Bin}}$ in (22). In both cases, the Jeffreys prior can be computed analytically: We obtain a constant Jeffreys prior for the likelihood model $\mathcal{M}^{\mathcal{N}}$ in (23) and $J(\theta) \propto |\theta(1 - \theta)|^{-1/2}$ for $\mathcal{M}^{\mathrm{Bin}}$ in (22).

We compare these two approaches to the JS estimator which assumes a normal prior $\pi_{\text{true}} = \mathcal{N}(\mu_\theta, \sigma_\theta^2)$, making it a *parametric* empirical Bayes method with only the parameters μ_θ and σ_θ^2 of the prior to be estimated. A wonderful derivation of the JS estimator and its application to the baseball data set can be found in Efron (2010, section 1.2), with the resulting estimates

$$\hat{\mu}_\theta = \overline{x} := \frac{1}{M}\sum_{m=1}^{M} x_m, \quad \hat{\sigma}_\theta^2 = \frac{B\sigma^2}{1-B}, \quad \hat{\theta}_{\text{JS},m} = \overline{x} + B(x_m - \overline{x}),$$

where, for $m=1, \ldots, M$, the estimates $\hat{\theta}_{\text{JS},m}$ approximate θ_m by the individual posterior means and

$$B := 1 - \frac{(M-3)\sigma^2}{\sum_{m=1}^{M}(x_m - \overline{x})^2}.$$

TABLE 1 Numerical comparison of four different estimates of the batting success rate θ of 18 major league baseball players based on their batting average x of the first 45 at bats early in the 1970 season (x = hits/number of at bats)

Name	$\hat{\theta}_{\text{MLE}} = x$	$\hat{\theta}_{\text{JS}}$	$\hat{\theta}_{\text{erp}}^{\mathcal{N}}$	$\hat{\theta}_{\text{erp}}^{\text{Bin}}$	θ
Clemente	0.400 = 18/45	0.294	0.303	0.298	0.346 = 127/367
Robinson	0.378 = 17/45	0.289	0.296	0.292	0.298 = 127/426
Howard	0.356 = 16/45	0.285	0.289	0.287	0.276 = 144/521
Johnstone	0.333 = 15/45	0.280	0.282	0.281	0.222 = 61/275
Berry	0.311 = 14/45	0.275	0.276	0.275	0.273 = 114/418
Spencer	0.311 = 14/45	0.275	0.276	0.275	0.270 = 126/466
Kessinger	0.289 = 13/45	0.270	0.269	0.270	0.265 = 155/586
Alvarado	0.267 = 12/45	0.266	0.264	0.265	0.210 = 29/138
Santo	0.244 = 11/45	0.261	0.258	0.260	0.269 = 137/510
Swaboda	0.244 = 11/45	0.261	0.258	0.260	0.230 = 46/200
Petrocelli	0.222 = 10/45	0.256	0.254	0.255	0.264 = 142/538
Rodriguez	0.222 = 10/45	0.256	0.254	0.255	0.226 = 42/186
Scott	0.222 = 10/45	0.256	0.254	0.255	0.303 = 132/435
Unser	0.222 = 10/45	0.256	0.254	0.255	0.264 = 73/277
Williams	0.222 = 10/45	0.256	0.254	0.255	0.330 = 195/591
Campaneris	0.200 = 9/45	0.252	0.249	0.251	0.285 = 159/558
Munson	0.178 = 8/45	0.247	0.245	0.247	0.316 = 129/408
Alvis	0.156 = 7/45	0.242	0.242	0.244	0.200 = 14/70
$\frac{\text{MSE}}{\text{MSE}(\hat{\theta}_{\text{MLE}})}$	1	0.312	0.312	0.310	(0)

Note: As illustrated in Figure 6, the three empirical Bayes methods perform very similarly, both in terms of the individual estimates $\hat{\theta}_{\text{JS}} \approx \hat{\theta}_{\text{erp}}^{\mathcal{N}} \approx \hat{\theta}_{\text{erp}}^{\text{Bin}}$ and in terms of the overall mean squared error (MSE), and strongly outperform the "overconfident" maximum likelihood estimate $\hat{\theta}_{\text{MLE}} = x$. The ground truth for θ is taken as the batting average of each player over the remaining season.

The results are given in Table 1, while Figure 6 illustrates the data as well as the various point estimates, together with the three estimated priors

$$\pi_{\mathrm{JS}} = \mathcal{N}(\hat{\mu}_\theta, \hat{\sigma}_\theta^2), \quad \pi_{\mathrm{erp}}^{\mathcal{N}} = \pi_{\mathrm{erp}}(\cdot \mid \mathcal{M}^{\mathcal{N}}), \quad \pi_{\mathrm{erp}}^{\mathrm{Bin}} = \pi_{\mathrm{erp}}(\cdot \mid \mathcal{M}^{\mathrm{Bin}})$$

(note that those are *not* density estimates in the classical sense, since the parameters θ_m, and not the measurements x_m, are samples from π_{true}).

All three approaches provide very similar results. The main reason for this is that, as is apparent from Figure 6, the assumption of a normal prior made by the James–Stein approach seems to be a meaningful parametric choice in this specific example, where the parameter θ represents the performance (in terms of batting success rates) of baseball players. If this was not the case, for example, if another parametrization was chosen by, say, measuring the performance on a logarithmic scale, our nonparametric approach is likely to outperform the James–Stein estimator. In fact, our approach would provide *exactly* the same results for any reparametrization of the model. To the best of our knowledge, this is a unique feature among all empirical Bayes methods.

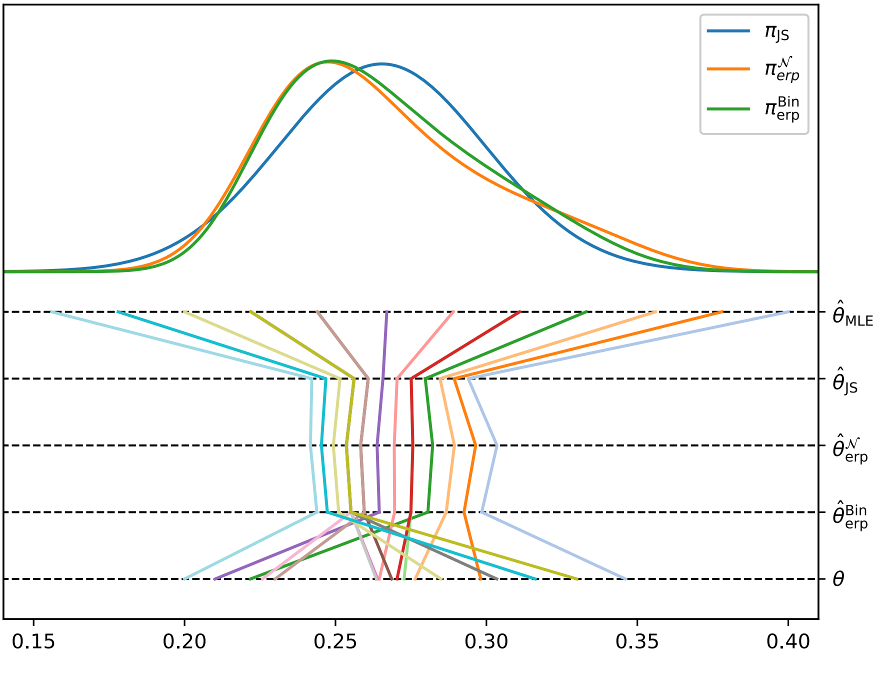

FIGURE 6 The empirical reference priors $\pi_{\mathrm{erp}}^{\mathcal{N}}$ and $\pi_{\mathrm{erp}}^{\mathrm{Bin}}$ and the corresponding posterior mean estimates $\hat{\theta}_m^{\mathrm{erp}}$ for the likelihood models given by (22) and (23) compared with the maximum likelihood estimates $\hat{\theta}_m^{\mathrm{MLE}}$ and James–Stein estimates $\hat{\theta}_m^{\mathrm{JS}}$ (and prior π_{JS}). While the results of the empirical reference prior approach and the JS approach are comparable, the normal prior π_{JS} clearly fails to be invariant under reparametrizations. Hence, the quality of the JS estimates could change drastically if, say, the players' performance was measured on a logarithmic scale. The "shrinkage toward the mean" effect of empirical Bayes estimates is well observable. The first impulse, that a smaller shrinkage would provide better results, is deceptive—this would reduce the (squared) error for some players, but drastically increase the (squared) error for others, a modification that is likely to increase the overall MSE [Color figure can be viewed at wileyonlinelibrary.com]

7 | THE MULTIPARAMETER CASE $d>1$

In the case of several parameters, the reference prior π_{ref} is no longer defined as the prior which maximizes the missing information, but by the sequential scheme presented in Berger and Bernardo (1992). This scheme applies the procedure described in Section 4 successively to the conditional priors $\pi(\theta_\delta|\theta_1, \dots, \theta_{\delta-1})$, $\delta = 1, \dots, d$, after a convenient ordering of the parameters. This leads us to three possible generalizations of empirical reference priors to the multiparameter case. The construction from Section 4, in particular the definition of the empirical reference prior, will be referred to as the *one-parameter construction*.

(A) Adopt the definitions from Section 4 exactly as they are. In the asymptotically normal case given by Condition 1, this corresponds to the optimization problem (16),

$$\pi_{\text{erp}} = \arg\max_\pi \ \log L(\pi) - \gamma D_{\text{KL}}(\pi \| J),$$

where $J(\theta) = \sqrt{|\det i(\theta)|}$ denotes the (arbitrarily scaled) Jeffreys prior. This construction provides an extension of the Jeffreys prior, *not* of reference priors. The reasons why reference priors are favored over Jeffreys prior in dimension $d>1$ are marginalization paradoxes and inconsistencies of the latter, see Bernardo (2005) and references therein. It is yet unclear in how far these arguments are valid in the presence of data. Hence, this approach might be justified in the empirical Bayes framework.

(B) In light of (16) and with the intention of generalizing reference priors, replace J by π_{ref} in the penalty term:

$$\pi_{\text{erp}} = \arg\max_\pi \ \log L(\pi) - \gamma D_{\text{KL}}(\pi \| \pi_{\text{ref}}).$$

This yields an extension of reference priors and agrees with the one-parameter construction in the asymptotically normal case (Condition 1), but not necessarily in the general case.

(C) Define a sequential scheme similar to the one used for reference priors. For simplicity, we will restrict the presentation to the case of $d=2$ parameters θ_1, θ_2, $\Theta = \Theta_1 \times \Theta_2$, but the construction can easily be extended to any number of parameters. As is common practice for reference priors, the parameters have to be ordered by "inferential importance". We will perform similar steps to the ones in Bernardo (2005, section 3.8). Note that, as in the case of reference priors, this scheme lacks objectivity since it requires an ordering of the parameters, which is a heuristic element and not unambiguous in many applications.

Algorithm 3.

(i) For every (fixed) θ_1, the one-parameter algorithm yields the conditional empirical reference prior $\pi_{\text{erp}}(\theta_2|\theta_1) = \pi_{\text{erp}}(\theta_2|\theta_1, \mathcal{M}, \mathcal{P}, X)$.

(ii) By integrating out parameter θ_2 we obtain the one-parameter model $\mathcal{M}_1$ given by

$$p(x|\theta_1) = \int_{\Theta_2} p(x|\theta_1, \theta_2)\ \pi_{\text{erp}}(\theta_2|\theta_1)\ \mathrm{d}\theta_2.$$

Apply the one-parameter construction to $\mathcal{M}_1$ to obtain the marginal empirical reference prior $\pi_{\text{erp}}(\theta_1) = \pi_{\text{erp}}(\theta_1|\mathcal{M}, \mathcal{P}, X)$.

(iii) The desired empirical reference prior is defined by $\pi_{\text{erp}}(\theta_1, \theta_2) = \pi_{\text{erp}}(\theta_1)\pi_{\text{erp}}(\theta_2|\theta_1)$.

(A) and (B) are straightforward generalizations of the theory presented in Section 4. It would be interesting to analyze the connection between the approaches (B) and (C).

Remark 2. Note that while all three of the above options are invariant under transformations of the measurement x, only (A) truly guarantees invariance under transformations of the parameter $\theta = (\theta_1, \dots, \theta_d)$, which it inherits from the invariance of the log-likelihood (up to an additive constant, Theorem 1) and of the Jeffreys prior. The other two options (B) and (C) depend on the particular choice of how the parameters (or, rather, the parameter components) $\theta_1, \dots, \theta_d$ are ordered. Hence invariance is only given under reparametrizations that transform each component separately,

$$\varphi(\theta) = (\varphi_1(\theta_1), \dots, \varphi_d(\theta_d)),$$

under the assumptions that their ordering is not altered.

8 CONCLUSION

We successfully applied the approach for the construction of reference priors to determine a transformation invariant penalty term for MPLE, which favors noninformativity of the prior, namely the missing information $\mathcal{I}[\pi \mid \mathcal{M}^k]$, $k \to \infty$. This interaction of objective Bayesian analysis and empirical Bayes methods results in a consistent and informative prior estimate, which we termed the empirical reference prior π_{erp}. The distinctive feature of π_{erp} is its invariance under reparametrization, which is, to the best of our knowledge, unique among all empirical Bayes methodologies.

The smoothing parameter γ tunes the amount of information contained in the prior: The data, represented by the marginal likelihood $L(\pi)$, yields information about the distribution of θ, but maximizing $L(\pi)$ alone overfits the data. The penalty term $\Phi(\pi)$, on the other hand, favors noninformative priors. We performed this trade-off by likelihood cross-validation which we also proved to be invariant under transformations (Proposition 1).

Besides invariance, our method has further favorable properties such as compatibility with sufficient statistics and, under the assumption of asymptotic normality, strict concavity of the resulting optimization problem (16). So far, our approach lacks an explicit formula for the empirical reference prior.[3]

We applied our methodology to a synthetic data set to illustrate the invariance property of the empirical reference prior as well as to a real-life example, the famous baseball data set of Efron and Morris (1975) collected to demonstrate the advantages of the James–Stein estimator, a parametric empirical Bayes method which we compare our method to. In the latter case, both approaches gave nearly identical results, the most natural explanation being that the parametric choice of the prior used in the James–Stein approach appears to be a good approximation in this specific example with this specific parametrization. Our nonparametric empirical reference prior approach is likely to provide better results in situations where the "true" prior π_{true} cannot be well-approximated by a Gaussian. Our code for the numerical computations is available at https://github.com/axsk/ObjectiveEmpiricalBayes.jl.

[3]An explicit formula for reference priors (Berger et al., 2009, theorem 7) exists, so far, only in the one-parameter case $d = 1$ and under rather restrictive assumptions. Furthermore, it requires the numerical approximations of integrals in the possibly high-dimensional measurement space $\mathcal{X}$ as well as the computation or at least some estimate of the limit $k \to \infty$.

The generalization of our approach to several dimensions is not unambiguous and has been discussed in Section 7.

ACKNOWLEDGEMENT

The authors thank Ingmar Schuster for countless enlightening discussions on the topic.

ORCID

Ilja Klebanov https://orcid.org/0000-0002-1911-4281

REFERENCES

Berger, J. O., & Bernardo, J. M. (1992). On the development of the reference prior method. *Bayesian Statistics*, *4*, 35–60.

Berger, J. O., Bernardo, J. M., & Sun, D. (2009). The formal definition of reference priors. *The Annals of Statistics*, *37*(2), 905–938. https://doi.org/10.1214/07-AOS587

Bernardo, J. M. (1979. Retrieved from). Reference posterior distributions for Bayesian inference. *Journal of the Royal Statistical Society: Series B (Methodological)*, *41*(2), 113–147. https://links.jstor.org/sici?sici=0035-9246(1979)41:2<113:RPDFBI>2.0.CO;2-P&origin=MSN

Bernardo, J. M. (2005). *Reference analysis*. In *Bayesian thinking: Modeling and computation, volume 25 of handbook of statistics* (pp. 17–90). Amsterdam, Netherlands: Elsevier/North-Holland.

Carlin, B. P., & Louis, T. A. (1996). *Bayes and empirical Bayes methods for data analysis, volume 69 of monographs on statistics and applied probability*. London, UK: Chapman & Hall.

Casella, G. (1985). An introduction to empirical Bayes data analysis. *The American Statistician*, *39*(2), 83–87. https://doi.org/10.2307/2682801

Clarke, B. S., & Barron, A. R. (1994). Jeffreys' prior is asymptotically least favorable under entropy risk. *Journal of Statistical planning and Inference*, *41*(1), 37–60. https://doi.org/10.1016/0378-3758(94)90153-8

Cover, T. M., & Thomas, J. A. (2006). *Elements of information theory* (2nd ed.). Hoboken, NJ: Wiley-Interscience John Wiley & Sons.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977. Retrieved from). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, *39*(1), 1–38. https://links.jstor.org/sici?sici=0035-9246(1977)39:1<1:MLFIDV>2.0.CO;2-Z\protect\LY1\textbraceleft& \protect\LY1\textbracerightorigin=MSN

Efron, B. (2010). *Large-scale inference: Empirical Bayes methods for estimation, testing, and prediction, volume 1 of institute of mathematical statistics (IMS) monographs*. Cambridge, MA: Cambridge University Press. https://doi.org/10.1017/CBO9780511761362

Efron, B., & Morris, C. (1975. Retrieved from). Data analysis using Stein's estimator and its generalizations. *Journal of the American Statistical Association*, *70*(350), 311–319. http://www.jstor.org/stable/2285814

Good, I. J. (1963). Maximum entropy for hypothesis formulation, especially for multidimensional contingency tables. *The Annals of Mathematical Statistics*, *34*, 911–934. https://doi.org/10.1214/aoms/1177704014

Good, I. J., & Gaskins, R. A. (1971). Nonparametric roughness penalties for probability densities. *Biometrika*, *58*, 255–277. https://doi.org/10.2307/2334515

Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proc Roy Soc Lond Ser A*, *186*, 453–461. https://doi.org/10.1098/rspa.1946.0056

Jeffreys, H. (1961). *Theory of probability* (3rd ed.). Oxford, UK: Clarendon Press.

Johnson S. G. *The NLopt nonlinear-optimization package*. Retrieved from http://ab-initio.mit.edu/nlopt

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical statistics*, *22*, 79–86. https://doi.org/10.1214/aoms/1177729694

Laird, N. (1978). Nonparametric maximum likelihood estimation of a mixing distribution. *Journal of the American Statistical Association*, *73*(364), 805–811. https://doi.org/10.1080/01621459.1978.10480103

Lindsay, B. G. (1995). Mixture models: Theory, geometry and applications. *NSF-CBMS Regional Conference Series in Probability and Statistics*, *5*, 1–163. http://www.jstor.org/stable/4153184

Maritz, J. S., & Lwin, T. (1989). *Empirical Bayes methods, volume 35 of monographs on statistics and applied probability* (2nd ed.). London, UK: Chapman & Hall.

McLachlan, G. J., & Krishnan, T. (2008). *The EM algorithm and extensions. Wiley series in probability and statistics* (2nd ed.). Hoboken, NJ: Wiley-Interscience, John Wiley & Sons.

Neal, M. L., & Kerckhoffs, R. (2009). Current progress in patient-specific modeling. *Briefings in Bioinformatics*, *11*(1), 111–126. https://doi.org/10.1093/bib/bbp049

Robbins H. 1956. *An empirical Bayes approach to statistics*. Proceedings of the 3rd Berkeley Symposium on Mathematical Statistics and Probability, 1954–1955 (vol. I, pp. 157–163). Berkeley and Los Angeles: University of California Press.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*, 379–423–623–656. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Shore, J. E., & Johnson, R. W. (1980). Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy. *IEEE Transactions on Information Theory*, *26*(1), 26–37. https://doi.org/10.1109/TIT.1980.1056144

Silverman, B. W. (1982). On the estimation of a probability density function by the maximum penalized likelihood method. *The Annals of Statistics*, *10*(3), 795–810. https://doi.org/10.1214/aos/1176345872

Silverman, B. W. (1986). *Density estimation for statistics and data analysis. Monographs on statistics and applied probability*. London, UK: Chapman & Hall.

Svanberg, K. (2001/02). A class of globally convergent optimization methods based on conservative convex separable approximations. *SIAM Journal on Optimization*, *12*(2), 555–573. https://doi.org/10.1137/S1052623499362822

van der Vaart A. W. *Asymptotic statistics*, volume 3 of *Cambridge series in statistical and probabilistic mathematics*. Cambridge University Press, Cambridge, UK: 1998.

van Erven, T., & Harremoës, P. (2014). Rényi divergence and Kullback-Leibler divergence. *IEEE Transactions on Information Theory*, *60*(7), 3797–3820. https://doi.org/10.1109/TIT.2014.2320500

How to cite this article: Klebanov I, Sikorski A, Schütte C, Röblitz S. Objective priors in the empirical Bayes framework. *Scand J Statist*. 2020;1–22. https://doi.org/10.1111/sjos.12485