Bayes and Empirical Bayes Approaches with Different Data Structures

Agostino Ruta

Contents

Abstract

Empirical Bayes methods have played a significant role in Statistics over the past decades. Many techniques emerged around the central idea of a compromise between the Bayesian and frequentist approaches. This dual nature of Empirical Bayes estimation implies uncertainty from a theoretical point of view. As stated by Efron (2019), *"Empirical Bayes methods, though of increasing use, still suffer from an uncertain theoretical basis, enjoying neither the safe haven of Bayes theorem nor the steady support of frequentist optimality"*. The objective of this thesis is thus to explore the properties of various Empirical Bayes estimation techniques that have evolved. The central distinction that guides this work lies between classic Empirical Bayes (EB), which includes the G-modeling and F-modeling approaches, and the use of empirical Bayes ingredients in Bayesian learning, which will be referred to as "Empirical Bayes in Bayes" (EBIB). In this latter case, the data structure does not necessarily envisage large-scale parallel experiments, as in classic EB, and there is no true prior law. Although debatable, recent results prove that the EBIB posterior distribution may be a computationally convenient approximation of a genuine Bayesian posterior law. The original contribution of the thesis is to explore these results further and develop their use in sparse regression. An extensive simulation study is conducted to give a more concrete sense of higher-order asymptotic approximation properties of the EB posterior distribution and is used to perform shrinkage regression on both real and simulated datasets.

1 Introduction and review of the literature

1.1 Introduction

The thesis discusses the concepts and uses of Empirical Bayes (EB) procedures across different data structures. The origin of the EB approach to Statistics traces back to the 1950's and it is still the object of a rich and growing literature. However, as noted by Efron (2019), "*considering the enormous gains potentially available from empirical Bayes methods, the effects on statistical practice have been somewhat underwhelming*". Initially, this was due to a relative scarcity of appropriate data. Indeed, as we detail in the thesis, EB techniques require large numbers of parallel experiments to be effective; and this kind of large-scale parallel inference problems were not common until the 1990s. The situation changed dramatically with the introduction of microarrays and the related explosion of large-scale inferential challenges posed by the genome project, where EB methods were revealed to be extremely powerful. Modern scientific technology and the current richness of big data would call for their further expansion. However, their potentialities seem still underdeveloped. As noted again by Efron (2019), "*Empirical Bayes has suffered from a philosophical identity problem. Not firmly attached to either frequentism or Bayesianism, expositions of empirical Bayes typically hover uncertainly around the middle.*" In fact, a clean formulation of the "EB problem" can be given from the Bayesian approach (Deely and Lindley, 1981). However, part of the literature still hesitates to entirely take a Bayesian approach, so that EB is still regarded as something in between, that uses techniques from both the frequentist and the Bayesian frameworks, with a

theoretical basis that remains somehow confused.

The first aim of the thesis is to present recent developments in the literature on the classic EB procedures. These consist in the f-modeling and the g-modeling approaches and are apt in case of large-scale parallel experiments. There is however another common use of Empirical Bayes ideas. Here, the setting is in fact different. The data structure is not necessarily given by (large-scale) parallel experiments: the setting is purely Bayesian. In these cases, a so-called (Petrone, Rizelli, and Rousseau, 2024) Empirical Bayes in Bayes (EBIB) approach is apt. This consists of using a data-driven choice of the prior hyperparameters λ, in an empirical Bayes fashion, making it a somewhat controversial yet commonly used method. The common sense is that EBIB would lead to a "better" choice of the hyperparameters for finite sample size, and would anyway be close to a genuine Bayesian inference asymptotically. However, these common beliefs are kind of vague, and general theoretical properties of EBIB are elusive. Thus, the second and main aim of the thesis is to explore recent results that give a rigorous formulation and prove precise approximation properties. These unravel around the central idea that the EBIB posterior can be interpreted as an approximation of a genuine Bayesian posterior distribution; in fact, it may be a (computationally convenient) approximation of a fully Bayesian posterior distribution. While, in regular models, the latter is actually the case at first-order asymptotic (Petrone, Rousseau, and Scricciolo, 2014), a finer analysis reveals that, in fact, the EBIB posterior is a faster approximation of the Bayesian posterior law that uses the given class of priors π_λ, $\lambda \in \Lambda$, and expressed the most information on the true model's parameter θ (Petrone, Rizelli, and Rousseau,

2024).

An original contribution of the thesis is the application of these very recent results in Bayesian hierarchical LASSO and Bayesian hierarchical RIDGE models onto multiple simulated datasets. The aim is to explore these theoretical approximation properties and, particularly, to investigate to what extent the higher-order approximation properties remain perceptible for increasing sample size.

The structure of the thesis is as follows. This initial section provides an overview of the literature, focusing on the differences between possible approaches. Section 2 extensively explores the G-modeling and F-modeling approaches within the classic Empirical Bayes framework. Within this section, various fundamental estimators falling under this category will be discussed as examples. Section 3 explores the EBIB approach through multiple simulation studies that span the entire chapter.

1.2 Evolution of the classic Empirical Bayes literature

Despite an extensive literature on the Empirical Bayes approaches, it is somehow challenging to provide a unique definition. However, one common feature emerges: these approaches exploit observations that, from a Bayesian or frequentist point of view, may appear unrelated but that can improve the estimate for the case of interest. Therefore, this approach is often considered as a compromise between frequentist and Bayesian methodologies. In the 40s, some initial results started using the information in parallel experiments to model the prior distribution. One of these early examples is "On

the Correct Use of Bayes' Formula" (Von Mises, 1942). In this work, the prior distribution for the parameter of interest is termed as "*overall distribution*" since it considers "*the distribution of θ-values within the total mass of samples, not regarding what the values of x are in each case*". Another early example is "The Relation Between the Number of Species and the Number of Individuals in a Random Sample of an Animal Population" (Fisher, Corbet, and Williams, 1943). Those embryonal examples capture the main idea of the EB approaches yet are quite far from it.

The approach, and the same term "empirical Bayes", were introduced in "An Empirical Bayes Approach to Statistics" (Robbins, 1956). In this work, Robbins addresses the estimation problem of the model's parameter θ for a discrete random variable $X \mid \theta \sim p(x \mid \theta) = P(X = x \mid \tilde{\theta} = \theta)$, where $\tilde{\theta}$ is a random variable with a *prior* distribution function π. (The data X may be the sufficient statistic of a random sample). With quadratic loss, the optimal estimate of θ is the Bayesian estimate

$$E_G(\tilde{\theta} \mid x) = \frac{\int \theta \; p(x \mid \theta) d\pi(\theta)}{\int p(x \mid \theta) d\pi(\theta)}.$$

This estimate relies on the prior distribution π being known. Otherwise, the Bayesian estimate could not be computed. However, suppose one has k replicates of the experiment, each described by the same distribution $p(x \mid \theta)$ but with replication-specific parameters. This consists of Lindley's formulation for the "Empirical Bayes problem" where the data structure consists of the sequence $\{(X_i, \tilde{\theta}_i);\ i = 1, 2, ...\}$ of random variables. The $\tilde{\theta}_i$ are *iid* with common distribution function $G : \tilde{\theta}_i \overset{iid}{\sim} G(\theta)$. Then, for each parallel

experiment j, it is $X_j|\tilde{\theta}_j = \theta_j \sim p(x|\theta_k)$.

Suppose now one is interested in estimating the parameter in a single experiment, say θ_k of the k^{th} experiment but with the prior distribution being unknown. Within the experiment, the same framework previously described can be applied. This implies that $\theta_k \sim G(\theta)$ and that $X_k \mid \tilde{\theta}_k = \theta_k \sim p(x \mid \theta_k)$ so that the Bayesian estimate $E_G(\tilde{\theta}_k \mid x_k)$ depends on the on the unknown $G(\theta)$. Robbins' proposal is to use observations $x_1, \ldots, x_{k-1}$ from the similar experiments to *estimate* $G(\theta)$. More precisely, to estimate the Bayesian solution $E_G(\tilde{\theta}_k \mid x_k)$. Informally, this gives a Bayesian estimator where the prior distribution is obtained empirically from the data: an *empirical Bayes* estimator.

Remark. Robbins does not accept the subjective interpretation of probability that is proper of the Bayesian approach, which would imply that the prior distribution cannot be *unknown*. The prior expresses the researcher's information on θ, formalized through probability, and the difficulty may be to elicit such information, but there is no *physical* true unknown prior law. Rather, as we will discuss further in section 2, the matter is that different researchers may have different states of information, expressed through different priors.

Example 1.1. *Suppose one has the same data structure previously described; that is, k replicates of an experiment, each with replication-specific parameters and described by the same distribution* $X_k|\tilde{\theta}_k = \theta_k \sim Poisson(\theta_k)$. *The parameters* $\tilde{\theta}_1, ..., \tilde{\theta}_k$ *are i.i.d. samples from a common unknown latent distribution G so that* $\tilde{\theta}_j \overset{iid}{\sim} G(\theta)\ \forall j \in \{1, ..., k\}$. *The Bayesian estimate for* θ_k

is:

$$\begin{aligned} E[\theta_k|x_k] &= \int_\Theta \theta_k \, g(\theta_k \mid x_k) d\theta_k = \int_\Theta \theta_k \frac{p(x_k|\theta_k) g(\theta_k)}{p(x_k)} d\theta_k \\ &= \int_\Theta \theta_k \frac{1}{p(x_k)} \frac{\theta_k^{x_k} exp\{-\theta_k\}}{x_k!} g(\theta_k) d\theta_k = (x_k+1) \int_\Theta \frac{\frac{\theta_k^{x_k+1} exp\{-\theta_k\}}{(x_k+1)!} p(\theta_k)}{p(x_k)} d\theta_k \\ &= (x_k+1) \int_\Theta \frac{p(x_k+1|\theta_k) p(\theta_k)}{p(x_k)} d\theta_k = \frac{(x_k+1) p(x_k+1)}{p(x_k)} \end{aligned}$$

Here, the observations $x_1, ..., x_{k-1}$ of the other experiments can be used to estimate the Bayesian solution:

$$\hat{E}[\theta_k|x_k] = \int_\Theta \theta_k \, g(\theta_k|x_k) = \frac{(x_k+1)\hat{p}(x_k+1)}{\hat{p}(x_k)} \tag{1}$$

where the unknown terms $p(x_k) = P(X = x_k)$ and $p(x_k+1) = P(X = x_{k+1})$ are regarded as the non-parametric probabilities of observing respectively x_k and x_{k+1}. These are estimated by the relative frequencies across experiments, i.e. the ratio of successes over the total number of cases.

The historical relevance of this example is due to two intertwined reasons. First, it proposes a Bayesian procedure that does not require the specification of the prior. Second, the estimation of the parameter of interest is now a function of data from other parallel and, in principle, independent, experiments. This is because to estimate the marginal probabilities, the observations from parallel experiments become necessary. Therefore, this approach deviates from the classic Bayesian one since the prior distribution is not determined subjectively based on prior information but empirically through the information coming from the parallel cases.

Another disruptive result for the statistical history is the James-Stein estimator (James and Stein, 1960). This estimator was initially developed as a means to further reduce the MSE compared to the maximum likelihood estimator. It was only later in "Combining Possibly Related Estimation Problems" (Efron and Morris, 1973), that this estimator was interpreted as an Empirical Bayes one. More in detail, the setting of this case is similar to the previous one. We have replicates of the experiment, generating a sequence $\{(x_i, \theta_i);\ i = 1, 2, ..., n\}$ where $\theta_1, ..., \theta_n$ are a non-observable random sample from a distribution $G(\theta)$. Robbins initially introduced the *nonparametric* Empirical Bayes, where no parametric form is assumed for G. In this case, instead, Efron and Morris assume a parametric form for the latent distribution G, namely $g(\theta) = N(0; A)$. The non-observable θ_i sampled from the latent distribution will then constitute the parameters for the observable X_i, so that $X_i|\theta_i \sim N(\theta_i; \sigma^2)$. If A and σ^2 were known, the posterior distribution of θ_i given x_i would be:

$$g(\theta_i|x_i) = N\left(\frac{\sigma^2}{A+\sigma^2}0 + \frac{A}{A+\sigma^2}x_i; \frac{A\sigma^2}{A+\sigma^2}\right) = N\left(x_i\left(1 - \frac{\sigma^2}{A+\sigma^2}\right); \frac{A\sigma^2}{A+\sigma^2}\right).$$

The Bayesian estimate $\hat{\theta}^{BAYES} = x_i\left(1 - \frac{\sigma^2}{A+\sigma^2}\right)$ obtained is a shrinkage of the ML estimate $\hat{\theta}^{MLE} = x_i$. This is a direct consequence of having the prior distribution centered at 0. The shrinkage term $\left(1 - \frac{\sigma^2}{A+\sigma^2}\right)$ is a function of the variances, or better the uncertainties, around the data and the prior distribution. Indeed, as the variance of the data σ^2 goes to 0, the data observed is more and more reliable. As a consequence, the Bayesian estimate will suggest an estimate with a lower shrinkage. Conversely, as the variance

of the data increases with respect to the variance of the prior, the shrinkage will be bigger as the observation is less reliable. Up to now, this estimator relies upon σ^2 and A being known. If this were not the case, and A were unknown, it would be possible to replace the term $\frac{\sigma^2}{A+\sigma}$ with its unbiased estimate $\frac{\sigma^2(N-2)}{S}$ with $S = \sum_{i=1}^{k}(x_i - \overline{x})^2$. The resulting estimator for the parameter θ_i will be:

$$\hat{\theta}_i^{JS} = \left(1 - \frac{\sigma^2(N-2)}{S}\right) x_i.$$

The estimator obtained is an empirical Bayes one since the estimate for θ_i depends on the other unrelated observations through S.

Many other examples paved the way for the development of the Empirical Bayes approach. For example, "The probability of an unobserved outcome of an experiment" (Robbins, 1968). Some more recent examples are cited by Casella (1985) in "An Introduction to Empirical Bayes Data Analysis" such as "Empirical Bayes Methods Applied to Estimating Fire Alarm Probabilities" by Carter and Rolph (1974) or "Estimation in Parallel Randomized Experiments" (Rubin, 1981).

1.3 Categories of Empirical Bayes approaches

The vast heterogeneity of the classic Empirical Bayes examples forces a distinction between the corresponding approaches. The most traditional is the one between parametric and non-parametric Empirical Bayes. The first case is the one in which the distribution for the latent parameters $\theta_1, ..., \theta_n$ exists but is left generic. An instance of the non-parametric Empirical Bayes

approach is Robbin's formula. Conversely, the James-Stein estimator is an example of parametric empirical Bayes since to the latent distribution it is assigned a parametric structure.

Another more recent distinction distinguishes the g-modeling from the f-modeling approach (Efron, 2014). In the first case, the modeling happens through the x-scale. One example of the f-modeling approach is Robbin's formula. There, it is not necessary to even specify a latent distribution for the parameters $\theta_1, ..., \theta_n$ to obtain the empirical Bayes estimation of the Bayesian estimate $E(\theta_i \mid x_i)$. In the case of g-modeling, the specification of a parametric form for the latent distribution is fundamental. This is because the parallel observations are used to estimate the hyperparameters of the latent distribution λ. This will in turn be used to provide a better prior specification $g(\theta|\hat{\lambda})$. In other words, the modeling happens through the θ-scale. The distinction followed in this thesis is the one between f-modeling and g-modeling.

1.4 Classic Empirical Bayes and Empirical Bayes in Bayes

Regardless of the distinction used, the key fundamental feature that justifies a classic Empirical Bayes approach is the existence of a latent distribution for $\theta_1, ..., \theta_n$. From a Bayesian perspective, this is equivalent to exchangeability in the sequence of the latent parameters (Gelman et al., 2007; Ghosh, Delampady, and Samanta, 2007) or comparability between the observations. In other words, the apparently unrelated parallel experiments carry useful

information for improved inference in a single experiment. Tukey indeed refers to the Empirical Bayes approach with the evocative term "borrowing strength procedure". Similarly, Efron refers to this approach as "learning from the experience of the others" (Brillinger, 2002).

Hence, a classic empirical Bayes approach is justified whenever the realizations belong to the same family. However, identifying a statistical family is far from straightforward. Already Miller in "Simultaneous Statistical Inference" warned us by saying that the choice of what constitutes a family *"takes leave of mathematics and must be guided by subjective judgement"* (Miller, 1981). Similarly, Barnard commented on the James-Stein estimator *"The difficulty here is to know what problems are to be combined together why should not all our estimation problems be lumped together into one grand Melee?"* (Efron and Morris, 1973).

If the observations do not truly belong to the same family, there are immediate consequences for the validity of the approach. This is because there would be no more a latent distribution among the parameters but, at most, a prior distribution. Those two objects are fundamentally different. Indeed, a latent distribution is objective and, in that case, each $\theta_1, ..., \theta_n$ is an extraction from it. Conversely, the Bayesian prior is subjective and is used to model the uncertainty around θ. Therefore, in this case, θ is unknown but fixed: it is not a random variable even though, to it, it is assigned a probability distribution. Using empirical Bayes procedures in the latter case is different from the classic EB approach of Robbins and Efron. Thus, to avoid confusion, it will be referred to as Empirical Bayes in Bayes (EBIB) (Petrone, Rizelli, and Rousseau, 2024). In this case, the framework is entirely

Bayesian: there are no parallel experiments, thus no true latent distribution. The distribution assigned to θ will be a (subjective) prior distribution and no longer a latent (objective) one. Therefore, the EBIB approach reduces to an empirical estimation of the prior hyperparameter λ. This clashes with the Bayes theorem and yet it is a common practice that carries some optimality properties.

2 Classic Empirical Bayes Methods

2.1 G-modeling approach

The G-modeling represents the most common type of Empirical Bayes procedure. It can be applied in case of repeated sampling from the sequence $\{((X_{1,i}, ..., X_{n,i}), \theta_i);\ i = 1, 2, ..., k\}$ where $\theta_1, ..., \theta_k$ are unobservable samples from the latent distribution $g(\theta|\lambda)$. The unobservable sample will be the parameter for the observable observations so that $x_{1,i}, ..., x_{n,i}|\theta_i \sim f(\underline{x}_i|\theta_i)$. The core concept behind the g-modeling is that the conditional distributions of the observations $\underline{x}_1, ..., \underline{x}_k$ belong to the same family and share common traits captured by the hyperparameter λ of the common latent distribution of the θ_i. For example, λ may capture the overall success of medical treatment while $\theta_1, ..., \theta_n$ capture the success in single hospitals. Here emerges the main difference with the f-modeling approach: specifying a form of the latent distribution G is necessary because it will be used as if it were a prior distribution. This allows the exploitation of information from other observations as if they were prior knowledge of the phenomenon. For example, if λ were known, the

posterior distribution of θ_k would be $g(\theta_k|\underline{x}_k, \lambda) \propto f(\underline{x}_k|\theta_k)g(\theta_k|\lambda)$. However, λ is unknown and will be estimated using the observations from the other parallel experiments $x_{1,1}, ..., x_{n,k-1}$. The first possible way to estimate $\hat{\lambda}$ is through Marginal Maximum Likelihood (MML). The EB approach obtained this way is termed Direct Empirical Bayes. In this case, estimating $\hat{\lambda}$ will depend on integrating out the unobserved $\theta_1, ..., \theta_{k-1}$:

$$\begin{aligned}
&\hat{\lambda}_{MMLE} = \arg\max_{\lambda} f(x_{1,1}, \ldots, x_{n,k-1}|\lambda) \\
= \ &\arg\max_{\lambda} \int_{\Theta_{1:k-1}} f(x_{1,1}, \ldots, x_{n,k-1}|\theta_1, \ldots, \theta_{k-1}) g(\theta_1, \ldots, \theta_{k-1}|\lambda) d\Theta_{1:k-1} \\
= \ &\arg\max_{\lambda} \int_{\Theta_{1:k-1}} \prod_{i=1}^{k-1} \prod_{l=1}^{n} f(x_{il}|\theta_i) g(\theta_i|\lambda) d\Theta_{1:k-1}.
\end{aligned}$$

The estimate $\hat{\lambda}$ obtained will then be used as a hyperparameter for the prior distribution of θ_k that is, in turn, necessary to obtain the posterior $g(\theta_k|\underline{x}_1, ..., \underline{x}_{k-1}, \hat{\lambda}) \propto f(\underline{x}_k|\theta_k)g(\theta_k|\hat{\lambda})$.

2.2 F-modeling approach

The Empirical Bayes alternative to the g-modeling is f-modeling. Here, the modeling occurs within the x scale. That is, the prior distribution is determined directly through the marginal distribution of the parallel observations, bypassing a direct specification of a prior distribution. This implies that, compared to the g-modeling approach, the f-modeling does not require a parametric specification of a latent distribution but only its existence. This, in turn, significantly simplifies the computations. One may for example think about the computational simplicity of Robbin's formula (1). Still, this comes

at the cost that the result is just an estimate, or better, the empirical Bayes estimation of the Bayesian estimate. Conversely, the g-modeling approach would guarantee an entire posterior distribution and so, more information. Another difference highlighted in "Empirical Bayes estimation: When does g-modeling beat f-modeling in theory (and in practice)?" (Y. Shen and Wu, 2022) is that the f-modeling directly depends on parallel observations. It is therefore less robust to outliers and, in general, less performing in the case of heavy-tailed data. Finally, if on the one hand, the f-modeling requires simpler computations, on the other, it is not necessarily easy to interpret and generalize as the g-modeling alternative.

2.2.1 Extensions: continuous case

Up to now, the f-modeling examples treated in the introduction seem to be distribution-specific. It is for example not straightforward to extend Robbins' formula (1) to any other distribution outside the Poisson. The objective of this section is then the one to provide a unique estimator $\hat{E}[\theta_i|x_i]$ for the Bayesian estimate $E[\theta_i|x_i]$ within a parametric framework. To do so, I will assume that the density of X_i, conditionally on θ_i, belongs to the exponential family

$$f(x_i|\theta_i) = t(x_i)\exp\{b(\theta_i)R(x_i) + h(\theta_i)\}.$$

Two cases will be developed in continuous and discrete sample space. Here I consider the continuous case.

Result 1. *If the sample space $\mathcal{X}$ is continuous and $f(x_i|\theta_i) = t(x_i)\exp\{b(\theta_i)R(x_i)+$*

$h(\theta_i)\}$, *then*

$$E[b(\theta_i)|x_i] = -\frac{\frac{\delta}{\delta x_i} g(x_i)}{g(x_i)\frac{\delta}{\delta x_i} R(x_i)} + \frac{\frac{\delta}{\delta x_i}\log f(x_i)}{\frac{\delta}{\delta x_i} R(x_i)}.$$

Proof. It is that:

$$\begin{aligned}
\frac{\partial}{\partial x_i}\log f(x_i) &= \frac{\frac{\partial}{\partial x_i}\int f(x_i|\theta_i)g(\theta_i)d\theta_i}{f(x_i)} = \frac{\int \frac{\partial}{\partial x_i} t(x_i)\exp\{b(\theta_i)R(x_i)+h(\theta_i)\}}{f(x_i)} \\
&= \int \frac{\frac{\partial}{\partial x_i}[t'(x_i)+t(x_i)b(\theta_i)R'(x_i)]\exp\{b(\theta_i)R(x_i)+h(\theta_i)\}}{f(x_i)} d\theta_i \\
&= R'(x_i)\int b(\theta_i)g(\theta_i|x_i)d\theta_i + \frac{t'(x)}{t(x)}\int g(\theta_i|x_i)d\theta_i \\
&= R'(x_i)E[\theta_i|x_i] + \frac{t'(x)}{t(x)}.
\end{aligned}$$

This implies

$$E[b(\theta_i)|x_i] = -\frac{\frac{\delta}{\delta x_i} t(x_i)}{t(x_i)\frac{\delta}{\delta x_i} R(x_i)} + \frac{\frac{\delta}{\delta x_i}\log\{f(x_i)\}}{\frac{\delta}{\delta x_i} R(x_i)}.$$

□

Interpretation The equality obtained allows us to estimate $E[b(\theta_i)|x_i]$. Therefore, in the cases in which $b(\theta_i)$ is linear in θ_i it is immediate to derive the term $E[\theta_i|x_i]$. The elements $t(x_i)$ and $R(x_i)$ of the right-hand side are known because of the parametric structure assigned to $f(x_i|\theta_i)$. Conversely, the term $f(x_i)$ is unknown and needs to be estimated through the other observations $x_1, ..., x_n$. Therefore, the final result will be

$$\hat{E}[b(\theta_i)|x_i] = -\frac{\frac{\delta}{\delta x_i} t(x_i)}{t(x_i)\frac{\delta}{\delta x_i} R(x_i)} + \frac{\frac{\delta}{\delta x_i}\log\{\hat{f}(x_i)\}}{\frac{\delta}{\delta x_i} R(x_i)}.$$

Example - Tweedie's Formula Let $X_i \mid \theta \sim N(\theta, \sigma^2)\ \forall i = 1, ..., n$ with σ^2 known and the prior $g(\theta)$ be left generic. Then $t(x_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{x_i^2}{2\sigma^2}\}$, $R(x_i) = x_i$ and $b(\theta_i) = \frac{\theta_i}{\sigma^2}$. Therefore

$$E[\frac{\theta_i}{\sigma^2}|x_i] = \frac{\frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{x_i^2}{2\sigma^2}\}\{-\frac{2x_i}{2\sigma^2}\}}{\frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{x_i^2}{2\sigma^2}\}} + \frac{\delta}{\delta x_i} \log\{f(x_i)\}.$$

This implies:

$$E[\theta_i|x_i] = x_i + \sigma^2 \frac{\delta}{\delta x_i} \log\{f(x_i)\}.$$

The result obtained is Tweedie's formula. To compute an estimate of $E[\theta_i|x_i]$, it is necessary to fit $f(x_i)$ since

$$\hat{E}[\theta_i|x_i] = x_i + \sigma^2 \frac{\delta}{\delta x_i} \log\{\hat{f}(x_i)\}.$$

In other words, the function $f(x)$ is unknown but can be fitted using all the observations $x_1, ..., x_n$.

Example - James-Stein Estimator A prior distribution can simplify the previous example of non-parametric estimation of $f(x_i)$. For instance, let $f(x_i|\theta_i) = N(x_i; \theta_i, \sigma^2)$ and $g(\theta_i) = N(\theta_i; 0, A)$. Therefore $f(x_i) = N(x_i, 0, \sigma^2 + A) = N(0; V)$. This implies that

$$\frac{\delta}{\delta x_i} \log\{f(x_i)\} = \frac{\delta}{\delta x_i}\{\log\{\frac{1}{\sqrt{2\pi V}}\} - \frac{1}{2V}x_i^2\} = -V^{-1}x_i.$$

While, as before, $t(x_i) = \frac{1}{\sqrt{2\pi}} \exp\{-\frac{x_i^2}{2}\}$, $R(x_i) = x_i$ and $b(\theta_i) = \theta_i$. This implies

$$E[b(\theta_i)|x_i] = \frac{x_i}{\sigma^2} - V^{-1}x_i,$$

so that $E[\theta_i|x_i] = x_i - \sigma^2 V^{-1} x_i$. It is now possible to replace the unknown V^{-1} with its unbiased estimate $\frac{N-2}{S}$ with $S = \sum_{i=1}^{k}(x_i - \overline{x})^2$. In this case, one obtains $E[\hat{\theta_i}|x_i] = (1 - \sigma^2 \frac{N-2}{S})x_i$ that is the James-Stein estimator for θ_i.

2.2.2 Extensions: discrete case

If the sample space is discrete, the result will be different.

Result 2. *If the sample space $\mathcal{X}$ is discrete and $f(x|\theta_i) = t(x)exp\{b(\theta_i)R(x)+ h(\theta_i)\}$, then*

$$E[\exp\{b(\theta_i)[R(x_i+1) - R(x_i)]\}|x_i] = -\frac{t(x_i)}{t(x_i+1)} + \frac{f(x_i+1)}{f(x_i)}.$$

Proof. It is that

$$\begin{aligned}
f(x_{i+1}) &= \int f(x_{i+1}|\theta_i)g(\theta_i)d\theta_i \\
&= \int t(x_{i+1})\exp\{b(\theta_i)R(x_{i+1}) + h(\theta_i)\}g(\theta_i)d\theta_i \\
&= \int t(x_{i+1})\exp\{b(\theta_i)[R(x_{i+1}) - R(x_i)] + b(\theta_i)R(x_i) + h(\theta_i)\}g(\theta_i)d\theta_i \\
&= \frac{t(x_{i+1})}{t(x_i)}f(x_i)\int \exp\{b(\theta_i)[R(x_{i+1}) - R(x_i)]\}\frac{t(x_i)\exp\{b(\theta_i)R(x_i) + h(\theta_i)\}g(\theta_i)}{f(x_i)}d\theta_i \\
&= \frac{t(x_{i+1})}{t(x_i)}f(x_i)E[\exp\{b(\theta_i)[R(x_{i+1}) - R(x_i)]\}|x_i].
\end{aligned}$$

This implies that

$$E[\exp\{b(\theta_i)[R(x_i+1) - R(x_i)]\}|x_i] = -\frac{t(x_i)}{t(x_i+1)} + \frac{f(x_i+1)}{f(x_i)}.$$

□

Interpretation The equality obtained allows a simple estimation of the expectation of the posterior distribution of the term $\exp\{b(\theta_i)[R(x_i+1)-R(x_i)]\}$. Therefore, in the cases in which $\exp\{b(\theta_i)[R(x_i+1)-R(x_i)]\}$ is linear in θ_i, it is possible to derive the term $E[\theta_i|x_i]$. The elements $t(x_i)$ and $R(x_i)$ of the right-hand side are known because of the parametric structure assigned to $f(x_i|\theta_i)$. Conversely, the term $f(x_i)$ is unknown and must be estimated through the other observations $x_1, ..., x_n$. Therefore, the final estimate will be

$$\hat{E}[\exp\{b(\theta_i)[R(x_i+1)-R(x_i)]\}|x_i] = \frac{t(x_i)}{t(x_i+1)}\frac{\hat{f}(x_i+1)}{\hat{f}(x_i)}.$$

Without further assumptions, it is possible to estimate $f(x_i)$ and $f(x_i+1)$ as the ratio between the number of successes over the total cases. The next few examples will show how this generalizes some Empirical Bayes results.

Example - Robbins Formula Let $f(x_i|\theta_i) = Po(x_i;\theta_i)$ so that $f(x_i|\theta_i) = \frac{\theta_i^{x_i}\exp\{-\theta_i\}}{x_i!} = \frac{\exp\{x_i\log(\theta_i)-\theta_i\}}{x_i!}$. This implies $t(x_i) = \frac{1}{x_i!}$, $h(\theta_i) = -\theta_i$, $R(x_i) = x_i$ and $b(\theta_i) = \log(\theta_i)$. Since $R(x_i+1)-R(x_i) = 1$, it is that

$$E[\exp\{b(\theta_i)[R(x_i+1)-R(x_i)]\}|x_i] = E[\theta_i|x_i].$$

This will be equal to

$$\frac{t(x_i)}{t(x_i+1)}\frac{f(x_i+1)}{f(x_i)} = \frac{(x_i+1)!}{x_i!}\frac{f(x_i+1)}{f(x_i)} = (x_i+1)\frac{f(x_i+1)}{f(x_i)}.$$

In short, it was obtained

$$E[\theta_i|x_i] = (x_i + 1)\frac{f(x_i + 1)}{f(x_i)},$$

that is equal to Robbin's formula (1). The corresponding estimator will be

$$\hat{E}[\theta_i|x_i] = (x_i + 1)\frac{\hat{f}(x_i + 1)}{\hat{f}(x_i)}.$$

Example - Geometric Distribution Let $f(x_i|\theta_i) = Ge(x_i, \theta_i)$ so that $f(x_i|\theta_i) = \theta_i(1 - \theta_i)^{x_i} = \exp\{x_i \log(1 - \theta_i) + \log(\theta_i)\}$. This implies that $t(x_i) = 1$, $h(\theta_i) = \log(\theta_i)$ and $b(\theta_i) = \log(1 - \theta_i)$. Since $R(x_i) = x_i$, we have that $R(x_i + 1) - R(x_i) = 1$. Then

$$E[\exp\{b(\theta_i)[R(x_i + 1) - R(x_i)]\}|x_i] = E[1 - \theta_i|x_i].$$

This is equal to

$$\frac{t(x_i)}{t(x_i + 1)}\frac{f(x_i + 1)}{f(x_i)} = \frac{f(x_i + 1)}{f(x_i)}.$$

In short,

$$E[1 - \theta_i|x_i] = \frac{f(x_i + 1)}{f(x_i)} \Rightarrow \hat{E}[\theta_i|x_i] = 1 - \frac{\hat{f}(x_i + 1)}{\hat{f}(x_i)},$$

as obtained in "Smooth Empirical Bayes Estimation for One-Parameter Discrete Distributions" (Maritz, 1966). Indeed, the traditional empirical Bayes

estimation for this case is

$$\begin{aligned} E[1-\theta_i|x_i] &= \int (1-\theta_i)\frac{f(x_i|\theta_i)g(\theta_i)}{f(x_i)}d\theta_i \\ &= \frac{\int (1-\theta_i)\theta_i(1-\theta_i)^{x_i}g(\theta_i)d\theta_i}{f(x_i)} \\ &= \frac{\int \theta_i(1-\theta_i)^{x_i+1}g(\theta_i)d\theta_i}{f(x_i)} \\ &= \frac{f(x_{i+1})}{f(x_i)}. \end{aligned}$$

So that the same result is obtained:

$$E[\hat{\theta_i}|x_i] = 1 - \frac{\hat{f}(x_i+1)}{\hat{f}(x_i)}.$$

3 Empirical Bayes in Bayes

This chapter aims to provide a comprehensive exploration of the Empirical Bayes in Bayes (EBIB) approach and its approximation properties. In section 3.1 I will highlight its features and elaborate on the theoretical differences with the classic EB approach. These distinctions will be further examined in the subsequent sections using both real and simulated datasets. Specifically, section 3.2 will explore the theoretical approximation properties of the EBIB approach through the lens of a hierarchical Bayesian LASSO. In section 3.4 the exploration will be carried out through a hierarchical Bayesian RIDGE model.

3.1 Theoretical differences between EB and EBIB

The application context of the EBIB differs significantly from the classic EB one. Instead of multiple parallel experiments, the EBIB involves repeated sampling from the same conditional model $X|\theta \sim f(\cdot|\theta)$. Here, θ is an unknown fixed value and not a sample from a latent distribution $G(\theta)$. In this context, the prior $\pi(\theta|\lambda)$ on θ can only be interpreted as the formalization of prior information on θ. If in the classic EB case, there exists a true latent distribution, in the EBIB case the prior distribution $\pi(\theta|\lambda)$ is subjective. In other words, the prior distribution used in this framework loses the frequentist interpretation it had in the classic EB one.

The EBIB approach also differs from a fully Bayesian one. The latter consists of modeling the uncertainty around the hyperparameter λ through a hyperprior specification. Conversely, the EBIB approach consists of choosing

empirically a value for the hyperparameter λ. The standard approach uses its marginal maximum likelihood estimate (MMLE). A range of alternative approximations for computing the MMLE is outlined in "Learning from a lot: Empirical Bayes in high-dimensional prediction settings." (Wiel, Beest, and Münch, 2017). These comprehend the "Laplace EB" that maximizes the marginal likelihood through a Laplace approximation, the "Variational Bayes EB" that maximizes the marginal likelihood through a Variational Bayes approach, and the MCMC Empirical Bayes. In this latter case, $\hat{\lambda}$ is estimated with extractions from the MCMC sampler. A good example of this is the Monte Carlo-Expectation maximization (MCEM) algorithm proposed by Levine and Casella (Levine and Casella, 2001). The statistical validity of estimating a hyperparameter of a hierarchical model through the MCEM can be found in the "Empirical Bayes Gibbs sampling" (Casella, 2001). Casella also proposed an EB estimation of the shrinkage hyperparameter of the Laplace prior for Bayesian (sparse) regression, or "Bayesian Lasso" (Park and Casella, 2008). This will be explored and implemented in section 3.2. The empirical estimation of the hyperparameters is not supported by the frequentist theory nor by the Bayesian theory. Still, EBIB is commonly used by practitioners and researchers as a computationally convenient approximation of a fully Bayesian approach that would assign a hyperprior on λ. The underlying idea is that (1) for small or moderate sample size, the EB selection provides a "better" choice of the prior hyperparameters and (2) for large sample size, the resulting EB and Bayesian posterior distributions will substantially agree.

However, there was no systematic theory to support these common beliefs. A

methodological approach is suggested by Petrone, Rousseau, and Scricciolo (2014) and more recently by Petrone, Rizelli, and Rousseau (2024) who formalize the problem in terms of *merging* of the Bayesian and the EB posterior distributions. On this basis, they study the approximation properties of the EBIB posterior distribution.

3.2 Hierarchical Bayesian LASSO

Linear regression is a common modelization in many fields. The model is given by the equation

$$\underset{n\times 1}{Y} = \underset{n\times k}{X}\ \underset{k\times 1}{\underline{\beta}} + \underset{n\times 1}{\varepsilon}. \tag{2}$$

Some datasets however present the challenging situation of a high number of non-significant regressors k. In these cases, sparse models become necessary. In those, a shrinkage component drives many parameters towards 0. This class of models reduces the problem's dimensionality and allows us to find a solution. An example of a sparse model is the LASSO estimation. Here, the coefficients are estimated as

$$\hat{\underline{\beta}} = \underset{\underline{\beta}}{\operatorname{argmin}} \left(\sum_{i=1}^{n} (y_i - \sum_{j=1}^{k} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{k} |\beta_j| \right).$$

Therefore, the coefficient λ determines the weight assigned to the absolute value of the coefficients. In other words, it represents the shrinkage since it penalizes high absolute values for the coefficients. In a Bayesian perspective, sparse models can be obtained by assigning a prior distribution centered at 0 to the coefficients $\underline{\beta}$. Some examples are the centred Normal or the

Horseshoe prior distributions. It was noticed (Tibshirani, 1996) that the LASSO regression can be obtained by assigning a Laplace prior centered at 0 to the $\underline{\beta}$ coefficients. This is developed by Park and Casella (2008) in "The Bayesian LASSO", which is the setting we focus on. Consider a regression model

$$Y|X, \underline{\beta}, \sigma^2 \sim N(X\underline{\beta}; \sigma^2 I_n)$$

where Y is the $n \times 1$ vector of independent observations to be predicted, X is an $n \times k$ matrix or predictors and $\underline{\beta}$ is the $k \times 1$ vector of coefficients. σ^2 represents the variance of the errors. The uncertainty around σ^2 is dealt with by setting an uninformative prior so that

$$\sigma^2 \sim \pi(\sigma^2) \propto \frac{1}{\sigma^2}.$$

In turn, the uncertainty around $\underline{\beta}$ is modeled with a prior specification *conditional* on parameters $\tau_1^2, ..., \tau_k^2$ and σ^2. In particular

$$\underline{\beta}|\tau_1^2, ..., \tau_k^2, \sigma^2 \sim N(\underline{0}_k; \sigma^2 D_\tau),$$

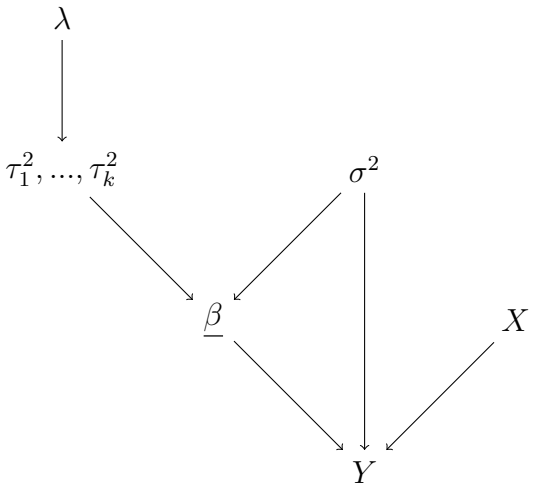
where $D_\tau = diag(\tau_1^2, ..., \tau_k^2)$. In turn, the uncertainty around the parameters $\tau_1^2, ..., \tau_k^2$ is assuming each τ_i^2 to be distributed independently from the others and as an Exponential. Therefore, the joint distribution of $\tau_1^2, ..., \tau_k^2$ is

$$\tau_1^2, ..., \tau_k^2|\lambda = \prod_{i=1}^{k} \frac{\lambda^2}{2} e^{-\lambda^2 \frac{\tau_i^2}{2}}.$$

In short, the model is:

$$\begin{cases} Y|X, \underline{\beta}, \sigma^2 \sim N(X\underline{\beta}, \sigma^2 I_n) \\ \sigma^2 \sim \pi(\sigma^2) \propto \frac{1}{\sigma^2} \\ \underline{\beta}|\tau_1^2, ..., \tau_k^2, \sigma^2 \sim N(\underline{0}_k, \sigma^2 D_\tau) \\ \tau_1^2, ..., \tau_k^2|\lambda \sim \prod_{i=1}^k \frac{\lambda^2}{2} e^{-\lambda^2 \frac{\tau_i^2}{2}} \end{cases} \tag{3}$$

and the corresponding Bayesian network is:



The choice of this model is convenient since the hyperparameter λ coincides with the shrinkage parameter of a LASSO model. This because integrating out $\tau_1^2, ..., \tau_k^2$ one obtain a conditional prior on $\underline{\beta}$ of form (Park and Casella, 2008):

$$f(\underline{\beta}|\sigma^2) = \prod_{j=1}^{k} \frac{\lambda}{2\sqrt{\sigma^2}} \exp\left(-\frac{\lambda|\beta_j|}{\sqrt{\sigma^2}}\right).$$

To compute the posterior distribution for $\underline{\beta}$ and σ^2, it is necessary to deal with the choice of the shrinkage parameter λ. This can be done in two

ways. First, the "fully Bayesian approach", which is the genuine Bayesian approach that includes uncertainty in the choice of λ by assigning it a prior distribution. Alternatively, one could follow the EBIB approach and estimate empirically λ.

3.2.1 Fully Bayesian model

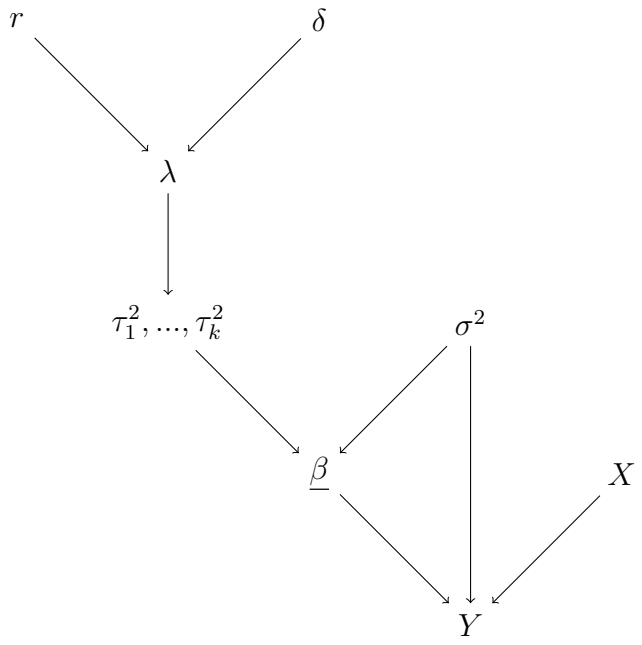
A standard, Bayesian way to model the uncertainty around λ is to assign a prior distribution to it. For example, in (Park and Casella, 2008), a Gamma prior with parameters r and δ is assigned to λ^2 so that:

$$f(\lambda^2|X, r, \delta) = \frac{\delta^r}{\Gamma(r)}(\lambda^2)^{r-1} \exp\{-\delta\lambda^2\}.$$

With the inclusion of the hyperprior on λ, the model (3) is extended to:

$$\begin{cases} Y|X, \underline{\beta}, \sigma^2 \sim N(X\underline{\beta}, \sigma^2 I_n) \\ \sigma^2 \sim \pi(\sigma^2) \propto \frac{1}{\sigma^2} \\ \underline{\beta}|\tau_1^2, ..., \tau_k^2, \sigma^2 \sim N(\underline{0}_k, \sigma^2 D_\tau) \\ \tau_1^2, ..., \tau_k^2|\lambda \sim \prod_{i=1}^k \frac{\lambda^2}{2} e^{-\lambda^2 \frac{\tau_i^2}{2}} \\ \lambda^2|X, r, \delta \sim \frac{\delta^r}{\Gamma(r)}(\lambda^2)^{r-1} \exp\{-\delta\lambda^2\} \end{cases} \tag{4}$$

The corresponding Bayesian Network will be:



The joint posterior of $\underline{\beta}, \tau_1^2, \ldots, \tau_k^2, \lambda^2, \sigma^2$ will then be

$$
\begin{aligned}
& f(\underline{\beta}, \tau_1^2, \ldots, \tau_k^2, \lambda^2, \sigma^2 | Y, X, r, \delta) \propto \\
& (2\pi)^{-\frac{n}{2}} |\sigma^2 I_n|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2\sigma^2} (Y - X\underline{\beta})'(Y - X\underline{\beta}) \right\} \\
& \times (2\pi)^{-\frac{k}{2}} |\sigma^2 D_\tau|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2\sigma^2} \underline{\beta}'(D_\tau)^{-1} \underline{\beta} \right\} \\
& \times \frac{1}{\sigma^2} \prod_{i=1}^{k} \frac{\lambda^2}{2} \exp\left\{ -\lambda^2 \frac{\tau_i^2}{2} \right\} \frac{\delta^r}{\Gamma(r)} (\lambda^2)^{r-1} \exp\left\{ -\delta\lambda^2 \right\}.
\end{aligned}
$$

The posterior distribution can be approximated by Gibbs sampling. To implement it, one must derive the full conditionals. In particular,

- The full conditional distribution $f(\underline{\beta} | \tau_1^2, \ldots \tau_k^2, \sigma^2, \lambda, Y, X, r, \delta)$ is

$$
\begin{aligned}
& f(\underline{\beta} | \tau_1^2, \ldots \tau_k^2, \sigma^2, \lambda, Y, X, r, \delta) \propto f(\underline{\beta}, \tau_1^2, \ldots \tau_k^2, \sigma^2, \lambda | Y, X, r, \delta) \\
\propto \quad & \exp\{ -\frac{1}{2\sigma^2 (X'X + (D_\tau)^{-1})^{-1}} (\underline{\beta}'\underline{\beta} - 2\underline{\beta}(X'X + (D_\tau)^{-1})^{-1} X'Y) \}.
\end{aligned}
$$

This is the kernel of $N_n((X'X + (D_\tau)^{-1})^{-1}X'Y; \sigma^2(X'X + (D_\tau)^{-1})^{-1})$. Setting $A = X'X + (D_\tau)^{-1}$, the previous result can be simplified to $N_n(A^{-1}(X'Y); \sigma^2 A^{-1})$.

- The full conditional distribution $f(\sigma^2|\underline{\beta}, \tau_1^2, ...\tau_k^2, \lambda, Y, X, r, \delta)$ is

$$
\begin{aligned}
&f(\sigma^2|\underline{\beta}, \tau_1^2, \ldots, \tau_k^2, \lambda^2, Y, X, r, \delta) \propto f(\sigma^2, \underline{\beta}, \tau_1^2, \ldots, \tau_k^2, \lambda^2|Y, X, r, \delta) \\
&\propto (\sigma^2)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma^2}(Y - X\underline{\beta})'(Y - X\underline{\beta})\right\} \\
&\times (\sigma^2)^{-\frac{k}{2}} \exp\left\{-\frac{1}{2\sigma^2}\underline{\beta}'(D_\tau)^{-1}\underline{\beta}\right\}(\sigma^2)^{-1} \\
&= (\sigma^2)^{-(\frac{n}{2}+\frac{k}{2}+1)} \exp\left\{-\frac{1}{\sigma^2}\frac{(Y - X\underline{\beta})'(Y - X\underline{\beta}) + \underline{\beta}'(D_\tau)^{-1}\underline{\beta}}{2}\right\}.
\end{aligned}
$$

  This result is the kernel of $InvGamma(\frac{n}{2} + \frac{k}{2}; \frac{(Y - X\underline{\beta})'(Y - X\underline{\beta}) + \underline{\beta}'(D_\tau)^{-1}\underline{\beta}}{2})$.

- The full conditional distribution $f(\tau_j^2|\underline{\beta}, \tau_1^2, ...\tau_{j-1}^2, \tau_{j+1}^2 ...\tau_k^2, \lambda^2, Y, X, r, \delta)$ for a generic $j \in \{1, ..., k\}$ is

$$
\begin{aligned}
&f(\tau_j^2|\underline{\beta}, \tau_1^2, \ldots, \tau_{j-1}^2, \tau_{j+1}^2, \ldots, \tau_k^2, \lambda^2, Y, X, r, \delta) \\
&\propto f(\sigma^2, \underline{\beta}, \tau_1^2, \ldots, \tau_k^2, \lambda^2|Y, X, r, \delta) \\
&\propto \exp\left\{-\frac{1}{2\sigma^2}\underline{\beta}'(D_\tau)^{-1}\underline{\beta}\right\}\frac{\lambda^2}{2}\exp\left\{-\lambda^2\frac{\tau_i^2}{2}\right\} \\
&\propto \exp\left\{-\frac{1}{2\sigma^2}\frac{\beta_j^2}{\tau_j^2}\right\}\frac{\lambda^2}{2}\exp\left\{-\lambda^2\frac{\tau_i^2}{2}\right\}.
\end{aligned}
$$

  Considering $\frac{1}{\tau_j^2}$, we recognize the kernel of an $InvGaussian(\sqrt{\frac{\lambda^2\sigma^2}{\beta_j^2}}; \lambda^2)$.

- Finally, the full conditional distribution $f(\lambda^2|\underline{\beta}, \tau_1^2, ..., \tau_k^2, Y, X, r, \delta))$ is

$$
\begin{aligned}
&f(\lambda^2|\underline{\beta}, \tau_1^2, \ldots, \tau_k^2, Y, X, r, \delta) \propto f(\sigma^2, \underline{\beta}, \tau_1^2, \ldots, \tau_k^2, \lambda^2|Y, X, r, \delta) \\
&\propto \lambda^{2k} \exp\left\{-\lambda^2 \sum_{i=1}^{k} \frac{\tau_i^2}{2}\right\} (\lambda^2)^{r-1} \exp\left\{-\delta\lambda^2\right\} \\
&= (\lambda^2)^{k+r-1} \exp\left\{-\lambda^2 \left(\sum_{i=1}^{k} \frac{\tau_i^2}{2} + \delta\right)\right\}.
\end{aligned}
$$

  This is the kernel of $Gamma(k + r; \sum_{i=1}^{k} \frac{\tau_i^2}{2} + \delta))$.

Once the full conditional distributions of the parameters of interest are specified, it is possible to implement the corresponding Gibbs sampler to sample from the posterior distributions. The steps to be followed to obtain it are:

1. Choose a feasible set of starting points
   $\theta^{(0)} = (\underline{\beta}^{(0)}, (\sigma^2)^{(0)}, (\tau_1^2)^{(0)}, ..., (\tau_k^2)^{(0)}, (\lambda^2)^{(0)})$.

2. At time t (and so for a current value
   $\theta^{(t-1)} = (\underline{\beta}^{(t-1)}, (\sigma^2)^{(t-1)}, (\tau_1^2)^{(t-1)}, ..., (\tau_k^2)^{(t-1)}, (\lambda^2)^{(t-1)}))$ new values are sampled, until convergence, from the full conditional distributions

   - $\underline{\beta}^{(t)}$ from $f(\underline{\beta}|(\sigma^2)^{(t-1)}, (\tau_1^2)^{(t-1)}, ..., (\tau_k^2)^{(t-1)}, (\lambda^2)^{(t-1)})$,
   - $(\sigma^2)^{(t)}$ from $f(\sigma^2|\underline{\beta}^{(t)}, (\tau_1^2)^{(t-1)}, ..., (\tau_k^2)^{(t-1)}, (\lambda^2)^{(t-1)})$,
   - $\frac{1}{(\tau_1^2)^{(t)}}$ from $f(\frac{1}{\tau_1^2}|\underline{\beta}^{(t)}, (\sigma^2)^{(t)}, (\tau_2^2)^{(t-1)}, ..., (\tau_k^2)^{(t-1)}, (\lambda^2)^{(t-1)})$,

     $\vdots$

   - $\frac{1}{(\tau_k^2)^{(t)}}$ from $f(\frac{1}{\tau_k^2}|\underline{\beta}^{(t)}, (\sigma^2)^{(t)}, (\tau_1^2)^{(t)}, ..., (\tau_{k-1}^2)^{(t)}, (\lambda^2)^{(t-1)})$,
   - $(\lambda^2)^{(t)}$ from $f(\lambda^2|\underline{\beta}^{(t)}, (\sigma^2)^{(t)}, (\tau_1^2)^{(t)}, ..., (\tau_k^2)^{(t)},)$.

Given the full conditional distributions previously derived, the Gibbs sampler is obtained by iterative sampling from

- $\underline{\beta}^{(t)}$ from $N_n(A^{-1}(X'Y); \sigma^2 A^{-1})$,
- $(\sigma^2)^{(t)}$ from $InvGamma(\frac{n}{2} + \frac{k}{2}; \frac{(Y - X\underline{\beta})'(Y - X\underline{\beta}) + \underline{\beta}'(D_\tau)^{-1}\underline{\beta}}{2})$,
- $\frac{1}{(\tau_1^2)^{(t)}}$ from $InvGaussian(\sqrt{\frac{\lambda^2\sigma^2}{\beta_1^2}}; \lambda^2)$,

$$\vdots$$

- $\frac{1}{(\tau_k^2)^{(t)}}$ from $InvGaussian(\sqrt{\frac{\lambda^2\sigma^2}{\beta_k^2}}; \lambda^2)$,
- $(\lambda^2)^{(t)}$ from $Gamma(k + r; \sum_{i=1}^{k} \frac{\tau_i^2}{2} + \delta))$.

The sampler is then obtained in **R** through the code provided in section 5.1.

3.2.2 Empirical Bayes model

The empirical Bayes model differs from the previous one because of the estimation of the parameter λ. In particular, Casella (2001) suggested an intriguing Empirical Bayes approach for the previous hierarchical model where the Gibbs sampler is complemented by an Expectation-Maximization (EM) procedure. This is obtained by adding steps 3 (E-step) and 4 (M-step) to the initial Gibbs sampler:

1. Choose a feasible set of starting points
   $\theta^{(0)} = (\underline{\beta}^{(0)}, (\sigma^2)^{(0)}, (\tau_1^2)^{(0)}, ..., (\tau_k^2)^{(0)})$ and an initial value for $(\lambda^2)^{(0)}$.

2. At the time t (and so for a current value
   $\theta^{(t-1)} = (\underline{\beta}^{(t-1)}, (\sigma^2)^{(t-1)}, (\tau_1^2)^{(t-1)}, ..., (\tau_k^2)^{(t-1)})$ and with $(\lambda^2)^{(t-1)})$ new values are sampled, until convergence, from the full conditional distributions

- $\underline{\beta}^{(t)}$ from $f(\underline{\beta}|(\sigma^2)^{(t-1)}, (\tau_1^2)^{(t-1)}, ..., (\tau_k^2)^{(t-1)}, (\lambda^2)^{(t-1)})$,
- $(\sigma^2)^{(t)}$ from $f(\sigma^2|\underline{\beta}^{(t)}, (\tau_1^2)^{(t-1)}, ..., (\tau_k^2)^{(t-1)}, (\lambda^2)^{(t-1)})$,
- $\frac{1}{(\tau_1^2)^{(t)}}$ from $f\left(\frac{1}{(\tau_1^2)}|\underline{\beta}^{(t)}, (\sigma^2)^{(t)}, \tau_2^{(t-1)}, ..., \tau_k^{(t-1)}, (\lambda^2)^{(t-1)}\right)$,

$$\vdots$$

- $\frac{1}{(\tau_k^2)^{(t)}}$ from $f\left(\frac{1}{(\tau_k^2)}|\underline{\beta}^{(t)}, (\sigma^2)^{(t)}, (\tau_1^2)^{(t)}, ..., (\tau_{k-1}^2)^{(t)}, (\lambda^2)^{(t-1)}\right)$.

Given the full conditional distributions previously derived, the Gibbs sampler is obtained by iterative sampling

- $\underline{\beta}^{(t)}$ from $N_n\left(A^{-1}(X'Y); (\sigma^2)^{(t-1)}A^{-1}\right)$,
- $(\sigma^2)^{(t)}$ from Inv Gamma $\left(\frac{n}{2} + \frac{k}{2}; \frac{(Y-X\underline{\beta})'(Y-X\underline{\beta})+\underline{\beta}'(D_\tau)^{-1}\underline{\beta}}{2}\right)$,
- $\frac{1}{(\tau_1^2)^{(t)}}$ from Inv Gaussian $\left(\sqrt{\frac{\lambda^2\sigma^2}{\beta_1^2}}; \lambda^2\right)$,

$$\vdots$$

- $\frac{1}{(\tau_k^2)^{(t)}}$ from Inv Gaussian $\left(\sqrt{\frac{\lambda^2\sigma^2}{\beta_k^2}}; \lambda^2\right)$.

3. (Expectation step) The expected log-likelihood for λ is proportional to $k \log \lambda^2 - \frac{\lambda^2}{2}\sum_{i=1}^{k} E[\tau_i^2]$. Since the generic term $E[\tau_i^2]$ is unknown, it will be replaced with the average of the corresponding Gibbs sampler extractions conditionally on the data X and Y and the previous value $\lambda^{(t-1)}$: $E[\tau_i^2|X, Y, \lambda^{(t-1)}]$.

4. (Maximization step) The approximation of the expected log-likelihood is maximized. The closed-form solution obtained will be the new value

for $\lambda^{(t)}$:

$$\begin{aligned}\lambda^{(t)} &= \arg\max_{\lambda} \left(k \log \lambda^2 - \lambda^2 \frac{\sum_{i=1}^{k} E[\tau_i^2 | X, Y, \lambda^{(t-1)}]}{2} \right) \\ &= \sqrt{\frac{2k}{\sum_{i=1}^{k} E[\tau_i^2 | X, Y, \lambda^{(t-1)}]}}\end{aligned}$$

5. Repeat steps 2, 3, and 4 until convergence is reached

The sampler is then obtained in **R** through the code provided in section 5.2.

3.2.3 A real data application

The Gibbs samplers outlined in sections 3.2.1 and 3.2.2, and their corresponding implementations described in section 5.1 and 5.2 are applied to the pre-loaded Diabetes dataset in **R**. This is used both by Efron (2010) in "Large scale inference: empirical Bayes methods for estimation, testing and prediction" and by Park and Casella (2008) in "The Bayesian lasso". It consists of 11 variables, 10 of which are predictors for the progression of diabetes in one year. A first illustrative table of the variables is:

Table 1: Summary Statistics for the Diabetes dataset variables

Variable	**Mean**	**Standard Deviation**	**Minimum**	**Median**	**Maximum**
age	48.518	13.109	19.000	50.000	79.000
sex	0.468	0.499	0.0000	0.00000	1.000
bmi	26.376	4.418	18.000	25.7000	42.200
map	94.647	13.831	62.000	93.000	133.000
tc	189.140	34.608	97.000	186.000	301.000
ldl	115.439	30.413	41.600	113.000	242.400
hdl	49.788	12.934	22.0000	48.00000	99.000
tch	4.070	1.290	2.0000	4.00000	9.090
ltg	4.641	0.522	3.258	4.620	6.107
glu	91.260	11.496	58.000	91.000	124.000
y	152.133	77.093	25.000	140.500	346.000

The implementation of the Gibbs sampler for the fully Bayesian model will follow the procedure in "The Bayesian lasso" (Park and Casella, 2008). Therefore, the dataset will be standardized, and the prior for λ^2 is set to be a $Gamma(1, 1.78) = Exp(1.78)$ as presented in Figure 1.

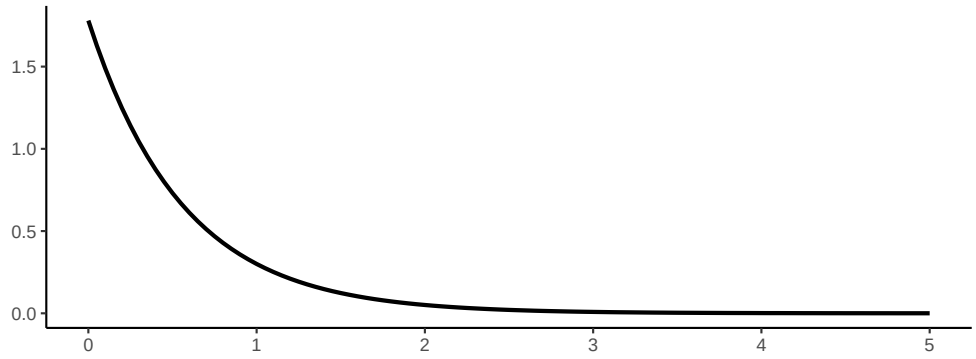


Figure 1: Prior distribution of $\lambda \sim Exp(1.78)$

The implementation of the Gibbs sampler with 10.000 iterations and 1.000 burn-in gives MCMC samples from the posterior distribution of λ as presented in Figure 2. For replicability, the **R** code is run setting the seed "123".

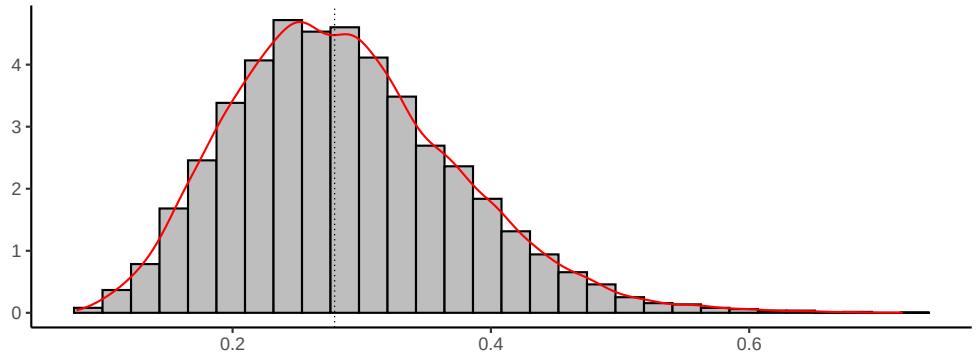


Figure 2: Samples from the posterior distribution of λ resulting from the Gibbs sampler of the fully Bayesian model. The vertical dotted line represents the median value of the extractions from the posterior distribution of λ. Its value is of 0.2789857.

The implementation of the Gibbs sampler for the EB model does not require a prior specification of λ that is instead obtained through the Expectation-Maximization algorithm. Its iterations are shown in Figure 3. For the applicability, the **R** command is run setting the seed "123".

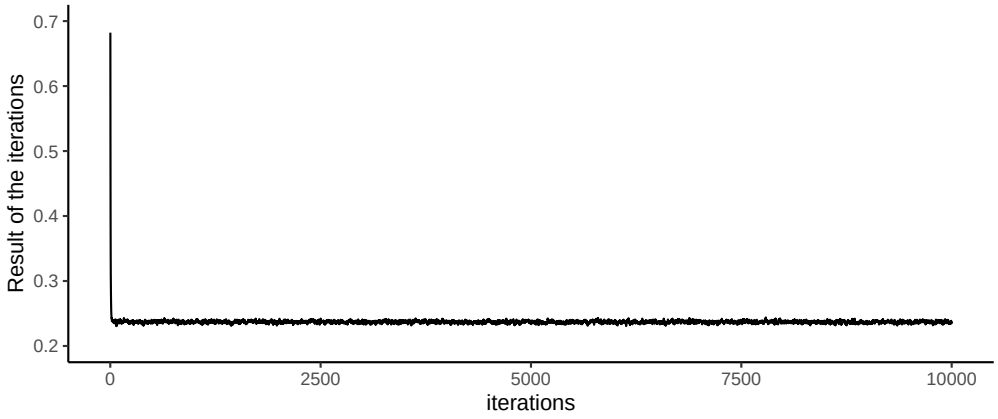


Figure 3: Evolution of the $\lambda^{(t)}$ through the EM-algorithm with an initial $\lambda^{(0)} = 1$.

The implementation of the Gibbs sampler with 10.000 iterations and 1.000 burn-in gives MCMC samples from the posterior distribution of λ as presented in 4. For replicability, the **R** code is run setting the seed "123".

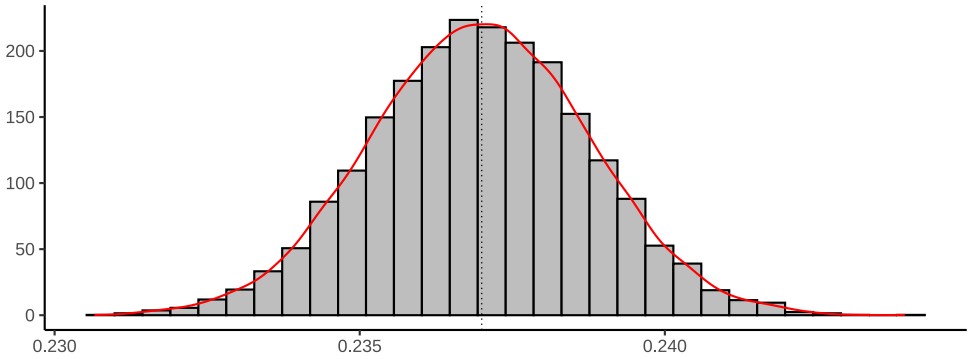


Figure 4: Samples from the posterior distribution of λ resulting from the Gibbs sampler of the EB model. The vertical dotted line represents the median value of the extractions from the posterior distribution of λ. Its value is of 0.2369713.

Therefore, the fully Bayesian approach will suggest a greater shrinkage than the EB approach.

3.3 Simulation study: higher-order merging

The simulation study aims to provide a more concrete understanding of the theoretical properties of the EBIB posterior distribution. Previous findings by Petrone, Rousseau, and Scricciolo (2014) and the ones by Petrone, Rizelli, and Rousseau (2024) foresee what we expect in these simulations. That is, for a small dataset, the EBIB choice of the hyperparameter will be more accurate. As the sample size increases, the EBIB posterior will substantially agree with the Bayesian one. Therefore, although not rigorous, the EBIB is proposed as a computationally attractive approximation of the fully Bayesian approach. This is because it is a fast approximation of the oracle posterior and, if this is non-degenerate, the EBIB posterior will merge with the corresponding fully Bayesian alternative. So, three will be the posteriors compared in this section: the EBIB posterior, the Bayesian posterior with λ fixed, and the oracle posterior.

The EBIB posterior is the one resulting from the model implemented in section 3.2.2. The Bayesian model with fixed λ implementation will instead be described in section 3.3.1. Finally, section 3.3.2 will detail how to obtain the oracle posterior.

3.3.1 Bayesian model with fixed λ

The EBIB posterior distributions will be compared to the ones of a Bayesian model with λ fixed. This represents another computationally simpler alternative to the fully Bayesian model described in section 3.2.1. This class of posteriors will be obtained for each example setting $\lambda = 3$. This value does

not have a precise meaning, but it is chosen to represent a poor choice of the hyperparameter. However, in this setting, λ is more than a generic hyperparameter. It indeed captures the shrinkage effects on the coefficients of interest. Therefore, setting $\lambda = 3$ will imply an automatic shrinkage of the equivalent OLS estimates. To obtain the posterior distributions in this case, it is possible to implement the same hierarchical model as (4), just with no sampling step for λ, which is now fixed.

3.3.2 Oracle Bayes posterior

Petrone, Rousseau, and Scricciolo (2014) proved how, under certain regularity conditions, the EBIB posterior is a fast approximation of the oracle posterior. This is the posterior distribution one would obtain by implementing the same Gibbs sampler of the model (4) where the hyperparameter λ is set to be fixed and equal to the oracle value λ^*. The interpretation of "oracle" in this work is not the one of Efron (2019) in "Bayes, Oracle Bayes and Empirical Bayes". This would consist of a classic EB estimation with the number of parallel classes of experiments $k \to \infty$ and so such that $\lambda^* = \lim_{k\to\infty} \hat{\lambda}_{MMLE}$. Instead, in this framework, there exists no true prior $g(\theta|\lambda)$ and so, in turn, there exists no true value of value for λ. Therefore, the notion of "oracle" is intended in the sense of Petrone, Rousseau, and Scricciolo (2014) in "Bayes and empirical Bayes: Do they merge?". Here, the oracle value λ^* is set to be the value that maximizes the prior density at β_{true}:

$$\lambda^* = \arg\max_{\lambda} f(\beta_{true}|\lambda).$$

With $f(\beta_{true}|\lambda)$ being the prior for the set of parameters β_{true}. In this sense, the oracle prior $f(\beta_{true}|\lambda^*)$ corresponds to the prior that would be expressed by the Bayesian researcher who has the most information on the true β_{true}. In these settings, Petrone, Rousseau, and Scricciolo (2014) showed that, if $\sup_{\lambda} f(\underline{\beta}_{true}|\lambda) < \infty$ and under regularity conditions, the MMLE $\hat{\lambda}_n$ converges to the oracle value $\lambda^* = \arg\max_{\lambda} f(\beta_{true}|\lambda)$. In addition, the posterior distribution $f(\underline{\beta}_{true}|\hat{\lambda}_n, Y, X)$ merges strongly with the oracle Bayesian posterior $f(\underline{\beta}_{true}|\lambda^*, Y, X)$. This merging may fail in the cases in which $\sup_{\lambda} f(\underline{\beta}_{true}|\lambda) = \infty$. The model previously derived is obtained assuming a Laplace prior specification so that

$$f(\beta_{true}|\lambda) = \prod_{i=1}^{k} \frac{\lambda}{2\sigma} e^{-\frac{\lambda|\beta_j|}{\sigma}}$$

It is now possible to derive the oracle value as

$$\lambda^* = \arg\max_{\lambda} f(\beta_{true}|\lambda) = \arg\max_{\lambda} \prod_{i=1}^{k} \frac{\lambda}{2\sigma} e^{-\frac{\lambda|\beta_j|}{\sigma}} = \sum_{j=1}^{k} \log\left(\frac{\lambda}{2\sigma}\right) - \frac{\lambda|\beta_j|}{\sigma}.$$

This implies $\lambda^* = \frac{k\sigma}{\sum_{j=1}^{k}|\beta_j|}$. The final model obtained will therefore be of form:

$$\begin{cases} Y|X, \underline{\beta}, \sigma^2 \sim N(X\underline{\beta}, \sigma^2 I_n) \\ \sigma^2 \sim \pi(\sigma^2) \propto \frac{1}{\sigma^2} \\ \underline{\beta}|\tau_1^2, ..., \tau_k^2, \sigma^2 \sim N(\underline{0}_k, \sigma^2 D_\tau) \\ \tau_1^2, ..., \tau_k^2|\lambda \sim \prod_{i=1}^k \frac{\lambda^2}{2} e^{-\lambda^2 \frac{\tau_i^2}{2}} \\ \lambda = \frac{k\sigma}{\sum_{j=1}^{k} |\beta_j|} \end{cases} \tag{5}$$

3.3.3 Simulation scenarios

The EBIB and Bayesian procedures will be compared through multiple simulation studies. These, aim to explore and showcase the theoretical results regarding the asymptotic properties of the EBIB posterior distribution to have a better sense of these properties. In general, the EBIB posterior will asymptotically agree with *any* Bayesian posterior distribution at a first order. However, merging is expected to be quicker between the EB posterior and the oracle-Bayes posterior distribution. To obtain a posterior distribution for the Oracle Bayes, a synthetic dataset is needed where the true coefficients are known. Let us consider the regression model 2. The corresponding design matrix $X = [x_{i,j}]$ will be generated as follows:

$$x_{i,j} \overset{iid}{\sim} Unif(-10, 10) \; \forall \; i = \{1, \ldots, n\}; \; \forall \; j = \{1, \ldots, k\}.$$

Then, the observations $\underline{y}$ are generated for a given value of coefficients $\underline{\beta} = \beta_1, ..., \beta_k$ so that:

$$\underline{y} = X\underline{\beta} + \underline{\varepsilon},$$

where $\underline{\varepsilon}$ is the vector of residuals so that $\varepsilon_i \overset{iid}{\sim} N(0; \sigma^2)$. In the following examples (see Examples 3.3.4, 3.3.5, 3.3.6, 3.3.7), σ^2 is set to be equal to 1. Conversely, the number of coefficients k, their value $\underline{\beta}$ and the sample size n will vary. For each data structure considered, the posterior distributions resulting from the oracle Bayes, EBIB, and actual Bayes approach will be compared.

3.3.4 Example 1: Dense model

The data structure in this example is dense: no true coefficient is equal to 0. The true parameters are set to be $\beta_{true} = (\beta_1, \beta_2, \beta_3, \beta_4) = (-1, 1, 3, 2)$. The only difference between the models applied to this dataset is the choice of the shrinkage parameter. In particular, the model based on an oracle prior will set $\lambda^* = \frac{k\sqrt{\sigma^2}}{\sum_{j=1} |\beta_j|} = \frac{4}{7}$. This will represent a lower shrinkage compared to the Bayesian alternative of choosing $\lambda = 3$. Finally, the EBIB model will set λ_{EBIB} through the Expectation-Maximization algorithm specified in section 3.2.2. Figure 5 shows how the iterations soon reach a stationary level of roughly 1.55:

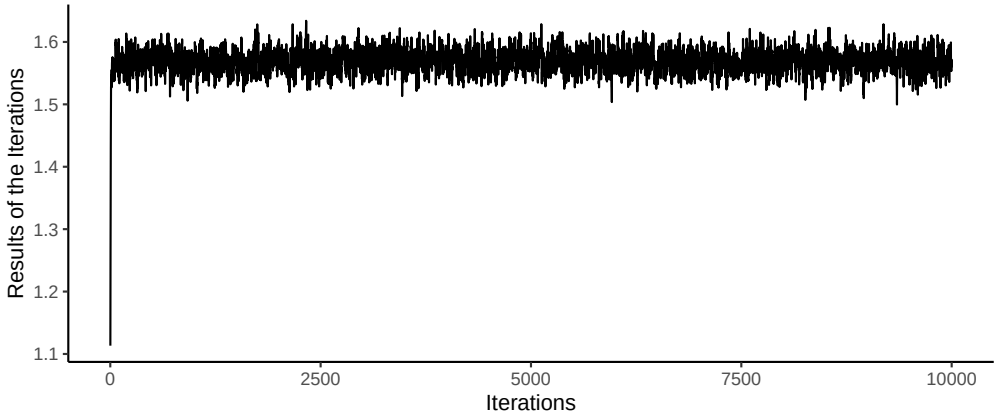


Figure 5: Example 1. Evolution of the $\lambda^{(t)}$ through the EM-algorithm with an initial value $\lambda^{(0)} = 1$

Different values of the λ coefficient will imply different shrinkages of the posterior distribution. The Bayesian estimation with $\lambda = 3$ will shrink the most the posterior distribution and, since the true coefficients are different from 0, it will have a flatter posterior. The EBIB posterior will instead be closer to the oracle posterior since its shrinkage coefficient of $\lambda_{EBIB} \approx 1.55$ is closer to the oracle value of $\lambda^* = 4/7$. Finally, the oracle posterior will be the one most spiked around the true coefficients. This is because the true value of the coefficients is different from 0 and the likelihood of the data reflects that. The posterior distributions of each coefficient resulting from the three models are shown in Figure 6.

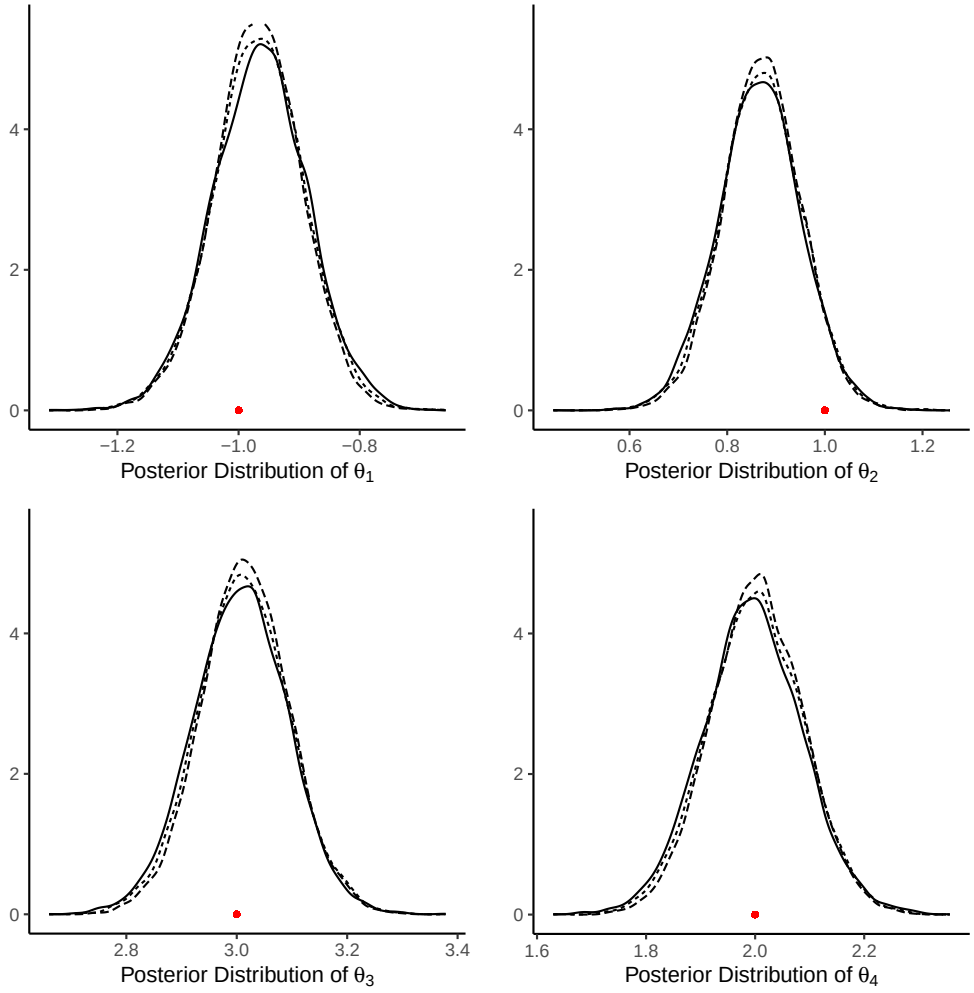


Figure 6: Example 1. The solid line represents the Bayesian posterior with $\lambda = 3$. The dotted line represents the Empirical Bayes posterior with $\hat{\lambda} \approx 1.55$ estimated through the EM algorithm. The dashed line represents the oracle Bayes posterior with $\lambda^* = 4/7$.

3.3.5 Example 2: Mixed model

In this example, we consider again a fairly simple regression model, where however some variables may not be statistically significant. More specifically,

here we fix the true parameters as $\underline{\beta} = (\beta_1, \beta_2, \beta_3, \beta_4) = (-1, 1, 3, 0)$. Despite this, the conditions for the strong merging between the oracle Bayes and the EBIB posterior distribution are satisfied.

In this case, the hyperparameter λ is common for all the coefficients since $\beta_1, ..., \beta_k \overset{iid}{\sim} Lapl\left(0; \frac{\sqrt{\sigma^2}}{\lambda}\right)$. This implies that the oracle λ^* is bounded satisfying the conditions for the strong merging between the oracle Bayes and the EBIB posterior distribution are satisfied. If however, one used a coefficient-specific hyperparameter so that $\beta_j \sim Lapl\,(0, \tau_j)$, where the role of the hyperparameters is played by the τ_j's, then the oracle value for τ_4 would be zero, which gives an oracle prior for β_4 degenerate on zero. The resulting EB posterior would be degenerate, too, and does not merge with any (non-degenerate) Bayesian posterior. Using a further stage in the prior specification, namely a prior on the τ_j, the Laplace prior "regularizes" the model, avoiding degenerate cases, and the results on merging of the EB and the Bayesian posterior distributions hold. A similar and more evident case is the example in section 3.5.3. In that case, the significance of just one coefficient guarantees non-degenerate oracle posteriors for all the coefficients. Since however the oracle posterior is non-degenerate, so will the EBIB posterior. In turn, this implies that the EBIB posterior will merge asymptotically with any Bayesian posterior distribution.

As said, the shrinkage parameter in the Bayesian model is set in all examples to be equal to 3. The oracle value for the shrinkage will be determined through the same closed-form solution since the prior assigned to the coefficients is the same as the previous example. In this case, the sets of coefficients

chosen will determine a value of $\lambda^* = \frac{k\sqrt{\sigma^2}}{\sum_{j=1}^{k} |\beta_j|} = \frac{4}{5}$. Finally, the EBIB model will set λ_{EBIB} through the Expectation-Maximization algorithm specified in section 3.2.2. Figure 7 shows how the iterations soon reach a stationary level of roughly 0.9:

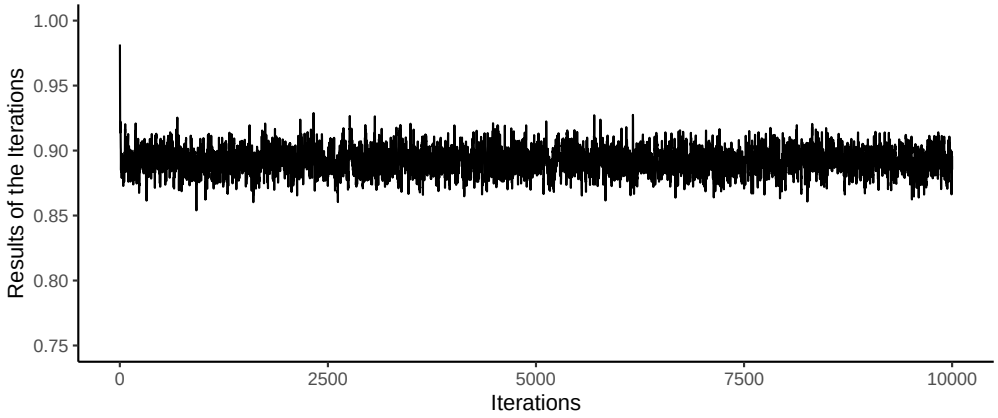


Figure 7: Example 2: Evolution of the $\lambda^{(t)}$ through the EM-algorithm with an initial $\lambda^{(0)} = 1$.

As in the previous example, Bayesian estimation with $\lambda = 3$ will shrink the posterior distributions the most towards 0. However, in this case, the EBIB approach will suggest a value of the shrinking coefficient $\lambda_{EBIB} \approx 0.9$ closer to the oracle value $\lambda^* = 0.8$. Therefore, in this case, the EBIB posterior will almost overlap with the oracle Bayesian posterior as the different estimation of the shrinking parameter is the only difference between the models. The posterior distributions of each coefficient are shown in 8.

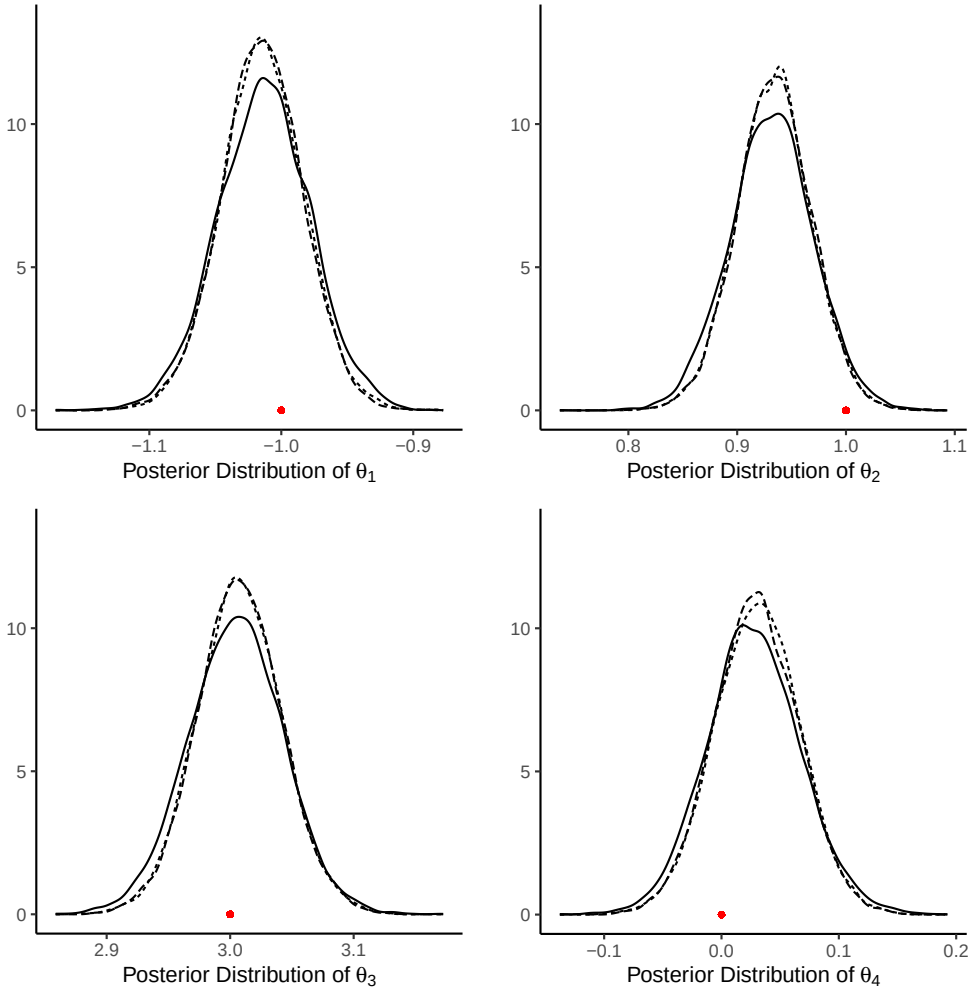


Figure 8: Example 2: The solid line represents the Bayesian posterior distribution with $\lambda = 3$. The dotted line represents the Empirical Bayes posterior distribution with $\hat{\lambda} \approx 0.9$ estimated through the EM algorithm. The dashed line represents the oracle Bayes posterior distribution with $\lambda^* = 4/5$.

3.3.6 Example 3: Sparse model

The third case aims to assess the performance of the EBIB approach in the case of an almost sparse dataset. In other words, there will be many

non-relevant coefficients alongside the only relevant one. Specifically, the true set of coefficients chosen will be set to be $\underline{\beta}_{true} = (\beta_1, ..., \beta_{19}, \beta_{20}) = (0, ..., 0, 1)$. This implies that a greater shrinkage is likely to enhance the model's performances since, pushing to 0 the estimates, will improve the performance of 19 estimators out of 20. Indeed, the shrinkage parameter with which the oracle posteriors are determined will be $\lambda^* = \frac{k\sqrt{\sigma^2}}{\sum_{j=1}^{k} |\beta_j|} = 20$. Similarly to before, the Bayesian alternative will impose a lower shrinkage by setting $\lambda = 3$. Finally, the EBIB alternative will set the shrinkage based on the EM algorithm specified in section 3.2.2. Figure 9 shows how the iterations soon reach a stationary level or roughly 10:

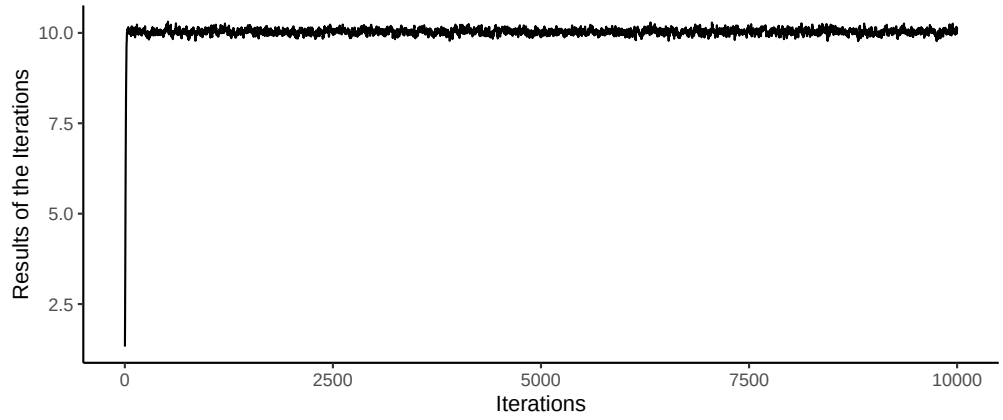


Figure 9: Example 2: Evolution of the $\lambda^{(t)}$ through the EM-algorithm with an initial $\lambda^{(0)} = 1$.

In this case, the Bayesian model with $\lambda = 3$ will shrink the posterior the least toward 0. The values of $\lambda_{EBIB} \approx 10$ and $\lambda^* = 20$ will be closer one to the other. The substantial differences in the shrinkage will be reflected in significant variations in the posterior distributions, as illustrated in figure

10.

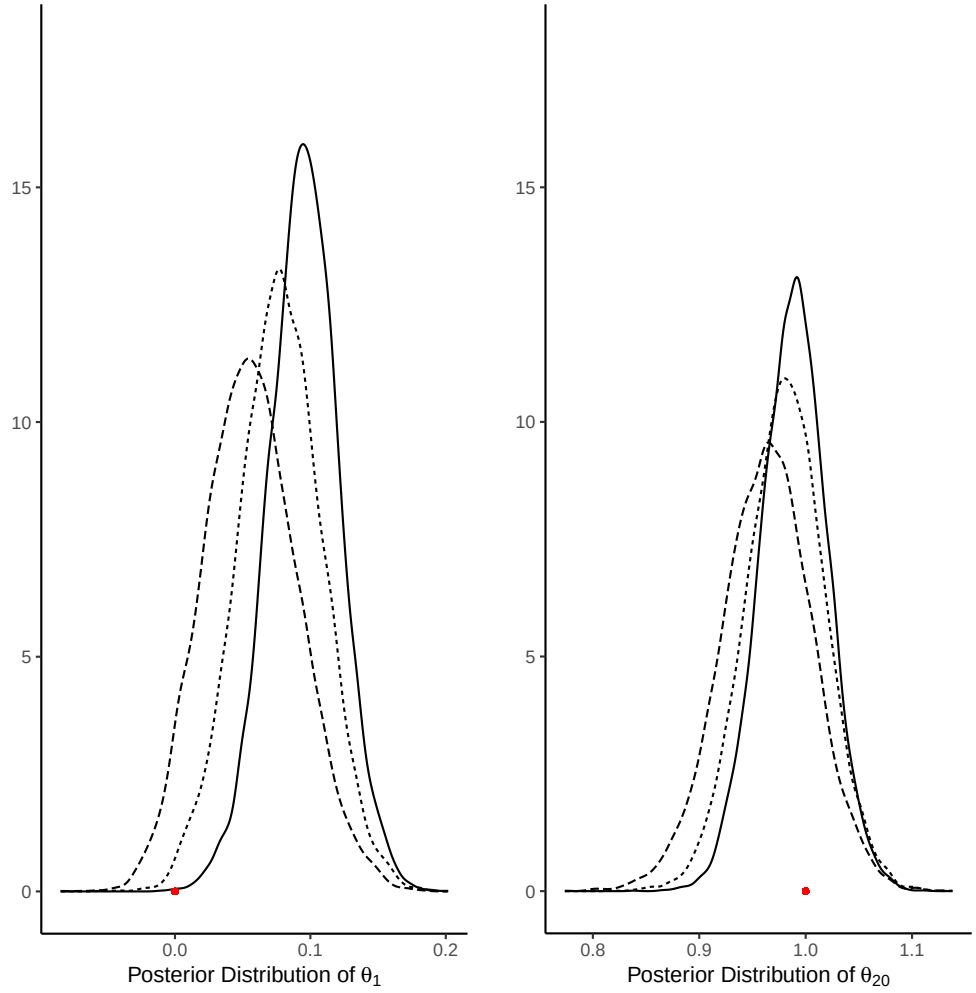


Figure 10: Example 3: The solid line represents the Bayesian posterior distribution with $\lambda = 3$. The dotted line represents the Empirical Bayes posterior distribution with $\hat{\lambda}$ estimated through the EM algorithm. The dashed line represents the oracle Bayes posterior distribution with $\lambda^* = 20$.

Even in this particular case, the EB posterior distribution is closer to the oracle Bayes one compared to the Bayesian alternative with $\lambda = 3$.

3.3.7 Example 4: Sparse model with varying sample size

The fourth and final case aims to see how the performance of the EBIB evolves as the number of observations increases. To achieve this, I will maintain the same coefficients as the previous example but vary the number of observations. The alternatives three alternatives with $n = 20$, $n = 50$, and $n = 100$ will be simulated. Figure 11 shows how, as n increases, the EM algorithm will suggest a value closer and closer to the oracle value $\lambda^* = 20$.

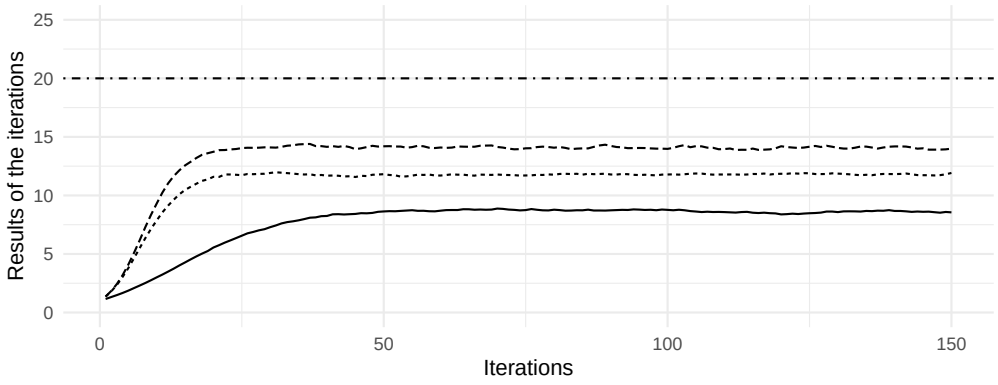


Figure 11: Example 4: The solid line represents the iterations of the EM algorithm in the case of $n = 20$. The dotted line represents the the iterations of the EM algorithm in case of $n = 50$. The dashed line represents the the iterations of the EM algorithm in case of $n = 100$. The horizontal dot-dashed line represents the oracle value of $\lambda^* = 20$.

Therefore, as the data increases, the prior of the EBIB model will be closer to the Oracle one. Hence, the faster merging to the oracle Bayes posterior compared to the Bayesian alternative with $\lambda = 3$. On the other hand, as the data increases, the likelihood of the data will dominate the prior distributions

reducing the relevance of a different prior specification. In other words, the differences between the models rely on the way the hyperparameter λ is set and this, in turn, determines different prior distributions. As long as the dataset is limited, the prior distributions will have a relevant weight for the posterior distribution. However, as the data increases, the differences in the priors will be washed away resulting in similar posterior distributions as plotted in Figure 12.

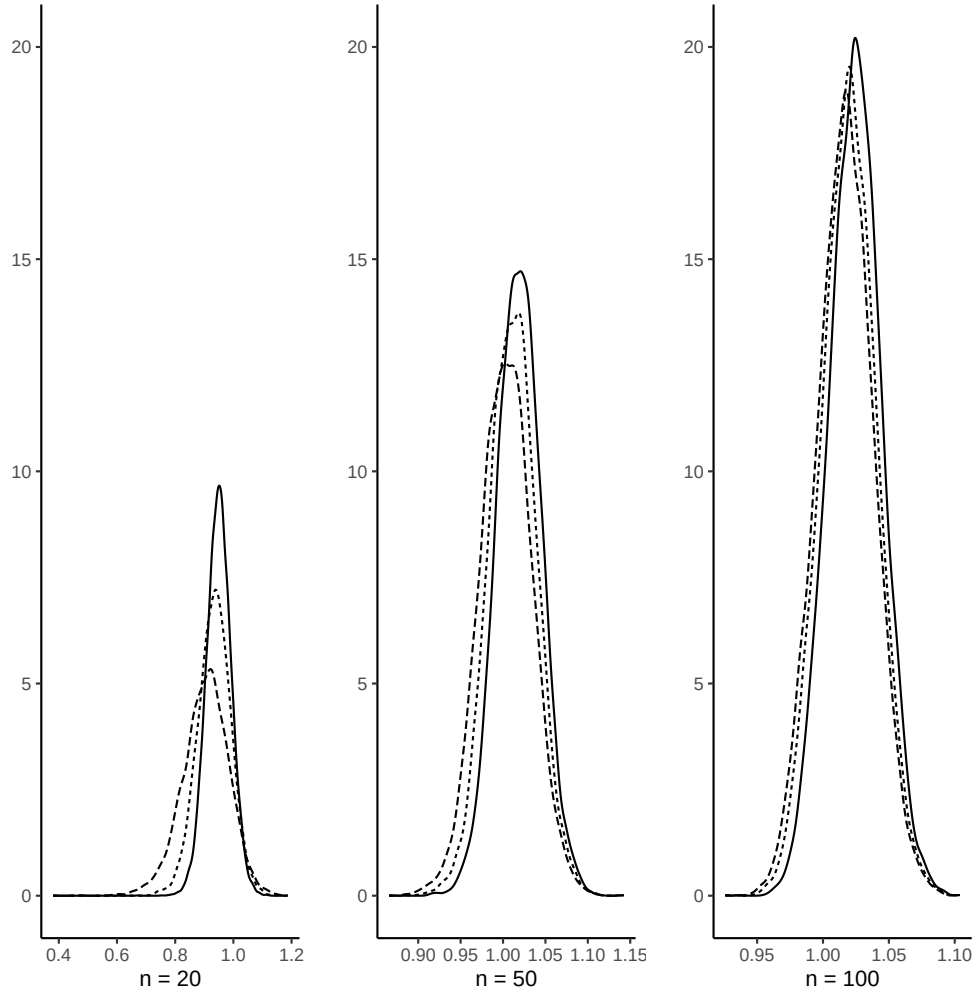


Figure 12: Example 4: Posterior densities for the parameter θ_{20} whose true value is 1, as the sample size changes. In all three graphs, the solid line represents the Bayesian posterior distribution with $\lambda = 3$. The dotted line is the Empirical Bayes posterior distribution with $\hat{\lambda}$ estimated through the EM algorithm. The dashed line represents the oracle Bayes posterior distribution density with $\lambda^* = 20$.

3.4 Hierarchical Bayesian RIDGE

This section will extend the simulation study to the hierarchical RIDGE model. This represents a sparse model one could use as an alternative to the LASSO model seen in section 3.2. The RIDGE model, similarly to the LASSO regression, shrinks the parameters towards 0 reducing the dimensionality of the problem. The difference is that, in this case, the coefficients are estimates as

$$\underline{\hat{\beta}} = \underset{\underline{\beta}}{\operatorname{argmin}} \left(\sum_{i=1}^{n} (y_i - \sum_{j=1}^{k} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{k} \beta_j^2 \right) = (X'X + \lambda I_k)X'Y.$$

Also here, the λ term represents how the magnitude of the coefficients is penalized. However, in this case, it is squared and therefore a different interpretation can be attached to it. In a Bayesian perspective, the RIDGE regression can be obtained by assigning a Normal prior centered at 0. The starting point is the regression model 2. So that

$$Y|X, \underline{\beta}, \sigma^2 \sim N_n(X\underline{\beta}; \sigma^2 I_n),$$

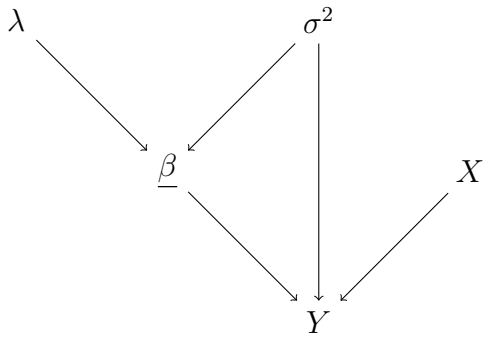
where Y is the $n \times 1$ vector of observations to be predicted, X is an $n \times k$ matrix or predictors and $\underline{\beta}$ is the $k \times 1$ vector of coefficients. σ^2 represents the variance of the errors. The observations are assumed to be independent and so the variance term σ^2 is multiplied to the identity matrix of order n. The uncertainty around σ^2 will be dealt with by setting an uninformative prior so that

$$\sigma^2 \sim \pi(\sigma^2) \propto \frac{1}{\sigma^2}.$$

In turn, the way in which the uncertainty around $\underline{\beta}$ is modeled, represents the first difference with the LASSO model since

$$\underline{\beta}|\lambda, \sigma^2 \sim N\left(\underline{0}, \frac{\sigma^2}{\lambda} I_k\right).$$

In short, the corresponding Bayesian network will be:



In short, the basic model is:

$$\begin{cases} Y|X, \underline{\beta}, \sigma^2 \sim N(X\underline{\beta}, \sigma^2 I_n) \\ \sigma^2 \sim \pi(\sigma^2) \propto \frac{1}{\sigma^2} \\ \underline{\beta}|\lambda, \sigma^2 \sim N\left(\underline{0}, \frac{\sigma^2}{\lambda} I_k\right) \end{cases} \tag{6}$$

The different models used to compute the posterior of $\underline{\beta}$ and σ^2, the different models will estimate λ differently. A fully Bayesian approach would assign a prior to λ. Conversely, the EBIB approach would estimate it empirically.

3.4.1 Fully Bayesian model

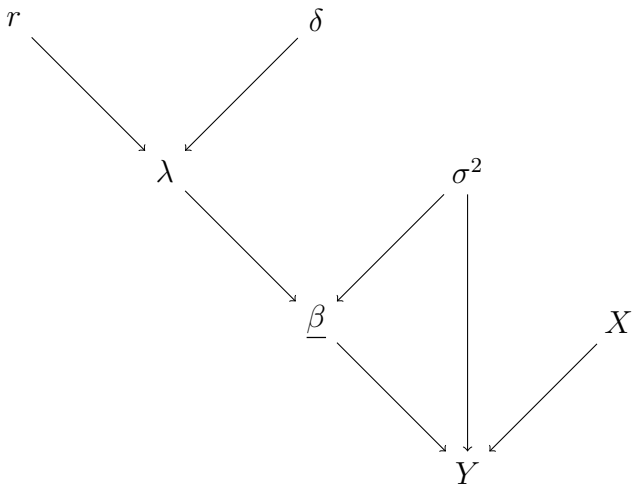
As in section 3.2.1, the standard way to approach this hierarchical model is to assign a prior distribution to λ to model the uncertainty around it. For the conjugacy, the prior assigned to λ is

$$\lambda|X, r, \delta \sim Gamma(\delta; r).$$

It is then possible to extend 6 to:

$$\begin{cases} Y|X, \underline{\beta}, \sigma^2 \sim N(X\underline{\beta}, \sigma^2 I_n) \\ \sigma^2 \sim \pi(\sigma^2) \propto \frac{1}{\sigma^2} \\ \underline{\beta}|\lambda, \sigma^2 \sim N\left(\underline{0}, \frac{\sigma^2}{\lambda} I_k\right) \\ \lambda|X, r, \delta \sim Gamma(\delta; r) \end{cases} \tag{7}$$

The corresponding Bayesian Network will be:



It is now possible to implement the Gibbs sampler to sample from the posterior distributions. The joint posterior distribution of $\underline{\beta}, \sigma^2$ and λ will be

$$
\begin{aligned}
&f(\underline{\beta}, \sigma^2, \lambda | Y, X, r, \delta) = \\
&(2\pi)^{-\frac{n}{2}} |\sigma^2 I_n|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2\sigma^2}(Y - X\underline{\beta})'(Y - X\underline{\beta})\right\} \\
&\times (2\pi)^{-\frac{k}{2}} \left|\frac{\sigma^2}{\lambda} I_k\right|^{-\frac{1}{2}} \exp\left\{-\frac{\lambda}{2\sigma^2}\underline{\beta}'\underline{\beta}\right\} \\
&\times \frac{1}{\sigma^2}\frac{\delta^r}{\Gamma(r)}(\lambda)^{r-1} \exp\left\{-\delta\lambda\right\}.
\end{aligned}
$$

The posterior distributions can be approximated by Gibbs sampling. To implement it, one must the derive the full conditionals. In particular,

- The full conditional distribution $f(\underline{\beta}|\sigma^2, \lambda, Y, X, r, \delta)$ is

$$
\begin{aligned}
&f(\underline{\beta}|\sigma^2, \lambda, Y, X, r, \delta) \\
&\propto \exp\left\{-\frac{1}{2\sigma^2(X'X + \lambda I_k)^{-1}}\left(\underline{\beta}'\underline{\beta} - 2\underline{\beta}'\frac{X'Y}{X'X + \lambda I_k}\right)\right\}.
\end{aligned}
$$

  This is the kernel of $N((X'X + \lambda I_k)^{-1}X'Y; \sigma^2(X'X + \lambda I_k)^{-1})$
  $= N(A^{-1}X'Y; \sigma^2 A^{-1})$ with $A = X'X + \lambda I_k$.

- The conditional distribution $f(\sigma^2|\underline{\beta}, \lambda, Y, X, r, \delta)$ is

$$
\begin{aligned}
&f(\sigma^2|\underline{\beta}, \lambda, Y, X, r, \delta) \\
&\propto (\sigma^2)^{-(\frac{n+k}{2}+1)} \exp\left\{\frac{(Y - X\underline{\beta})'(Y - X\underline{\beta}) + \lambda\underline{\beta}'\underline{\beta}}{2\sigma^2}\right\}.
\end{aligned}
$$

  This is the kernel of $Inv\ Gamma\left(\frac{n+k}{2}; \frac{(Y-X\underline{\beta})'(Y-X\underline{\beta})+\lambda\underline{\beta}'\underline{\beta}}{2}\right)$.

- Finally, the conditional distribution $f(\lambda|\sigma^2, \underline{\beta}, Y, X, r, \delta)$ is

$$f(\lambda|\sigma^2, \underline{\beta}, Y, X, r, \delta) \quad \propto \quad \lambda^{\frac{k}{2}+r-1} \exp\left\{-\lambda\left(\frac{\underline{\beta}'\underline{\beta}}{2\sigma^2}+\delta\right)\right\}.$$

  This is the kernel of $Gamma\left(\frac{k}{2}+r; \delta+\frac{\underline{\beta}'\underline{\beta}}{2\sigma^2}\right)$.

Once the full conditional distributions of the parameters of interest are specified, it is possible to implement the corresponding Gibbs sampler to sample from the posterior distributions. The steps to be followed to obtain it are:

1. Choose a feasible set of starting points $\theta^{(0)} = (\underline{\beta}^{(0)}, (\sigma^2)^{(0)}, (\lambda^2)^{(0)})$.

2. At time t (and so for a current value $\theta^{(t-1)} = (\underline{\beta}^{(t-1)}, (\sigma^2)^{(t-1)}, (\lambda)^{(t-1)})$) new values are sampled, until convergence, from the full conditional distributions

   - $\underline{\beta}^{(t)}$ from $f(\underline{\beta}|(\sigma^2)^{(t-1)}, (\lambda)^{(t-1)})$,
   - $(\sigma^2)^{(t)}$ from $f(\sigma^2|\underline{\beta}^{(t)}, (\lambda)^{(t-1)})$,
   - $(\lambda)^{(t)}$ from $f(\lambda|\underline{\beta}^{(t)}, (\sigma^2)^{(t)})$.

   Given the full conditional distributions previously derived, the Gibbs sampler is obtained by iterative sampling from

   - $\underline{\beta}^{(t)}$ from $N_n(A^{-1}(X'Y); \sigma^2 A^{-1})$,
   - $(\sigma^2)^{(t)}$ from $Inv\ Gamma\left(\frac{n+k}{2}; \frac{(Y-X\underline{\beta})'(Y-X\underline{\beta})+\lambda\underline{\beta}'\underline{\beta}}{2}\right)$,
   - $(\lambda)^{(t)}$ from $Gamma\left(\frac{k}{2}+r; \delta+\frac{\underline{\beta}'\underline{\beta}}{2\sigma^2}\right)$.

The sampler is then obtained in **R** through the code provided in section 5.4.

3.4.2 Empirical Bayes model

In the empirical Bayes estimation, λ is estimated through the expectation-maximization algorithm. This, joint with the Gibbs sampler, implies an MCEM algorithm similar to the one developed in section 3.2.2. This is obtained by adding steps 3 (E-step) and 4 (M-step) to the original Gibbs sampler:

1. Choose a feasible set of starting points $\theta^{(0)} = (\underline{\beta}^{(0)}, (\sigma^2)^{(0)})$ and an initial value for $(\lambda)^{(0)}$.

2. At the time t (and so for a current value $\theta^{(t-1)} = (\underline{\beta}^{(t-1)}, (\sigma^2)^{(t-1)})$ and with $(\lambda)^{(t-1)}$) new values are sampled, until convergence, from the full conditional distributions

   - $\underline{\beta}^{(t)}$ from $f(\underline{\beta}|(\sigma^2)^{(t-1)}, (\lambda)^{(t-1)})$,
   - $(\sigma^2)^{(t)}$ from $f(\sigma^2|\underline{\beta}^{(t)}, (\lambda)^{(t-1)})$.

   Given the full conditional distributions previously derived, the Gibbs sampler is obtained by iterative sampling

   - $\underline{\beta}^{(t)}$ from $N_n(A^{-1}(X'Y); \sigma^2 A^{-1})$,
   - $(\sigma^2)^{(t)}$ from $Inv\ Gamma\left(\frac{n+k}{2}; \frac{(Y-X\underline{\beta})'(Y-X\underline{\beta})+\lambda\underline{\beta}'\underline{\beta}}{2}\right)$,

3. (Expectation step) The expected log-likelihood for $\frac{k}{2}\log\lambda - \frac{\lambda}{2}E\left[\frac{\underline{\beta}'\underline{\beta}}{\sigma^2}\right]$. Since the term $E\left[\frac{\underline{\beta}'\underline{\beta}}{\sigma^2}\right]$ is unknown, it will be replaced with the average of the corresponding Gibbs sampler extractions conditionally on the data X and Y and the previous value $\lambda^{(t-1)}$: $E\left[\frac{\underline{\beta}'\underline{\beta}}{\sigma^2}\middle| X, Y, \lambda^{(t-1)}\right]$

4. (Maximization step) The approximation of the expected log-likelihood is maximized. The closed-form solution obtained will be the new value for $\lambda^{(t)}$:

$$\begin{aligned}\lambda^{(t)} &= \arg\max_{\lambda}\left(\frac{k}{2}\ln(\lambda)-\frac{\lambda}{2}E\left[\frac{\underline{\beta}'\underline{\beta}}{\sigma^2}|X,Y,\lambda^{(t-1)}\right]\right)\\ &= \frac{k}{E\left[\frac{\underline{\beta}'\underline{\beta}}{\sigma^2}|X,Y,\lambda^{(t-1)}\right]}\end{aligned}$$

5. Repeat steps 2, 3, and 4 until convergence is reached

The sampler is then obtained in **R** through the code provided in section 5.5.

3.4.3 A real data application

The Gibbs sampler outlined in sections 3.3.2 and 3.4.1 and their corresponding implementations described in sections 5.4 and 5.5 are again applied to the standardized Diabetes dataset described in section 3.2.3.

To implement the Gibbs sampler for the fully Bayesian model, the prior of λ is set to be a $Gamma(1,2) = Exp(2)$ as presented in figure 13.

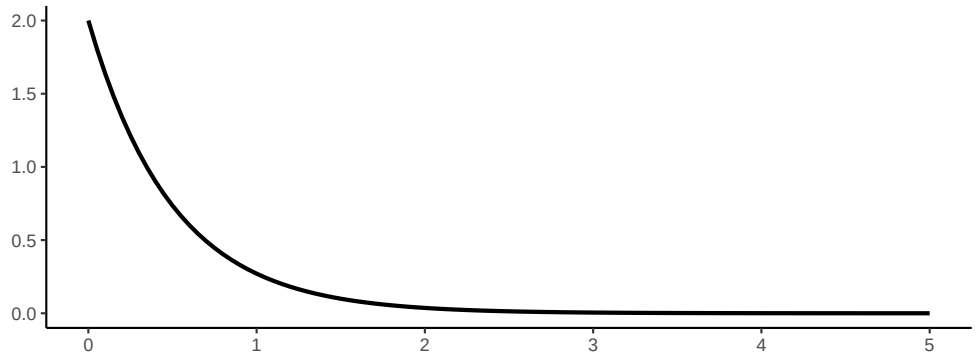


Figure 13: Prior distribution of $\lambda \sim Exp(2)$

The implementation of the Gibbs sampler with 10.000 iterations and 1.000 burn-in gives MCMC samples from the posterior distribution of λ as presented in Figure 14. For replicability, the **R** code is run setting the seed "123".

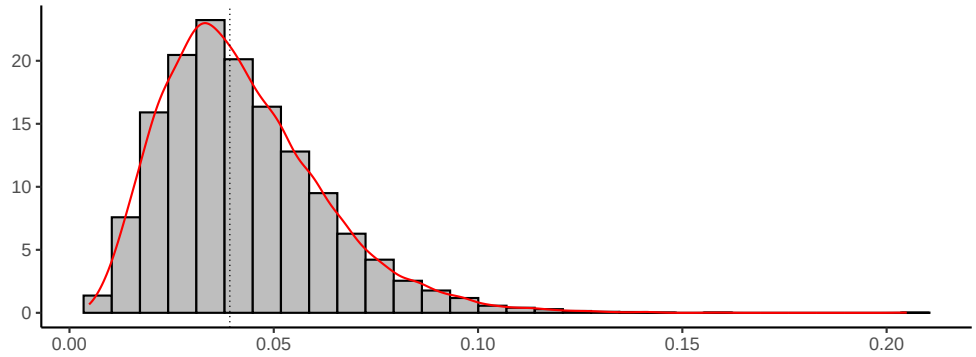


Figure 14: Samples from the posterior distribution of λ resulting from the Gibbs sampler of the fully Bayesian model. The vertical dotted line represents the median value of the extractions from the posterior distribution of λ. Its value is of 0.03924627.

The implementation of the Gibbs sampling for the EB model does not require a prior specification of λ that is instead obtained through the Expectation-Maximization algorithm. Its iterations are shown in figure 15. For replicability, the **R** command is run setting the seed "123".

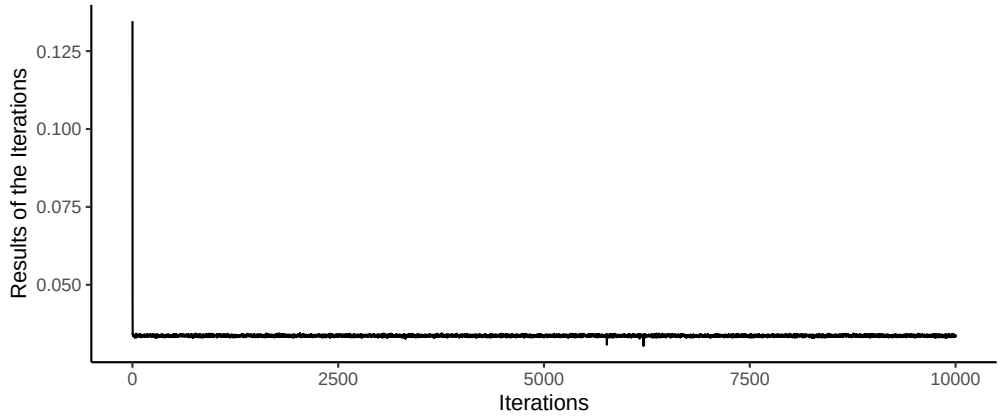


Figure 15: Evolution of the $\lambda^{(t)}$ through the EM-algorithm with an initial $\lambda^{(0)} = 1$

The implementation of the Gibbs sampler with 10.000 iterations and 1.000 burn-in gives MCMC samples from the posterior distribution of λ as presented in Figure 16. For replicability, the **R** code is run setting the seed "123".

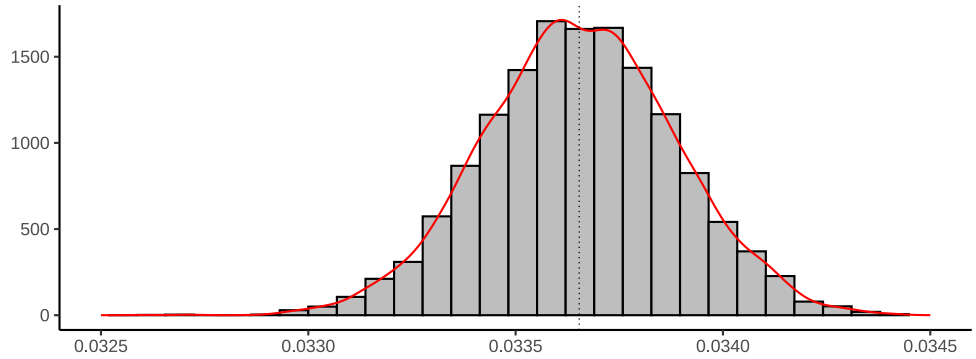


Figure 16: Samples from the posterior distribution of λ resulting from the Gibbs sampler of the fully Bayesian model. The vertical dotted line represents the median value of the extractions from the posterior distribution of λ. Its value is 0.03365336.

To give a term of comparison, the EB approach will suggest a lower shrinkage compared to the fully Bayesian one. The differences between the two models can be seen in Figure 17.

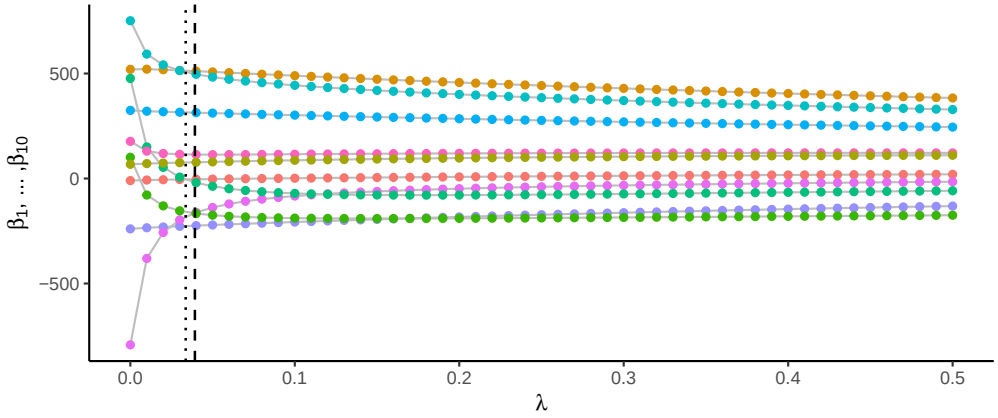


Figure 17: The horizontal and curved lines represent the evolution of the $\underline{\beta}$ coefficients as the shrinkage parameter increases. The dotted vertical line represents the shrinkage value suggested by the Empirical Bayes approach. The dashed vertical line represents the shrinkage value suggested by the classic Bayesian approach.

3.5 Simulation study: Sparse regression with Bayesian RIDGE

The simulation study conducted in this section aims to explore examples where the strong merging may fail. Indeed, if $\underline{\beta}_{true} = \underline{0}$, then $\sup_{\lambda} f(\underline{\beta}_{true}|\lambda, \sigma^2) \to \infty$ implying a possible failure of strong merging. We will consider models similar to those discussed in section 3.3 but, as argued, we expect different results. The example presented in section 3.5.1 explores the regular case of $\underline{\beta}_{true} \neq \underline{0}$ where strong merging holds. In contrast, the second example of the section 3.5.2 explores the degenerate case of $\underline{\beta}_{true} = \underline{0}$ where merging may be unsuccessful. Finally, the third example of the section 3.5.3 investigates

the mixed case of only a subset of true coefficients equal to 0.

In this section, we will compare the EBIB posteriors with those resulting from a Bayes model with λ fixed and the oracle Bayes posterior. The posterior of the Bayes model with λ fixed is obtained through the hierarchical model (6) just setting $\lambda = 3$:

$$\begin{cases} Y|X, \underline{\beta}, \sigma^2 \sim N(X\underline{\beta}, \sigma^2 I_n) \\ \sigma^2 \sim \pi(\sigma^2) \propto \frac{1}{\sigma^2} \\ \underline{\beta}|\lambda, \sigma^2 \sim N\left(\underline{0}, \frac{\sigma^2}{\lambda} I_k\right) \\ \lambda = 3 \end{cases} \tag{8}$$

Even in this case, the Oracle Bayes approach starts from 6 and assigns a fixed value to λ. The difference lies in how the oracle value λ is determined. Since $f(\beta_{true}|\lambda) = \prod_{i=1}^{k} \frac{1}{\sqrt{2\frac{\sigma^2}{\lambda}}} e^{-\frac{\lambda\beta_j^2}{2\sigma^2}}$, the oracle value is:

$$\lambda^* = \arg\max_{\lambda} f(\beta_{\text{true}}|\lambda) = \arg\max_{\lambda} \sum_{j=1}^{k} \frac{1}{2}\log(\lambda) - \frac{\lambda\beta_j^2}{2\sigma^2} = \frac{k\sigma^2}{\sum_{j=1}^{k}\beta_j^2}$$

The corresponding hierarchical model obtained will be of the form:

$$\begin{cases} Y|X, \underline{\beta}, \sigma^2 \sim N(X\underline{\beta}, \sigma^2 I_n) \\ \sigma^2 \sim \pi(\sigma^2) \propto \frac{1}{\sigma^2} \\ \underline{\beta}|\lambda, \sigma^2 \sim N\left(\underline{0}, \frac{\sigma^2}{\lambda} I_k\right) \\ \lambda = \frac{k\sigma^2}{\sum_{j=1}^{k}\beta_j^2} \end{cases} \tag{9}$$

The simulation of the data structures will follow section 3.3.3.

3.5.1 Example 5: Regular case

The data structure in this example is characterized by no coefficient equal to 0. To ensure comparability, the data structure as the one of 3.3.4. Therefore, the true parameters are set to be $\beta_{true} = (\beta_1, \beta_2, \beta_3, \beta_4) = (-1, 1, 3, 2)$. With this data structure, the EBIB posterior will be compared to the actual Bayesian, and the Oracle Bayes posteriors. The shrinkage parameter in the actual Bayesian model is set to 3. The model based on an oracle prior will set $\lambda^* = \frac{k\sigma^2}{\sum_{j=1}^{k}\beta_j^2} = \frac{4}{15}$. This results in a lower shrinkage compared to the Bayesian alternative, which selects $\lambda = 3$. Finally, the EBIB approach will set λ_{EBIB} through the Expectation-Maximization algorithm detailed in section 3.4.2. This algorithm will return a median value of 0.165447. Therefore, the Bayesian estimation with $\lambda = 3$ will shrink the most the posterior distribution and, since the true coefficients are different from 0, it will have a flatter posterior. The EBIB posterior will instead be closer to the oracle posterior since the shrinkage coefficient is closer to the oracle value of 4/15. Finally, the oracle posterior will be the most spiked around the true coefficients. This is because the true value of the coefficients is different from 0 and the likelihood of the data reflects that. The posterior distributions of each coefficient resulting from the three models are shown in figure 18.

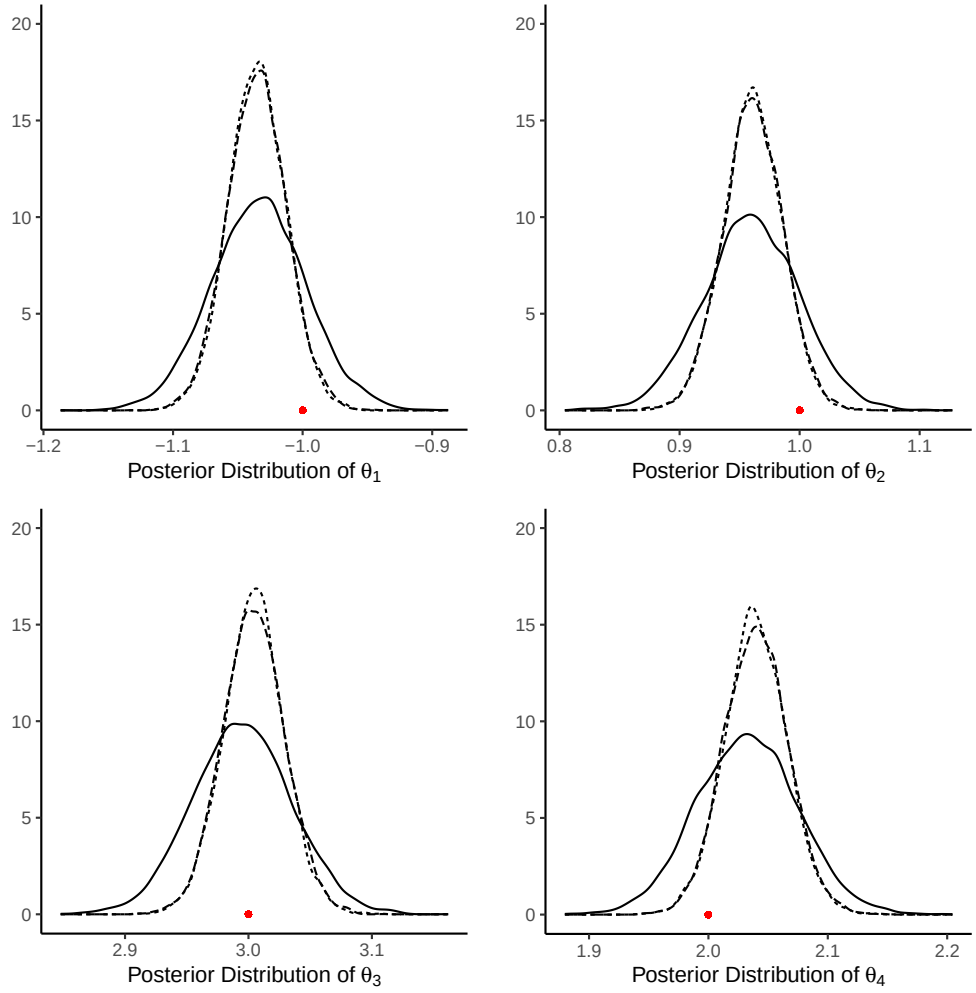


Figure 18: Example 5. The solid line represents the Bayesian posterior with $\lambda = 3$. The dotted line represents the Empirical Bayes posterior with $\hat{\lambda} \approx 0.165$ estimated through the EM algorithm. The dashed line represents the oracle Bayes posterior with $\lambda^* = 4/15$

3.5.2 Example 6: Degenerate case

In this example, I consider a data structure consisting of only non-relevant coefficients so that $\underline{\beta} = (0, 0, 0, 0)$. This type of dataset may prevent strong

merging. This is because $\sup_{\lambda} f(\underline{\beta}_{true}|\lambda, \sigma^2) \to \infty$ and $\lambda^* = \frac{k\sigma^2}{\sum_{j=1}^{k} \beta_j^2} \to \infty$. In other words, the oracle shrinkage parameter will go to ∞ since, by doing so, it will shrink all the estimates towards 0: their true value. This is the equivalent of having as prior distribution a degenerate one in 0 so that, $\forall j \in \{1, ..., k\}$ it is

$$f(\beta_{j,\text{true}}|\lambda, \sigma^2) = \begin{cases} \infty & \text{if } \beta_j = 0 \\ 0 & \text{otherwise} \end{cases}. \tag{10}$$

As in the previous examples, the actual Bayesian approach will set $\lambda = 3$. Finally, the EBIB model will set λ_{EBIB} through the Expectation-Maximization algorithm specified in section 3.4.2. In this case, the EM algorithm returns a sequence of iterations with high variability and values higher than the previous examples. In this case, the oracle λ^* leads to a degenerate prior for $\beta_1, ..., \beta_4$. The EM algorithm used to identify λ^{EBIB} will still converge to this value. Therefore, the EBIB posterior will not approximate any Bayesian posterior.

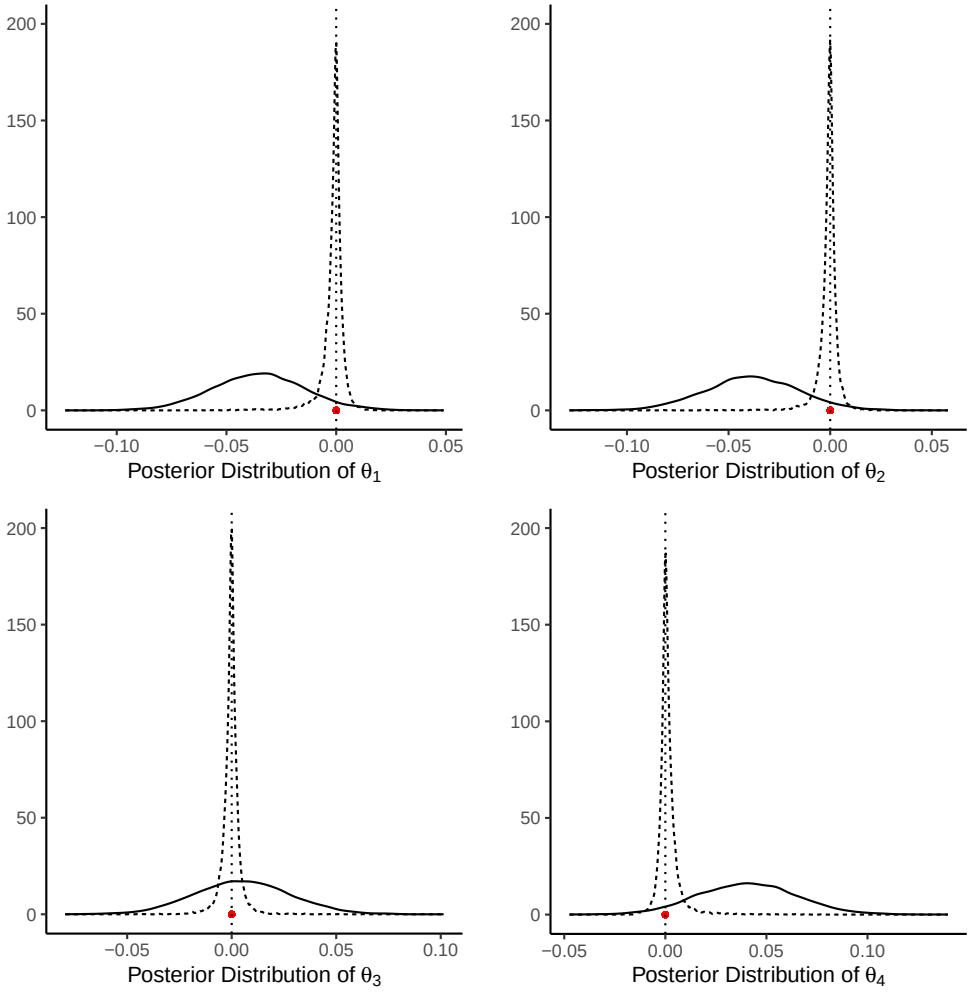


Figure 19: Example 6. The solid line represents the Bayesian posterior with $\lambda = 3$. The dotted line represents the Empirical Bayes posterior. The dashed vertical line represents the oracle Bayes posterior with $\lambda^* = \infty$

3.5.3 Example 7: Mixed case

In this example, the data structure used will be mixed. That is, it will be composed of some coefficients equal to 0 and some others are set to be different. For comparability, the same dataset of 3.3.6 is used. In this case, the set

of coefficients chosen is set to be $\underline{\beta}_{true} = (\beta_1, ..., \beta_{20}) = (0, ..., 0, 1)$. With this data structure, we expect (based on the results of 2014), that merging will hold only for the posterior of the coefficient different from 0. This is because $\sup_{\lambda} f(\beta_{1,...,19}|\lambda, \sigma^2) \to \infty$ while $\sup_{\lambda} f(\beta_{20}|\lambda, \sigma^2) < \infty$. In this case, a high value of the oracle λ^* is expected since it would improve the performance of 19 estimates out of 20. Still, unlike the previous example the denominator won't go to 0 and so $\lambda^* < \infty$. More in detail, the closed form solution derived for the RIDGE model leads us to $\lambda^* = \frac{k\sigma^2}{\sum_{j=1}^{k} \beta_j^2} = 20$. As in the previous examples, the actual Bayesian approach will set $\lambda = 3$. Finally, the EBIB model will set λ_{EBIB} through the Expectation-Maximization algorithm specified in section 3.4.2. This algorithm will return a median value of 16.02616. Therefore, the Bayesian posterior will shrink the Bayesian posterior with $\lambda = 3$ will shrink the Bayesian posterior the least. Conversely, the oracle posterior and the EBIB posterior will have a similar level of shrinkage. Therefore, as shown in figure 20, their posteriors will be closer to each other.

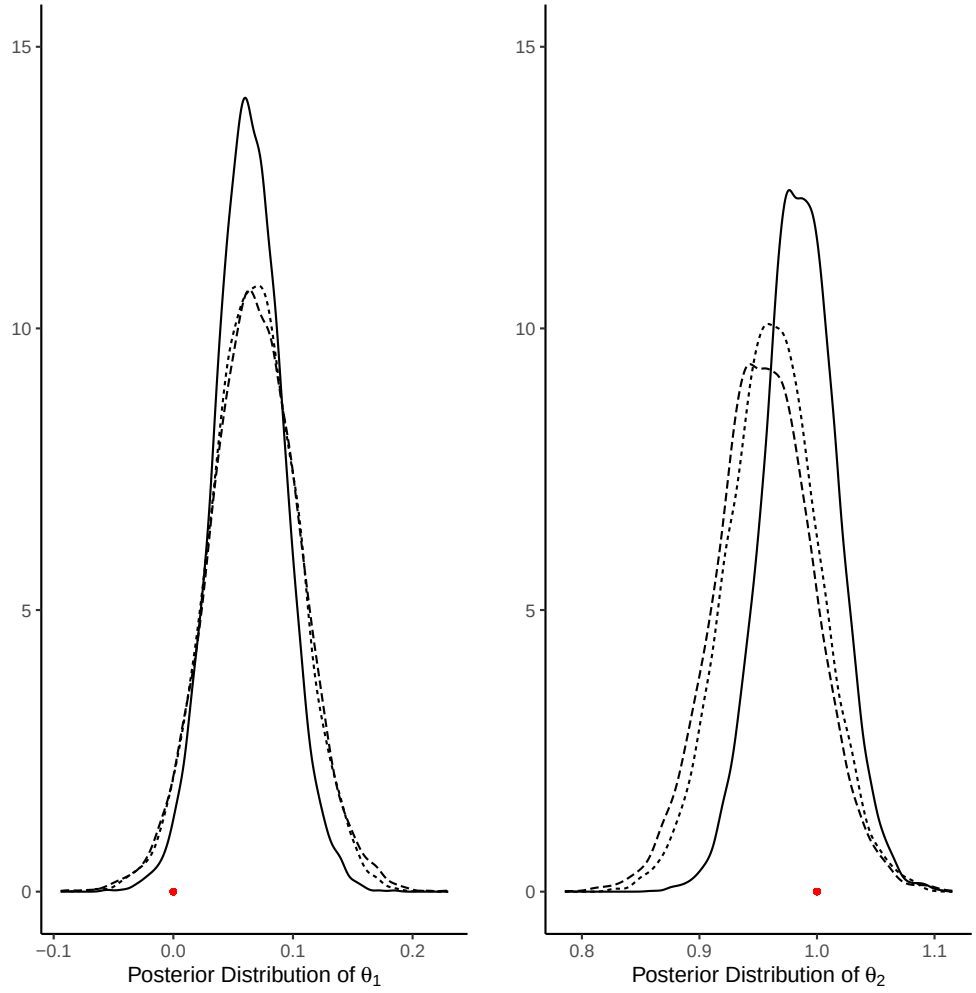


Figure 20: Example 7. The solid line represents the Bayesian posterior with $\lambda = 3$. The dotted line represents the Empirical Bayes posterior with $\hat{\lambda} \approx 16.026$ estimated through the EM algorithm. The dashed line represents the oracle Bayes posterior with $\lambda^* = 20$

This case is related to the example described in section 3.3.6. Indeed, the structure of the hierarchical model with a common shrinkage parameter implies that it is enough to have one coefficient different from 0 to avoid degen-

eracy. Yet, it is possible to structure the hierarchical model with coefficient-specific hyperparameters for the prior so that $\beta_j \sim N(0, \tau_j)$. In this case, the EBIB posteriors of the coefficients whose true value would be 0 will degenerate while the others won't.

4 Conclusions

This thesis explored two classes of approaches that are both referred as "Empirical Bayes" in the literature, although they are, in essence, very different: the classic empirical Bayes approach as paved by Robbins and Efron, and the so-called 'empirical Bayes in Bayes (EBIB)'. Methodologically, the difference depends on the data structure. For datasets consisting of a large set of parallel experiments, with the same statistical model but experiment-specific parameters as discussed in section 1.2, the classic EB is appropriate. Conversely, the EBIB approach is used when dealing with repeated sampling from the same conditional model, with no frequentist randomness of the model's parameters as outlined in section 3.1. In practice, the implementation of the parametric EB and the EBIB approaches is similar. Both estimate a hyperparameter $\hat{\lambda}$. However, the different data structures imply differences in the estimation procedure and the interpretation of λ itself. On one hand, the parametric EB estimates using the observations from parallel experiments. Therefore, the latent distribution is a means to transfer information from apparently unrelated observations. Conversely, the EBIB approach estimates λ using the observations coming from the same model.

In the classic EB approach, a true latent distribution exists, and one samples

$\theta_1, ..., \theta_k$ from it. Consequently, the dataset itself is characterized by a superstructure similar to what a Bayesian statistician would implement to model uncertainty. The latent distribution is then used as a substitute for the prior despite the two objects being inherently different. The latent distribution that plays the role of the prior in a specific experiment is the reason why the EB approach, in its early stages, attracted criticism from a frequentist perspective, and the EB estimation of the prior from the data is questionable from a Bayesian perspective. As noted, "the EB approach is essentially non-Bayesian in the sense of not involving subjective probabilities" (Maritz, 2018). The subsequent "philosophical identity problem" (Efron, 2019) is, however, compensated by the numerous advantages it offers. As summarized in the paper "Innovations in Bayes and empirical Bayes methods: estimating parameters, populations and ranks" (Louis and W. Shen, 1999) "Empirical Bayes methods can produce more valid, efficient and informative statistical evaluations than those based on traditional methods" . Those advantages are however limited to the peculiar dataset this approach is bound to.

The framework of the EBIB approach is fully Bayesian since there does not exist a true latent distribution; the prior distribution models the incomplete information around θ. This approach is commonly used, but is questionable since its estimation procedure would imply "using the data twice" (Carlin and Louis, 2000): first, for the prior estimation and a second time in the likelihood, to compute the posterior distribution. On the other hand, many complex and computational issues force an empirical estimation of the hyperparameter. This contributes to the widespread use of the EBIB approach. In the second part of the thesis, I conducted multiple simulation studies to

explore the properties of the EBIB posterior distribution. As proved in recent work (Petrone, Rousseau, and Scricciolo, 2014), (Petrone, Rizelli, and Rousseau, 2024), the EBIB posterior is a fast approximation of the oracle posterior, which is the posterior distribution of the Bayesian researcher that uses the given class of priors and has the most information on the true θ. This higher-order merging property was explored in sparse regression through Bayesian-LASSO and in a Bayesian-RIDGE model. The results of this extensive simulation study confirmed the theoretical behaviour of the EBIB posterior distribution, but also provide more detailed hints on the rate of convergence. In the cases observed, the EBIB posterior is shown to be a fast approximation of the oracle posterior. In non-degenerate cases, it asymptotically merges with the "corresponding" fully Bayesian posterior distribution; however, for small and moderate sample size, it is shown to actually approximate the oracle Bayesian posterior density, which is restricted to the *given class of priors*. In this sense, the EBIB posterior density is not to be compared with the 'fully Bayesian' posterior density, which uses a different and more structured hierarchical prior. The deeper problem of choosing the form of the prior is a fundamental problem in Bayesian statistics, but it is beyond this study.

5 R Code

This section provides the R code built and used to run the Gibbs samplers and the augmented Gibbs samplers of the section 3.

5.1 Hiearchical Bayesian LASSO, fully Bayes: Gibbs sampling

The R code for the Gibbs sampler from the posterior distribution in the hierarchical Bayesian LASSO with a 'fully Bayesian' prior specification in section 3.2.1 is

```
Gibbs_sampler_1 = function(n_iterations, burn_in, X, Y, r, delta){
  n = length(X[,1])
  k = length(X[1,])

  beta = matrix(data = NA, nrow = n_iterations, ncol = k)
  sigma_2 = matrix(data = NA, nrow = n_iterations, ncol = 1)
  tau_2 = matrix(data = NA, nrow = n_iterations, ncol = k)
  lambda_2 = rep(NA, n_iterations)

  beta[1,] = runif(length(X[1,]))
  sigma_2[1] = runif(1)
  tau_2[1,] = runif(length(X[1,]))
  lambda_2[1] = runif(1)

  for (i in 2:n_iterations){
    #Extraction of the Beta coefficient from its full conditional
    Diagonal_taus = diag(tau_2[i-1,], nrow = length(X[1,]), ncol =
    ↪  length(X[1,]))
    A = t(X) %*% X + inv(Diagonal_taus)
    mean = inv(A) %*% t(X) %*% (Y)
    var = sigma_2[i-1] * inv(A)
    var[lower.tri(var)] = t(var)[lower.tri(var)]
    beta[i,] = rmvnorm(1, mean, var)

    #Extraction of the Sigma coefficient from its full conditional
    shape_parameter = (n+k)/2
```

```
    scale_parameter = (t(Y - (X %*% beta[i,])) %*% (Y - (X %*%
    ↪  beta[i,])) + t(beta[i,]) %*% inv(Diagonal_taus) %*%
    ↪  beta[i,])/2
    sigma_2[i] = rinvgamma(1, shape_parameter, scale_parameter)

    #Extraction of the Taus coefficient from its full conditional
    for (j in 1:k){
    parameter_1 = sqrt((lambda_2[i-1] * sigma_2[i]) /
    ↪  ((beta[i,j])^2))
    parameter_2 = lambda_2[i-1]
    tau_2[i,j] = (rinvGauss(1, parameter_1, parameter_2))^(-1)
    }

    #Extraction of the Lambda coefficient from its full
    ↪  conditional
    lambda_2[i] = rgamma(1, k+r, sum(tau_2[i,])/2 + delta)
  }
  #Out of all the coefficients extracted the Lambdas and the Betas
  ↪  are
  #returned. The result will then be a list composed of a vector
  ↪  of Lambda
  #and a matrix of Betas.
  result = list(sqrt(lambda_2[burn_in:n_iterations]),
  ↪  beta[burn_in:n_iterations,])
  return(result)
}
```

5.2 Hiearchical Bayesian LASSO, empirical Bayes: EM within Gibbs sampling

The Gibbs sampler from the posterior distribution in the hierarchical Bayesian LASSO with an 'empirical Bayesian' prior specification in section 3.4.2 is implemented through three separate functions in **R**. The first one is used to obtain the term $\lambda^{(t)} = \sum_{i=1}^{k} E[(\tau_i^2)^{(t)}|X, Y, \lambda^{(t-1)}]$. This is done through an additional Gibbs sampler that is run with fixed $\lambda = \lambda^{(t-1)}$ to obtain the posterior distributions for $\tau_1^2, ..., \tau_k^2$ which, in turn, are used to approximate

the E-step in the running EM algorithm.

```
EM_1 = function(n_iterations, burn_in, X, Y, lambda) {
  n = nrow(X)
  k = ncol(X)

  beta = matrix(data = NA, nrow = n_iterations, ncol = k)
  sigma_2 = numeric(n_iterations)
  tau_2 = matrix(data = NA, nrow = n_iterations, ncol = k)

  beta[1, ] = runif(k)
  sigma_2[1] = runif(1)
  tau_2[1, ] = runif(k)


  for (i in 2:n_iterations) {
    # Beta
    Diagonal_taus = diag(tau_2[i - 1, ], nrow = k, ncol = k)
    A = t(X) %*% X + solve(Diagonal_taus)
    mean = solve(A) %*% t(X) %*% Y
    var = sigma_2[i - 1] * solve(A)
    var[lower.tri(var)] = t(var)[lower.tri(var)]
    beta[i, ] = rmvnorm(1, mean, var)

    # Sigma
    shape_parameter = (n + k) / 2
    scale_parameter = (t(Y - (X %*% beta[i, ])) %*% (Y - (X %*%
    ↪  beta[i, ])) + t(beta[i, ]) %*% solve(Diagonal_taus) %*%
    ↪  beta[i, ]) / 2
    sigma_2[i] = rinvgamma(1, shape_parameter, scale_parameter)

    # Taus
    for (j in 1:k) {
      parameter_1 = sqrt((lambda^2 * sigma_2[i]) / ((beta[i,
      ↪  j])^2))
      parameter_2 = lambda^2
      extraction = rinvGauss(1, parameter_1, parameter_2)

      while (extraction <= 0) {
        extraction <- rinvGauss(1, parameter_1, parameter_2)
        }

      tau_2[i, j] = extraction^(-1)
      }
```

```
  }
  result = mean(tau_2[burn_in:n_iterations, ])
  return(result)
}
```

This procedure is then iterated for a number of times equal to the extractions. By doing so, one obtains the vector $\lambda^{(1)}, ..., \lambda^{(10.000)}$ of iterations of the EM algorithm. The corresponding **R** code will be:

```
EM_2 = function(n_iterations_1, n_iterations_2, X, Y,
↪  initial_lambda) {
  k = ncol(X)
  LAMBDA = numeric(n_iterations_1)
  LAMBDA = matrix(NA, nrow = n_iterations_1, ncol = 1)
  LAMBDA[1] = sqrt((2) / EM_1(n_iterations_2, 0, X = X, Y = Y,
  ↪  lambda = initial_lambda))

  for (i in 2:n_iterations_1) {
    denominator = EM_1(n_iterations_2, 0, X = X, Y = Y, lambda =
    ↪  LAMBDA[i - 1])
    LAMBDA[i] = sqrt(2/mean(denominator))
  }
  return(LAMBDA)
}
```

To obtain the extractions from the posterior distributions of interest, an additional Gibbs sampler is run. In this case, the iterations of λ are not sampled from its full conditional but the iterations of the EM algorithm are considered. The corresponding **R** code will be:

```
Complementary_Gibbs_Sampler = function(X, Y, evolution_lambda,
↪  burn_in) {
  k = ncol(X)
  n = nrow(X)
  n_iterations_3 = length(evolution_lambda)

  beta = matrix(data = NA, nrow = n_iterations_3, ncol = k)
```

```
  sigma_2 = matrix(data = NA, nrow = n_iterations_3, ncol = 1)
  tau_2 = matrix(data = NA, nrow = n_iterations_3, ncol = k)

  beta[1, ] = runif(k)
  sigma_2[1] = runif(1)
  tau_2[1, ] = runif(k)

  for (i in 2:n_iterations_3) {
    # Beta
    Diagonal_taus = diag(tau_2[i - 1, ], nrow = k, ncol = k)
    A = t(X) %*% X + solve(Diagonal_taus)
    mean = solve(A) %*% t(X) %*% Y
    var = sigma_2[i - 1] * solve(A)
    var[lower.tri(var)] = t(var)[lower.tri(var)]
    beta[i, ] = rmvnorm(1, mean, var)

    # Sigma
    shape_parameter = (n + k) / 2
    scale_parameter = (t(Y - (X %*% beta[i, ])) %*% (Y - (X %*%
    ↪  beta[i, ])) + t(beta[i, ]) %*% solve(Diagonal_taus) %*%
    ↪  beta[i, ]) / 2
    sigma_2[i] = rinvgamma(1, shape_parameter, scale_parameter)

    # Taus
    for (j in 1:k) {
      parameter_1 = sqrt((evolution_lambda[i - 1]^2 * sigma_2[i])
      ↪  / ((beta[i, j])^2))
      parameter_2 = evolution_lambda[i - 1]^2
      tau_2[i, j] = (rinvGauss(1, parameter_1, parameter_2))^(-1)
    }
  }
  return(beta[burn_in:n_iterations_3, ])
}
```

5.3 Hiearchical Bayesian LASSO, Bayesian model with fixed λ and oracle posterior

To obtain the posterior distributions outlined in sections 3.3.1 and 3.3.2 the complementary Gibbs sampler 5.2 is used. However, in these cases the

iterations of λ used are not the iterations of the EM algorithm. Indeed, to obtain the oracle posterior, the vector $(\lambda^*, ..., \lambda^*)$ is used with λ^* being the oracle value of λ. A similar approach but with a different value is used for the Bayesian model with fixed λ.

5.4 Hiearchical Bayesian RIDGE, fully Bayes: Gibbs sampling

The **R** code for the Gibbs sampler from the posterior distribution in the hierarchical Bayesian RIDGE with a 'fully Bayesian' prior specification in section 3.4.1 is:

```
Gibbs_sampler_4 = function(n_iterations, burn_in, X, Y, r, delta){
  n = length(X[,1])
  k = length(X[1,])

  beta = matrix(data = NA, nrow = n_iterations, ncol = 10)
  sigma_2 = matrix(data = NA, nrow = n_iterations, ncol = 1)
  lambda = matrix(data = NA, nrow = n_iterations, ncol = 1)

  beta[1,] = rep(1, length(X[1,]))
  sigma_2[1] = 1
  lambda[1] = 1

  for (i in 2:n_iterations){
    #Beta
    A = t(X) %*% X + lambda[i-1] * diag(1, nrow = k, ncol = k)
    mean = inv(A) %*% t(X) %*% (Y)
    var = sigma_2[i-1] * inv(A)
    beta[i,] = rmvnorm(1, mean, var)

    #Sigma
    shape_parameter = (n+k)/2
    scale_parameter = (t(Y - (X %*% beta[i,])) %*% (Y - (X %*%
    ↪  beta[i,])) + lambda[i-1,] * t(beta[i,]) %*%  beta[i,])/2
    sigma_2[i] = rinvgamma(1, shape_parameter, scale_parameter)

```

```
    #Lambda
    lambda[i] = rgamma(1, (k/2)+r, (t(beta[i,]) %*% beta[i,])/(2 *
    ↪  sigma_2[i]) + delta)
  }
  result = list(lambda[burn_in:n_iterations],
  ↪  beta[burn_in:n_iterations,])
  return(result)
}
```

5.5 Hiearchical Bayesian RIDGE, empirical Bayes: EM within Gibbs sampling

The Gibbs sampler from the posterior distribution in the hierarchical Bayesian RIDGE with an 'empirical Bayesian' prior specification in section 3.2.1 is implemented through three separate functions in **R**. The first one is used to obtain the term $\lambda^{(t)} = E\left[\frac{\beta'\beta}{\sigma^2}|X, Y, \lambda^{(t-1)}\right]$. This is done through an additional Gibbs sampler that is run with fixed $\lambda = \lambda^{(t-1)}$.

```
EM_1_ridge = function(n_iterations, burn_in, X, Y, lambda) {
  n = nrow(X)
  k = ncol(X)

  beta = matrix(data = NA, nrow = n_iterations, ncol = k)
  sigma_2 = numeric(n_iterations)

  beta[1, ] = runif(k)
  sigma_2[1] = runif(1)


  for (i in 2:n_iterations) {
    # Beta
    A = t(X) %*% X + lambda * diag(1, nrow = k, ncol = k)
    mean = inv(A) %*% t(X) %*% (Y)
    var = sigma_2[i-1] * inv(A)
    beta[i,] = rmvnorm(1, mean, var)

    # Sigma
```

```
    shape_parameter = (n + k) / 2
    scale_parameter = (t(Y - (X %*% beta[i, ])) %*% (Y - (X %*%
    ↪  beta[i, ])) + lambda * t(beta[i, ]) %*% beta[i, ]) / 2
    sigma_2[i] = rinvgamma(1, shape_parameter, scale_parameter)

  }
  result = colMeans(beta[burn_in:n_iterations,
  ↪  ]*beta[burn_in:n_iterations, ]/sigma_2)
  return(result)
}
```

This procedure is then iterated for a number of times equal to the extractions. By doing so, one obtains the vector $\lambda^{(1)}, ..., \lambda^{(10.000)}$ of iterations of the EM algorithm. The corresponding **R** code will be:

```
EM_2_ridge = function(n_iterations_1, n_iterations_2, X, Y,
↪  initial_lambda) {
  k = ncol(X)
  LAMBDA = numeric(n_iterations_1)
  LAMBDA = matrix(NA, nrow = n_iterations_1, ncol = 1)
  result = EM_1_ridge(n_iterations_2, 0, X = X, Y=Y, lambda =
  ↪  initial_lambda)
  LAMBDA[1] = k/sum(result)

  for (i in 2:n_iterations_1) {
    result = EM_1_ridge(n_iterations_2, 0, X = X, Y = Y, lambda =
    ↪  LAMBDA[i - 1])
    LAMBDA[i] = k/sum(result)
    print(i)
  }
  return(LAMBDA)
}
```

To obtain the extractions from the posterior distributions of interest, an additional Gibbs sampler is run. In this case, the iterations of λ are not sampled from its full conditional but the iterations of the EM algorithm are considered. The corresponding **R** code will be:

```r
Complementary_Gibbs_Sampler_RIDGE = function(X, Y,
↪  evolution_lambda, burn_in) {
  k = ncol(X)
  n = nrow(X)
  n_iterations_3 = length(evolution_lambda)

  beta = matrix(data = NA, nrow = n_iterations_3, ncol = k)
  sigma_2 = matrix(data = NA, nrow = n_iterations_3, ncol = 1)

  beta[1, ] = runif(k)
  sigma_2[1] = runif(1)

  for (i in 2:n_iterations_3) {
    # Beta
    A = t(X) %*% X + evolution_lambda[i - 1] * diag(1, nrow = k,
    ↪  ncol = k)
    mean = solve(A) %*% t(X) %*% Y
    var = sigma_2[i - 1] * solve(A)
    var[lower.tri(var)] = t(var)[lower.tri(var)]
    beta[i, ] = rmvnorm(1, mean, var)

    # Sigma
    shape_parameter = (n + k) / 2
    scale_parameter = (t(Y - (X %*% beta[i, ])) %*% (Y - (X %*%
    ↪  beta[i, ])) + evolution_lambda[i - 1] * t(beta[i, ]) %*%
    ↪  beta[i, ]) / 2
    sigma_2[i] = rinvgamma(1, shape_parameter, scale_parameter)

  }
  return(beta[burn_in:n_iterations_3, ])
}
```

5.6 Hiearchical Bayesian RIDGE, Bayesian model with fixed λ and oracle posterior

To obtain the posterior distributions outlined in section 3.5, the complementary Gibbs sampler 5.5 is used. However, in these cases, the iterations of λ

used are not the iterations of the EM algorithm. Indeed, to obtain the oracle posterior, the vector $(\lambda^*, ..., \lambda^*)$ is used with λ^* being the oracle value of λ. A similar approach but with a different value is used for the Bayesian model with fixed λ.

References

Brillinger, David R. (2002). "John Wilder Tukey (1915–2000)". In: *Notices Of The American Mathematical Society – AMS* 49.2, pp. 193–201.

Carlin, Bradley P. and Thomas A. Louis (2000). "Empirical Bayes: Past, present and future". In: *Journal of the American Statistical Association* 95.452, pp. 1286–1289.

Carter, Grace M. and John E. Rolph (1974). "Empirical Bayes methods applied to estimating fire alarm probabilities". In: *Journal of the American Statistical Association* 69.348, pp. 880–885.

Casella, George (1985). "An Introduction to Empirical Bayes Data Analysis". In: *The American Statistician* 39.2, pp. 83–87.

— (2001). "Empirical Bayes Gibbs sampling". In: *Biostatistics* 2.4, pp. 485–500.

Deely, Jhon and Dennis Lindley (1981). "Bayes Empirical Bayes". In: *Journal of the American Statistical Association* 76.376, pp. 833–841.

Efron, Bradley (2010). *Large-Scale inference: Empiricl Bayes Methods for Estimation, Testing and Prediction*. Stanford University Press.

— (2014). "Two Modeling Strategies for Empirical Bayes Estimation". In: *Statistical Science* 29, pp. 285–301.

Efron, Bradley (2019). "Bayes, oracle Bayes and empirical Bayes". In: *Statistical science* 34.2, pp. 177–201.

Efron, Bradley and Carl Morris (1973). "Combining possibly related estimation problems". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 35.3, pp. 379–402.

Fisher, Ronald A., A. Steven Corbet, and Carrington B. Williams (1943). "The Relation Between the Number of Species and the Number of Individuals in a Random Sample of an Animal Population". In: *The Journal of Animal Ecology* 12.1, pp. 42–58.

Gelman, Andrew et al. (2007). *Bayesian Data analysis. Third edition.*

Ghosh, Jayanta K., Mohan Delampady, and Tapas Samanta (2007). *An introduction to Bayesian analysis. Theory and methods.* Springer.

James, William and Charles Stein (1960). *Estimation with quadratic loss.*

Levine, Richard A. and George Casella (2001). "Implementations of the Monte Carlo EM algorithm". In: *Journal of Computational and Graphical Statistics* 10.3, pp. 422–439.

Louis, Thomas A. and Wei Shen (1999). "Innovations in Bayes and empirical Bayes methods: estimating parameters, populations and ranks". In: *Statistics in medicine* 18.17-18, pp. 2493–2505.

Maritz, Johannes S. (1966). "Smooth Empirical Bayes Estimation for One-Parameter Discrete Distributions". In: *Biometrika* 53.3/4, pp. 417–429.

— (2018). *Empirical Bayes methods with applications.* CRC Press.

Miller, Rupert G. (1981). *Simultaneous Statistical Inference.* Springer.

Park, Trevor and George Casella (2008). "The Bayesian lasso". In: *Journal of the American Statistical Association* 103.482, pp. 681–686.

Petrone, Sonia, Stefano Rizelli, and Judith Rousseau (2024). "Empirical Bayes in Bayesian learning: understanding a common practice". In: *arXiv*.

Petrone, Sonia, Judith Rousseau, and Catia Scricciolo (2014). "Bayes and empirical Bayes: do they merge?" In: *Biometrika* 101.2, pp. 285–302.

Robbins, Herbert (1956). "An Empirical Bayes Approach to Statistics". In: *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability* 1, pp. 157–163.

— (1968). "Estimating the Total Probability of the Unobserved Outcomes of an Experiment". In: *The Annals of Mathematical Statistics* 39.1, pp. 256–257.

Rubin, Donald B. (1981). "Estimation in parallel randomized experiments". In: *Journal of Educational Statistics* 6.4, pp. 377–401.

Shen, Yandi and Yihong Wu (2022). "Empirical Bayes estimation: When does g-modeling beat f-modeling in theory (and in practice)?" In: *arXiv preprint arXiv:2211.12692*.

Tibshirani, Robert (1996). "Regression shrinkage and selection via the lasso". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 58.1, pp. 267–288.

Von Mises, Richard (1942). "On the correct use of Bayes' formula". In: *The annals of mathematical statistics* 13.2, pp. 156–165.

Wiel, Mark A. van de, Dennis E. te Beest, and Magnus Münch (2017). "Learning from a lot: Empirical Bayes in high-dimensional prediction settings". In: *arXiv preprint arXiv:1709.04192*.