Bayesian Statistics III/IV (MATH3341/4031) Michaelmas term, 2021

Handout 16: Empirical Bayes

Lecturer & author: Georgios P. Karagiannis georgios.karagiannis@durham.ac.uk

Aim

To be able to perform Empirical Bayes analysis, as well as to extend previously introduces concepts in the Empirical Bayes framework.

Basic reading list:

- Berger, J. O. (2013). Statistical decision theory and Bayesian analysis. Springer.
- Casella, G. (1985). An introduction to empirical Bayes data analysis. The American Statistician, 39(2), 83-87.
- Robbins, H. (1964). The empirical Bayes approach to statistical decision problems. The Annals of Mathematical Statistics, 35(1), 1-20.

1 About

Note 1. Empirical Bayes (EB) is a style of statistical analysis –hybrid of (fully) Bayesian and Frequentics statistics.

Note 2. Suppose we have a set of n observations $y = (y_1, ..., y_n) \in \mathcal{Y}$ drawn from sampling distribution $F(\cdot|\theta)$ with pmf/pdf $f(\cdot|\theta)$ labeled by the unknown k-dimensional paramerter $\theta \in \Theta$. If we had fully specified the prior $\theta \sim \Pi(\cdot)$ with pdf/pmf $\pi(\cdot)$ then the Bayesian model would be

$$\begin{cases} y|\theta & \sim F(\cdot|\theta) \\ \theta & \sim \Pi(\cdot) \end{cases} \tag{1}$$

with posterior and prior predictive pdf/pmf

$$\pi(\theta|y) = \frac{f(y|\theta)\pi(\theta)}{g(y)} \; ; \qquad g(y) = \int_\Theta f(y|\theta)\pi(\theta)\mathrm{d}\theta$$

Note 3. Assume that the modeler does not desire to exactly specify a prior distribution Π in the Bayesian model (1). This may be due to the luck of prior information that could justify the use of a pacific prior distribution (similar to the Hierarchical Bayes), or due to desire to reduce the parametric space for computation reasons.

Note 4. Empirical Bayes supports it is illusory to try to model the imprecision of prior $\Pi(\cdot)$ by adding several layers of conditional distributions (such as in Hierarchical Bayes) when the first level is already partially or completely unknown.

Note 5. Empirical Bayes aims at “recovering” part or the whole of the ’unknown’ prior distribution $\Pi(\cdot)$ of θ from the observables y. This violates the Bayesian paradigm, and is the reason why EB is not considered as a fully Bayesian approach but rather a hybrid of Bayes and Frequentist statistics.

Note 6. Given that the statistical model $f(y|\theta)$ is fully specified, the core idea underpinning the implementation of the Empirical Bayesian approach is that the ’unknown’ prior distribution $\Pi(\cdot)$ of θ can be recovered from the observables y by utilizing Frequentist statistics inferential techniques on the marginal pdf/pmf $g(\cdot)$ which can be seen as marginal likelihood of the observables y.

Note 7. Recall: According to Bayesian model (2), the marginal distribution $G(\cdot)$ with pdf/pmf

$$g(y) = \int_\Theta f(y|\theta)\mathrm{d}\Pi(\theta) = \begin{cases} \int_\Theta f(y|\theta)\pi(\theta)\mathrm{d}\theta \\ \sum_{\forall\theta\in\Theta} f(y|\theta)\pi(\theta) \end{cases} , \tag{2}$$

1. aims at representing the distribution according to which the observables y occurred. It is how You modeled the unknown real data generation process $R(\cdot)$; hence why $G(\cdot)$ is also called prior predictive distribution.

2. indicates whether the statistical model $F(\cdot|\theta)$ is useful. Eg., it is unlikely that the Bayesian model could actually predict the accrued y if $g(y)$ is small; hence it is not representative to the real generating process $R(\cdot)$; hence it is not a useful model.

Note 8. Empirical Bayes is traditionally categorized as Non-parametric Empirical Bays (NPEB), and Parametric Empirical Bayes (PEB) approach, discussed in Sections 2 and 3.

Note 9. EB approaches are useful in the following appealing problem setting. Assume the data $y = (y_1, ..., y_n)$ consists of n independent components $\{y_i\}$ where $y_i \sim F(\cdot|\theta_i)$, and assume that the unknown $\theta = (\theta_1, ..., \theta_n)$ consists of n components $\{\theta_i\}$ following a common prior distribution $\theta_i \sim \Pi(\cdot)$; i.e.

$$y_i|\theta_i \sim F(\cdot|\theta_i), \text{ and } \theta_i \sim \Pi(\cdot) \text{ for } i = 1, ..., n.$$

- $\{y_i\}_{i=1}^n$ could be current observations of $\{\theta_i\}_{i=1}^n$ and interest lies on learning all of $\{\theta_i\}_{i=1}^n$.

- $\{y_i\}_{i=1}^n$ could be past observations of $\{\theta_i\}_{i=1}^n$ where interest lies in learning $\Pi(\cdot)$ for use on future θ_i.

Notice that,

$$g(y) = \int_{\Theta} \prod_{i=1}^{n} f(y_i|\theta_i)\pi(\theta_i)\mathrm{d}\theta_i = \prod_{i=1}^{n} \int_{\theta} f(y_i|\theta_i)\pi(\theta_i)\mathrm{d}\theta_i = \prod_{i=1}^{n} g(y_i) \tag{3}$$

$$\pi(\theta|y) \propto \prod_{i=1}^{n} f(y_i|\theta_i)\pi(\theta_i) \propto \prod_{i=1}^{n} \pi(\theta_i|y_i) \tag{4}$$

where $g(y_i) = \int_\theta f(y_i|\theta_i)\pi(\theta_i)\mathrm{d}\theta$, and $\pi(\theta_i|y_i) \propto f(y_i|\theta_i)\pi(\theta_i)$.

2 Non parametric Empirical Bayes

Remark 10. Non-parametric Empirical Bayes (NEB) works as follows. The problem is as in Note 9: to draw inference on θ_i given realizations $y_i \sim F(\cdot|\theta_i)$ where F is known, and under the assumption that $\theta_i \overset{\text{iid}}{\sim} \Pi(\cdot)$ for $i = 1, ..., n$ where $\Pi(\cdot)$ is completely unspecified.

Note 11. The general is to derive an approximation $\hat{\pi}(\cdot)$ of prior $\pi(\cdot)$ by using (2) and by estimating the marginal likelihood $g(y)$ from the data $\{y_i\}_{i=1}^n$. Then plug in the proxy $\hat{\pi}(\cdot)$ to the posterior distribution

$$\hat{\pi}(\theta_i|y_i) \propto f(y_i|\theta_i)\,\hat{\pi}(\theta_i)$$

and make Bayesian inference/prediction as if $\hat{\pi}(\cdot)$ was the specified prior.

Remark 12. A simpler trick is the following. Assume it is desired to use the Bayes rule $\delta(y)$ for point estimation, credible intervals, hypothesis testing, etc.... The NEB rule (or estimate) $\delta^{\text{EB}}(y)$ can be computed as:

1. Find a representation of the Bayes rule as $\delta(y) = \psi(y, \xi(g))$ where ψ and ξ are known functionals.

2. Produce an estimate $\hat{g} := \hat{g}(y)$ of $g(y)$ by using the observations. Eg., histograms, kernel density estimations, etc...

3. Compute $\delta^{\text{EB}}(y) = \psi(y, \xi(\hat{g}))$ by pluging-in the estimate $\hat{g}$ in the original Bayes rule.

Example 13. Suppose $\{y_i\}_{i=1}^n$ are iid such that $y_i|\theta_i \sim \text{Poi}(\theta_i)$ for $i = 1, ..., n$. Assume that $\{\theta_i\}_{i=1}^n$ are iid from a common prior with pdf $\pi(\cdot)$. Compute the NEB estimator of θ_n, under the quadratic loss function, by using the empirical density estimate to recover the marginal likelihood $g(y)$.

1. The Bayes rule is the posterior mean

$$\begin{aligned}\delta(y) =& \mathrm{E}_{\Pi}\left(\theta_n|y_1,...,y_n\right) = \mathrm{E}_{\Pi}\left(\theta_n|y_n\right) = \int_{\Theta}\theta_n\pi\left(\theta_n|y_n\right)\mathrm{d}\theta_n = \int_{\Theta}\theta_n\frac{f\left(y_n|\theta_n\right)\pi\left(\theta_n\right)}{g(y_n)}\mathrm{d}\theta_n \\ =& \int_{\Theta}\theta_n\frac{\left[\theta_n^{y_n}\exp\left(-\theta_n\right)/y_n!\right]\pi\left(\theta_n\right)}{g(y_n)}\mathrm{d}\theta_n \;=\; (y_n+1)\,\frac{\int_{\Theta}f\left(y_n+1|\theta_n\right)\pi\left(\theta_n\right)\mathrm{d}\theta_n}{g(y_n)} \\ =& (y_n+1)\,\frac{g(y_n+1)}{g(y_n)} \end{aligned} \tag{5}$$

2. As each y_i is discrete in $\{0,1,2,...\}$, the pmf of y_i can be estimated with the relative frequency, i.e.

$$\hat{g}_0(j) = \begin{cases}\frac{1}{n}\sum_{i=1}^{n}1_{\{j\}}(y_i) & \text{if } j\in\{0,1,2,...\}\\ 0 & \text{else}\end{cases}$$

3. Hence, by substitution I get

$$\delta^{\mathrm{EB}}(y) = (y_n+1)\,\frac{\hat{g}_0(y_n+1)}{\hat{g}_0(y_n)} = (y_n+1)\,\frac{\sum_{i=1}^{n}1_{\{y_n+1\}}(y_i)}{\sum_{i=1}^{n}1_{\{y_n\}}(y_i)}$$

Remark 14. NEB is not always easy to be applied, eg., it may not be possible to find a representation $\delta(y) = \psi\left(y,\xi\left(g\right)\right)$.

Remark 15. NEB requires a large n to be able to recover finer features of Π which is not always the case. In Example 13, $n\to\infty$ guaranties the success of estimate $\hat{g}_0(j)$.

3 Parametric Empirical Bayes

Note 16. In Parametric Empirical Bayes (PEB), the prior distribution is modeled as a parametric family/class of distributions. Consequently, the task of recovering the whole of the unknown prior reduces to the task of learning (from the data) the unknown parameters of prior by a chosen Frequency learning technique.

Note 17. Below, we discuss some popular classes of prior, as well as learning techniques for the hyper-parameters of the priors.

3.1 Some restricted classes of prior

Below is a number of restricted classes of priors.

Note 18. Priors of a given functional form: Assume the class of priors

$$\mathscr{P} = \{\Pi : \pi(\theta) := \pi(\theta|\phi) > 0, \forall\phi\in\Phi\}$$

The pdf/pmf of the prior distribution belongs to a parametric family where the choice of the prior reduces to the choice of ϕ .

Note 19. Mixture priors with unknown hyper-parameters $\phi = (\varpi,\chi)$

$$\mathscr{P} = \left\{\Pi : \pi(\theta|\phi) = \sum_{\ell}\varpi_{\ell}\pi\left(\theta|\chi_{\ell}\right),\;\forall\varpi,\forall\chi\right\}$$

where $\{\pi\left(\theta|\chi_{\ell}\right)\}$ are known pmf/pdf.

Note 20. Priors close to an elicited prior: Assume the ϵ-contaminated class of priors

$$\mathscr{P} = \{\Pi : \pi(\theta) := (1-\epsilon)\pi_0(\theta) + \epsilon q(\theta),\;\forall q\in\mathscr{Q}\}$$

where $\epsilon \in [0,1]$. Here, we elicit the base prior π_0 by assuming that eny prior close to π_0 is reasonable. Hence, ϵ reflects how close we feel that π must be to π_0 with respect to a class $\mathscr{Q}$ of possible contamination.

3.2 Some prior selection approaches

Below we re-formulate a number of known Frequentist learning approaches used in PEB.

The ML-II approach

Note 21. Suppose $\mathscr{P}$ is a class of priors under consideration. Then $\hat{\pi} \in \mathscr{P}$ is the ML-II prior (type II maximum likelihood prior) if

$$g(y|\hat{\pi}) = \sup_{\pi \in \mathscr{P}} \{g(y|\pi)\}.$$

Example 22. (Cont. Example 13) Assume $y_i|\theta_i \overset{\text{iid}}{\sim} \text{Poi}(\theta_i)$ and $\theta_i \overset{\text{iid}}{\sim} \text{Ex}(\lambda)$ for $i = 1, ..., n$. Compute the PEB estimator of θ_n under the quadratic loss function, by using ML-II.

Solution. The marginal likelihood is $g(y) = \prod_{i=1}^{n} g_0(y_i)$ where

$$g_0(y_i) = \int \left[\frac{1}{y_i!}\theta_i^{y_i} \exp(-\theta_i)\right] [\lambda \exp(-\lambda y_i)] \, \mathrm{d}\theta = \lambda / (\lambda + 1)^{y_i+1}$$

Then $\hat{\lambda} = \arg\max_{\forall \lambda} (g(y)) = 1/\bar{y}$ by optimization. So (5) implies that $\delta^{\text{EB}}(y) = \frac{\bar{y}}{\bar{y}+1}(y_n + 1)$.

The method of moments approach

Note 23. It relates prior moments to moments of the marginal distribution

$$\text{E}_G\left(y^k\right) = \bar{h}_k, \quad \bar{h}_k = \frac{1}{n}\sum_{i=1}^{n}(y_i)^k$$

for $k = 1, ..., K$ where K is large enough to learn ϕ.

Example 24. Assume a Bayesian model

$$\begin{cases} y_i|p_i & \sim \text{Bin}(m, p_i) \\ p_i & \sim \text{Be}(\phi, \phi) \end{cases}$$

for $i = 1, ..., n$ and when $\phi > 0$. Find the posterior of p via PEB where ϕ is learned by method of moments.

Solution. The model is such that in Note 9. By Bayesian theorem and setting $\phi := \hat{\phi}$, it is obviously

$$p_i|y_i \sim \text{Be}\left(y_i + \hat{\phi}, m - y_i + \hat{\phi}\right)$$

where $\hat{\phi}$ is found via MoM on the marginal likelihood. The marginal likelihood is $g(y) = \prod_{i=1}^{n} g(y_i)$ where

$$g(y_i) == \int \binom{m}{y_i} p_i^{y_i}(1-p_i)^{m-y_i} \frac{1}{B(\phi,\phi)} p_i^{\phi-1}(1-p_i)^{\phi-1} \mathrm{d}\theta_i = \binom{m}{y_i} \frac{B(y_i+\phi, m-y_i+\phi)}{B(\phi,\phi)}$$

It is $\text{E}_G(y_i) = \text{E}_\Pi(\text{E}_F(y_i|p_i)) = m\frac{a}{a+a} = \frac{m}{2}$ which is not useful. Then

$$\text{Var}_G(y_i) = \text{E}_\Pi(\text{Var}_F(y_i|p_i)) + \text{Var}_\Pi(\text{E}_F(y_i|p_i)) = \text{E}_\Pi(mp_i(1-p_i)) + \text{Var}_\Pi(mp_i) = ... = \frac{m(2\phi+m)}{4(2\phi+1)}$$

I can estimate $s_y^2 = \frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y})^2$. Then, after matching the moments

$$\text{Var}_G(y_i) = s_y^2 \iff \frac{m\left(2\hat{\phi}+m\right)}{4\left(2\hat{\phi}+1\right)} = s_y^2 \iff \hat{\phi} = \frac{m^2 - 2s_y^2}{8s_y^2 - 2m}$$

The material in this box is not discussed in the lecture, and it is not needed to be read for the exams.

The distance approach

Note 25. In order to recover the prior $\pi(\cdot)$, we seek to find an estimate $\hat{\pi}(\cdot)$ such that

$$g(y|\hat{\pi}) = \int_{\Theta} f(y|\theta)\,\hat{\pi}(\theta)\,\text{d}\theta$$

is as close as possible to the empirical estimate $\hat{g}(y)$ of $g(\cdot)$ by some distance let's say dist $(\cdot,\cdot)$; ie.,

$$\hat{\pi}(\theta) = \arg\min_{\forall\hat{\pi}}\left(\text{dist}\left(\hat{g}(y)\middle|\middle|g(y|\hat{\pi})\right)\right)$$

Example 26. If you consider KL divergence, then

$$\hat{\pi}(\theta) = \arg\min_{\forall\hat{\pi}}\left(\text{KL}\left(\hat{g}(y)\middle|\middle|g(y|\hat{\pi})\right)\right) = \arg\min_{\forall\hat{\pi}}\left(\text{E}_{\hat{G}(y)}\left(\log\left(\frac{\hat{g}(y)}{g(y|\hat{\pi})}\right)\right)\right) = \arg\max_{\forall\hat{\pi}}\left(\text{E}_{\hat{G}(y)}\left(\log\left(g(y|\hat{\pi})\right)\right)\right)$$

where $\text{E}_{\hat{G}(y)}(\log(g(y|\hat{\pi}))) = \int \log(g(y|\hat{\pi}))\,\text{d}\hat{G}(y)$.
If I assume that θ is discrete such as $\Theta \in \{\theta_1, ..., \theta_k\}$ where $\pi_j = \pi(\{\theta = \theta_j\})$ for $j = 1, ..., k$ it is

$$\{\hat{\pi}_j\} = \arg\max_{\forall\hat{\pi}}\left(\text{E}_{\hat{G}(y)}\left(\log\left(\sum_{j=1}^{k} f(y|\theta_j)\,\hat{\pi}_j\right)\right)\right)$$

If I assume $\hat{g}(t) = \frac{1}{n}\sum_{i=1}^{n} 1_{\{y_i\}}(t)$ then it is

$$\{\hat{\pi}_j\} = \arg\max_{\forall\hat{\pi}}\left(\frac{1}{n}\sum_{i=1}^{n}\left(\log\left(\sum_{j=1}^{k} f(y_i|\theta_j)\,\hat{\pi}_j\right)\right)\right)$$

Example 27. [1-way ANOVA] Consider the following 1-way ANOVA problem where the factor has n levels and each level has the same number of repeatations m. Namely, $y_{i,j} = \theta_i + \epsilon_{i,j}$ where θ_i is the effect of the ith level and $\epsilon_{i,j} \sim \text{N}(0, \sigma^2)$ is the error in the j the repeatation, for $j = 1, ..., m$, and $i = 1, ..., n$. Assume $\theta_i \sim \text{N}(\mu, \tau^2)$ for $i = 1, ..., n$. Assume σ^2 is known, while μ and τ^2 are unknown.

1. Compute the EB estimator of $\{\theta_i\}$ under the square loss where μ and τ^2 are learned via ML-II, and show it can be written in the form $\theta_i = \varpi\mu + (1-\varpi)\,y_i$.
2. Compute the EB estimator of $\{\theta_i\}$ under the square loss where μ and τ^2 are learned by MoM
3. Compute the EB estimator of $\{\theta_i\}$ by constructing unbiased estimates for ϖ and μ . Hint: if $\xi \sim \chi_v^2$then $\text{E}(1/\xi) = 1/(v-2)$
4. If we would ignore the prior on θ, the MLE of θ_i would be $\hat{\theta}_i = y_i$ for all i. Compare it with EB estimates.

Solution. Assume that $y_i = \frac{1}{m}\sum_j y_{i,j}$ and $\sigma_m^2 = \sigma^2/m$. The Bayesian model is

$$\begin{cases} y_i|\theta_i & \sim \text{N}\left(\theta_i, \sigma_m^2\right),\ i = 1, ..., n \\ \theta_i & \sim \text{N}\left(\mu, \tau^2\right),\ i = 1, ..., n \end{cases}$$

By using Bayesian theorem, the posterior distribution of $\theta = (\theta_1, ..., \theta_n)^\top$ is

$$\theta_i|y_i \sim \mathrm{N}\left(\frac{\sigma_m^2}{\sigma_m^2+\tau^2}\mu + \frac{\tau^2}{\sigma_m^2+\tau^2}y_i, \frac{\sigma_m^2\tau^2}{\sigma_m^2+\tau^2}\right) \equiv \mathrm{N}\left(\varpi\mu + (1-\varpi)\, y_i, \sigma_m^2\,(1-\varpi)\right)$$

where $\varpi = \frac{\sigma_m^2}{\sigma_m^2+\tau^2}$. Hence the Bayesian estimator is $\theta_i = \varpi\mu + (1-\varpi)\, y_i$.

1. The marginal posterior is $y_i \overset{\text{iid}}{\sim} \mathrm{N}\left(\mu, \sigma_m^2+\tau^2\right)$ by using standard Normal pdf properties. Essentially this is a regression problem, $y \sim \mathrm{N}\left(1\mu, I_n\left(\sigma_m^2+\tau^2\right)\right)$ with log likelihood

   $$\log(g(y)) \overset{\text{calcul...}}{=} \cdots = -\frac{n}{2}\log\left(\sigma_m^2+\tau^2\right) - \frac{ns^2}{2\left(\sigma_m^2+\tau^2\right)} - \frac{n\left(\bar{y}-\mu\right)^2}{2\left(\sigma_m^2+\tau^2\right)} + \text{const}$$

   where $s^2 = \frac{1}{n}\sum_i\left(y_i-\bar{y}\right)^2$. It is maximized for $\hat{\mu}_{\text{ML-II}} = \bar{y}$ regardless τ^2. Then at $\hat{\mu}_{\text{ML-II}} = \bar{y}$, it is

   $$0 = \frac{\mathrm{d}}{\mathrm{d}\tau}\log(g(y))\Big|_{\tau=\hat{\tau}} = -\frac{n}{2\left(\sigma_m^2+\hat{\tau}^2\right)} + \frac{ns^2}{2\left(\sigma_m^2+\hat{\tau}^2\right)}$$

   which implies $\hat{\tau}_{\text{ML-II}} = \max\left(s-\sigma_m^2\right)$. Hence

   $$\hat{\theta}_i^{\text{EB}} = \hat{\varpi}_{\text{ML-II}}\bar{y} + \left(1-\hat{\varpi}_{\text{ML-II}}\right)y_i, \quad \text{where} \quad \hat{\varpi}_{\text{ML-II}} = \frac{\sigma_m^2}{\sigma_m^2+\max\left(s-\sigma_m^2\right)}$$

2. It is

   $$\begin{cases}\mathrm{E}(\bar{y}) = \bar{y} \\ \mathrm{Var}(\bar{y}) = \frac{n}{n-1}s^2\end{cases} \iff \begin{cases}\hat{\mu} = \bar{y} \\ \sigma_m^2+\hat{\tau}^2 = \frac{n}{n-1}s^2\end{cases} \iff \begin{cases}\hat{\mu}_{\text{MoM}} = \bar{y} \\ \hat{\tau}_{\text{MoM}}^2 = \frac{n}{n-1}s^2-\sigma_m^2\end{cases}$$

   Hence

   $$\hat{\theta}_i^{\text{EB}} = \hat{\varpi}_{\text{MoM}}\bar{y} + \left(1-\hat{\varpi}_{\text{MoM}}\right)y_i, \quad \text{where} \quad \hat{\varpi}_{\text{MoM}} = \frac{n-1}{n}\frac{\sigma_m^2}{s^2}$$

3. From Cohran's theorem I know that $\bar{y} \sim \mathrm{N}\left(\mu, \frac{1}{n}(\sigma_m^2+\tau^2)\right)$, $n\frac{s^2}{\sigma_m^2+\tau^2} \sim \chi_{n-1}^2$, and that $\bar{y}$and s^2 are independent. Hence

   $$\mathrm{E}(\bar{y}) = \mu, \quad \text{and} \qquad\qquad \mathrm{E}\left(\frac{(n-3)\sigma_m^2}{ns^2}\right) = \frac{\sigma_m^2}{\sigma_m^2+\tau^2}$$

   So

   $$\hat{\theta}_i^{\text{EB}} = \hat{\varpi}\bar{y} + (1-\hat{\varpi})\,y_i, \quad \text{where} \quad \hat{\varpi} = \frac{(n-3)\sigma_m^2}{ns^2}$$

4. The EB estimate of θ_i uses information from all y_i's which can be beneficial in many cases –FYI this is know as Stain's effect.

4 Discussions

Remark 28. Hierarchical Bayesian model is simply a special type of Bayesian model, where

$$\begin{cases} y|\theta & \sim f(y|\theta) \\ \theta|\phi & \sim \pi(\theta|\phi) \\ \phi & \sim \pi(\phi) \end{cases} \tag{6}$$

for $\phi = (\phi_1, ..., \phi_{m-1})$, and ϕ_m fixed hyper-parameter.

$$\pi^{\text{FB}}(\theta|y,\phi) = \int_{\Phi} \pi(\theta|y,\phi)\pi(\phi|y)\text{d}\phi$$

with $\pi(\phi|y) \propto f(y|\phi)\pi(\phi)$

$$\pi^{\text{EB}}(\theta|y,\phi) = \pi(\theta|y,\hat{\phi})$$

with

$$\hat{\phi} = \hat{\phi}(y)$$

implying that $\pi(\phi) = 1_{\hat{\phi}=\phi}(\phi)$. Consequently Empirical Bayes tends to produce overconfident inference, (e.g. narrower credible set, smaller standard errors, etc...) by ignoring the uncertainty about the hyper-parameter ϕ.

Remark 29. Fully Bayesian methods may be quite computationally demanding in modern applications with Big Data and High dimensional unknown parameters –at least in 2021. In many such cases (but not always), Empirical Bayes is able to produce inference numerical results close to those that the fully Bayesian methods would produce. Hence EB could be use as an 'approximation' of the fully Bayesian inference or as an alternative but with caution.

Remark 30. (Just for your information) In theory, it has been proven that empirical Bayes when implemented in statistical problems of the form as in Note 9, can produce decision procedures (estimates, CI, etc...) δ_n^{EB} whose Bayes risk is equal to that of fully Bayes ones δ^{FB} in the limit as $n \to \infty$; i.e

$$\lim_{n\to\infty} r\left(\pi, \delta_n^{\text{EB}}\right) = r\left(\pi, \delta^{\text{FB}}\right),\ \forall\pi$$

Namely, one can do as well asymptotically as the Bayes rule, knowing nothing about the prior of θ. We will not go so deep here,for more info see: Robbins, H. (1964).

Question 31. *For your practice try to do the Exercise 80 from the Exercise sheet.*