1 Hierarchical Bayes and Empirical Bayes. MLII Method.

Hierarchical Bayes and Empirical Bayes are related by their goals, but quite different by the methods of how these goals are achieved. The attribute *hierarchical* refers mostly to the modeling strategy, while *empirical* is referring to the methodology. Both methods are concerned in specifying the distribution at prior level, hierarchical via Bayes inference involving additional degrees of hierarchy (hyperpriors and hyperparameters), while empirical Bayes is using data more directly.

In expanding Bayesian models and inference to more complex problems, going beyond the simple likelihood-prior-posterior scheme, a hierarchy of models may be needed. The parameter(s) of interest considered are entering the model via their "realizations" which are modeled similarly as they were "measurements." The common name *parameter population distribution* is indicative of the nature of the approach.

1.1 Hierarchical Bayesian Analysis

Hierarchical Bayesian Analysis is a convenient representation of a Bayesian model, in particular the prior π, via a conditional hierarchy of so called hyper-priors $\pi_1, \ldots, \pi_{n+1}$,

$$\pi(\theta) = \int \pi_1(\theta|\theta_1)\pi_2(\theta_1|\theta_2)\ldots\pi_n(\theta_{n-1}|\theta_n)\pi_{n+1}(\theta_n)\, d\theta_1 d\theta_2 \ldots d\theta_n. \tag{1}$$

Operationally, the model

$$[x|\theta] \sim f(x|\theta),\ [\theta|\theta_1] \sim \pi_1(\theta|\theta_1),\ [\theta_{n-1}|\theta_n] \sim \pi_n(\theta|\theta_1),\ [\theta_n] \sim \pi_{n+1}(\theta_n). \tag{2}$$

is equivalent to the model

$$[x|\theta] \sim f(x|\theta),\ [\theta] \sim \pi(\theta),$$

as the inference on θ is concerned. Notice that in the hierarchy of data, parameters and hyperparameters,

$$X \longrightarrow \boxed{\theta} \longrightarrow \theta_1 \longrightarrow \theta_2 \ldots \longrightarrow \theta_n$$

X and θ_i are independent, given θ. That means,

$$[X|\theta, \theta_1, \ldots] \stackrel{d}{=} [X|\theta], \qquad [\theta_i|\theta, X] \stackrel{d}{=} [\theta_i|\theta],$$

where $\stackrel{d}{=}$ is equality in distribution. The joint distribution $[X, \theta, \theta_1, \ldots, \theta_n]$ which by definition is

$$[X, \theta, \theta_1, \ldots, \theta_n] = [X|\theta, \theta_1, \ldots, \theta_n]\ [\theta|\theta_1, \ldots, \theta_n]\ [\theta_1|\theta_2, \ldots, \theta_n]\ \ldots\ [\theta_{n-1}|\theta_n]\ [\theta_n]$$

can be represented as

$$[X, \theta, \theta_1, \ldots, \theta_n] = [X|\theta][\theta|\theta_1]\ [\theta_1|\theta_2]\ \ldots\ [\theta_{n-1}|\theta_n]\ [\theta_n],$$

thus, to fully specify the model, only "neighbouring" conditionals $[X|\theta], [\theta|\theta_1],\ [\theta_1|\theta_2],\ \ldots,\ [\theta_{n-1}|\theta_n]$ and the "closure" distribution $[\theta_n]$ are needed.

Why then decompose the prior, as in (1) and use the model (2). Here are some of the reasons:

• Modeling requirements may lead to the hierarchy in the prior. For example Bayesian models in meta analysis;

• The prior information may be separated into the structural part and the subjective/noninformative part at higher level of hierarchy;

• Robustness and objectiveness – "let the data talk about the hyperparameters;"

• Calculational issues (utilizing hidden mixtures, mixture priors, missing data, MCMC format).

Sometimes it is not calculatingly feasible to carry out the analysis by reducing the sequence of hyperpriors (1) to a single prior (2).

Rather, Bayes rule is obtained (by using Fubini's theorem) as repeated integral with respect to more convenient conditional distributions. Here is the result involving the model (f), conditional prior ($\pi_1(\theta|\theta_1)$), and hyperprior $\pi_2(\theta_1)$).

Suppose the hierarchical model is given as $[X|\theta] \sim f(x|\theta)$, $[\theta|\theta_1] \sim \pi_1(\theta|\theta_1)$, and $[\theta_1] \sim \pi_2(\theta_1)$, then the posterior distribution can be written as

$$\pi(\theta|x) = \int_{\Theta_1} \pi(\theta|x, \theta_1)\pi(\theta_1|x)d\theta_1.$$

The densities under the integral are $\pi(\theta|x, \theta_1) = \frac{f(x|\theta)\pi_1(\theta|\theta_1)}{m_1(x|\theta_1)}$, and $\pi(\theta_1|x) = \frac{m_1(x|\theta_1)\pi_2(\theta_1)}{m(x)}$, where $m_1(x|\theta_1) = \int_\Theta f(x|\theta)\pi_1(\theta|\theta_1)d\theta$ is the marginal likelihood, and $m(x) = \int_{\Theta_1} m_1(x|\theta_1)\pi_2(\theta_1)d\theta_1$ marginal.

Now, for any function of the parameter h,

$$E^{\theta|x}h(\theta) = E^{\theta_1|x}[E^{\theta|\theta_1,x}h(\theta)]. \tag{3}$$

Example: Suppose you pool n people about their favorite presidential candidate and X of them favor the candidate A. The likelihood is $[X|p] \sim \mathcal{B}(n, p)$, and the proportion p is the patrameter of interest. You believe that proportion is close to 1/2, but not quite sure. The appropriate prior on p would be Beta $\mathcal{B}e(k, k),\ k \in \mathbb{N}$ (see Figure 1(a)); it is symmetric about 1/2, but you are reluctant to specify the natural number k. Thus, $[p|k] \sim \mathcal{B}e(k, k)$. Finally, you put a hyperprior on k, $\pi_2(k) \propto \frac{1}{k(2k-1)}$. The hyperprior probability mass function is in fact

$$\pi_2(k) = \frac{1}{2\log(2)k(2k-1)},\ k = 1, 2, \ldots.$$

What is the Bayes estimator for p if $n = 20$ and $X = 12$.

The unconditional prior on p is

$$\pi(p) = \sum_{k=1}^{\infty} \frac{p^{k-1}(1-p)^{k-1}}{B(k,k)} \frac{1}{2\log(2)k(2k-1)} = \frac{1-|1-2p|}{4\log(2)p(1-p)}. \tag{4}$$

This prior, depicted in Figure 1(b), is symmetric about 1/2.

The posterior does not have a finite form (can be expressed in terms of special functions) and the Bayes rule has to be found by numerical integration.

$$\delta_\pi(12) = \frac{\int_0^1 p\, f(12|p)\pi(p)dp}{\int_0^1 f(12|p)\pi(p)dp} = 0.581368.$$

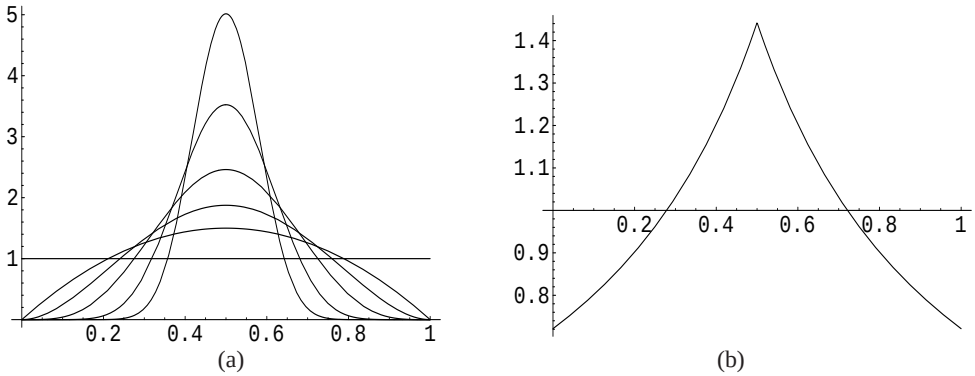


Figure 1: Beta(k, k) densities for $k = 1, 2, 3, 5, 10,$ and 20. The unconditional prior $\pi(p)$ from (4).

Mathematica programming is particularly convenient in this case, and the integrals have exact solution given n and X. In our case the exact result is 288241/495798.

The above approach to produce the Bayes estimator is direct, the unconditional prior is obtained as mixture of Betas. Thus we utilized transition from equations in (1) to prior in (2) Now we consider the same problem by utilizing hierarchical Bayes model discussed previously and equation (3).

The hierarchical Bayes approach is as follows. The Bayes rule is

$$\delta_\pi(x) = E^{k|x} E^{p|k,x} p,$$

where the expectation $E^{k|x}$ is taken with respect to $\pi_2(k|x)$ and $E^{p|k,x}$ with respect to $\pi_1(p|k,x)$.

The distribution for $[p|x,k]$ is again Beta, $\pi_1(p|k,x) \propto p^x(1-p)^{n-x} \cdot p^{k-1}(1-p)^{k-1}$, with parameters $x+k$ and $n-x+k$. The expectation of p is $\frac{x+k}{(x+k)+(n-x+k)} = \frac{x+m}{2k+n}$.

The distribution $\pi_2(k|x) = \frac{m_1(x|k)\pi_2(k)}{m(x)}$ does not have a finite form.

$m_1(x|k) = \int_0^1 f(x|p)\pi_1(p|k)dp$ is Beta-Binomial, given by

$$m_1(x|k) = \frac{\Gamma(2k)}{\Gamma(k)\Gamma(k)} \cdot \frac{\Gamma(n+1)}{\Gamma(x+1)\Gamma(n-x+1)} \frac{\Gamma(k+x)\Gamma(k+n-x)}{\Gamma(2k+n)}.$$

Of course, $m_1(x|k)$ can be simplified a bit more. Now, the marginal $m(x)$ is $\sum_{k=1}^\infty m_1(x|k)\frac{1}{2\log(2)k(2k-1)}$ and it does not have a finite form although can be expressed in terms of hypergeometric pFq special functions. [1] However,

$$\begin{aligned}
\delta_\pi(12) &= E^{k|x} E^{p|k,x} p = E^{k|x} \frac{x+k}{2k+n} \\
&= \frac{\sum_{k=1}^\infty \frac{x+k}{2k+n} m_1(x|k)\frac{1}{2\log(2)k(2k-1)}}{\sum_{k=1}^\infty m_1(x|k)\frac{1}{2\log(2)k(2k-1)}} \\
&= \frac{x({}_3F_2([-1/2, n-x, 1+x], [1/2+n/2, 1+n/2], 1) - 1)}{n({}_3F_2([-1/2, n-x, x], [1/2+n/2, n/2], 1) - 1)},
\end{aligned}$$

is a ratio of two hypergeometric pFq functions with argument $z = 1$, thus representing an infinite sum of ratios of Gamma functions. This ratio is easily numerically evaluated, see mathematica notebook. The result

[1] The hypergeometric pFq function is defined as $pFq([a_1, \ldots, a_p], [b_1, \ldots, b_q], z) = \sum_{m=0}^\infty \frac{(a_1)_k \ldots (a_p)_k}{(b_1)_k \ldots (b_q)_k} \frac{z^k}{k!}$, where $(a)_n = a(a+1)\ldots(a+n-1) = \frac{\Gamma(a+n)}{\Gamma(a)}$.

for $n = 20$ and $X = 12$ is $\delta_\pi(12) = 288241/495798$. Notice slight shrinkage toward 1/2 compared with the MLE estimator $\hat{p} = 12/20 = 0.6$.

You may have noticed that even in the simple hierarchy, as in the previous example the Bayesian analysis may not be computationally easy. This is true even if we have perfect conjugate structure at different levels of hierarchy and normal models. The following Theorem is adapted from Berger (1985) and illustrated on IQ adventures of our old friend Jeremy. Berger (1985), Section 4.6 pages 180–195 contains an excellent account on hierarchical models with detailed proofs. The following model, in addition to its educational value, could be quite is useful as a modeling tool, if one believes in normality at various stages of the hierarchy.

Assume that $\boldsymbol{X} = (X_1, \ldots, X_p)'$ is a p-dimensional observation, with the multivariate normal likelihood $f(\boldsymbol{x}|\boldsymbol{\theta}) \sim \mathcal{MVN}_p(\boldsymbol{\theta}, \sigma_f^2 I)$. Parameter $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_t)'$ is of interest and positive scalar σ_f^2 is assumed known. The first stage prior on $\boldsymbol{\theta}$, $\pi_1(\boldsymbol{\theta}|\mu_\pi, \sigma_\pi^2)$, is multivariate normal $\mathcal{MVN}_p(\boldsymbol{\mu}, \sigma_\pi^2 I)$, where $\boldsymbol{\mu} = \mu_\pi \cdot \mathbf{1}$ and $\mathbf{1}$ is $p \times 1$ vector of 1's. The hyperparameter in this case is two-dimensional, $\boldsymbol{\lambda} = (\mu_\pi, \sigma_\pi)$. To complete the model, assume $\pi_2(\boldsymbol{\lambda}) = \pi_{21}(\mu_\pi)\, \pi_{22}(\sigma_\pi^2)$ with $\pi_{21}(\mu_\pi) \sim \mathcal{N}(\beta, \tau^2)$ and appropriate $\pi_{22}(\sigma_\pi^2)$. Of course β, τ^2 and possible parameters in $\pi_{22}(\sigma_\pi^2)$ are assumed known, i.e., the hierarchy stops here.

The following theorem gives an explicit (up to a univariate numerical integration) Bayes estimator of $\boldsymbol{\theta}$ and its covariance matrix. The

Theorem 1.1 *The posterior mean is*

$$\boldsymbol{\delta}_\pi(\boldsymbol{x}) = E^{\boldsymbol{\theta}|\boldsymbol{x}}\boldsymbol{\theta} = E^{\sigma_\pi^2|\boldsymbol{x}}\boldsymbol{\delta}^*(\boldsymbol{x}, \sigma_\pi^2).$$

where

$$\boldsymbol{\delta}^*(\boldsymbol{x}, \sigma_\pi^2) = \boldsymbol{x} - \frac{\sigma_f^2}{\sigma_f^2 + \sigma_\pi^2}(\boldsymbol{x} - \bar{\boldsymbol{x}}\mathbf{1}) - \frac{\sigma_f^2 + \sigma_\pi^2}{\sigma_f^2 + \sigma_\pi^2 + p\tau^2}(\bar{\boldsymbol{x}} - \beta)\mathbf{1}. \tag{5}$$

The posterior covariance matrix is

$$\boldsymbol{V} = E^{\sigma_\pi^2|\boldsymbol{x}}\left[\sigma_f^2 I - \frac{\sigma_f^4}{\sigma_\pi^2 + \sigma_f^2} I + \frac{\sigma_f^4 \tau^2}{(\sigma_\pi^2 + \sigma_f^2)(p\tau^2 + \sigma_\pi^2 + \sigma_f^2)} J + (\boldsymbol{\delta}^*(\boldsymbol{x}, \sigma_\pi^2) - \boldsymbol{\delta}_\pi(\boldsymbol{x}))(\boldsymbol{\delta}^*(\boldsymbol{x}, \sigma_\pi^2) - \boldsymbol{\delta}_\pi(\boldsymbol{x}))'\right]$$

The distribution for $[\sigma_\pi^2|\boldsymbol{x}]$, *important for the above expectations, satisfies*

$$\pi_{22}(\sigma_\pi^2|\boldsymbol{x}) \propto \frac{\tau \exp\left\{-\frac{1}{2}\left[\frac{s^2}{\sigma_\pi^2 + \sigma_f^2} + \frac{p(\bar{\boldsymbol{x}} - \beta)^2}{p\tau^2 + \sigma_\pi^2 + \sigma_f^2}\right]\right\}}{(\sigma_\pi^2 + \sigma_f^2)^{(p-1)/2}(p\tau^2 + \sigma_\pi^2 + \sigma_f^2)^{1/2}}\, \pi_{22}(\sigma_\pi^2),$$

where $s^2 = \sum_{i=1}^p (x_i - \bar{\boldsymbol{x}})^2$.

The expectation with respect to $[\sigma_\pi^2|\boldsymbol{x}]$ needs to be carried out numerically. It is important that the marginal posterior $\pi_{22}(\sigma_\pi^2|\boldsymbol{x})$ is finite. This is true whenever the "closure" prior $\pi_{22}(\sigma_\pi^2)$ is bounded and $p \geq 3$.

Exercise: Suppose that Jeremy, IQ concerned fellow, has taken 5 IQ tests in the last 5 years and have obtained a vector score $\boldsymbol{X} = (102, 112, 96, 109, 98)$. Assume that each measurement $X_i \sim \mathcal{N}(\theta_i, 80),\ i =$

$1, \ldots, 5$ and θ_i's represent realizations of a random variable representing Jeremy's true ability. Unlike before, we assume that his true ability randomly changes in time, however, the underlying law from which such time-varying abilities are generated is common. Thus, $\theta_i \sim \mathcal{N}(\mu_\pi, \sigma_\pi^2)$, $i = 1, \ldots, 5$. Finally the model is closed by assuming that $\mu_\pi \sim \mathcal{N}(110, 120)$, and $\pi_{2,2}(\sigma_\pi^2) = 1$. Find Bayes' estimator of $\boldsymbol{\theta}$ and the covariance matrix of the estimate.

Solution: The estimator for $\boldsymbol{\theta}$ can be found coordinatewise. The Theorem 1.1 is directly applicable with: $\sigma_f^2 = 80$, $\beta = 110$,$\tau^2 = 120$, and $p = 5$. The result (see MATHEMATICA program `jeremy.nb` on the web site) is:

$$\hat{\boldsymbol{\theta}} = (104.645, 110.239, 101.288, 108.561, 102.407)$$

and

$$\boldsymbol{V} = \begin{pmatrix} 51.1158 & 8.1635 & 5.28252 & 7.62331 & 5.64264 \\ 8.1635 & 62.1226 & 2.63978 & 14.6078 & 4.48102 \\ 5.28252 & 2.63978 & 51.6211 & 3.4326 & 6.33962 \\ 7.62331 & 14.6078 & 3.4326 & 57.2654 & 4.8295 \\ 5.64264 & 4.48102 & 6.33962 & 4.8295 & 50.8602 \end{pmatrix}.$$

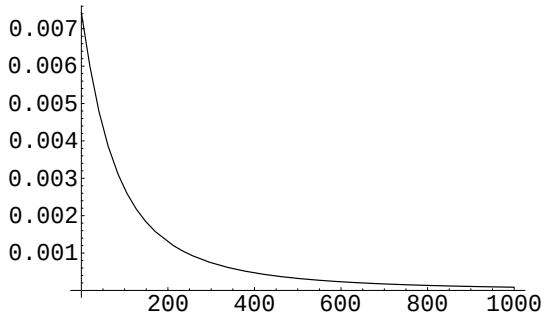


Figure 2: Marginal posterior $\pi_{22}(\sigma_\pi^2|\boldsymbol{x})$ when $\pi_{22}(\sigma_\pi^2) = 1$ in the IQ example.

Figure 2 shows $\pi_{22}(\sigma_\pi^2|\boldsymbol{x})$ when $\pi_{22}(\sigma_\pi^2) = 1$. Notice that even when the closure prior $\pi_{22}(\sigma_\pi^2)$ is improper, the marginal posterior is proper density, although quite flat.

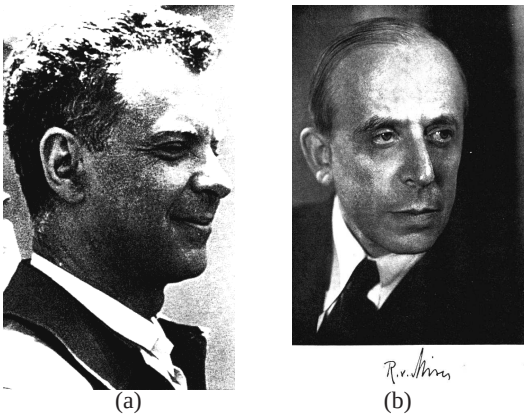
1.2 Empirical Bayes. ML II Method

Empirical Bayes has several formulations. Original formulation of empirical Bayes assumes that past values of X_i and corresponding parameter θ_i are known to the statistician who then on basis of current observation X_{n+1} tries to make inference on unobserved θ_{n+1}. Of course, the parameters θ_i are seldom known. However, it may be assumed that the past (and current) θ's are realizations from the same unknown prior distribution.

Empirical Bayes is an approach to inference in which the observations are used to select the prior, usually via the marginal distribution. Once the prior is specified, the inference proceed in a standard Bayesian fashion. The use of data to estimate the prior in addition to subsequent use for the inference in empirical Bayes is criticized by subjectivists who consider the prior information exogenous to observations. The repeated use of data is also loaded with perils since it can underestimate modeling errors. Any data is going to be complacent with a model which used the same data to specify some of its features.

An excellent and comprehensive monograph on empirical Bayes methodology is Maritz and Lwin (1989).

Empirical Bayes and compound decision problems were introduced by Robbins in the 1950's, but first implicit empirical Bayes approach goes back to Von Mises (1943).



(a) (b)

Figure 3: Herbert Robbins, 1915–2001, Richard von Mises, 1883–1953

Von Mises was considering the problem of testing drinking water for a particular bacteria contamination. In a single experiment 5 small batches of water are taken. The experiment is *positive* if at least one batch is positive, i.e., contains at least one bacteria. Denote by θ the probability that a single batch contains at least one bacteria. Let X be the number of positive batches among 5 taken.

Here

$$m(x) = \int \binom{5}{x} \theta^x (1-\theta)^{5-x} \pi(\theta) d\theta,$$

and $\pi(\theta)$ is not specified. Von Mises was interested in $P(0 \leq \theta \leq \theta_0 | X = 0)$, for an appropriate θ_0. For the "rest of the story" and solution consult Von Mises (1943).

We modify Von Mises' problem to an exercise and provide step-by-step solution. Suppose that in 16 independent experiments, each containing 5 batches of water the following number of batches was found positive for bacteria:

Experiment	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
# of positive	0	0	1	0	2	0	1	1	3	2	0	1	0	1	2	0

Consider these 16 measurements historic and coming from $\mathcal{B}(5, \theta_i)$ distributions, where θ_i's are different. For example the experiments may have been conducted at different places. Take 17th experiment and measure X_{17}.

Assume now that $\theta_i,\ i = 1, \ldots, 17$ have $\mathcal{B}eta(\alpha, \beta)$ distribution. Evaluate $P(0 \leq \theta_{17} \leq 0.2 | X_{17} = 0)$. Here are the steps in the solution:

(i) We find the marginal for X_i in the experiment i. The result is Beta-Binomial,

$$P(X_i = k | \alpha, \beta) = \binom{5}{k} \frac{B(k+\alpha, 5-k+\beta)}{B(\alpha, \beta)},\ k = 0, \ldots, 5.$$

What is critical here, this distribution is free of variable variables θ_i (Pun not intended).

(ii) Estimate α and β in the marginal. The method of moments is most straightforward. The theoretical moments $E(X_i) = 5\alpha/(\alpha+\beta)$ and $E(X_i^2) = \frac{5a(5a+5+b)}{(a+b)(a+b+1)}$ are replaced by the empirical moments based on historic data, $m_1 = \frac{1}{16}\sum_{i=1}^{16} X_i = 0.875$ and $m_2 = \frac{1}{16}\sum_{i=1}^{16} X_i^2 = 1.625$, and equations solved with respect to α and β. Solutions are

$$\hat{\alpha} = \frac{m_1(m_2 - 5m_1)}{m_1(4m_1+5) - 5m_2} = 3.5 \quad \hat{\beta} = \frac{(m_1-5)(m_2-5m_1)}{m_1(4m_1+5)-5m_2} = 16.5.$$

(iii) Express $P(0 \le \theta_{17} \le 0.2 | X_{17} = 0)$ in terms of incomplete Beta functions with estimated hyperparameters $\hat{\alpha}$ and $\hat{\beta}$. For example, in MATHEMATICA:

```
Beta[0.2, 3.5, 16.5]/Beta[3.5, 16.5]=0.658454.
```

Unlike the von Mises example, the original Robbins formulation of empirical Bayes was non-parametric. In the following example the Bayes rule with respect to an unknown prior will be expressed in terms of marginal distribution. Nonparametric empirical Bayes follows and uses the historic data to estimate the marginal distribution in nonparametric fashion. The estimated marginal distributions are then plugged in the formal Bayes rule.

Example: Assume that $X|\theta \sim \mathcal{P}oi(\theta)$ and that $\pi(\theta)$ is to be specified. The Bayes rule is

$$\delta_\pi(x) = \frac{\int_0^\infty \frac{\theta^{x+1}}{x!} e^{-\theta}\pi(\theta)d\theta}{\int_0^\infty \frac{\theta^x}{x!} e^{-\theta}\pi(\theta)d\theta} = \frac{(x+1)\int_0^\infty \frac{\theta^{x+1}}{(x+1)!} e^{-\theta}\pi(\theta)d\theta}{\int_0^\infty \frac{\theta^x}{x!} e^{-\theta}\pi(\theta)d\theta} = (x+1)\frac{m_\pi(x+1)}{m_\pi(x)},$$

i.e., the Bayes rule is depends on the model only via the marginal (prior predictive) distribution $m_\pi(x)$.

Let $X_i|\theta_i \sim \mathcal{P}oi(\theta_i), i = 1, \ldots, n+1$, and all θ_i are parameters having the same distribution.

We are interested in estimating θ_{n+1}, using X_{n+1} and $X_1, \ldots, X_n$. The estimator $\frac{X_1+\cdots+X_{n+1}}{n+1}$ cannot be used since underlying θ's are different. Thus the MLE for θ_{n+1} is X_{n+1}. The empirical Bayes rule is $\delta_\pi(x_{n+1}) = \delta_\pi(x_{n+1}; x_1, \ldots, x_n) = (x_{n+1}+1)\frac{\hat{m}_n(x+1)}{1/n+\hat{m}_n(x_{n+1})}$, where $\hat{m}_n(x_{n+1} : x_1, \ldots, x_n) = \frac{\#\{x_i,\, i=1,\ldots,n | x_{n+1}=x_i\}}{n}$. The performance of the estimator could be improved if the estimators of marginals $\hat{m}_n$ are smoothed.

The matlab file `eb.m` demonstrates the NPEB estimator. Description of this m-file and simulations in it will be added soon. In the meanwhile, please take a look at matlab file, since it is annotated.

Now, consider the same setup, $X_i|\theta_i \sim \mathcal{P}oi(\theta_i)$, but this time the prior distribution of θ's is assumed known up to a hyperparameter. Assume that θ_i's have exponential distribution with density $\pi(\theta_i|\lambda) = \lambda e^{-\lambda\theta_i}\mathbf{1}(\theta_i \ge 0), \lambda \in \mathbb{R}^+$. A Negative Binomial is a marginal distribution for Poisson likelihood and Gamma prior. If the shape parameter in the Gamma prior is 1 then the Negative Binomial becomes Geometric distribution (Handout 0),

$$m_\pi(x_i) = \left(\frac{1}{1+\lambda}\right)^{x_i} \frac{\lambda}{1+\lambda}.$$

The MLE estimator of λ, based on $x_1, \ldots, x_n$ is $\hat{\lambda}_{\text{MLE}} = \frac{1}{\bar{x}}$. Now, the Bayes estimator of θ_{n+1} is $\frac{x_{n+1}+1}{\lambda+1}$ because of the Poisson/Gamma conjugate structure. The parametric empirical Bayes estimator is $\delta(x_{n+1}; x_1, \ldots, x_n) = \frac{\bar{x}}{\bar{x}+1}(x_{n+1}+1)$.

When $n \to \infty$, $\hat{m}_n(x) \to m_\pi(x)$, in probability, and empirical Bayes rule is consistent,

$$\delta_\pi(x; x_1, \ldots, x_n) \to \delta_\pi(x).$$

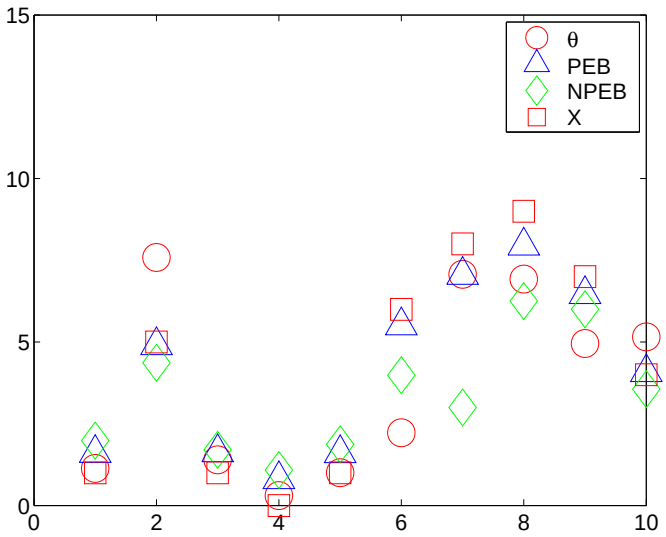


Figure 4: Comparison of MLE, NPEB, and PEB. The circle is the "true" parameter θ.

Exercise: If $X|\theta \sim \mathcal{G}eom(1-\theta)$ with $f(x|\theta) = \theta^x(1-\theta),\ x = 0, 1, 2, \ldots$ The prior $\pi(\theta)$ is unknown. Show $\delta_\pi(x) = \frac{m_\pi(x+1)}{m_\pi(x)}$. For $X|\theta \sim \mathcal{NB}(n, 1-\theta) = \frac{\Gamma(n+x)}{\Gamma(x+1)\Gamma(n)}\theta^x(1-\theta)^n$, the Bayes rule is $\delta_\pi(x) = \frac{x+1}{n+x}\frac{m_\pi(x+1)}{m_\pi(x)}$.

James Stein Estimator and its EB Justification. Consider the estimation of $\boldsymbol{\theta}$ in a model $\boldsymbol{X} \sim \mathcal{MVN}_p(\boldsymbol{\theta}, I)$ under squared error loss $L(\boldsymbol{\theta}, \boldsymbol{a}) = \sum_i(\theta_i - a_i)^2$.

For $p = 1$ and 2, estimator $\hat{\boldsymbol{\theta}} = \boldsymbol{X}$ is admissible (as unique minimax), i.e., no estimator has uniformly better risk. However, for $p \geq 3$ $\boldsymbol{X}$ is neither unique minimax nor admissible. A better estimator is

$$\delta_{JS}(\boldsymbol{X}) = \left(1 - \frac{p-2}{\sum_{i=1}^p X_i^2}\right)\boldsymbol{X},$$

known as James-Stein estimator.

The empirical Bayes justification for $\delta_{JS}(\boldsymbol{X})$ is provided next.

Suppose that $\boldsymbol{\theta}$ has a prior distribution

$$\boldsymbol{\theta} \sim \mathcal{MVN}(0, \tau^2 I),$$

where hyperparameter τ^2 is not known and will be estimated from the sample, $\boldsymbol{X}$ in this case.

The Bayes rule, under squared error loss is

$$\delta_B(\boldsymbol{X}) = \frac{\tau^2}{1+\tau^2}\boldsymbol{X} = \left(1 - \frac{1}{1+\tau^2}\right)\boldsymbol{X}.$$

Marginal (prior predictive) distribution for $\boldsymbol{X}$ is $\mathcal{MVN}_p(0, (1+\tau^2)I)$. For such $\boldsymbol{X}$, the random variable

$$T = \frac{\sum_{i=1}^p X_i^2}{1+\tau^2}$$

has χ^2_p distribution. Then,

$$\begin{aligned} E\frac{1}{T} &= \int_0^\infty \frac{1}{t}\,\frac{t^{p/2-1}e^{-t/2}}{\Gamma(p/2)2^{p/2}}\,dt \\ &= \frac{\Gamma(p/2-1)}{\Gamma(p/2)}\,\frac{1}{2}\int_0^\infty \frac{t^{p/2-2}e^{-t}}{\Gamma(p/2-1)2^{p/2-1}}dt \\ &= \frac{1}{p/2-1}\cdot\frac{1}{2}\cdot 1 \\ &= \frac{1}{p-2}. \end{aligned}$$

Thus

$$E\frac{1+\tau^2}{\sum_{i=1}^p X_i^2} = \frac{1}{p-2} \quad \Rightarrow \quad E\frac{p-2}{\sum_{i=1}^p X_i^2} = \frac{1}{1+\tau^2}.$$

Therefore, the method-of-moments estimator of $\frac{1}{1+\tau^2}$ is $\frac{p-2}{\sum_{i=1}^p X_i^2}$, which yields an empirical Bayes estimator

$$\delta_{EB}(\boldsymbol{X}) = \left(1 - \frac{p-2}{\sum_{i=1}^p X_i^2}\right)\boldsymbol{X}.$$

1.2.1 ML II

We already have seen the spirit of ML II method in Parametric Empirical Bayes. The ML II approach was proposed by I. J. Good, a statistician that was in the team who broke German code in the WWII. The idea is to mimic the maximum likelihood estimation at the marginal level: Select a prior π that maximizes $m_\pi(x)$, given the data.

Figure 5: I. J. Good, born 1916

Exercise: Suppose that Bayesian have chosen a prior π_0 but wants to look at all priors close to π_0. One such family of priors, "close" to π_0 is contamination family,

$$\Gamma = \{(1-\epsilon)\pi_0 + \epsilon q, \;\; q \in Q\}.$$

Suppose that Q is a family of *all* distributions. Determine empirical Bayes choice.

Hint: Observe that for $\pi \in \Gamma$ the marginal is $m_\pi(x) = (1-\epsilon)m_0(x) + \epsilon m_q(x)$. Also, for any model $f(x|\theta)$ a weighted average is smaller than mode, i.e.,

$$\int f(x|\theta)q(\theta)d\theta \leq f(x|\hat{\theta}_{\text{MLE}}) = \int f(x|\theta)\delta(\hat{\theta}_{\text{MLE}}),$$

where $\delta(\hat{\theta}_{\text{MLE}})$ is a point mass distribution concentrated at θ_{MLE}. Thus, the empirical Bayes choice is $\pi(\theta) = (1-\epsilon)\pi_0(\theta) + \epsilon\delta(\hat{\theta}_{\text{MLE}})$.

2 Exercises

1. Assume $X|\theta$ is exponential $\mathcal{E}(1/\theta)$ with density $f(x|\theta) = \frac{1}{\theta}e^{-x/\theta},\ x \geq 0$. Let F be the cdf corresponding to f. Assume a prior on θ, $\pi(\theta)$.

Let $m_\pi(x) = \int_\Theta f(x|\theta)\pi(\theta)d\theta$ be the marginal and $M_\pi(x) = \int_0^x m_\pi(t)dt$ be the corresponding c.d.f.

(a) Show that $\theta = \frac{1-F(x|\theta)}{f(x|\theta)}$.

(b) Show that Bayes estimator of θ with respect π is $\delta(x) = \frac{1-M_\pi(x)}{m_\pi(x)}$.

[Hint. You will need to use a version of Fubini's theorem (Tonelli theorem) and change order of integration. Tonelli theorem allows for change when integrands are nonnegative.]

(c) Suppose you observe $X_i|\theta_i \sim \mathcal{E}(1/\theta_i), i = 1, \ldots, n+1$. Explain how would you estimate θ_{n+1} in the empirical Bayes fashion, using the result in (b).

2. Assume $[X|\theta] \sim \mathcal{N}(\theta, 1)$ and $[\theta|\mu, \tau^2] \sim \mathcal{N}(\mu, \tau^2)$.

(a) Find the marginal for X.

(b) What are the moment matching estimators of μ and τ^2, if the sample $X_i \sim \mathcal{N}(\theta_i, 1),\ i = 1, \ldots, n$, is available.

[Hint. Find the moments of the marginal and be careful, the estimator of the variance need to be nonnegative.]

(c) Propose an empirical Bayes estimator of θ based on the considerations in (b).

(d) What modifications in (b) are needed if you use MLE II estimator of μ and τ^2.

[Sol. moment matching. Marginal is $\mathcal{N}(\mu, 1+\tau^2)$. $\hat{\mu} = \bar{X}$ and estimator of $1+\tau^2$ is s^2. So $\hat{\tau}^2 = \max\{0, s^2 - 1\}$. For MLE $\hat{\tau}^2 = \max\{0, \frac{n-1}{n}s^2 - 1\}$.

3. If the data $X_i \sim f(x_i|\theta)$ can not be reduced by a sufficient statistics, then so called pseudo-Bayes approach is possible. Let T be an estimator of θ for which distribution $g(t|\theta)$ is known and π the adopted prior. Instead of finding the Bayes rule

$$\delta_\pi(x_1, \ldots, x_n) = \frac{\int_\Theta \theta \prod_{i=1}^n f(x_i|\theta)\pi(\theta)d\theta}{\int_\Theta \prod_{i=1}^n f(x_i|\theta)\pi(\theta)d\theta}$$

one finds the pseudo-Bayes rule as

$$\delta_\pi^*(t) = \frac{\int_\Theta t g(t|\theta)\pi(\theta)d\theta}{\int_\Theta g(t|\theta)\pi(\theta)d\theta}.$$

Suppose that you have model

$$X_i \sim \mathcal{N}(\theta, 1),$$
$$\theta \sim \mathcal{N}(\mu, \tau^2).$$

For n large, the distribution of the sample median $m = \text{Med}(X_1, X_2, \ldots, X_n)$, is approximately $\mathcal{N}(\theta, \frac{\pi}{2n})$.
Write down pseudo-Bayes estimator of θ using the median estimator of θ and its normal approximation.

4. Another way to justify Stein shrinkage estimator is to mimic Exercise 2. Suppose you have the model

$$X \sim \mathcal{MVN}_p(\boldsymbol{\theta}, I),$$
$$\boldsymbol{\theta} \sim \mathcal{MVN}(0, \tau^2 I).$$

(i) Find the marginal distribution.
(ii) Show that in (i) the MLE of τ^2 is

$$\hat{\tau}^2 = \begin{cases} \frac{\sum x_i^2}{p} - 1, & \sum x_i^2 > p \\ 0, & \text{else} \end{cases}$$

(iii) Replacing the MLE in the Bayes estimator

$$\delta_\pi(\boldsymbol{x}) = \frac{\tau^2 \boldsymbol{x}}{1 + \tau^2}$$

the *truncated* James-Stein estimator is obtained,

$$\delta_{\text{EB}}(\boldsymbol{x}) = \left(1 - \frac{p}{\sum x_i^2}\right)_+ \boldsymbol{x},$$

where $(x)_+ = \max\{x, 0\}$ is the positive part of x.

References

[1] Berger, J. (1985). *Statistical Decision Theory and Bayesian Analysis*, Second Edition, Springer Verlag.

[2] Maritz and Lwin (1989) *Empirical Bayes Methods*, Second Edition, Chapman and Hall, New York,

[3] Robert, C. (2001). *Bayesian Choice*, Second Edition, Springer Verlag.

[4] Von Mises, R. (1943). On the correct use of Bayes formula. *Ann. Math. Statist.*, **13**, 156–165.

Appendix

Mathematica code for finding the hierarchical Bayes estimator in the example about the political pool.

```
Integrate[
p Gamma[n + 1]/(Gamma[x + 1] Gamma[n - x + 1]) p^x (1 - p)^(n - x)
(1 - Abs[1 - 2p])/(4 Log[2] p (1 - p)) /. {n -> 20, x -> 12} , {p, 0, 1}]/
Integrate[
Gamma[n + 1]/(Gamma[x + 1] Gamma[n - x + 1]) p^x (1 - p)^(n - x)
 (1 - Abs[1 - 2p])/(4 Log[2] p (1 - p)) /. {n ->20, x -> 12}  , {p, 0, 1}] // N
```