the likelihood function $L(\boldsymbol{\theta}|y) = f(y|\boldsymbol{\theta})$. This involves taking derivatives of L with respect to $\boldsymbol{\theta}$ and yields the maximum likelihood estimate. To draw statistical conclusions, one has to derive the distributional properties of the maximum likelihood estimate (MLE) analytically (often, by assuming a sufficiently large sample), or obtain them via simulations (e.g., bootstrapping).

In contrast, in the Bayesian approach we consider $\boldsymbol{\theta}$ as a random quantity which has some prior distribution $\pi(\boldsymbol{\theta}|\boldsymbol{\eta})$, where the vector $\boldsymbol{\eta}$ is referred to as "hyper-parameter." Then, to perform statistical inference with regard to $\boldsymbol{\theta}$, we obtain the posterior distribution of $\boldsymbol{\theta}$ given the data and the hyper-parameter, $p(\boldsymbol{\theta}|y, \boldsymbol{\eta})$, which is proportional to $f(y|\boldsymbol{\theta}) \cdot \pi(\boldsymbol{\theta}|\boldsymbol{\eta})$ (namely, the distribution function, $f(y|\boldsymbol{\theta})$, times the prior distribution, $\pi(\boldsymbol{\theta}|\boldsymbol{\eta})$.) To obtain the exact form of $p(\boldsymbol{\theta}|y, \boldsymbol{\eta})$ we *marginalize* or average out over the possible values of $\boldsymbol{\theta}$, and use Bayes' rule

$$p(\boldsymbol{\theta}|y,\boldsymbol{\eta}) = \frac{f(y|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\boldsymbol{\eta})}{\int f(y|\mathbf{u})\pi(\mathbf{u}|\boldsymbol{\eta})d\mathbf{u}} = \frac{f(y|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\boldsymbol{\eta})}{m(y|\boldsymbol{\eta})}. \tag{5}$$

With the posterior distribution in hand, inference with regard to $\boldsymbol{\theta}$ is readily available without further derivations. However, obtaining the posterior distribution analytically is hard, unless the prior distribution is a conjugate of $f(y|\boldsymbol{\theta})$. Numerical integration via MCMC simulations is a viable option in many cases, but it may be computationally infeasible if the dimension of the vector $\boldsymbol{\theta}$ is large (as is the case in regression models involving many predictors.). Another problem with the Bayesian approach is the specification of the prior distribution and choosing the hyper-parameter. For example, George and Foster (2000) show that some priors which may be perceived as objective may (a) be impractical for large p because the MCMC sampling will cover a very small portion of the support of the posterior distribution, and, (b) yield results which are dominated by the prior, rather than the data. In addition, as stated in Carlin and Louis (2000), assessing the convergence of MCMC simulations is far from being easy.

3.2 | Empirical Bayes

The EB approach can be seen as a bridge between the frequentist and Bayesian approaches. Starting with a Bayesian hierarchical model, the critical EB step is to use the marginal distribution $m(y|\boldsymbol{\eta})$ in the denominator of Equation (5) to obtain an estimate for $\boldsymbol{\eta}$, through (typically) maximum likelihood estimation:

$$\hat{\boldsymbol{\eta}} = \arg\max_{\boldsymbol{\eta}} m(y|\boldsymbol{\eta}). \tag{6}$$

Then, the analysis proceeds with $p(\boldsymbol{\theta}|y,\hat{\boldsymbol{\eta}})$ as an approximation for the posterior distribution of $\boldsymbol{\theta}$. Importantly, this estimate of $\boldsymbol{\eta}$ is based on the data. In contrast, using a fully Bayesian approach, one would either choose an arbitrarily value (but usually with the intention to do so in a noninformative way so as to avoid introducing bias), or put yet another prior to specify a distribution of the hyper-parameter. The difference between EB and the fully Bayesian approaches was summarized by Efron (2014) as follows: "the essential empirical Bayes task: learning an appropriate prior distribution from ongoing statistical experience, rather than knowing it by assumption."

The hierarchical modeling approach which EB inherits from the Bayesian framework leads to "shrinkage estimation" and "information borrowing"—terms which we feel are best understood through a famous example (Carlin & Louis, 2010; Efron & Morris, 1973). Suppose that for some known σ^2 and τ^2 we assume the following hierarchical model:

$$\begin{aligned} Y_i \mid \theta_i &\sim N(\theta_i,\sigma^2) \\ \theta_i \mid \eta &\sim N(\eta,\tau^2). \end{aligned}$$

Note that in this case θ_i and η are scalars. Furthermore, notice that each Y_i is assumed to have a different mean. It can be shown that the Y_is are marginally i.i.d. with a normal distribution,

$$Y_i \mid \eta \sim N\left(\eta,\sigma^2+\tau^2\right). \tag{7}$$

The common prior distribution of all the θ_is allows us to estimate their overall mean, η, simply by calculating the sample mean of the marginal distribution:

$$\hat{\eta} = \bar{y} = \frac{1}{p}\sum_{i=1}^{p} y_i.$$

The critical EB step is to plug in this estimate in Equation (5) and obtain the posterior distribution:

$$\theta_i \mid y_i,\hat{\eta} \sim N\left(B\hat{\eta}+(1-B)y_i(1-B)\sigma^2\right)$$

where $B = \sigma^2/(\sigma^2 + \tau^2)$. Thus, the EB estimator of θ_i is

$$\hat{\theta}_i^{JS} = B\hat{\eta}+(1-B)y_i. \tag{8}$$