Hierarchical & Empirical Bayesian Analysis

Robert L. Wolpert
Department of Statistical Science
Duke University, Durham, NC, USA

1 Empirical Bayes

The Bayesian approach to making inference about the parameter θ of a random sample $\{X_i\} \stackrel{\text{iid}}{\sim} f_\theta(x)$ from a parametric family $\mathcal{P} = \{f_\theta(x) : \theta \in \Theta\}$ begins by selecting a prior distribution $\pi(\theta)$. One "objective" approach is to let the data help with this prior selection. First, an example.

Normal Example

Let $\{X_i\} \stackrel{\text{iid}}{\sim} \mathsf{No}(\mu_i, 1)$ and let $\mu_i \stackrel{\text{iid}}{\sim} \mathsf{No}(\theta, \tau^2)$ for some indeterminant prior mean θ and variance τ^2. The *marginal* distribution for the $\{X_i\}$ is

$$f(x) = \int_\Theta f_\mu(x)\, \pi(\mu)\, d\mu,$$

easily shown to be $X_i \sim \mathsf{No}(\theta, \tau^2 + 1)$. For large enough sample size p we should expect that θ and τ^2 would be close to their MLEs $\hat\theta = \bar X_p$ and $\hat\tau^2 = \left[\frac{1}{n}\sum (X_i - \bar X_p)^2 - 1\right]_+$ (this used to be called "type-two maximum likelihood", but one seldom hears that phrase nowadays). *Empirical Bayes* inference proceeds by doing ordinary Bayesian analysis with prior distribution $\pi = \mathsf{No}(\hat\theta, \hat\tau^2)$, as if this had been the choice all along.

Conditional on θ and τ^2, the Bayes posterior distribution of $\vec\mu$ is $\mathsf{No}(M, V)$ with mean and variance

$$M = \mathsf{E}[\vec\mu \mid \vec X, \theta, \tau^2] = \frac{\tau^2}{1+\tau^2}\vec X + \frac{1}{1+\tau^2}\theta \qquad V = \mathsf{V}[\vec\mu \mid \vec X, \theta, \tau^2] = \frac{\tau^2}{1+\tau^2}$$

for sample-size $n = 1$, so the squared-error Bayes risk of the posterior mean is

$$r = \mathsf{E}\sum_{i=1}^p (\mu_i - M_i)^2 = \sum_{i=1}^p V_i = \frac{p\tau^2}{1+\tau^2}$$

Marginally $X_i \sim \mathsf{No}(\theta, \tau^2+1)$ so $(X_i - \theta)/\sqrt{\tau^2+1}$ are iid $\mathsf{No}(0,1)$ and $\|X - \theta\|^2/(\tau^2+1) \sim \chi^2_p$ and $\|X - \bar X_p\|^2/(\tau^2+1) \sim \chi^2_{p-1}$. The expected inverse of any $Y \sim \mathsf{Ga}(\alpha, \beta)$ random variable is $\mathsf{E}[1/Y] = \beta/(\alpha - 1)$ for $\alpha > 1$, and in particular for χ^2 variables we have $\mathsf{E}\|X - \theta\|^{-2} = 1/[(p-2)(1+\tau^2)$ and $\mathsf{E}\|X - \bar X_p\|^{-2} = 1/[(p-3)(1+\tau^2)$, so

$$\mathsf{E}\left[1 - \frac{p-2}{\|X-\theta\|^2}\right] = \frac{\tau^2}{1+\tau^2} = \mathsf{E}\left[1 - \frac{p-3}{\|X - \bar X\|^2}\right]$$

Estimating $\tau^2/(1+\tau^2)$ by $1-(p-2)/\|X-\theta\|^2$ in the expression for M leads to the James-Stein estimator

$$\delta_{\text{JS}}(X) = \left[1 - \frac{p-2}{\|X-\theta\|^2}\right]\vec{X} + \left[\frac{p-2}{\|X-\theta\|^2}\right]\theta = \vec{X} + \frac{p-2}{\|X-\theta\|^2}(\theta - \vec{X})$$

that shrinks towards θ in $p > 2$ dimensions (most authors follow Stein in shrinking towards $\theta = 0$), while estimating it by $1-(p-3)/\|X-\bar{X}\|^2$ leads to a related estimator

$$\delta_{\text{JS}-\bar{X}}(X) = \vec{X} + \frac{p-3}{\|X-\bar{X}_p\|^2}(\bar{X}_p - \vec{X})$$

in dimensions four or more. One can show (and Young & Smith do) that the Bayes risk of δ_{JS} is

$$r(\tau, \delta_{\text{JS}}) = p - \frac{p-2}{\tau^2+1} = r(\tau, \delta_\tau^\star) + \frac{2}{\tau^2+1},$$

exceeding the risk of the Bayes estimator $\delta_\tau^\star$ for a known value of τ^2 by an amount $2/(\tau^2+1)$ that may be interpreted as the price for having to estimate τ^2 from the data.

Binomial Example

Let $X_i \overset{\text{ind}}{\sim} \mathsf{Bi}(n_i, p_i)$ be independent Binomial random variables, the numbers of successes in known numbers n_i of trials with possibly different success probabilities $\{p_i\}$, and assign conjugate Beta prior probability distribution $\{p_i\} \overset{\text{iid}}{\sim} \mathsf{Be}(\alpha, \beta)$. For specified values α, β of the hyperparameters, the posterior distribution of p_i given $\mathbf{x} = \{X_i\}$ would be $p_i \mid \mathbf{x} \sim \mathsf{Be}(\alpha_i^\star, \beta_i^\star)$ for $\alpha_i^\star = \alpha + x_i$ and $\beta_i^\star = \beta + n_i - x_i$, with mean $\mathsf{E}[p_i \mid \mathbf{x}] = \alpha_i^\star/(\alpha_i^\star + \beta_i^\star)$— but what if we don't wish to specify α, β? The marginal distribution of each X_i is the "beta-binomial" distribution with pmf

$$\begin{aligned} m_i(x) &= \int_0^1 \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \binom{n_i}{x} p^{\alpha+x-1}(1-p)^{\beta+n-x}\,dp \\ &= \frac{\Gamma(\alpha+\beta) n_i! \Gamma(\alpha+x)\Gamma(\beta+n_i-x)}{\Gamma(\alpha)\Gamma(\beta) x! (n_i-x)! \Gamma(\alpha+\beta+n_i)} \end{aligned} \tag{1}$$

with marginal mean $\mathsf{E}[X_i] = n_i\alpha/\beta$ and marginal variance $\mathsf{V}[X_i] = n_i\alpha\beta/(\alpha+\beta)(\alpha+\beta+1) + n_i^2\alpha/\beta^2$. Using either Method of Moments with these means and variances, or maximizing $\sum \log m_i(x_i)$ from (1), we can find data-dependent estimates $\hat{\alpha}$, $\hat{\beta}$. With these in hand the estimated binomial means become

$$\bar{\theta}_i = \frac{\hat{\alpha}+x_i}{\hat{\alpha}+\hat{\beta}+n_i} = \left\{\frac{n_i}{\hat{\alpha}+\hat{\beta}+n_i}\right\}\frac{x_i}{n_i} + \left\{\frac{\hat{\alpha}+\hat{\beta}}{\hat{\alpha}+\hat{\beta}+n_i}\right\}\frac{\hat{\alpha}}{\hat{\alpha}+\hat{\beta}}$$

shrunk from the MLE x_i/n_i towards an overall mean $\hat{\alpha}/(\hat{\alpha}+\hat{\beta})$.

Poisson Example

Let $X_i \overset{\text{ind}}{\sim} \mathsf{Po}(\theta_i)$ be independent Poisson-distributed random variables, and assign independent $\{\theta_i\} \overset{\text{iid}}{\sim} \mathsf{Ga}(\alpha, \beta)$ prior distributions to the Poisson means. The marginal distributions for the $\{X_i\}$

are negative binomial, with pmf

$$
\begin{aligned}
m_i(x) &= \int_0^\infty \frac{\theta^x}{x!}\, e^{-\theta} \quad \times \quad \frac{\beta^\alpha \theta^{\alpha-1}}{\Gamma(\alpha)} e^{-\beta\theta}\, d\theta \\
&= \frac{\Gamma(\alpha+x)}{\Gamma(\alpha)\, x!} \left(\frac{\beta}{\beta+1}\right)^\alpha \left(\frac{1}{\beta+1}\right)^x, \qquad (2)
\end{aligned}
$$

with mean $\mathsf{E}[X_i] = \alpha/\beta$ and variance $\alpha(\beta+1)/\beta^2$. By either using MOM with these moments or maximizing $\sum \log m_i(x)$ from (2), we can find estimates $\hat{\alpha}$, $\hat{\beta}$ for the hyper-parameters and, using them, find EB estimates of the Poisson means

$$
\bar{\theta}_i = \frac{\hat{\alpha} + x_i}{\hat{\beta} + 1}
$$

that are shrunk from the MLE x_i toward a common value $\hat{\alpha}/\hat{\beta}$.

2 Hierarchical Bayes

An alternative to *estimating* the values of "hyper-parameters" like θ and τ^2 above is to *model* uncertainty about them and through a Bayesian prior distribution. To simplify the presentation let's introduce the *precision* parameter $\lambda = 1/\tau^2$. A conjugate hierarchical model for the data of Section (1) would be

$$
\begin{aligned}
X_i \mid \mu_i &\sim \mathsf{No}(\mu_i, 1) \\
\mu_i \mid \theta, \lambda &\sim \mathsf{No}(\theta, \lambda^{-1}) \\
\theta, \lambda &\sim \lambda^{\alpha-1} \beta^\alpha e^{-\beta\lambda},
\end{aligned}
$$

an improper prior for *a priori* independent $\theta \sim \mathsf{Un}(\mathbb{R})$ and $\lambda \sim \mathsf{Ga}(\alpha, \beta)$.

Notes still in-progress. Next steps: evaluate available conditional distributions, discuss MCMC approach to learning about the $\{\mu_i\}$s, contrast with the EB approach above. Another example: Re-parametrize the NB of (2) by $p = \beta/(\beta+1)$; fix α; and use $p \sim \mathsf{Be}(a, b)$ hyper-prior distribution. Discuss MCMC evaluation in hierarchical models. Perhaps discuss Morris correction and Robbins Miracle. Mention PEB has lower ensemble risk; better frequentist risk than the MLE despite its UMVUE properties. Make connection with Stein.

3 Bayesian Forecasting

Next steps: Make forecasts; illustrate with Pareto model for volcanic eruption durations.