Foundations of Statistical Inference

J. Berestycki & D. Sejdinovic

Department of Statistics
University of Oxford

MT 2019

Chapter 11: Empirical Bayes

Empirical Bayes

Bayes estimators have good risk properties (for example, the posterior mean is usually admissible for quadratic loss).

However, Bayes estimators may be hard to compute, particularly for hierarchical models.

In Empirical Bayes, we use hierarchical Bayesian reasoning to derive estimators, but with a particular strategy to simplify hierarchical models.

Empirical Bayes

Recall the setup for Bayesian inference for hierarchical models.

$$\begin{aligned} X &\sim f(x;\theta) \\ \theta &\sim \pi(\theta;\psi) \\ \psi &\sim g(\psi) \end{aligned}$$

Our prior for θ has a parameter ψ which also has a prior. The posterior is

$$\pi(\theta, \psi|x) \propto L(\theta; x)\pi(\theta; \psi)g(\psi)$$

If we want minimum risk for quadratic loss we should use the posterior mean:

$$\hat{\theta} = \int \int \theta\pi(\theta, \psi|x)d\theta d\psi$$

Empirical Bayes

Recall the setup for Bayesian inference for hierarchical models.

$$\begin{aligned} X &\sim f(x;\theta) \\ \theta &\sim \pi(\theta;\psi) \\ \psi &\sim g(\psi) \end{aligned}$$

Our prior for θ has a parameter ψ which also has a prior. The posterior is

$$\pi(\theta, \psi|x) \propto L(\theta; x)\pi(\theta; \psi)g(\psi)$$

If we want minimum risk for quadratic loss we should use the posterior mean:

$$\hat{\theta} = \int\int \theta\pi(\theta, \psi|x)d\theta d\psi$$

Empirical Bayes

Recall the setup for Bayesian inference for hierarchical models.

$$\begin{aligned} X &\sim f(x;\theta) \\ \theta &\sim \pi(\theta;\psi) \\ \psi &\sim g(\psi) \end{aligned}$$

Our prior for θ has a parameter ψ which also has a prior. The posterior is

$$\pi(\theta, \psi|x) \propto L(\theta; x)\pi(\theta; \psi)g(\psi)$$

If we want minimum risk for quadratic loss we should use the posterior mean:

$$\hat{\theta} = \int \int \theta\pi(\theta, \psi|x)d\theta d\psi$$

Empirical Bayes

Empirical Bayes (EB)

The EB approach is to avoid doing ψ-integrals by replacing ψ with a point estimate $\hat{\psi}$, derived from the data, and consider the model

$$\begin{aligned} X &\sim f(x;\theta) \\ \theta &\sim \pi(\theta;\hat{\psi}) \end{aligned}$$

This EB approximation to the full posterior 'chops off' a layer of the hierarchy. The reduced model has posterior

$$\hat{\pi}(\theta|x) \propto L(\theta;x)\pi(\theta;\hat{\psi}),$$

and a Bayes estimator $\hat{\theta}_{EB}$ is calculated using $\hat{\pi}(\theta|x)$. For example, for quadratic loss,

$$\hat{\theta} = \int \theta \hat{\pi}(\theta|x) d\theta.$$

Empirical Bayes

Empirical Bayes (EB)

The EB approach is to avoid doing ψ-integrals by replacing ψ with a point estimate $\hat{\psi}$, derived from the data, and consider the model

$$\begin{aligned} X &\sim f(x;\theta) \\ \theta &\sim \pi(\theta;\hat{\psi}) \end{aligned}$$

This EB approximation to the full posterior 'chops off' a layer of the hierarchy. The reduced model has posterior

$$\hat{\pi}(\theta|x) \propto L(\theta;x)\pi(\theta;\hat{\psi}),$$

and a Bayes estimator $\hat{\theta}_{EB}$ is calculated using $\hat{\pi}(\theta|x)$. For example, for quadratic loss,

$$\hat{\theta} = \int \theta\hat{\pi}(\theta|x)d\theta.$$

Empirical Bayes

We still need an estimator for ψ. There are several choices.

We can use the MMLE $\hat{\psi} = \arg\max_{\psi} p(x|\psi)$ for ψ in the marginal likelihood

$$p(x|\psi) = \int L(\theta; x)\pi(\theta; \psi)d\theta.$$

Moment-matching estimators are also used: e.g. choose $\hat{\psi}$ such that $\pi(\theta; \hat{\psi})$ has the same mean and variance as the sample mean and variance of the MLEs of θ_i.

Empirical Bayes

We still need an estimator for ψ. There are several choices.

We can use the MMLE $\hat{\psi} = \arg\max_{\psi} p(x|\psi)$ for ψ in the marginal likelihood

$$p(x|\psi) = \int L(\theta; x)\pi(\theta; \psi)d\theta.$$

Moment-matching estimators are also used: e.g. choose $\hat{\psi}$ such that $\pi(\theta; \hat{\psi})$ has the same mean and variance as the sample mean and variance of the MLEs of θ_i.

Empirical Bayes

We still need an estimator for ψ. There are several choices.

We can use the MMLE $\hat{\psi} = \arg\max_{\psi} p(x|\psi)$ for ψ in the marginal likelihood

$$p(x|\psi) = \int L(\theta; x)\pi(\theta; \psi)d\theta.$$

Moment-matching estimators are also used: e.g. choose $\hat{\psi}$ such that $\pi(\theta; \hat{\psi})$ has the same mean and variance as the sample mean and variance of the MLEs of θ_i.

Beta-binomial example

Meta-analysis of study of tumors in rodents.

Previous experiments:

0/20	0/20	0/20	0/20	0/20	0/20	0/20	0/19	0/19	0/19
0/19	0/18	0/18	0/17	1/20	1/20	1/20	1/20	1/19	1/19
1/18	1/18	2/25	2/24	2/23	2/20	2/20	2/20	2/20	2/20
2/20	1/10	5/49	2/19	5/46	3/27	2/17	7/49	7/47	3/20
3/20	2/13	9/48	10/50	4/20	4/20	4/20	4/20	4/20	4/20
4/20	10/48	4/19	4/19	4/19	5/22	11/46	12/49	5/20	5/20
6/23	5/19	6/22	6/20	6/20	6/20	16/52	15/47	15/46	9/24

Current experiment:
4/14

Table 5.1 *Tumor incidence in historical control groups and current group of rats, from Tarone (1982). The table displays the values of* $\frac{y_j}{n_j}$*: (number of rats with tumors)/(total number of rats).*

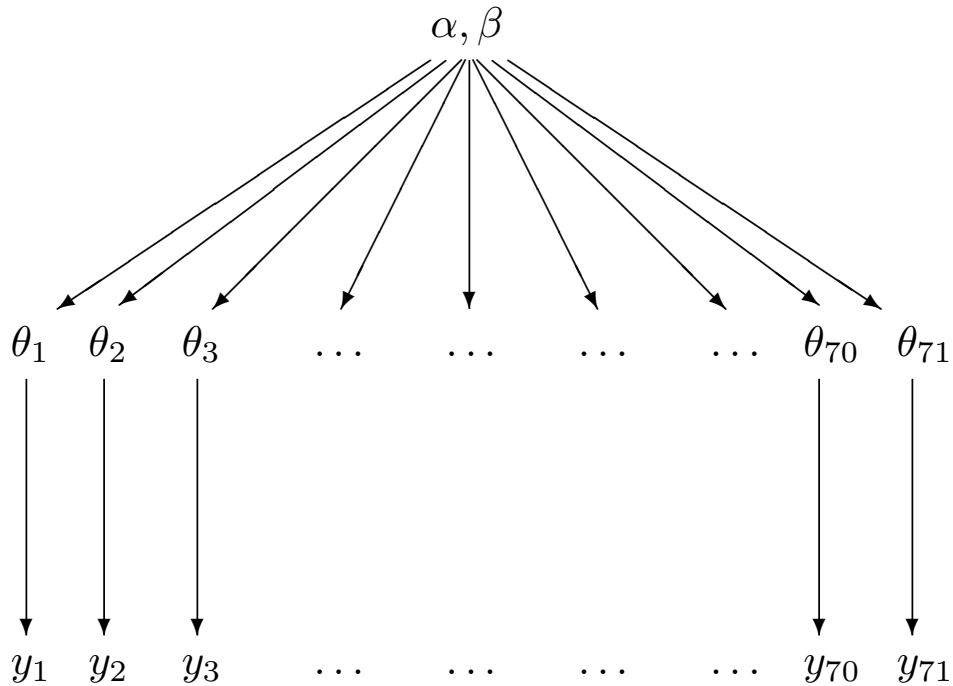
Assume # tumors $\sim$ Bin(n, θ). New experiment $n = 14$ and $Y = 4$. MLE is $4/14 = 0.286$. With a conjugate prior $\theta \sim$ Beta(α, β), posterior is

$$p(\theta|y) = Beta(\alpha + 4, \beta + 10).$$

E.g. for Jeffreys prior $Beta(1/2, 1/2)$, posterior mean is $4.5/15 = 0.3$.

Beta-binomial example

Meta-analysis of study of tumors in rodents.



Assume # tumors $\sim$ Bin(n, θ). New experiment $n = 14$ and $Y = 4$. MLE is $4/14 = 0.286$. With a conjugate prior $\theta \sim$ Beta(α, β), posterior is

$$p(\theta|y) = Beta(\alpha + 4, \beta + 10).$$

E.g. for Jeffreys prior $Beta(1/2, 1/2)$, posterior mean is $4.5/15 = 0.3$.

Beta-binomial example

Empirical Bayes using moment matching:

1. Compute MLEs Y_i/n_i for previous experiments, $i = 1, \ldots, 70$.
2. Compute their sample mean and sample variance: $m = 0.136$, $v = 0.0106$.
3. Pick $\hat{\alpha}, \hat{\beta}$ such that $Beta(\alpha, \beta)$ has "matched moments", i.e. mean $\alpha/(\alpha + \beta) = m$, and variance $\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} = v$. Get $\hat{\alpha} = 1.4, \hat{\beta} = 8.6$.

Posterior is now:

$$p(\theta|y) \sim Beta(5.4, 18.6).$$

Posterior mean is 0.225, lower than $4/14 = 0.286$.

James-Stein estimator as an EB estimator

Data $x_i \sim \mathcal{N}(\theta_i, 1)$, $i = 1, \ldots, p$ (so one observation x_i for each parameter θ_i). The MLE for θ_i is simply $\hat{\theta}_{MLE,i} = x_i$.

Suppose the prior is $\theta_i \sim \mathcal{N}(0, \tau^2)$.

If we knew τ we would have (simple completing the square exercise)

$$\theta_i | (x_i, \tau) \sim \mathcal{N}\left(\frac{x_i \tau^2}{1+\tau^2}, \frac{\tau^2}{1+\tau^2} \right).$$

James-Stein estimator as an EB estimator

Data $x_i \sim \mathcal{N}(\theta_i, 1)$, $i = 1, \ldots, p$ (so one observation x_i for each parameter θ_i). The MLE for θ_i is simply $\hat{\theta}_{MLE,i} = x_i$.

Suppose the prior is $\theta_i \sim \mathcal{N}(0, \tau^2)$.

If we knew τ we would have (simple completing the square exercise)

$$\theta_i | (x_i, \tau) \sim \mathcal{N}\left(\frac{x_i \tau^2}{1+\tau^2}, \frac{\tau^2}{1+\tau^2}\right).$$

James-Stein estimator as an EB estimator

To get an estimate for τ we compute the marginal distribution for X_i given τ, which is $X_i \sim \mathcal{N}(0, \tau^2 + 1)$. The (unconstrained) MMLE for τ is then $\hat{\tau}^2 = \frac{1}{p}\sum_{i=1}^{p} X_i^2 - 1$, and this gives

$$\begin{aligned} \hat{\theta}_{EB,i} &= \frac{X_i \hat{\tau}^2}{1 + \hat{\tau}^2} \\ &= \left(1 - \frac{p}{\sum_i X_i^2}\right) X_i \end{aligned}$$

which is the James-Stein estimator

$$\hat{\theta}_{JS,i} = \left(1 - \frac{a}{\sum_i X_i^2}\right) X_i, \qquad \text{with } a = p.$$

Note: This is not the minimum risk JS estimator for quadratic loss (for which $a = p - 2$), but it already strictly dominates the MLE for all θ. (the JS estimator with $a = p - 2$ can be recovered using a method of moments estimator for τ. See Young and Smith, Section 3.5)

James-Stein estimator as an EB estimator

To get an estimate for τ we compute the marginal distribution for X_i given τ, which is $X_i \sim \mathcal{N}(0, \tau^2 + 1)$. The (unconstrained) MMLE for τ is then $\hat{\tau}^2 = \frac{1}{p}\sum_{i=1}^{p} X_i^2 - 1$, and this gives

$$\begin{aligned} \hat{\theta}_{EB,i} &= \frac{X_i \hat{\tau}^2}{1 + \hat{\tau}^2} \\ &= \left(1 - \frac{p}{\sum_i X_i^2}\right) X_i \end{aligned}$$

which is the James-Stein estimator

$$\hat{\theta}_{JS,i} = \left(1 - \frac{a}{\sum_i X_i^2}\right) X_i, \qquad \text{with } a = p.$$

Note: This is not the minimum risk JS estimator for quadratic loss (for which $a = p - 2$), but it already strictly dominates the MLE for all θ. (the JS estimator with $a = p - 2$ can be recovered using a method of moments estimator for τ. See Young and Smith, Section 3.5)

Example: Poisson

Data $x_i \sim$ Poisson(θ_i), $i = 1, \ldots, n$ (so one observation x_i for each parameter θ_i). The MLE for θ_i is simply $\hat{\theta}_{MLE,i} = x_i$. Construct an EB estimator for quadratic loss.

Suppose the prior for θ_i's is iid Exponential(λ) i.e. $\pi(\theta_i|\lambda) = \lambda e^{-\lambda\theta_i}$.

$$\begin{aligned} p(x_i \mid \lambda) &= \int_0^\infty \frac{e^{-\theta_i}\theta_i^{x_i}}{x_i!}\lambda e^{-\lambda\theta_i} d\theta_i \\ &= \left(\frac{1}{1+\lambda}\right)^{x_i} \frac{\lambda}{1+\lambda} \end{aligned}$$

$\Rightarrow$ given λ the x_is are marginally iid Geometric$(\lambda/(1+\lambda))$ with mean $\frac{1-p}{p} = \lambda^{-1}$.

The MMLE of λ based on $x_1, \ldots, x_n$ is $\widehat{\lambda} = 1/\bar{x}$, where $\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$.

Example: Poisson

Data $x_i \sim$ Poisson(θ_i), $i = 1, \ldots, n$ (so one observation x_i for each parameter θ_i). The MLE for θ_i is simply $\hat{\theta}_{MLE,i} = x_i$. Construct an EB estimator for quadratic loss.

Suppose the prior for θ_i's is iid Exponential(λ) i.e. $\pi(\theta_i|\lambda) = \lambda e^{-\lambda\theta_i}$.

$$\begin{aligned} p(x_i \mid \lambda) &= \int_0^\infty \frac{e^{-\theta_i}\theta_i^{x_i}}{x_i!}\lambda e^{-\lambda\theta_i} d\theta_i \\ &= \left(\frac{1}{1+\lambda}\right)^{x_i} \frac{\lambda}{1+\lambda} \end{aligned}$$

$\Rightarrow$ given λ the x_is are marginally iid Geometric$(\lambda/(1+\lambda))$ with mean $\frac{1-p}{p} = \lambda^{-1}$.

The MMLE of λ based on $x_1, \ldots, x_n$ is $\widehat{\lambda} = 1/\bar{x}$, where $\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$.

Example: Poisson

Data $x_i \sim$ Poisson(θ_i), $i = 1, \ldots, n$ (so one observation x_i for each parameter θ_i). The MLE for θ_i is simply $\hat{\theta}_{MLE,i} = x_i$. Construct an EB estimator for quadratic loss.

Suppose the prior for θ_i's is iid Exponential(λ) i.e. $\pi(\theta_i|\lambda) = \lambda e^{-\lambda\theta_i}$.

$$\begin{aligned} p(x_i \mid \lambda) &= \int_0^\infty \frac{e^{-\theta_i}\theta_i^{x_i}}{x_i!}\lambda e^{-\lambda\theta_i} d\theta_i \\ &= \left(\frac{1}{1+\lambda}\right)^{x_i} \frac{\lambda}{1+\lambda} \end{aligned}$$

$\Rightarrow$ given λ the x_is are marginally iid Geometric$(\lambda/(1+\lambda))$ with mean $\frac{1-p}{p} = \lambda^{-1}$.

The MMLE of λ based on $x_1, \ldots, x_n$ is $\widehat{\lambda} = 1/\bar{x}$, where $\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$.

Example: Poisson

Data $x_i \sim$ Poisson(θ_i), $i = 1, \ldots, n$ (so one observation x_i for each parameter θ_i). The MLE for θ_i is simply $\hat{\theta}_{MLE,i} = x_i$. Construct an EB estimator for quadratic loss.

Suppose the prior for θ_i's is iid Exponential(λ) i.e. $\pi(\theta_i|\lambda) = \lambda e^{-\lambda\theta_i}$.

$$\begin{aligned} p(x_i \mid \lambda) &= \int_0^\infty \frac{e^{-\theta_i}\theta_i^{x_i}}{x_i!}\lambda e^{-\lambda\theta_i} d\theta_i \\ &= \left(\frac{1}{1+\lambda}\right)^{x_i} \frac{\lambda}{1+\lambda} \end{aligned}$$

$\Rightarrow$ given λ the x_is are marginally iid Geometric$(\lambda/(1+\lambda))$ with mean $\frac{1-p}{p} = \lambda^{-1}$.

The MMLE of λ based on $x_1, \ldots, x_n$ is $\widehat{\lambda} = 1/\bar{x}$, where $\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$.

Example: Poisson

Now, under the EB simplification, set $\lambda = \hat{\lambda}$, so that

$$\hat{\pi}(\theta|x) \propto L(\theta; x)\pi(\theta|\hat{\lambda}) \;=\; \prod_{i=1}^{n} e^{-\theta_i}\theta_i^{x_i}\hat{\lambda}e^{-\hat{\lambda}\theta_i}$$

and we recognise $\theta_i|x \sim \Gamma(x_i + 1, \hat{\lambda} + 1)$ in this EB approximation. This leads to an estimator

$$\begin{aligned}\hat{\theta}_{EB,i} &= \int \theta_i \hat{\pi}(\theta_i|x)d\theta_i \\ &= \frac{x_i + 1}{\hat{\lambda} + 1} = \bar{x}\frac{x_i + 1}{\bar{x} + 1}\end{aligned}$$

We can rewrite this

$$\hat{\theta}_{EB,i} = x_i + \frac{\bar{x}}{\bar{x} + 1}(\bar{x} - x_i)$$

showing that this EB estimator shrinks the estimates towards the common mean.

Example: Poisson

Now, under the EB simplification, set $\lambda = \hat{\lambda}$, so that

$$\hat{\pi}(\theta|x) \propto L(\theta; x)\pi(\theta|\hat{\lambda}) \; = \; \prod_{i=1}^{n} e^{-\theta_i} \theta_i^{x_i} \hat{\lambda} e^{-\hat{\lambda}\theta_i}$$

and we recognise $\theta_i|x \sim \Gamma(x_i + 1, \hat{\lambda} + 1)$ in this EB approximation. This leads to an estimator

$$\begin{aligned} \hat{\theta}_{EB,i} &= \int \theta_i \hat{\pi}(\theta_i|x) d\theta_i \\ &= \frac{x_i + 1}{\hat{\lambda} + 1} = \bar{x} \frac{x_i + 1}{\bar{x} + 1} \end{aligned}$$

We can rewrite this

$$\hat{\theta}_{EB,i} = x_i + \frac{\bar{x}}{\bar{x} + 1}(\bar{x} - x_i)$$

showing that this EB estimator shrinks the estimates towards the common mean.

Example: Poisson

Now, under the EB simplification, set $\lambda = \hat{\lambda}$, so that

$$\hat{\pi}(\theta|x) \propto L(\theta; x)\pi(\theta|\hat{\lambda}) \;=\; \prod_{i=1}^{n} e^{-\theta_i}\theta_i^{x_i}\hat{\lambda}e^{-\hat{\lambda}\theta_i}$$

and we recognise $\theta_i|x \sim \Gamma(x_i + 1, \hat{\lambda} + 1)$ in this EB approximation. This leads to an estimator

$$\begin{aligned}\hat{\theta}_{EB,i} &= \int \theta_i \hat{\pi}(\theta_i|x)d\theta_i \\ &= \frac{x_i + 1}{\hat{\lambda} + 1} = \bar{x}\frac{x_i + 1}{\bar{x} + 1}\end{aligned}$$

We can rewrite this

$$\hat{\theta}_{EB,i} = x_i + \frac{\bar{x}}{\bar{x} + 1}(\bar{x} - x_i)$$

showing that this EB estimator shrinks the estimates towards the common mean.

Example: Poisson

Now, under the EB simplification, set $\lambda = \hat{\lambda}$, so that

$$\hat{\pi}(\theta|x) \propto L(\theta; x)\pi(\theta|\hat{\lambda}) \;=\; \prod_{i=1}^{n} e^{-\theta_i}\theta_i^{x_i}\hat{\lambda}e^{-\hat{\lambda}\theta_i}$$

and we recognise $\theta_i|x \sim \Gamma(x_i + 1, \hat{\lambda} + 1)$ in this EB approximation. This leads to an estimator

$$\begin{aligned}\hat{\theta}_{EB,i} &= \int \theta_i \hat{\pi}(\theta_i|x)d\theta_i \\ &= \frac{x_i + 1}{\hat{\lambda} + 1} = \bar{x}\frac{x_i + 1}{\bar{x} + 1}\end{aligned}$$

We can rewrite this

$$\hat{\theta}_{EB,i} = x_i + \frac{\bar{x}}{\bar{x} + 1}(\bar{x} - x_i)$$

showing that this EB estimator shrinks the estimates towards the common mean.

Example: Poisson

Now, under the EB simplification, set $\lambda = \hat{\lambda}$, so that

$$\hat{\pi}(\theta|x) \propto L(\theta; x)\pi(\theta|\hat{\lambda}) \;=\; \prod_{i=1}^{n} e^{-\theta_i}\theta_i^{x_i}\hat{\lambda}e^{-\hat{\lambda}\theta_i}$$

and we recognise $\theta_i|x \sim \Gamma(x_i + 1, \hat{\lambda} + 1)$ in this EB approximation. This leads to an estimator

$$\begin{aligned}\hat{\theta}_{EB,i} &= \int \theta_i \hat{\pi}(\theta_i|x)d\theta_i \\ &= \frac{x_i + 1}{\hat{\lambda} + 1} = \bar{x}\frac{x_i + 1}{\bar{x} + 1}\end{aligned}$$

We can rewrite this

$$\hat{\theta}_{EB,i} = x_i + \frac{\bar{x}}{\bar{x} + 1}(\bar{x} - x_i)$$

showing that this EB estimator shrinks the estimates towards the common mean.

Some surprising features of the MMLE

Consider the hierarchical model $X|\theta \sim f(\cdot|\theta)$, $\theta|\psi \sim \pi(\cdot|\psi)$. The marginal likelihood is

$$m(X|\psi) = \int_\Theta f(X|\theta)\pi(\theta|\psi)d\theta$$

maximum marginal likelihood estimator

$$\hat{\psi} = \text{argmax}_\psi m(X|\psi)$$

Simple example

$$X = (X_1, \cdots, X_n) \overset{iid}{\sim} \mathcal{N}(\mu, 1), \quad \mu \sim \mathcal{N}(\mu_0, \tau^2)$$

$$\bar{X}_n|\mu_0, \tau \sim \mathcal{N}(\mu_0, \tau^2 + 1/n)$$

- Fixed τ, optimization in μ_0

$$\hat{\mu}_0 = \bar{X}_n, \quad [\mu|X] \sim \mathcal{N}(\bar{X}_n, (n + 1/\tau^2)^{-1})$$

This posterior behaves similarly to a posterior with fixed μ_0, τ^2

Some surprising features of the MMLE

Consider the hierarchical model $X|\theta \sim f(\cdot|\theta)$, $\theta|\psi \sim \pi(\cdot|\psi)$. The marginal likelihood is

$$m(X|\psi) = \int_\Theta f(X|\theta)\pi(\theta|\psi)d\theta$$

maximum marginal likelihood estimator

$$\hat{\psi} = \text{argmax}_\psi m(X|\psi)$$

Simple example

$$X = (X_1, \cdots, X_n) \overset{iid}{\sim} \mathcal{N}(\mu, 1), \quad \mu \sim \mathcal{N}(\mu_0, \tau^2)$$

$$\bar{X}_n|\mu_0, \tau \sim \mathcal{N}(\mu_0, \tau^2 + 1/n)$$

- Fixed τ, optimization in μ_0

$$\hat{\mu}_0 = \bar{X}_n, \quad [\mu|X] \sim \mathcal{N}(\bar{X}_n, (n + 1/\tau^2)^{-1})$$

This posterior behaves similarly to a posterior with fixed μ_0, τ^2

Some surprising features of the MMLE

Consider the hierarchical model $X|\theta \sim f(\cdot|\theta)$, $\theta|\psi \sim \pi(\cdot|\psi)$. The marginal likelihood is

$$m(X|\psi) = \int_\Theta f(X|\theta)\pi(\theta|\psi)d\theta$$

maximum marginal likelihood estimator

$$\hat{\psi} = \text{argmax}_\psi m(X|\psi)$$

Simple example

$$X = (X_1, \cdots, X_n) \overset{iid}{\sim} \mathcal{N}(\mu, 1), \quad \mu \sim \mathcal{N}(\mu_0, \tau^2)$$

$$\bar{X}_n|\mu_0, \tau \sim \mathcal{N}(\mu_0, \tau^2 + 1/n)$$

- Fixed τ, optimization in μ_0

$$\hat{\mu}_0 = \bar{X}_n, \quad [\mu|X] \sim \mathcal{N}(\bar{X}_n, (n + 1/\tau^2)^{-1})$$

This posterior behaves similarly to a posterior with fixed μ_0, τ^2

Some surprising features of the MMLE

Consider the hierarchical model $X|\theta \sim f(\cdot|\theta)$, $\theta|\psi \sim \pi(\cdot|\psi)$. The marginal likelihood is

$$m(X|\psi) = \int_\Theta f(X|\theta)\pi(\theta|\psi)d\theta$$

maximum marginal likelihood estimator

$$\hat{\psi} = \text{argmax}_\psi m(X|\psi)$$

Simple example

$$X = (X_1, \cdots, X_n) \overset{iid}{\sim} \mathcal{N}(\mu, 1), \quad \mu \sim \mathcal{N}(\mu_0, \tau^2)$$

$$\bar{X}_n|\mu_0, \tau \sim \mathcal{N}(\mu_0, \tau^2 + 1/n)$$

- Fixed τ, optimization in μ_0

$$\hat{\mu}_0 = \bar{X}_n, \quad [\mu|X] \sim \mathcal{N}(\bar{X}_n, (n + 1/\tau^2)^{-1})$$

This posterior behaves similarly to a posterior with fixed μ_0, τ^2

- Fixed μ_0, optimization in τ

$$\ell_n(\mu_0, \tau^2) = -\frac{n(\bar{X}_n - \mu_0)^2}{2(n\tau^2+1)} - \frac{1}{2}\log(n\tau^2+1)$$

$$\frac{\partial \ell_n(\tau^2)}{\partial \tau^2} = \frac{n^2(\bar{X}_n - \mu_0)^2}{2(n\tau^2+1)^2} - \frac{n}{2(n\tau^2+1)} = 0 \;\Leftrightarrow\; \tau^2 = (\bar{X}_n - \mu_0)^2 - 1/n$$

  - If $(\bar{X}_n - \mu_0)^2 - 1/n > 0$ then $\hat{\tau}^2 = (\bar{X}_n - \mu_0)^2 - 1/n$
  - If $(\bar{X}_n - \mu_0)^2 - 1/n \leq 0$ then $\ell_n(\tau^2)$ is non increasing and $\hat{\tau}^2 = 0$

$$\hat{\tau}^2 = \left((\bar{X}_n - \mu_0)^2 - 1/n\right)_+$$

and $\hat{\mu} = \mu_0$ if $\hat{\tau}^2 = 0$ and if $\hat{\tau} > 0$

$$\hat{\mu} = \bar{X}_n\left(1 - \frac{1}{n(\bar{X}_n - \mu_0)^2}\right) + \frac{\mu_0}{n(\bar{X}_n - \mu_0)^2}$$

- Fixed μ_0, optimization in τ

$$\ell_n(\mu_0, \tau^2) = -\frac{n(\bar{X}_n - \mu_0)^2}{2(n\tau^2+1)} - \frac{1}{2}\log(n\tau^2+1)$$

$$\frac{\partial \ell_n(\tau^2)}{\partial \tau^2} = \frac{n^2(\bar{X}_n - \mu_0)^2}{2(n\tau^2+1)^2} - \frac{n}{2(n\tau^2+1)} = 0 \;\Leftrightarrow\; \tau^2 = (\bar{X}_n - \mu_0)^2 - 1/n$$

- If $(\bar{X}_n - \mu_0)^2 - 1/n > 0$ then $\hat{\tau}^2 = (\bar{X}_n - \mu_0)^2 - 1/n$
- If $(\bar{X}_n - \mu_0)^2 - 1/n \leq 0$ then $\ell_n(\tau^2)$ is non increasing and $\hat{\tau}^2 = 0$

$$\hat{\tau}^2 = \left((\bar{X}_n - \mu_0)^2 - 1/n\right)_+$$

and $\hat{\mu} = \mu_0$ if $\hat{\tau}^2 = 0$ and if $\hat{\tau} > 0$

$$\hat{\mu} = \bar{X}_n\left(1 - \frac{1}{n(\bar{X}_n - \mu_0)^2}\right) + \frac{\mu_0}{n(\bar{X}_n - \mu_0)^2}$$

Even more degenerate

- optimization in both τ and μ_0

$$\hat{\mu}_0 = \bar{X}_n \quad \text{and} \quad \hat{\tau}^2 = 0, \quad (\mu|X) \sim \delta_{(\bar{X}_n)}$$

where $\delta_{(a)}$ is the Dirac mass at a

Non-parametric EB

Assume only that the θ_i are iid from some distribution π. Use the data to estimate the prior or the marginal distribution directly.
(pioneered by Robbins (1950))

Model: $y_i|\theta_i \sim$ Poisson (θ_i) and $\theta_i \overset{iid}{\sim} \pi(\cdot)$

Square error loss $\Rightarrow$ Bayes estimator is posterior mean:

$$\begin{aligned}\hat{\theta}_i = \mathbb{E}[\theta_i|y_i] &= \int \theta\pi(\theta|y_i)d\theta \\ &= \frac{\int \left(\theta^{y_i+1}e^{-\theta}/y_i!\right)\pi(d\theta)}{\int \left(\theta^{y_i}e^{-\theta}/y_i!\right)\pi(d\theta)} \\ &= \frac{(y_i+1)p(y_i+1)}{p(y_i)}\end{aligned}$$

The Robbins method: $\hat{\theta}_i$ is directly estimable as

$$\hat{\theta}_i = \frac{(y_i+1)\hat{p}(y_i+1)}{\hat{p}(y_i)} = \frac{(y_i+1)[\#y's = (y_i+1)]}{[\#y's = y_i]}.$$

Non-parametric EB

Assume only that the θ_i are iid from some distribution π. Use the data to estimate the prior or the marginal distribution directly.
(pioneered by Robbins (1950))

Model: $y_i|\theta_i \sim$ Poisson (θ_i) and $\theta_i \overset{iid}{\sim} \pi(\cdot)$

Square error loss $\Rightarrow$ Bayes estimator is posterior mean:

$$\begin{aligned}\hat{\theta}_i = \mathbb{E}[\theta_i|y_i] &= \int \theta\pi(\theta|y_i)d\theta \\ &= \frac{\int \left(\theta^{y_i+1}e^{-\theta}/y_i!\right)\pi(d\theta)}{\int \left(\theta^{y_i}e^{-\theta}/y_i!\right)\pi(d\theta)} \\ &= \frac{(y_i+1)p(y_i+1)}{p(y_i)}\end{aligned}$$

The Robbins method: $\hat{\theta}_i$ is directly estimable as

$$\hat{\theta}_i = \frac{(y_i+1)\hat{p}(y_i+1)}{\hat{p}(y_i)} = \frac{(y_i+1)[\#y's = (y_i+1)]}{[\#y's = y_i]}.$$

Summary

1. Parametric EB: suppose θ_i iid $\pi(\theta|\psi)$ and evaluate ψ by $\hat{\psi}$ estimated from data.
   - Avoids integrating over hyperparameters in complex, e.g. hierarchical models.
   - In models with exchangeable parameters pulls the estimates towards the common mean.
   - Recovers James-Stein estimators as a special case.
   - But... some examples of degeneracy and potentially using the data twice (overestimation of precision).
2. Non-parametric EB: suppose θ_i iid $\pi(\cdot)$ and estimate $\hat{\pi}$ from data.