Handout 3
Empirical Bayes (continued). Computing Bayesian estimators

Instructor: Vira Semenova Note author: Vira Semenova

1 Empirical Bayes (ANOVA)(required)[1] .

James-Stein (revisited). Suppose $\bar{X} = (\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_p)$ has a p-variate normal distribution with known $\sigma_n^2 = \sigma^2/n$. That is

$$\bar{X}_i \sim N(\mu_i, \sigma^2/n) \quad i = 1, 2, \ldots, p \tag{1.1}$$
$$\mu_i \sim N(0, \tau^2) \quad i = 1, 2, \ldots, p \text{ independent} \tag{1.2}$$

The Bayesian estimator of θ is

$$\delta_i^B(\bar{X}) := \mathbb{E}\left[\theta \mid \bar{X}\right] = \frac{\tau^2}{\tau^2 + \sigma_n^2}\bar{X}_i, \quad i = 1, 2, \ldots, p.$$

The Bayesian estimator requires plugging in a value of τ^2. The empirical Bayesian agrees with the Bayes model but refuses to specify values of τ^2. Instead, it estimates τ^2. In the last lecture, we considered using MLE to estimate τ^2. For the today's lecture, we will consider an unbiased estimator of $\frac{\sigma^2}{\tau^2 + \sigma^2}$ instead. Define the James-Stein estimator as

$$\delta_i^{JS}(X) = \left(1 - \frac{(p-2)\sigma_n^2}{\|\bar{X}\|^2}\right)\bar{X}, \quad i = 1, 2, \ldots, p.$$

Recall that James-Stein was introduced as an estimator whose frequentist risk is smaller than MLE for all values of μ. However, in practice shrinkage to zero may be a poor choice if $\|\mathbb{E}[X]\|$ is very far from zero. Instead, we may want to estimate the shrinkage point from the data. The next example elaborates on this.

One-way Analysis of Variance (ANOVA). Consider the many means model

$$X_{ij} \sim N(\mu_i, \sigma^2) \quad j = 1, 2, \ldots, n \text{ independent}, \quad i = 1, 2, \ldots, p \tag{1.3}$$
$$\mu_i \sim N(\mu, \tau^2) \quad i = 1, 2, \ldots, p \text{ independent} \tag{1.4}$$

The goal is to estimate $\mu = (\mu_1, \mu_2, \ldots, \mu_p) \in \mathrm{R}^p$. Similar to (1.1)-(1.2), we postulate a common mean for μ_i, but unlike (1.2), we refuse to specify its exact value. The MLE (**unrestricted**) estimator of μ is a vector of group specific means

$$\bar{X}_i := n^{-1}\sum_{j=1}^{n} X_{ij}, \quad i = 1, 2, \ldots, p.$$

The unrestricted estimator does not require specification of the prior (1.4). The frequentist risk of MLE estimator of $p\sigma_n^2 = p/n\sigma^2$.

[1]This section is based on Lehmann and Casella, Chapter 4.6.

1. **Both μ and τ^2 known.** Next, consider imposing the prior distribution (1.4). This prior is similar to (1.2), except the prior mean μ may not be zero. Assuming both μ and σ^2 are known, the Bayes posterior mean for each i is

$$\delta_i^B(\bar{X}) = \frac{\sigma_n^2}{\sigma_n^2 + \tau^2}\mu + \frac{\tau^2}{\sigma_n^2 + \tau^2}\bar{X}_i, \quad i = 1, 2, \ldots, p,$$

which is a weighted average of MLE (unrestricted estimator) $\bar{X}$ and the common (known) mean μ.

2. μ **is unknown,** τ^2 **is known.** If μ is unknown, we replace it by MLE. The MLE of μ is the full sample mean

$$\bar{\bar{X}} := p^{-1}\sum_{i=1}^{p}\bar{X}_i = (np)^{-1}\sum_{i=1}^{p}\sum_{j=1}^{n}X_{ij}.$$

The Empirical Bayes estimator is

$$\delta_i^{EB}(\bar{X}) := \frac{\sigma_n^2}{\sigma_n^2 + \tau^2}\bar{\bar{X}} + \frac{\tau^2}{\sigma_n^2 + \tau^2}\bar{X}_i, \quad i = 1, 2, \ldots, p.$$

which is a weighted average of MLE (unrestricted estimator) $\bar{X}$ and the grand mean $\bar{\bar{X}}$ (restricted estimator).

3. **Both μ and τ^2 unknown.** The δ_i^{EB} takes the form

$$\delta_i^L(\bar{X}) := \bar{\bar{X}} + \left(1 - \frac{(p-3)\sigma_n^2}{p^{-1}\sum_{i=1}^{p}(\bar{X}_i - \bar{\bar{X}})^2}\right)(\bar{X}_i - \bar{\bar{X}}).$$

, which was first derived by Lindley (1962) and examined in detail by Efron and Morris (1972a 1972b, 1973a, 1973b).

ANOVA with a regression submodel. Shrinking to a common mean (1.4) may still be restrictive. If we have observed covariates, we may allow the shrinkage point to vary with observed covariates.

$$\begin{aligned} X_{ij} &\sim N(\mu_i, \sigma^2) \quad j = 1, 2, \ldots, n, \quad i = 1, 2, \ldots, p \text{independent} \\ \mu_i &\sim N(\alpha + \beta t_i, \tau^2) \quad i = 1, 2, \ldots, p \text{ independent} \end{aligned}$$

where $t = (t_1, t_2, \ldots, t_p)$ is a vector of observed characteristics. Take

$$\bar{t} := p^{-1}\sum_{i=1}^{p}t_i.$$

The Bayes estimator of μ_i is calculated assuming the parameters are known. The (partial) empirical Bayesian agrees with the Bayes model but refuses to specify values of α and β (but assumes τ^2 is known). Instead, we estimate α and β using MLE. This method has two steps:

1. Derive the likelihood function of $\bar{X}_i$ as a function of α and β. For each $i = 1, 2, \ldots, p$,

$$\bar{X}_i \sim N(\alpha + \beta t_i, \tau^2 + \sigma_n^2), \quad i = 1, 2, \ldots, n.$$

2. Specify the total likelihood function for $\bar{X} = (\bar{X}_1, \ldots, \bar{X}_p)$. It is equal to

$$\prod_{i=1}^{p} f_{\bar{X}_i}(\bar{x}_i \mid \alpha, \beta) \sim \prod_{i=1}^{p}\frac{1}{(\sqrt{2\pi\sigma_n^2})^p}\exp^{-(\bar{x}_i - \alpha - \beta t_i)^2/2\sigma_n^2}$$

3. The negative log likelihood is a convex function of α, β

$$\sum_{i=1}^{p}(\bar{x}_i - \alpha - \beta t_i)^2.$$

Taking FOC conditions gives the OLS estimators of α and β:

$$\widehat{\alpha} = \bar{\bar{X}} - \widehat{\beta}\bar{t}, \quad \widehat{\beta} := \frac{\sum_{i=1}^{p}(\bar{X}_i - \bar{X})(\bar{t}_i - \bar{t})}{\sum_{i=1}^{p}(\bar{t}_i - \bar{t})^2}.$$

1. **Both α and β and τ^2 known.** The Bayes estimator of μ_i is

$$\delta_i^B(\bar{X}) = \frac{\sigma_n^2}{\sigma_n^2 + \tau^2}(\alpha + \beta t_i) + \frac{\tau^2}{\sigma_n^2 + \tau^2}\bar{X}_i, \quad i = 1, 2, \ldots, p.$$

2. **α and β are unknown; τ^2 is known.** The empirical Bayes estimator is

$$\delta_i^{EB1}(\bar{X}) := \frac{\sigma_n^2}{\sigma_n^2 + \tau^2}(\widehat{\alpha} + \widehat{\beta} t_i) + \frac{\tau^2}{\sigma_n^2 + \tau^2}\bar{X}_i, \quad i = 1, 2, \ldots, p.$$

3. **Both α and β and τ^2 unknown.** Here, we replace $\frac{\sigma_n^2}{\sigma_n^2 + \tau^2}$ by an unbiased estimator. The empirical Bayes estimator is

$$\delta_i^{EB2}(\bar{X}) := \widehat{\alpha} + \widehat{\beta} t_i + \left(1 - \frac{(p-4)\sigma_n^2}{p^{-1}\sum_{i=1}^{p}(\bar{X}_i - \widehat{\alpha} - \widehat{\beta} t_i)^2}\right)(\bar{X}_i - \widehat{\alpha} - \widehat{\beta} t_i).$$

2 Computing Bayesian Estimators.

2.1 Acceptance-Rejection Sampling. (Required).

Posterior distribution is key to construct Bayes estimators. However, posterior distribution

$$\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{\int_{\widetilde{\theta}} f(x|\widetilde{\theta})\pi(\widetilde{\theta})d\widetilde{\theta}}$$

is often difficult to compute: denominator is intractable. Do not calculate the posterior. Simulate from the posterior! Generate a random sample $\theta_1, \theta_2, \ldots, \theta_B$ from $\pi(\theta \mid \text{data})$ B times.

$$\tfrac{1}{B}\textstyle\sum_{b=1}^{B}\theta_i \to \mathbb{E}\left[\theta \mid \text{data}\right]$$

Notation

- π - the target distribution, which we want to simulate from
- q - the distribution we CAN simulate from
- $\pi(x) = f(x)/k$ is known up to constant
- there exists c such that $f(x) \leq cq(x)$

Clever ways to get from q to π:

(1) Draw from q, but discard some draws (A-R)

(2) Draw from q. Discarding a draw means staying at the present (current) draw. Accepting a draw means moving to the proposed state (MCMC).

Goal is to simulate $\xi \sim \pi(x)$. But $\pi(x)$ has an intractable (cannot be evaluated) denominator. Acceptance-Rejection algorithm is

1. Draw $z \sim q(\cdot)$, $u \sim U[0,1]$ independently
2. If $u \leq \dfrac{f(z)}{cq(z)}$, accept $\xi = z$. Otherwise, discard the draw.

Let ξ be the first draw in the retained pool that is not rejected. The CDF of ξ is

$$\Pr(\xi \leq x) = \Pr\left(z_1 \leq x, u_1 \leq \frac{f(z_1)}{cq(z_1)}\right) \tag{2.1}$$

$$+ \Pr\left(\text{first draw rejected}, z_2 \leq x, u_2 \leq \frac{f(z_2)}{cq(z_2)}\right) + \quad \cdots + \Pr\left(k-1 \text{ draw rejected}, z_k \leq x, u_k \leq \frac{f(z_k)}{cq(z_k)}\right) + \ldots. \tag{2.2}$$

For each k, the draw (z_k, u_k) is independent of prior history of acceptance and rejections. That is,

$$\Pr\left(k-1 \text{ draw rejected}, z_k \leq x, u_k \leq \frac{f(z_k)}{cq(z_k)}\right) = \bar{\rho}^{k-1} \Pr\left(z_k \leq x, u_k \leq \frac{f(z_k)}{cq(z_k)}\right). \tag{2.3}$$

Furthermore, the draws (z_k, u_k) have identical distribution irrespective of prior history of acceptance and rejections. Therefore,

$$\begin{aligned}
\Pr(\xi \leq x) &= \Pr\left(z \leq x, u \leq \frac{f(z)}{cq(z)}\right)(1 + \bar{\rho} + \bar{\rho}^2 + \ldots) \\
&= \frac{1}{1-\bar{\rho}} \int_{-\infty}^{x} \frac{f(z)}{cq(z)} q(z) dz \\
&= \frac{1}{1-\bar{\rho}} \int_{-\infty}^{x} \frac{f(z)}{cq(z)} q(z) dz \\
&= \frac{1}{c(1-\bar{\rho})} \int_{-\infty}^{x} f(z) dz.
\end{aligned}$$

A-R algorithm is

1. Draw $z \sim q(\cdot)$, $u \sim U[0,1]$ independently
2. If $u \leq \dfrac{f(z)}{cq(z)}$, accept $\xi = z$. Otherwise, discard the draw.

Problems with A-R algorithm:

- if we choose c and $q(z)$ poorly, then $f(z)/cq(z)$ could be very small for many z
- small $f(z)/cq(z)$ means we have to reject many draws before we accept one

Difficult to choose c and $q(z)$ when we do not know much about $\pi(z)$. Rarely used in practice. A more sophisticated version of A-R is MCMC (Markov Chain Monte Carlo) sampling method, which is optional in this course.

2.2 Markov Chain Monte Carlo (Optional)

Same as in the earlier section, the goal is to simulate from the posterior distribution π but π is the target density that has no closed form because the denominator is intractable. We can only compute the numerator $f(x)$ where $\pi(x) = f(x)/k$. In Acceptance-Rejection method, each draw (z, u) is independent of the past draws. Relying on independence has substantially simplified the theoretical argument (see (2.3)). However, a cost of independence is that we cannot use past draws to decide where/how to sample the next draw, which may lead to inefficient (time-consuming) sampling.

A MCMC method sacrifices the independence property (2.3) in favor of using proposal distribution that depend on the past state. A sequence of draws from the proposal distribution is no longer independent, but is a Markov chain. Below, I define some basic quantities of a Markov chain.

Definition 1 (Markov chain). *A sequence $\{x_t\}$ is a first-order Markov chain if for any set A*

$$P(x_{t+1} \in A \mid x_t = x, x_{t-1}, \dots) = P(x_{t+1} \in A \mid x_t = x). \tag{2.4}$$

Definition 2 (Transition kernel). *The function*

$$P(x, A) := P(x_t \in A \mid x_{t-1} = x)$$

is a transition kernel. Let $q(x, y)$ be a proposal distribution of sampling the next state y given the current state x. The transition kernel $P(x, A)$ corresponding to $q(x, y)$ is

$$P(x, A) := \int_{y \in A} q(x, y) dy \tag{2.5}$$

Definition 3 (Invariant distribution). *A distribution π^* is an invariant distribution for the kernel $P(x, A)$ if*

$$\pi^*(y)dy = \int_R \pi^*(x) P(x, dy) dx. \tag{2.6}$$

With a large number of draws, the Markov chain converges to its invariant distribution. A classic Markov problem is to find π^* given the transition kernel $P(x, A)$. Our problem is reverse problem - to find transition kernel $P(x, A)$ so that the target π distribution is its invariant distribution:

$$\pi^* = \pi.$$

A sufficient condition condition for the distribution π to be invariant for the kernel $P(x, A)$ is to obey reversibility condition:

$$\pi(x)q(x, y) = \pi(y)q(y, x). \tag{2.7}$$

Lemma 1. *If $q(x, y)$ obeys (2.7), $\pi^* = \pi$ obeys (2.6) with $P(a, X)$ in (2.5).*

Proof. We need to check that π satisfies definition of invariant distribution. For any set A

$$\begin{aligned}
\int_R \pi^*(x) P(x, dy) dy dx &= \int_R \pi^*(x) q(x, y) dy dx = \int_R \pi^*(y) q(y, x) dy dx \\
&= \pi^*(y) \left(\int_R q(y, x) dx \right) dy \\
&= \pi^*(y) \cdot (1)\, dy.
\end{aligned}$$

□

Note that the condition (2.7) requires knowing $\pi(x)$ up to the denominator. Indeed, (2.7) holds if and only if

$$f(x)q(x, y) = f(y)q(y, x). \tag{2.8}$$

which we can verify. If we can find $q(x, y)$ such that (2.7) holds, then, sampling a Markov chain from the proposal distribution $q(y, x)$ gives the target distribution in the limit.

In most cases, the proposal distribution $q(x, y)$ may not obey (2.8). Does it mean we should discard $q(x, y)$? No! We can borrow some idea from Acceptance-Rejection method. Let

- x is the current draw
- y is the next candidate draw from $q(x, \cdot)$ (move $x \to y$)
- w.p. $r(x)$, accept y. w.p. $1 - r(x)$, stay at x and discard y (stay $x \to x$)

Our goal is to find $q(x, y)$ and $r(x)$ so that the limiting distribution of the chain is π.
For a given x, y, suppose (2.8) fails and $f(x)q(x, y) > f(y)q(y, x)$. Introduce ratio function $\alpha(x, y)$ is

$$\pi(x)q(x, y)\alpha(x, y) = \pi(y)q(y, x)\alpha(y, x),$$

Define

$$\alpha(x, y) = \min\left\{1, \frac{\pi(y)q(y, x)}{\pi(x)q(x, y)}\right\}.$$

Since $\alpha(y, x) < 1$, the probability of move is less than one

$$\int q(y, x)\alpha(y, x)dy = r(x) < 1.$$

Define the probability of stay at x is

$$1 - r(x) = 1 - \int q(y, x)\alpha(y, x)dy$$

$P(x, dy)$ is a valid transition kernel

$$P(x, dy) = q(y, x)\alpha(y, x)dy + r(x)\delta_x(dy)$$

Indeed, $\int_R P(x, dy) = 1$ since $\int_R q(y, x)\alpha(y, x)dy = 1 - r(x)$ and $r(x)\int_R \delta_x(dy) = r(x) \cdot 1$

Definition 4 (Metropolis-Hastings algorithm). *Given a draw x_t, the next draw x_{t+1} is generated as*

1. *Draw y from $q(x_t, \cdot)$*
2. *Calculate $\alpha(x_t, y) = \min\left\{1, \frac{f(y)q(y, x_t)}{f(x)q(x_t, y)}\right\}$*
3. *Draw $u \sim U[0, 1]$*
4. *If $u < \alpha(x_t, y)$, then $x_{t+1} = y$. Otherwise, $x_{t+1} = x_t$.*

Lemma 2. *The proposal distribution $q(y, x)\alpha(y, x)$ obeys reversibility condition for π (2.7).*

Proof.

$$\begin{aligned}
\int \pi(x)P(x, A)dx &= \int \left(\int_A p(x, y)dy\right)\pi(x)dx + \int (1 - r(x))\delta_x(A)\pi(x)dx \\
&= \int_A \int p(x, y)\pi(x)dxdy + \int_A (1 - r(x))\pi(x)dx \\
&= \int_A \int p(y, x)\pi(y)dxdy + \int_A (1 - r(x))\pi(x)dx \\
&= \int_A \pi(y)\left(\int p(y, x)dx\right)dy + \int_A (1 - r(x))\pi(x)dx \\
&= \int_A \pi(y)r(y)dy + \int_A (1 - r(x))\pi(x)dx = \pi(A)
\end{aligned}$$

□

Example 1 (Random walk chain). *Consider a proposal distribution with*

$$y = x + \epsilon, \quad \epsilon \sim N(0, \sigma^2)$$

whose proposal distribution is

$$q(x, y) = \phi((y - x)/\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp^{-(y-x)^2/2\sigma^2} = q(x, y)$$

The ratio function is

$$\alpha(x, y) = \min\left\{1, \frac{f(y)q(y, x)}{f(x)q(x, y)}\right\}$$

Since $N(0, 1)$ *is symmetric,* $q(y, x) = q(x, y)$ *and*

$$\alpha(x, y) = \min\left\{1, \frac{f(y)}{f(x)}\right\}.$$