Handout 2
Shrinkage in normal means model. Empirical Bayes

Instructor: Vira Semenova Note author: Vira Semenova

1 Many means model

Consider many means model[1]:

$$X_{kj} = \theta_k + \sigma\delta_{kj}, \quad k = 1, 2, \ldots, p, \quad j = 1, 2, \ldots, n, \tag{1.1}$$

where $\delta_{kj} \sim N(0, 1)$ for k and j i.i.d across k and j. Let $X = (X_{kj})_{k,j=1,1}^{p,n}$ be the matrix of observations, where j indicates observation index $j = 1, 2, \ldots, n$. The target parameter is the vector of population means

$$\theta = (\theta_1, \theta_2, \ldots, \theta_p) \in \mathrm{R}^p$$

while the variance σ^2 is assumed known. The likelihood of j' th observation $X_j := (X_{1j}, X_{2j}, \ldots, X_{pj})$ is

$$f_{X_j}(x_j \mid \theta) = \phi(x_j) := \frac{1}{\sqrt{2\pi(\sigma^2)^p}} \exp^{-\|x_j - \theta\|^2/2\sigma^2},$$

and the total likelihood is

$$f_X(x \mid \theta) = \prod_{j=1}^{n} f_{X_j}(x_j \mid \theta).$$

The decision set $\mathcal{A} = \Theta = \mathrm{R}^p$ and the loss function

$$L(a, \theta) = \|a - \theta\|^2 = \sum_{k=1}^{p} (a_k - \theta_k)^2.$$

The frequentist risk of a decision rule $\delta : \mathcal{X} \to \mathrm{R}^p$ is

$$R(\delta, \theta) = \mathbb{E}\left[L(\delta(X), \theta)\right] = \sum_{k=1}^{p} \mathbb{E}\left[(\delta_k(X) - \theta_k)^2\right].$$

The MLE decision rule is

$$\bar{X}_k := n^{-1} \sum_{j=1}^{n} X_{kj}, \quad k = 1, 2, \ldots, p, \quad \bar{X} = (\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_p). \tag{1.2}$$

The bias and variance of each coordinate $k = 1, 2, \ldots, p$ are

$$\mathbb{E}\left[\bar{X}_k\right] = \theta_k, \quad \mathrm{Var}\left(\bar{X}_k\right) = \sigma^2/n = \sigma_n^2$$

and the risk is

$$R(\bar{X}_k, \theta_k) = 0^2 + \sigma_n^2 = \sigma_n^2.$$

The frequentist risk of vector $\bar{X}$ is

$$R(\bar{X}, \theta) = \sum_{k=1}^{p} \sigma_n^2 = p\sigma_n^2.$$

MLE estimator has many credentials: it is minimum variance unbiased estimator, Bayes estimator with flat prior. However, it may perform relatively poorly if p is large relative to n.

[1]This material is largely taken from [Wasserman, 2006], Chapter 7. Unlike Chapter 7, we do not restrict $p = n$ in this handout.

Linear estimator with oracle shrinkage. Consider a class of linear estimators

$$\mathcal{F} = \left\{ \beta \bar{X}, \quad \beta \in \mathrm{R}, \quad \bar{X} \in \mathrm{R}^p \right\}. \tag{1.3}$$

Notice that each coordinate k is multiplied by the same constant β. The frequentist risk of $\beta \bar{X}_j$ is

$$R(\beta \bar{X}_k, \theta_k) = (\beta - 1)^2 \theta_k^2 + \beta^2 \sigma_n^2.$$

Summing across $k = 1, 2, \ldots, p$ gives

$$R(\beta \bar{X}, \theta) = \sum_{k=1}^{p} (\beta - 1)^2 \theta_k^2 + \beta^2 \sigma_n^2 = (\beta - 1)^2 \|\theta\|^2 + p\beta^2 \sigma_n^2.$$

The function $\beta \to R(\beta \bar{X}, \theta)$ is convex in β. Taking FOC gives the unique solution

$$\beta^* = \beta^*(p, n) = \frac{\|\theta\|^2}{\|\theta\|^2 + p\sigma_n^2} = 1 - \frac{p\sigma_n^2}{\|\theta\|^2 + p\sigma_n^2},$$

which is the minimizer. The oracle risk is

$$\min_{\beta \in \mathrm{R}} R(\beta \bar{X}, \theta) = \frac{p\sigma_n^2 \|\theta\|^2}{\|\theta\|^2 + p\sigma_n^2}.$$

Note that $\beta^* \bar{X}$ is not a feasible decision rule since $\|\theta\|^2$ is unknown. If $p = 1$ and $n \to \infty$, the variance is decaying

$$\sigma_n^2 = \sigma^2 / n \to 0$$

and the optimal estimator converges to MLE

$$\beta^*(1, n) \to 1, n \to \infty.$$

As a result, MLE is asymptotically optimal. When $p \neq n$, the optimal shrinkage coefficient may not $\to 1$. Replacing $\frac{p\sigma_n^2}{\|\theta\|^2 + p\sigma_n^2}$ by its unbiased estimate $\frac{(p-2)\sigma_n^2}{\|\bar{X}\|^2}$ gives **James-Stein** estimator

$$\delta^{JS} := \left(1 - \frac{(p-2)\sigma_n^2}{\|\bar{X}\|^2}\right) \bar{X}.$$

The proof of unbiasedness is given in [Wasserman, 2006], Chapter 7.

2 James-Stein: Frequentist perspective

Definition 1. *A decision rule δ' is inadmissible if there exists a "uniformly weakly better" decision rule*

$$R(\delta, \theta) \leq R(\delta', \theta) \quad \forall \theta \in \Theta, \quad R(\delta, \theta_0) < R(\delta', \theta_0) \text{ for some } \theta_0.$$

MLE is inadmissible.

Lemma 1 (Property of Normal PDF)**.** *For any $k = 1, 2, \ldots, p$, the normal CDF has the following derivative*

$$f'(X_k) = -\frac{(X_k - \theta_k) f(X_k)}{\sigma_n^2} = -\frac{(X_k - \theta_k)}{\sigma_n^2} \frac{1}{\sqrt{2\pi\sigma_n^2}} \exp^{-(X_k - \theta_k)^2 / 2\sigma_n^2}.$$

Therefore, for any function $g_k(X)$, integration by parts gives

$$\mathbb{E}\left[g_k(X) \frac{\bar{X}_k - \theta_k}{\sigma_n^2} \right] = -\mathbb{E}\left[\nabla_k g_k(X) \right], \quad k = 1, 2, \ldots, p.$$

Theorem 1. *MLE is inadmissible. In particular, $\delta^{JS}(X)$ dominates MLE $\delta(X) = \bar{X}$* [2]

Proof. Consider the decision rule

$$\delta^{JS}(X) := \bar{X} + g^{JS}(X),$$

where $g(X)$ is a shrinkage correction. James and Stein choose

$$g(X) = g^{JS}(X) = \frac{(p-2)\sigma_n^2}{\|\bar{X}\|^2}\bar{X} = \frac{(p-2)\sigma_n^2}{\sum_{k=1}^p \bar{X}_k^2}\bar{X}.$$

The risk of $\delta^{JS}(X)$ can be decomposed into 3 terms:

$$\begin{aligned}
R(\delta^{JS}, \theta) &= \sum_{k=1}^p \mathbb{E}\left[(\delta_k^{JS}(X) - \theta_k)^2\right] = \sum_{k=1}^p \mathbb{E}\left[(\delta_k^{JS}(X) \pm \bar{X}_k - \theta_k)^2\right] \\
&= \sum_{k=1}^p \mathbb{E}\left[(\delta_k^{JS}(X) - \bar{X}_k)^2\right] + 2\sum_{k=1}^p \mathbb{E}\left[(\delta_k^{JS}(X) - \bar{X}_k)(\bar{X}_k - \theta_k)\right] + \sum_{k=1}^p \mathbb{E}\left[(\theta_k - \bar{X}_k)^2\right] \\
&= \sum_{k=1}^p \mathbb{E}\left[(g_k^{JS})^2(X)\right] + 2\sum_{k=1}^p \mathbb{E}\left[g^{JS}(X)(\bar{X}_k - \theta_k)\right] + p\sigma_n^2 \\
&= S_1 + S_2 + p\sigma_n^2.
\end{aligned}$$

By definition of $g^{JS}(X)$,

$$S_1 = (p-2)^2(\sigma_n^2)^2\mathbb{E}\left[\frac{1}{\sum_{k=1}^p \bar{X}_k^2}\right].$$

By Lemma 1,

$$S_2 = -2\sigma_n^2\mathbb{E}\left[\sum_{k=1}^p \nabla_k g_k^{JS}(X)\right].$$

For each coordinate $k = 1, 2, \ldots, p$,

$$\nabla_k g_k^{JS}(X) = \frac{(p-2)\sigma_n^2}{\|\bar{X}\|^2} - \frac{(p-2)2\bar{X}_k^2\sigma_n^2}{\|\bar{X}\|^4}$$

Therefore,

$$S_2 = -2\sigma_n^2\mathbb{E}\left[\left(\frac{(p-2)p}{\|\bar{X}\|^2} - \frac{(p-2)2}{\|\bar{X}\|^2}\right)\right] = -2\sigma_n^2\mathbb{E}\left[\frac{(p-2)^2(\sigma_n^2)^2}{\|\bar{X}\|^2}\right]$$

Collecting the terms gives

$$\begin{aligned}
S_1 + S_2 + p\sigma_n^2 &= p\sigma_n^2 - \mathbb{E}\left[\frac{2(p-2)^2(\sigma_n^2)^2}{\|\bar{X}\|^2}\right] + \mathbb{E}\left[\frac{(p-2)^2(\sigma_n^2)^2}{\|\bar{X}\|^2}\right] \\
&= (p - \mathbb{E}\left[\frac{(p-2)^2(\sigma_n^2)^2}{\|\bar{X}\|^2}\right])\sigma_n^2 < p\sigma_n^2
\end{aligned}$$

$\square$

Theorem 2. *The James Stein decision rule $\delta^{JS}(X)$ is asymptotically minimax in the class $\mathcal{F}$*

$$\min_\beta R(\beta X, \theta) + 2\sigma_n^2 \geq R(\delta^{JS}, \theta) \geq \min_\beta R(\beta X, \theta)$$

[2] If not covered in the lecture, the proof is optional (not for exam).

3 James-Stein: Bayesian perspective

In this section, I motivate James-Stein from Bayesian perspective. Consider the following model with $p \geq 1$ and $n = 1$ and $\sigma^2 = 1$. The likelihood is

$$X \sim N(\theta, 1 \cdot I_p)$$

and the prior is $\pi(\theta) : \theta \sim N(0, \tau^2 I_p)$. Then, the posterior distribution follows from Bayes rule

$$\pi(\theta \mid X) = \frac{f_X(X \mid \theta)\pi(\theta)}{m(X)},$$

where $m(X) = \int_\Theta f_X(X \mid \theta)\pi(\theta)d\theta$ is the marginal density. Because $\pi(\theta \mid X)$ must integrate to 1 (as a density), one can derive the posterior distribution without calculating $m(X)$:

$$\pi(\theta \mid X) \sim N(\theta_p, \tau_p^2) := \left(\frac{\tau^2}{\tau^2+1}X, \frac{\tau^2}{\tau^2+1}I_p\right) \tag{3.1}$$

The posterior mean is

$$\theta_p = \mathbb{E}\left[\theta \mid X = x\right] = \frac{\tau^2}{\tau^2+1}X = \left(1 - \frac{1}{\tau^2+1}\right)X.$$

The posterior variance matrix is

$$\tau_p^2 I_p := \frac{\tau^2}{\tau^2+1}I_p$$

If τ is known, $\mathbb{E}\left[\theta \mid X = x\right]$ is a feasible decision rule.

An empirical Bayes approach proposes estimating τ using MLE. This approach has two steps:

1. Derive the marginal distribution $m(X)$ as a function of τ^2: Answer

   $$X \sim N(0, (\tau^2+1)I_p).$$

   The marginal density is the denominator constant in Bayesian update for the posterior $\pi(\theta \mid X)$. Rearranging Bayes rule gives

   $$m(X) = \frac{f_X(X \mid \theta)\pi(\theta)}{\pi(\theta \mid X)} \sim \frac{\exp^{-(X-\theta)'(X-\theta)/2}\exp^{-\theta'\theta/2\tau^2}}{\exp^{-(\theta-\theta_p)'(\theta-\theta_p)/2\tau_p^2}}.$$

   Because there are only exponents above, the density $m(X)$ must correspond to $N(0, \tau_{\max}^2 I_p)$ distribution. Therefore, it must be that

   $$\exp^{-X'X/2\tau_{\max}^2} = \frac{\exp^{-(X-\theta)'(X-\theta)/2}\exp^{-\theta'\theta/2\tau^2}}{\exp^{-(\theta-\theta_p)'(\theta-\theta_p)/2\tau_p^2}}, \tag{3.2}$$

   where θ_p and τ_p^2 are posterior mean and variance in (3.1).

   To find $\tau_{\max}^2$, I equate the coefficient near $X'X$ in LHS and RHS of (3.2). The coefficient in LHS of (3.2) is

   $$1/2\tau_{\max}^2. \tag{3.3}$$

   In RHS, $X'X$ appears in the numerator of (3.2) (the likelihood) and the posterior mean $\theta_p'\theta_p$. The coefficient in RHS of (3.2) is

   $$1/2\left(1 - \frac{1}{\tau_p^2}\left(\frac{\tau^2}{\tau^2+1}\right)^2\right) = 1/2(1 - \frac{\tau^2}{\tau^2+1}) = 1/2(\frac{1}{\tau^2+1}). \tag{3.4}$$

   Equating (3.3) and (3.4) gives

   $$\tau_{\max}^2 = \tau^2 + 1.$$

2. Find $\widehat{\tau^2}$ by maximizing log-likelihood:

$$\arg\max_{\tau^2} \log f_X(X \mid \tau^2) = \arg\max_{\tau^2} -p/2 \log(\tau^2+1) - \frac{1}{(\tau^2+1)} X'X.$$

FOC gives

$$p\frac{1}{\tau^2+1}\bigg|_{\widehat{\tau^2}_{\text{MLE}}} = \frac{1}{2(\tau^2+1)^2} X'X\bigg|_{\widehat{\tau^2}_{\text{MLE}}} \Rightarrow \widehat{\tau^2_{\text{MLE}}} + 1 = X'X/p = \|X\|^2/p.$$

Replacing $\frac{1}{\tau^2+1}$ by its MLE estimate $\frac{p}{\|X\|^2}$ almost gives James Stein (note that difference between $p-2$ and p):

$$\left(1 - \frac{p}{\|X\|^2}\right) X.$$

4 Empirical Bayes

Example 1 (Lehmann and Casella). *Suppose an insurance company observes the number of claims in a single year by auto-insurance policy holders. Number of claims in each group is n_k. For each policy holder j, we have 1 observation the number of claims X_j. We assume that $X_j \sim Poisson(\theta_j)$.*

$$\Pr(X_j = k|\theta_j) = \frac{e^{-\theta_j}\theta_j^k}{k!}$$

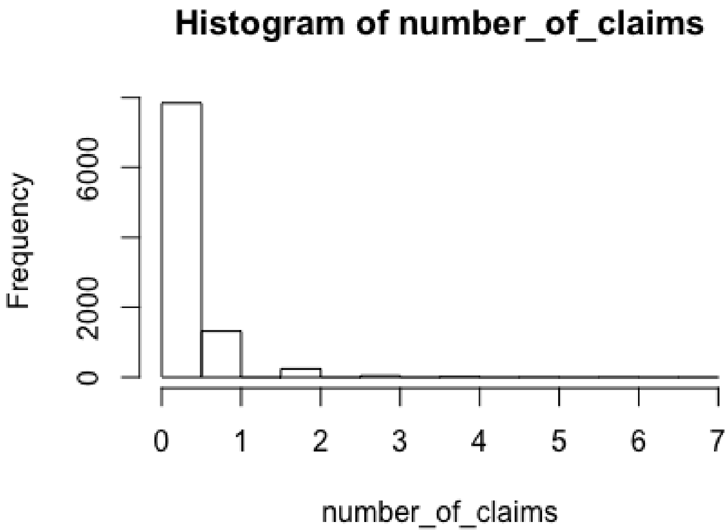

Figure 4.1: Table 6.1 from Lehmann and Casella

For each policy holder, MLE estimate

$$\widehat{\theta}_j = X_j$$

k	0	1	2	3	4	5	6	7
n_k	7840	1317	239	42	14	4	4	1
$\widehat{\theta}^{EB}$	0.168	0.363	0.527	1.333	1.429	6.000	1.750	-
$\widehat{\theta}^{MLE}$	0	1	2	3	4	5	6	7

Table 1: MLE vs Empirical Bayes estimator

is based on a single observation. Suppose we have a prior density, $\pi(\theta)$ for θ. Posterior mean is

$$\begin{aligned}\mathbb{E}\left[\theta|X=k\right] &= \frac{\int_0^\infty \theta f(k|\theta)\pi(\theta)d\theta}{\int_0^\infty f(k|\theta)\pi(\theta)d\theta} \\ &= \frac{\int_0^\infty \theta^{k+1}e^{-\theta}/k!\pi(\theta)d\theta}{\Pr(X=k)} \\ &= (k+1)\frac{\Pr(X=k+1)}{\Pr(X=k)}\end{aligned}$$

Prior has been concentrated out!. Robbins' empirical Bayes estimator is

$$\widehat{\theta}_k^{EB} = (k+1)\frac{n_{k+1}}{n_k}$$

We circumvented the choice of prior by replacing $\dfrac{\Pr(X=k+1)}{\Pr(X=k)}$ by its estimate n_k/n_{k+1}

$$\widehat{\theta}_k^{EB} = (k+1)\frac{n_{k+1}}{n_k}$$

Substantial shrinkage towards zero for positive values of k.

Example 2 (Lehmann and Casella, Example 6.2)**.** *Each patient group k of size n receives a group-specific treatment with success probability θ_k. Number of successes in each group is n_k. MLE estimate $\widehat{\theta}_k = n_k/n$. MLE ignores that patients come from the same pool. Bayesian approach proposes a hierarchy*

$$n_k \sim \textit{binomial}(\theta_k, n), \quad \theta_k \sim \textit{beta}(\alpha,\beta), \quad k=1,2,\ldots,K$$

The posterior mean estimator is

$$\delta^\pi(n_k) = \mathbb{E}\left[\widehat{\theta}_k|\alpha,\beta\right] = \frac{n_k+\alpha}{n_k+\alpha+(n-n_k)+\beta}$$

The Bayes estimator $\delta^\pi(n_k)$ depends on prior parameters α and β, which we do not know. What to do with α and β? Empirical Bayes: use maximum likelihood to estimate them!

$$\delta^{\widehat{\pi}}(\cdot) = \mathbb{E}\left[\theta_k|n_k,\widehat{a},\widehat{b}\right] = \frac{\widehat{a}+n_k}{\widehat{a}+\widehat{b}+n}$$

Empirical Bayes achieves almost-oracle performance, and is much better than MLE!

Summary. Bayes posterior mean often relies on the unknown components of prior $\pi(\theta)$

$$\mathbb{E}\left[\theta|X=x\right]$$

Empirical Bayes avoids using $\pi(\theta)$ using one of those options

- concentrate out $\pi(\theta)$ and replace it by marginal distribution (auto-insurance example)
- impose parametric structure $\pi(\theta)=\pi_{\alpha,\beta}(\theta)$ and estimate α,β using MLE
- nonparametrically estimate $\pi(\theta)$ (not considered)

Table 6.1. ***Bayes Risks for the Bayes, Empirical Bayes, and Unbiased Estimators of Example 6.2, where $K = 10$ and $n = 20$***

Prior Parameters		Bayes Risk		
a	b	δ^{π} of (6.7)	$\delta^{\hat{\pi}}$ of (6.9)	$\mathbf{x}/n$
2	2	.0833	.0850	.1000
6	6	.0721	.0726	.1154
20	20	.0407	.0407	.1220
3	1	.0625	.0641	.0750
9	3	.0541	.0565	.0865
30	10	.0305	.0326	.0915

Figure 4.2: Table 6.1 from Lehmann and Casella

References

[Wasserman, 2006] Wasserman, L. (2006). *All of Nonparametric Statistics.* Springer Texts in Statistics, New York, NY, USA.