

Stat 710: Mathematical Statistics

Lecture 9

Jun Shao

Department of Statistics
University of Wisconsin
Madison, WI 53706, USA

Lecture 9: Shrinkage estimators

Motivation for James-Stein estimators

The risk of $\delta_{c,r}$, $R_{\delta_{c,r}}(\theta) = p - (2r - r^2)(p - 2)^2 E(\|X - c\|^{-2})$, is smaller than p , the risk of X for every value of θ when $p \geq 3$ and $0 < r < 2$. X is inadmissible when $p \geq 3$.

Argument 1: shrink the observation toward a given point c

Suppose it were thought a priori likely, though not certain, that $\theta = c$. Then we might first test a hypothesis $H_0 : \theta = c$ and estimate θ by c if H_0 is accepted and by X otherwise.

The best rejection region has the form $\|X - c\|^2 > t$ for some constant $t > 0$ (see Chapter 6) so that we might estimate θ by

$$I_{(t,\infty)}(\|X - c\|^2)X + [1 - I_{(t,\infty)}(\|X - c\|^2)]c.$$

$\delta_{c,r}$ is a smoothed version of this estimator, since, for some function ψ ,

$$\delta_{c,r} = \psi(\|X - c\|^2)X + [1 - \psi(\|X - c\|^2)]c$$

Any estimator having this form is called a *shrinkage estimator*.

Lecture 9: Shrinkage estimators

Motivation for James-Stein estimators

The risk of $\delta_{c,r}$, $R_{\delta_{c,r}}(\theta) = p - (2r - r^2)(p - 2)^2 E(\|X - c\|^{-2})$, is smaller than p , the risk of X for every value of θ when $p \geq 3$ and $0 < r < 2$. X is inadmissible when $p \geq 3$.

Argument 1: shrink the observation toward a given point c

Suppose it were thought a priori likely, though not certain, that $\theta = c$. Then we might first test a hypothesis $H_0 : \theta = c$ and estimate θ by c if H_0 is accepted and by X otherwise.

The best rejection region has the form $\|X - c\|^2 > t$ for some constant $t > 0$ (see Chapter 6) so that we might estimate θ by

$$I_{(t,\infty)}(\|X - c\|^2)X + [1 - I_{(t,\infty)}(\|X - c\|^2)]c.$$

$\delta_{c,r}$ is a smoothed version of this estimator, since, for some function ψ ,

$$\delta_{c,r} = \psi(\|X - c\|^2)X + [1 - \psi(\|X - c\|^2)]c$$

Any estimator having this form is called a *shrinkage estimator*.

Argument 2: empirical Bayes estimator

Next, $\delta_{c,r}$ can be viewed as an empirical Bayes estimator (§4.1.2). In view of Example 2.25, a Bayes estimator of θ is of the form

$$\delta = (1 - B)X + Bc,$$

where c is the prior mean of θ and B involves prior variances. If $1 - B$ is “estimated” by $\psi(\|X - c\|^2)$, then δ_c is an empirical Bayes estimator.

James-Stein estimator δ_c

$\delta_c = \delta_{c,1}$ is better than any $\delta_{c,r}$ with $r \neq 1$, since the factor $2r - r^2$ is maximized at $r = 1$ for $0 < r < 2$.

To see that δ_c may have a substantial improvement over X in terms of risks, consider the special case where $\theta = c$.

Since $\|X - c\|^2$ has the chi-square distribution χ_p^2 when $\theta = c$,

$E\|X - c\|^{-2} = (p - 2)^{-1}$ and

$R_{\delta_{c,1}}(\theta) = p - (2r - r^2)(p - 1)^2 E(\|X - c\|^{-2}) = p - (p - 2)^2 / (p - 2) = 2$.

The ratio $R_X(\theta) / R_{\delta_c}(\theta)$ equals $p/2$ when $\theta = c$ and can be substantially larger than 1 near $\theta = c$ when p is large.

Argument 2: empirical Bayes estimator

Next, $\delta_{c,r}$ can be viewed as an empirical Bayes estimator (§4.1.2). In view of Example 2.25, a Bayes estimator of θ is of the form

$$\delta = (1 - B)X + Bc,$$

where c is the prior mean of θ and B involves prior variances. If $1 - B$ is "estimated" by $\psi(\|X - c\|^2)$, then δ_c is an empirical Bayes estimator.

James-Stein estimator δ_c

$\delta_c = \delta_{c,1}$ is better than any $\delta_{c,r}$ with $r \neq 1$, since the factor $2r - r^2$ is maximized at $r = 1$ for $0 < r < 2$.

To see that δ_c may have a substantial improvement over X in terms of risks, consider the special case where $\theta = c$.

Since $\|X - c\|^2$ has the chi-square distribution χ_p^2 when $\theta = c$,

$E\|X - c\|^{-2} = (p - 2)^{-1}$ and

$R_{\delta_{c,1}}(\theta) = p - (2r - r^2)(p - 1)^2 E(\|X - c\|^{-2}) = p - (p - 2)^2 / (p - 2) = 2$.

The ratio $R_X(\theta) / R_{\delta_c}(\theta)$ equals $p/2$ when $\theta = c$ and can be substantially larger than 1 near $\theta = c$ when p is large.

Minimaxity and admissibility of δ_c

Since X is minimax (Example 4.25), $\delta_{c,r}$ is minimax provided that $p \geq 3$ and $0 < r < 2$.

Unfortunately, the James-Stein estimator δ_c with any c is also inadmissible.

It is dominated by

$$\delta_c^+ = X - \min \left\{ 1, \frac{p-2}{\|X-c\|^2} \right\} (X-c)$$

see, for example, Lehmann (1983, Theorem 4.6.2).

This estimator, however, is still inadmissible.

An example of an admissible shrinkage estimator is provided by Strawderman (1971); see also Lehmann (1983, p. 304).

Although neither the James-Stein estimator δ_c nor δ_c^+ is admissible, it is found that no substantial improvements over δ_c^+ are possible (Efron and Morris, 1973).

Extension of Theorem 4.15 to $\text{Var}(X) = \sigma^2 D$

Consider the case where $\text{Var}(X) = \sigma^2 D$ with an unknown $\sigma^2 > 0$ and a known positive definite matrix D .

If σ^2 is known, then an extended James-Stein estimator is

$$\tilde{\delta}_{c,r} = X - \frac{(p-2)r\sigma^2}{\|D^{-1}(X-c)\|^2} D^{-1}(X-c).$$

Under the squared error loss, the risk of $\tilde{\delta}_{c,r}$ is (exercise)

$$\sigma^2 \left[\text{tr}(D) - (2r - r^2)(p-2)^2 \sigma^2 E(\|D^{-1}(X-c)\|^{-2}) \right].$$

When σ^2 is unknown, we assume that there exists a statistic S_0^2 such that S_0^2 is independent of X and S_0^2/σ^2 has the chi-square distribution χ_m^2 (see Example 4.27).

Replacing $r\sigma^2$ in $\tilde{\delta}_{c,r}$ by $\hat{\sigma}^2 = tS_0^2$ with a constant $t > 0$ leads to the following extended James-Stein estimator:

$$\tilde{\delta}_c = X - \frac{(p-2)\hat{\sigma}^2}{\|D^{-1}(X-c)\|^2} D^{-1}(X-c).$$

The risk of $\tilde{\delta}_c$

From the risk formula for $\tilde{\delta}_{c,r}$ and the independence of $\hat{\sigma}^2$ and X , the risk of $\tilde{\delta}_c$ (as an estimator of $\vartheta = EX$) is

$$\begin{aligned}R_{\tilde{\delta}_c}(\theta) &= E \left[E(\|\tilde{\delta}_c - \vartheta\|^2 | \hat{\sigma}^2) \right] \\&= E \left[E(\|\tilde{\delta}_{c,(\hat{\sigma}^2/\sigma^2)} - \vartheta\|^2 | \hat{\sigma}^2) \right] \\&= \sigma^2 E \left\{ \text{tr}(D) - [2(\hat{\sigma}^2/\sigma^2) - (\hat{\sigma}^2/\sigma^2)^2](p-2)^2 \sigma^2 \kappa(\theta) \right\} \\&= \sigma^2 \left\{ \text{tr}(D) - [2E(\hat{\sigma}^2/\sigma^2) - E(\hat{\sigma}^2/\sigma^2)^2](p-2)^2 \sigma^2 \kappa(\theta) \right\} \\&= \sigma^2 \left\{ \text{tr}(D) - [2tm - t^2 m(m+2)](p-2)^2 \sigma^2 \kappa(\theta) \right\},\end{aligned}$$

where $\theta = (\vartheta, \sigma^2)$ and $\kappa(\theta) = E(\|D^{-1}(X - c)\|^2)$.

Since $2tm - t^2 m(m+2)$ is maximized at $t = 1/(m+2)$, replacing t by $1/(m+2)$ leads to

$$R_{\tilde{\delta}_c}(\theta) = \sigma^2 \left[\text{tr}(D) - m(m+2)^{-1}(p-2)^2 \sigma^2 E(\|D^{-1}(X - c)\|^2) \right].$$

which is smaller than $\sigma^2 \text{tr}(D)$ (the risk of X) for any fixed θ , $p \geq 3$.

Example 4.27

Consider the general linear model

$$X = Z\beta + \varepsilon,$$

with $\varepsilon \sim N_p(0, \sigma^2)$, $p \geq 3$, and a full rank Z ,

Consider the estimation of $\vartheta = \beta$ under the squared error loss.

From Theorem 3.8, the LSE $\hat{\beta}$ is from $N(\beta, \sigma^2 D)$ with a known matrix $D = (Z^\tau Z)^{-1}$

$S_0^2 = SSR$ is independent of $\hat{\beta}$

S_0^2/σ^2 has the chi-square distribution χ_{n-p}^2 .

Hence, from the previous discussion, the risk of the shrinkage estimator

$$\hat{\beta} - \frac{(p-2)\hat{\sigma}^2}{\|Z^\tau Z(\hat{\beta} - c)\|^2} Z^\tau Z(\hat{\beta} - c)$$

is smaller than that of $\hat{\beta}$ for any β and σ^2 , where $c \in \mathcal{R}^p$ is fixed and $\hat{\sigma}^2 = SSR/(n-p+2)$.

Other shrinkage estimators

From the previous discussion, the James-Stein estimators improve X substantially when we shrink the observations toward a vector c that is near $\vartheta = EX$.

Of course, this cannot be done since ϑ is unknown.

One may consider shrinking the observations toward the mean of the observations rather than a given point;

that is, one may obtain a shrinkage estimator by replacing c in $\delta_{c,r}$ by $\bar{X}J_p$, where $\bar{X} = p^{-1} \sum_{i=1}^p X_i$ and J_p is the p -vector of ones.

However, we have to replace the factor $p - 2$ in $\delta_{c,r}$ by $p - 3$.

This leads to shrinkage estimators

$$X - \frac{p-3}{\|X - \bar{X}J_p\|^2} (X - \bar{X}J_p)$$

and

$$X - \frac{(p-3)\hat{\sigma}^2}{\|D^{-1}(X - \bar{X}J_p)\|^2} D^{-1}(X - \bar{X}J_p).$$

These estimators are better than X (and, hence, are minimax) when $p \geq 4$, under the squared error loss.

Other shrinkage estimators

The results discussed in this section for the simultaneous estimation of a vector of normal means can be extended to a wide variety of cases

- Brown (1966) considered loss functions that are not the squared error loss
- The results have also been extended to exponential families and to general location parameter families.
- Berger (1976) studied the inadmissibility of generalized Bayes estimators of a location vector
- Berger (1980) considered simultaneous estimation of gamma scale parameters
- Tsui (1981) investigated simultaneous estimation of several Poisson parameters
- See Lehmann (1983, pp. 320-330) for some further references.