Lecture 4: Simultaneous estimation and shrinkage estimators

Simultaneous estimation

Estimation of a p-vector ϑ of parameters (functions of θ) under the decision theory approach.
A vector-valued estimator $T(X)$ can be viewed as a decision rule taking values in the action space $\tilde{\Theta}$ (the range of ϑ).

Difference from estimating ϑ component-by-component

A single loss function $L(\vartheta, a)$, instead of p loss functions

Squared error loss

A natural generalization of the squared error loss is

$$L(\theta, a) = \|a - \vartheta\|^2 = \sum_{i=1}^{p} (a_i - \vartheta_i)^2,$$

where a_i and ϑ_i are the ith components of a and ϑ, respectively.

Many results for the case of a real-valued ϑ can be extended to simultaneous estimation in a straightforward manner: Unbiasedness and UMVUE, Bayes, Minimaxity

Admissibility

Results for admissibility in simultaneous estimation, however, are quite different.

A surprising result (Stein, 1956)

In estimating the vector mean $\theta = EX$ of a normally distributed p-vector X (Example 4.25), X is inadmissible under the squared error loss when $p \geq 3$, although X is the UMVUE and minimax estimator.
Since any estimator better than a minimax estimator is also minimax, there exist many (in fact, infinitely many) minimax estimators in Example 4.25 when $p \geq 3$, which is different from the case of $p = 1$ in which X is the unique admissible minimax estimator (Example 4.6 and Theorem 4.13).
For $p = 2$, Stein (1956) showed that X is admissible and minimax under the squared error loss.

James-Stein estimator

We start with the simple case where X is from $N_p(\theta, I_p)$ with an unknown $\theta \in \mathscr{R}^p$.

James and Stein (1961) proposed the following class of estimators of $\vartheta = \theta$ having smaller risks than X when the squared error loss is used and $p \geq 3$:

$$\delta_c = X - \frac{p-2}{\|X-c\|^2}(X-c),$$

where $c \in \mathscr{R}^p$ is fixed and the choice of c is discussed later.

Extended James-Stein estimators

For the purpose of generalizing the results to more complicated situations, we consider the following extension of the James-Stein estimator:

$$\delta_{c,r} = X - \frac{r(p-2)}{\|X-c\|^2}(X-c),$$

where $c \in \mathscr{R}^p$ and $r \in \mathscr{R}$ are known.

$\delta_c = \delta_{c,1}$

Motivation 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*.

Motivation 2: empirical Bayes estimator

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.
$1 - B$ is “estimated" by $\psi(\|X - c\|^2)$
$\delta_{c,r}$ can be viewed as an empirical Bayes estimator (§4.1.2).

Theorem 4.15 (Risks of shrinkage estimators)

Suppose that X is from $N_p(\theta, I_p)$ with $p \geq 3$. Then, under the squared error loss, the risks of the following shrinkage estimators of θ,

$$\delta_{c,r} = X - \frac{r(p-2)}{\|X-c\|^2}(X-c),$$

where $c \in \mathscr{R}^p$ and $r \in \mathscr{R}$ are known, are given by

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

- The risk of $\delta_{c,r}$ is smaller than p, the risk of X for every value of θ when $p \geq 3$ and $0 < r < 2$.
- $\delta_c = \delta_{c,1}$ is better than any $\delta_{c,r}$ with $r \neq 1$, since the factor $2r - r^2$ is maximized at $r = 1$ for $0 < r < 2$.

Proof

We only need to show the case of $c = 0$, since, if $Z = X - c$,

$$R_{\delta_{c,r}}(\theta) = E\|\delta_{c,r} - E(X)\|^2 = E\left\| \left[1 - \frac{r(p-2)}{\|Z\|^2}\right] Z - E(Z) \right\|^2.$$

Proof (continued)

Let $h(\theta) = R_{\delta_{0,r}}(\theta)$, $g(\theta) = p - (2r - r^2)(p-2)^2 E(\|X\|^{-2})$, and $\pi_\alpha(\theta) = (2\pi\alpha)^{-p/2} e^{-\|\theta\|^2/(2\alpha)}$, which is the p.d.f. of $N_p(0, \alpha I_p)$.
To show $g(\theta) = h(\theta)$, we first establish

$$\int_{\mathscr{R}^p} g(\theta)\pi_\alpha(\theta)d\theta = \int_{\mathscr{R}^p} h(\theta)\pi_\alpha(\theta)d\theta, \quad \alpha > 0.$$

Note that the distribution of X can be viewed as the conditional distribution of X given $\vec{\theta} = \theta$, where $\vec{\theta}$ has the Lebesgue p.d.f. $\pi_\alpha(\theta)$.

$$\begin{aligned}
\int_{\mathscr{R}^p} g(\theta)\pi_\alpha(\theta)d\theta &= p - (2r - r^2)(p-2)^2 E[E(\|X\|^{-2}|\vec{\theta})] \\
&= p - (2r - r^2)(p-2)^2 E(\|X\|^{-2}) \\
&= p - (2r - r^2)(p-2)/(\alpha+1),
\end{aligned}$$

where the expectation in the second line of the previous expression is w.r.t. the joint distribution of $(X, \vec{\theta})$ and the last equality follows from the fact that the marginal distribution of X is $N_p(0, (\alpha+1)I_p)$, $\|X\|^2/(\alpha+1)$ has the chi-square distribution χ_p^2 and $E(\|X\|^{-2}) = 1/[(p-2)(\alpha+1)]$.

Proof (continued)

Let $B = 1/(\alpha+1)$ and $\widehat{B} = r(p-2)/\|X\|^2$.

$$
\begin{aligned}
\int_{\mathscr{R}^p} h(\theta)\pi_\alpha(\theta)d\theta &= E\|(1-\widehat{B})X - \vec{\theta}\|^2 \\
&= E\{E[\|(1-\widehat{B})X - \vec{\theta}\|^2|X]\} \\
&= E\{E[\|\vec{\theta} - E(\vec{\theta}|X)\|^2|X] \\
&\quad + \|E(\vec{\theta}|X) - (1-\widehat{B})X\|^2\} \\
&= E\{p(1-B) + (\widehat{B}-B)^2\|X\|^2\} \\
&= E\{p(1-B) + B^2\|X\|^2 \\
&\quad -2Br(p-2) + r^2(p-2)^2\|X\|^{-2}\} \\
&= p - (2r-r^2)(p-2)B,
\end{aligned}
$$

where the fourth equality follows from the fact that the conditional distribution of $\vec{\theta}$ given X is $N_p\big((1-B)X, (1-B)I_p\big)$ and the last equality follows from $E\|X\|^{-2} = B/(p-2)$ and $E\|X\|^2 = p/B$.

Proof (continued)

This proves

$$\int_{\mathscr{R}^p} g(\theta)\pi_\alpha(\theta)d\theta = \int_{\mathscr{R}^p} h(\theta)\pi_\alpha(\theta)d\theta, \quad \alpha > 0.$$

$h(\theta)$ and $g(\theta)$ are expectations of functions of $\|X\|^2$, $\theta^\tau X$, and $\|\theta\|^2$.
Make an orthogonal transformation from X to Y such that
$Y_1 = \theta^\tau X/\|\theta\|$, $EY_j = 0$ for $j > 1$, and $\mathrm{Var}(Y) = I_p$.
Then $h(\theta)$ and $g(\theta)$ are expectations of functions of Y_1, $\sum_{j=2}^p Y_j^2$, and $\|\theta\|^2$.
Thus, both h and g are functions of $\|\theta\|^2$.
For the family of p.d.f.'s $\{\pi_\alpha(\theta) : \alpha > 0\}$, $\|\theta\|^2$ is a complete and sufficient "statistic".
Hence, $\int g(\theta)\pi_\alpha(\theta)d\theta = \int h(\theta)\pi_\alpha(\theta)d\theta$ and the fact that h and g are functions of $\|\theta\|^2$ imply that $h(\theta) = g(\theta)$ a.e. w.r.t. Lebesgue measure.
From Theorem 2.1, both h and g are continuous functions of $\|\theta\|^2$ and, therefore, $h(\theta) = g(\theta)$ for all $\theta \in \mathscr{R}^p$.
This completes the proof.

The improvement

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

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

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 shinkage 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).

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

Consider the case where $\mathrm{Var}(X) = \sigma^2 D$ 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)

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

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 $\widehat{\sigma}^2 = tS_0^2$ with a constant $t > 0$ leads to the following extended James-Stein estimator:

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

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

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

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

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

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

which is smaller than $\sigma^2\mathrm{tr}(D)$ (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 $\widehat{\beta}$ is from $N(\beta, \sigma^2 D)$ with a known matrix $D = (Z^\tau Z)^{-1}$
$S_0^2 = SSR$ is independent of $\widehat{\beta}$
S_0^2/σ^2 has the chi-square distribution χ^2_{n-p}.
Hence, from the previous discussion, the risk of the shrinkage estimator

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

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

Other shinkage 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

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

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.
- The idea of shrinkage has now been used in problems with high dimensions, such as LASSO.