

# Stat 710: Mathematical Statistics

## Lecture 8

Jun Shao

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

# Lecture 8: Simultaneous estimation and James-Stein estimators

## Simultaneous estimation

Estimation of a  $p$ -vector  $\vartheta$  of parameters (functions of  $\theta$ ) 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  $\vartheta$ ).

## Difference from estimating $\vartheta$ 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  $\vartheta_i$  are the  $i$ th components of  $a$  and  $\vartheta$ , respectively.

# Lecture 8: Simultaneous estimation and James-Stein estimators

## Simultaneous estimation

Estimation of a  $p$ -vector  $\vartheta$  of parameters (functions of  $\theta$ ) 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  $\vartheta$ ).

## Difference from estimating $\vartheta$ 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  $\vartheta_i$  are the  $i$ th components of  $a$  and  $\vartheta$ , respectively.

# Lecture 8: Simultaneous estimation and James-Stein estimators

## Simultaneous estimation

Estimation of a  $p$ -vector  $\vartheta$  of parameters (functions of  $\theta$ ) 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  $\vartheta$ ).

## Difference from estimating $\vartheta$ component-by-component

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

## Squared error loss

A natural generalization of the squared error loss is

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

where  $a_i$  and  $\vartheta_i$  are the  $i$ th components of  $\mathbf{a}$  and  $\vartheta$ , respectively.

Many results for the case of a real-valued  $\vartheta$  can be extended to simultaneous estimation in a straightforward manner.

## Unbiasedness and UMVUE

A vector-valued estimator  $T$  is called unbiased if and only if  $E(T) = \vartheta$  for all  $\theta \in \Theta$ .

If there is an unbiased estimator of  $\vartheta$ , then  $\vartheta$  is called estimable.

It can be seen that the result in Theorem 3.1 extends to the case of vector-valued  $\vartheta$  with any  $L$  strictly convex in  $a$ .

If the loss function is the squared error loss and  $T_i$  is a UMVUE of  $\vartheta_i$  for each  $i$ , then  $T = (T_1, \dots, T_p)$  is a UMVUE of  $\vartheta$ .

If there is a sufficient and complete statistic  $U(X)$  for  $\theta$ , then by Theorem 2.5 (Rao-Blackwell theorem),  $T$  must be a function of  $U(X)$  and is the unique best unbiased estimator of  $\vartheta$ .

Many results for the case of a real-valued  $\vartheta$  can be extended to simultaneous estimation in a straightforward manner.

## Unbiasedness and UMVUE

A vector-valued estimator  $T$  is called unbiased if and only if  $E(T) = \vartheta$  for all  $\theta \in \Theta$ .

If there is an unbiased estimator of  $\vartheta$ , then  $\vartheta$  is called estimable.

It can be seen that the result in Theorem 3.1 extends to the case of vector-valued  $\vartheta$  with any  $L$  strictly convex in  $a$ .

If the loss function is the squared error loss and  $T_i$  is a UMVUE of  $\vartheta_i$  for each  $i$ , then  $T = (T_1, \dots, T_p)$  is a UMVUE of  $\vartheta$ .

If there is a sufficient and complete statistic  $U(X)$  for  $\theta$ , then by Theorem 2.5 (Rao-Blackwell theorem),  $T$  must be a function of  $U(X)$  and is the unique best unbiased estimator of  $\vartheta$ .

## Example 4.22

Consider the general linear model

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

with assumption  $\varepsilon = N_n(0, \sigma^2 I_n)$  and a full rank  $Z$ . Let  $\vartheta = \beta$ .

An unbiased estimator of  $\beta$  is then the LSE  $\hat{\beta}$ .

From the proof of Theorem 3.7,  $\hat{\beta}$  is a function of the sufficient and complete statistic for  $\theta = (\beta, \sigma^2)$ .

Hence,  $\hat{\beta}$  is the unique best unbiased estimator of  $\vartheta$  under any strictly convex loss function.

In particular,  $\hat{\beta}$  is the UMVUE of  $\beta$  under the squared error loss.

## Bayes estimators

Bayes estimators are still defined to be Bayes actions considered as functions of  $X$ .

Under the squared error loss, the Bayes estimator is still the posterior mean (vector).

## Example 4.22

Consider the general linear model

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

with assumption  $\varepsilon = N_n(0, \sigma^2 I_n)$  and a full rank  $Z$ . Let  $\vartheta = \beta$ .

An unbiased estimator of  $\beta$  is then the LSE  $\hat{\beta}$ .

From the proof of Theorem 3.7,  $\hat{\beta}$  is a function of the sufficient and complete statistic for  $\theta = (\beta, \sigma^2)$ .

Hence,  $\hat{\beta}$  is the unique best unbiased estimator of  $\vartheta$  under any strictly convex loss function.

In particular,  $\hat{\beta}$  is the UMVUE of  $\beta$  under the squared error loss.

## Bayes estimators

Bayes estimators are still defined to be Bayes actions considered as functions of  $X$ .

Under the squared error loss, the Bayes estimator is still the posterior mean (vector).

## Example 4.23

Let  $X = (X_0, X_1, \dots, X_k)$  have the multinomial distribution given in Example 2.7.

Consider the estimation of the vector  $\theta = (p_0, p_1, \dots, p_k)$  under the squared error loss, and the Dirichlet prior for  $\theta$  that has the Lebesgue p.d.f.

$$\frac{\Gamma(\alpha_0 + \dots + \alpha_k)}{\Gamma(\alpha_0) \dots \Gamma(\alpha_k)} p_0^{\alpha_0 - 1} \dots p_k^{\alpha_k - 1} I_A(\theta),$$

where  $\alpha_j$ 's are known positive constants and

$$A = \{\theta : 0 \leq p_j, \sum_{j=0}^k p_j = 1\}.$$

It turns out that the Dirichlet prior is conjugate so that the posterior of  $\theta$  given  $X = x$  is also a Dirichlet distribution having the p.d.f. with  $\alpha_j$  replaced by  $\alpha_j + x_j$ ,  $j = 0, 1, \dots, k$ .

Thus, the Bayes estimator of  $\theta$  is  $\delta = (\delta_0, \delta_1, \dots, \delta_k)$  with

$$\delta_j(X) = \frac{\alpha_j + X_j}{\alpha_0 + \alpha_1 + \dots + \alpha_k + n}, \quad j = 0, 1, \dots, k.$$

## Minimaxity

The definition of minimax estimators applies without changes.

### Example 4.25

Let  $X$  be a sample from  $N_p(\theta, I_p)$  with an unknown  $\theta \in \mathcal{R}^p$ .

Consider the estimation of  $\theta$  under the squared error loss.

A modification of the proof of Theorem 4.12 with independent priors for  $\theta_j$ 's shows that  $X$  is a minimax estimator of  $\theta$  (exercise).

### Example 4.26

Consider Example 4.23.

If we choose  $\alpha_0 = \dots = \alpha_k = \sqrt{n}/(k+1)$ , then the Bayes estimator of  $\theta$  in Example 4.23 has constant risk.

Using the same argument in the proof of Theorem 4.11, we can show that this Bayes estimator is minimax.

The previous results for simultaneous estimation are fairly straightforward generalizations of those for the case of a real-valued  $\vartheta$ .

## Minimaxity

The definition of minimax estimators applies without changes.

### Example 4.25

Let  $X$  be a sample from  $N_p(\theta, I_p)$  with an unknown  $\theta \in \mathcal{R}^p$ .

Consider the estimation of  $\theta$  under the squared error loss.

A modification of the proof of Theorem 4.12 with independent priors for  $\theta_i$ 's shows that  $X$  is a minimax estimator of  $\theta$  (exercise).

### Example 4.26

Consider Example 4.23.

If we choose  $\alpha_0 = \dots = \alpha_k = \sqrt{n}/(k+1)$ , then the Bayes estimator of  $\theta$  in Example 4.23 has constant risk.

Using the same argument in the proof of Theorem 4.11, we can show that this Bayes estimator is minimax.

The previous results for simultaneous estimation are fairly straightforward generalizations of those for the case of a real-valued  $\vartheta$ .

## Minimaxity

The definition of minimax estimators applies without changes.

### Example 4.25

Let  $X$  be a sample from  $N_p(\theta, I_p)$  with an unknown  $\theta \in \mathcal{R}^p$ .

Consider the estimation of  $\theta$  under the squared error loss.

A modification of the proof of Theorem 4.12 with independent priors for  $\theta_i$ 's shows that  $X$  is a minimax estimator of  $\theta$  (exercise).

### Example 4.26

Consider Example 4.23.

If we choose  $\alpha_0 = \cdots = \alpha_k = \sqrt{n}/(k+1)$ , then the Bayes estimator of  $\theta$  in Example 4.23 has constant risk.

Using the same argument in the proof of Theorem 4.11, we can show that this Bayes estimator is minimax.

The previous results for simultaneous estimation are fairly straightforward generalizations of those for the case of a real-valued  $\vartheta$ .

## Minimaxity

The definition of minimax estimators applies without changes.

### Example 4.25

Let  $X$  be a sample from  $N_p(\theta, I_p)$  with an unknown  $\theta \in \mathcal{R}^p$ .

Consider the estimation of  $\theta$  under the squared error loss.

A modification of the proof of Theorem 4.12 with independent priors for  $\theta_j$ 's shows that  $X$  is a minimax estimator of  $\theta$  (exercise).

### Example 4.26

Consider Example 4.23.

If we choose  $\alpha_0 = \cdots = \alpha_k = \sqrt{n}/(k+1)$ , then the Bayes estimator of  $\theta$  in Example 4.23 has constant risk.

Using the same argument in the proof of Theorem 4.11, we can show that this Bayes estimator is minimax.

The previous results for simultaneous estimation are fairly straightforward generalizations of those for the case of a real-valued  $\vartheta$ .

## 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 (Example 4.25).

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.

## 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 (Example 4.25).

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  $\theta \in \mathcal{R}^p$ .

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

### Theorem 4.15

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 estimators of  $\theta$ ,

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

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

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

## James-Stein estimator

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

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

### Theorem 4.15

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 estimators of  $\theta$ ,

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

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

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

Let  $Z = X - c$ . Then

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

Hence, we only need to show the case of  $c = 0$ .

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  $N_p(0, \alpha I_p)$ .

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.  $\pi_\alpha(\theta)$ .

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

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  $\chi_p^2$  and  $E(\|X\|^{-2}) = 1/[(p - 2)(\alpha + 1)]$ .

## Proof (continued)

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

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

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

## Proof (continued)

This proves

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

$h(\theta)$  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  $\text{Var}(Y) = I_p$ .

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  $\theta \in \mathcal{R}^p$ .

This completes the proof.