Ensemble minimaxity of James-Stein estimators

Yuzo Maruyama, Larry Brown and Ed George

Mathematics & Informatics Center, University of Tokyo
Department of Statistics, Wharton School, University of Pennsylvania

coauthor Larry

coauthor Ed

Two parts

Estimation of a multivariate normal mean

- $\boldsymbol{X} \sim N_d(\boldsymbol{\theta}, \boldsymbol{\Sigma})$ heteroscedasticity

  $\boldsymbol{\Sigma} = \text{diag}(\sigma_1^2, \ldots, \sigma_d^2)$, with $\sigma_1^2 > \sigma_2^2 > \cdots > \sigma_d^2$

- $\boldsymbol{X} \sim N_d(\boldsymbol{\theta}, \boldsymbol{I})$ homoscedasticity

- Ensemble minimaxity of some James-Stein variants under loss $L(\boldsymbol{\delta}, \boldsymbol{\theta}) = \|\boldsymbol{\delta} - \boldsymbol{\theta}\|^2 = \sum_{i=1}^{d} (\delta_i - \theta_i)^2$

Problem setting

- Let $\boldsymbol{X} \sim N_d(\boldsymbol{\theta}, \boldsymbol{\Sigma})$ where

$$\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)^{\mathrm{T}}, \ \boldsymbol{\Sigma} = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_d^2)$$

- assume $\sigma_1^2 > \sigma_2^2 > \cdots > \sigma_d^2$
- estimation of $\boldsymbol{\theta}$ w.r.t.

$$L(\boldsymbol{\delta}, \boldsymbol{\theta}) = \|\boldsymbol{\delta} - \boldsymbol{\theta}\|^2 = \sum_{i=1}^{d} (\delta_i - \theta_i)^2$$

- The risk of $\boldsymbol{\delta}(\boldsymbol{X})$

$$R(\boldsymbol{\delta}, \boldsymbol{\theta}) = E\left[L(\boldsymbol{\delta}, \boldsymbol{\theta})\right]$$

Loss I

Why $\sum_{i=1}^{d}(\delta_i - \theta_i)^2$?

- Whether or not an estimator is minimax is tied to the particular loss function chosen
- For example, the scale invariant loss $\sum_{i=1}^{d}\frac{(\delta_i - \theta_i)^2}{\sigma_i^2}$ reduces the effect of components with larger variances

  $\Uparrow$ See Casella (1980,1985)

Loss II

- $L_0(\boldsymbol{\delta}, \boldsymbol{\theta}) = \sum_{i=1}^{d} (\delta_i - \theta_i)^2$ is a kind of least favorable among the class

$$\left\{ L_j(\boldsymbol{\delta}, \boldsymbol{\theta}) = \sum_{i=1}^{d} \frac{(\delta_i - \theta_i)^2}{\{\sigma_i^2\}^j} : 0 \leq j \leq 2 \right\}$$

- If an estimator among the class, which we will consider in this talk, is minimax under L_0, then minimaxity of the estimator under L_j for $0 < j \leq 2$ still holds

$\Uparrow$ See Maruyama & Strawderman (2005)

Original James-Stein

- The MLE $\boldsymbol{X}$ $\begin{cases} \text{the constant risk } \operatorname{tr}\boldsymbol{\Sigma} = \sum \sigma_i^2 \\ \text{minimax for any } p \text{ and any } \boldsymbol{\Sigma} \end{cases}$

  James and Stein (1961)
- In the homoscedastic case, $\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{I}_d$, or equivalently

$$\sigma_1^2 = \cdots = \sigma_d^2 = \sigma^2,$$

$$\left(1 - \frac{c\sigma^2}{\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}}\right)\boldsymbol{X} = \left(1 - \frac{c}{\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}}\right)\boldsymbol{X}$$

  for $c \in (0, 2(d-2))$ dominates $\boldsymbol{X}$ for $d \geq 3$

Review I

$\exists$ some literature discussing the minimax properties of shrinkage estimators under heteroscedasticity

$$\sigma_1^2 > \sigma_2^2 > \cdots > \sigma_d^2$$

Brown (1975)

- the James-Stein estimator $\left(1 - \frac{c}{\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}}\right)\boldsymbol{X}$ is not always minimax under heteroscedasticity
- Specifically, it is not minimax for any $c \in (0, 2(d-2))$

$$\text{when} \qquad 2\sigma_1^2 > \sum_{i=1}^{d} \sigma_i^2$$

Review II

Berger (1976)

- For $d \geq 3$ and any $\boldsymbol{\Sigma}$, minimaxity of

$$\left(\boldsymbol{I} - \boldsymbol{\Sigma}^{-1}\frac{c}{\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-2}\boldsymbol{X}}\right)\boldsymbol{X} \text{ for } c \in (0, 2(d-2))$$

(recall $\boldsymbol{\Sigma} = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_d^2)$ with $\sigma_1^2 > \sigma_2^2 > \cdots > \sigma_d^2$)

Casella (1980)

- the estimator $\left(\boldsymbol{I} - \boldsymbol{\Sigma}^{-1}\frac{c}{\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-2}\boldsymbol{X}}\right)\boldsymbol{X}$ is not desirable even if it is minimax
- Ordinary minimax estimators, typically shrink more on the coordinates with smaller variances

Review III

- From Casella's viewpoint, one of the most natural variant of the James-Stein estimator is

$$\left(\boldsymbol{I} - \boldsymbol{\Sigma}\frac{c}{\|\boldsymbol{X}\|^2}\right)\boldsymbol{X} \text{ for } c > 0,$$

  (recall $\boldsymbol{\Sigma} = \mathrm{diag}(\sigma_1^2, \dots, \sigma_d^2)$ with $\sigma_1^2 > \sigma_2^2 > \cdots > \sigma_d^2$)
  which shrink most on the coordinates with larger variances

  $\Uparrow$ not typically ordinary minimax
- We are going to save the shrinkage estimators above, by providing ensemble minimaxity

Ensemble risk I

- the Bayes risk with respect to the prior π

$$\bar{R}(\pi, \boldsymbol{\delta}) = E_{\pi}(R(\boldsymbol{\theta}, \boldsymbol{\delta})) = \int_{\mathbb{R}^d} R(\boldsymbol{\theta}, \boldsymbol{\delta})\pi(\mathrm{d}\boldsymbol{\theta})$$

- Efron and Morris (1971, 1972a, 1972b, 1973) addressed this problem from both the Bayes and empirical Bayes perspective

Ensemble risk II

- Especially, they considered a prior distribution

$$\boldsymbol{\theta} \sim N_d(\mathbf{0}, \tau \boldsymbol{I}_d) \text{ with } \tau \in (0, \infty)$$

- They used the term "ensemble risk" for $\bar{R}(\pi, \boldsymbol{\delta})$
- A set of ensemble risks $\{\bar{R}(\boldsymbol{\delta}, \tau) : \tau \in (0, \infty)\}$

$$\bar{R}(\boldsymbol{\delta}, \tau) = \int_{\mathbb{R}^d} R(\boldsymbol{\delta}, \boldsymbol{\theta}) \frac{1}{(2\pi\tau)^{d/2}} \exp\left(-\frac{\|\boldsymbol{\theta}\|^2}{2\tau}\right) \mathrm{d}\boldsymbol{\theta},$$

Ensemble risk III

Definition of ensemble minimaxity

the estimator $\boldsymbol{\delta}$ is ensemble minimax w.r.t. $\mathcal{P}_\star$

$$\Leftrightarrow \sup_{\tau\in(0,\infty)} \bar{R}(\boldsymbol{\delta},\tau) = \inf_{\boldsymbol{\delta}'} \sup_{\tau\in(0,\infty)} \bar{R}(\boldsymbol{\delta}',\tau)$$

c.f. $\boldsymbol{\delta}$ is said to be ordinary minimax

$$\Leftrightarrow \sup_{\boldsymbol{\theta}\in\Theta} R(\boldsymbol{\theta},\boldsymbol{\delta}) = \inf_{\boldsymbol{\delta}'} \sup_{\boldsymbol{\theta}\in\Theta} R(\boldsymbol{\theta},\boldsymbol{\delta}')$$

Ensemble risk IV

In our problem, $\boldsymbol{X}$ is still ensemble minimax with the constant risk $\sum \sigma_i^2$

$$\text{ensemble minimaxity if } \sup_{\tau \in (0,\infty)} \bar{R}(\boldsymbol{\delta}, \tau) = \sum \sigma_i^2$$

ordinary minimaxity $\Rightarrow$ ensemble minimaxity
$\not\Leftarrow$

Ensemble risk V

- As a matter of fact, Larry, in his unpublished manuscript, has already introduced the concept of ensemble minimaxity
- Here, we follow their spirit but propose a simpler and clearer approach for establishing ensemble minimaxity

Ensemble minimaxity I

(Our unpublished) paper

- A class of shrinkage estimators with general $\boldsymbol{G}$

$$\boldsymbol{\delta}_\phi = \left(\boldsymbol{I} - \boldsymbol{G}\frac{\phi(z)}{z}\right)\boldsymbol{x}, \begin{cases} z = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{G}\boldsymbol{\Sigma}^{-1}\boldsymbol{x} = \sum \frac{g_i x_i^2}{\sigma_i^2} \\ \boldsymbol{G} = \mathrm{diag}(g_1, \dots, g_d), 0 < g_i \le 1 \ \forall i \end{cases}$$

This talk $\boldsymbol{G} = \boldsymbol{\Sigma}/\sigma_1^2$

- a special class of shrinkage estimators

$$\boldsymbol{\delta}_\phi = \left(\boldsymbol{I} - \frac{\boldsymbol{\Sigma}}{\sigma_1^2}\frac{\phi(\|\boldsymbol{x}\|^2/\sigma_1^2)}{\|\boldsymbol{x}\|^2/\sigma_1^2}\right)\boldsymbol{x}$$

(recall $\boldsymbol{\Sigma} = \mathrm{diag}(\sigma_1^2, \dots, \sigma_d^2)$ with $\sigma_1^2 > \sigma_2^2 > \cdots > \sigma_d^2$)

Ensemble minimaxity II

Berger and Srinivasan (1978)

Given positive-definite $\boldsymbol{C}$ and non-singular $\boldsymbol{B}$, a necessary condition for an estimator of the form

$$\left(\boldsymbol{I} - \boldsymbol{B}\frac{\phi(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{C}\boldsymbol{x})}{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{C}\boldsymbol{x}}\right)\boldsymbol{x}$$

to be admissible is $\boldsymbol{B} \propto \boldsymbol{\Sigma}\boldsymbol{C}$

$\Uparrow$ which is satisfied by

$$\left(\boldsymbol{I} - \boldsymbol{G}\frac{\phi(z)}{z}\right)\boldsymbol{x}, \quad \left(\boldsymbol{I} - \frac{\boldsymbol{\Sigma}}{\sigma_1^2}\frac{\phi(\|\boldsymbol{x}\|^2/\sigma_1^2)}{\|\boldsymbol{x}\|^2/\sigma_1^2}\right)\boldsymbol{x}$$

Ensemble minimaxity III

Baranchik-type sufficient condition for minimaxity

For given $\boldsymbol{G}$ which satisfies

$$h(\boldsymbol{\Sigma}, \boldsymbol{G}) = 2\left(\frac{\sum g_i\sigma_i^2}{\max(g_i\sigma_i^2)} - 2\right) > 0,$$

$\left(\boldsymbol{I} - \boldsymbol{G}\frac{\phi(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{G}\boldsymbol{\Sigma}^{-1}\boldsymbol{x})}{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{G}\boldsymbol{\Sigma}^{-1}\boldsymbol{x}}\right)\boldsymbol{x}$, is ordinary minimax if

ϕ is non-decreasing and $0 \le \phi(\cdot) \le h(\boldsymbol{\Sigma}, \boldsymbol{G})$

techniques $\begin{cases} \text{Stein's identity} \\ \sum a_i y_i^2 \le \max_i a_i \sum y_i^2 \end{cases}$

Ensemble minimaxity IV

Berger (1976)

For any given $\boldsymbol{\Sigma}$,

$$\max_{\boldsymbol{G}} h(\boldsymbol{\Sigma}, \boldsymbol{G}) = 2(d-2), \quad \arg\max_{\boldsymbol{G}} h(\boldsymbol{\Sigma}, \boldsymbol{G}) = \sigma_d^2 \boldsymbol{\Sigma}^{-1}$$

Is $\boldsymbol{G} = \sigma_d^2 \boldsymbol{\Sigma}^{-1} = \mathrm{diag}\left(\frac{\sigma_d^2}{\sigma_1^2}, \dots, 1\right)$ the right choice?

(recall $\boldsymbol{\Sigma} = \mathrm{diag}(\sigma_1^2, \dots, \sigma_d^2)$ with $\sigma_1^2 > \sigma_2^2 > \cdots > \sigma_d^2$)

Casella (1980)

More shrinkage on higher variance corresponds to the descending order $g_1 > \cdots > g_d$

Ensemble minimaxity V

Our choice $\boldsymbol{G} = \boldsymbol{\Sigma}/\sigma_1^2 \Rightarrow$ the descending order!

(recall $\boldsymbol{\Sigma} = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_d^2)$ with $\sigma_1^2 > \sigma_2^2 > \cdots > \sigma_d^2$)

- $h(\boldsymbol{\Sigma}, \boldsymbol{G}) = 2 \left(\dfrac{\sum \sigma_i^4}{\sigma_1^4} - 2 \right)$
- For $\sigma_1^2 >> \sigma_2^2 >, \ldots,$, $h(\boldsymbol{\Sigma}, \boldsymbol{G})$ is typically negative, which implies that a sufficient condition for ordinary minimaxity by Baranchik is empty

$\Uparrow$ We are going to save the shrinkage estimators $\boldsymbol{\delta}_\phi = \left(\boldsymbol{I} - \dfrac{\boldsymbol{\Sigma}}{\sigma_1^2} \dfrac{\phi(\|\boldsymbol{x}\|^2/\sigma_1^2)}{\|\boldsymbol{x}\|^2/\sigma_1^2} \right) \boldsymbol{x}$ under this situation

Theorem of ensemble minimaxity

Assumption $\begin{cases} \phi : & \geq 0, \ \nearrow, \ \text{concave} \\ \phi(z)/z : & \searrow \end{cases}$

$\Uparrow$ Baranchik, extra

Then

$$\boldsymbol{\delta}_\phi = \left(\boldsymbol{I} - \frac{\boldsymbol{\Sigma}}{\sigma_1^2} \frac{\phi(\|\boldsymbol{x}\|^2/\sigma_1^2)}{\|\boldsymbol{x}\|^2/\sigma_1^2} \right) \boldsymbol{x}$$

is ensemble minimax if

$$\phi\left(d\frac{\sigma_d^2 + \tau}{\sigma_1^2} \right) \leq 2(d-2)\frac{\sigma_d^2 + \tau}{\sigma_1^2 + \tau} \quad \forall \tau \in (0, \infty)$$

$\Uparrow$ Upperbound of ϕ like Baranchik's condition, but including τ

Sketch of the Proof I

- Note $\theta_i|x_i \sim N\left(\frac{\tau}{\tau+\sigma_i^2}x_i, \frac{\tau\sigma_i^2}{\tau+\sigma_i^2}\right)$ and $x_i \sim N(0, \tau+\sigma_i^2)$,
  $\Uparrow$ $\theta_1|x_1, \ldots, \theta_d|x_d$ are independent and $x_1, \ldots, x_d$ are independent
- Then the Bayes risk

$$\bar{R}(\boldsymbol{\delta}_\phi, \tau) = \sum_{i=1}^{d} E_{\boldsymbol{\theta}} E_{\boldsymbol{x}|\boldsymbol{\theta}}\left[\left\{\left(1 - \frac{\sigma_i^2}{\sigma_1^2}\frac{\phi(z)}{z}\right)x_i - \theta_i\right\}^2\right],\ z = \frac{\sum x_i^2}{\sigma_1^2}$$

$\Downarrow$

$$\bar{R}(\boldsymbol{\delta}_\phi, \tau) - \sum \sigma_i^2 = E_{\boldsymbol{x}}\left[-2\sum_{i=1}^{d}\frac{\sigma_i^4 x_i^2}{\sigma_1^2(\tau+\sigma_i^2)}\frac{\phi(z)}{z} + \frac{\sum_{i=1}^{d}\sigma_i^4 x_i^2}{\sigma_1^4}\frac{\phi^2(z)}{z^2}\right]$$

Sketch of the Proof II

- Let $w_i = \dfrac{x_i^2}{\sigma_i^2 + \tau}$, $w = \sum_{i=1}^d w_i$ and $t_i = \dfrac{w_i}{w}$ for $i = 1, \ldots, d$.
- Then w and $\boldsymbol{t} = (t_1, \ldots, t_d)^{\mathrm{T}}$ are mutually independent

$$w = \sum_{i=1}^d w_i \sim \chi_d^2, \quad \boldsymbol{t} \sim \mathrm{Dirichlet}(1/2, \ldots, 1/2)$$

- With the notation, we have

$$x_i^2 = w t_i(\sigma_i^2 + \tau) \text{ and } z = \frac{1}{\sigma_1^2} \sum_{i=1}^d x_i^2 = \frac{w}{\sigma_1^2} \sum_{i=1}^d t_i \left(\sigma_i^2 + \tau\right)$$

Sketch of the Proof III

- Then, after some inequalities including Jensen's inequality, the correlation inequality

$$\mathrm{E}[f(X)g(X)] \geq \mathrm{E}[f(X)]\mathrm{E}[g(X)] \text{ if } f \nearrow,\ g \nearrow$$

and

$$\sum(\sigma_i^2 + \tau)t_i\sigma_i^4 \leq (\sigma_1^2 + \tau)\sum t_i\sigma_i^4,$$

we have the result

- No Stein's identity is used!

Example motivated by Stein (1956) I

Let $\phi(z) = \dfrac{c_1 z}{c_2 + z} \quad c_1 > 0$ and $c_2 \geq 0$

Then $\left(\boldsymbol{I} - \boldsymbol{\Sigma}\dfrac{c_1}{c_2\sigma_1^2 + \|\boldsymbol{x}\|^2}\right)\boldsymbol{x} \quad c_1 > 0$ and $c_2 \geq 0$

Stein (1956)

Under $\boldsymbol{\Sigma} = \boldsymbol{I}_d$, Stein (1956) suggested that there exist estimators dominating $\boldsymbol{x}$ among a class of estimators $\left(1 - \dfrac{c_1}{c_2 + \|\boldsymbol{x}\|^2}\right)\boldsymbol{x}$ for small c_1 and large c_2

Example motivated by Stein (1956) II

- $\phi(z) = \dfrac{c_1 z}{c_2 + z} \quad c_1 > 0$ and $c_2 \geq 0$
- Note $\phi(z) \geq 0$, $\nearrow$, concave and $\phi(z)/z \searrow$

$$\frac{c_1 d(\sigma_d^2 + \tau)/\sigma_1^2}{c_2 + d(\sigma_d^2 + \tau)/\sigma_1^2} \leq 2(d-2)\frac{\sigma_d^2 + \tau}{\sigma_1^2 + \tau} \quad \forall \tau \in (0, \infty)$$

which is equivalent to

$$d\tau \left\{2(d-2) - c_1\right\} + 2(d-2)\sigma_1^2 \left\{c_2 - d\left(\frac{c_1}{2(d-2)} - \frac{\sigma_d^2}{\sigma_1^2}\right)\right\} \geq 0$$

Example motivated by Stein (1956) III

the estimator $\left(\boldsymbol{I} - \boldsymbol{\Sigma}\frac{c_1}{c_2\sigma_1^2 + \|\boldsymbol{x}\|^2}\right)\boldsymbol{x}$

1. ensemble minimax if

$$0 < c_1 \leq 2(d-2) \text{ and } c_2 \geq \max\left(0, d\left(\frac{c_1}{2(d-2)} - \frac{\sigma_d^2}{\sigma_1^2}\right)\right)$$

2. ordinary minimax if

$$\underbrace{\sum \frac{\sigma_i^4}{\sigma_1^4} - 2}_{<0 \text{ if } \sigma_1^2 >> \sigma_2^2} > 0 \text{ and } c_1 \leq 2\left(\sum \frac{\sigma_i^4}{\sigma_1^4} - 2\right)$$

Example motivated by Stein (1956) IV

An interesting case: $c_1 = c_2 = d - 2$

the James-Stein variant $\left(\boldsymbol{I} - \boldsymbol{\Sigma} \frac{d-2}{(d-2)\sigma_1^2 + \|\boldsymbol{x}\|^2} \right) \boldsymbol{x}$

the i-th shrinkage factor $\Uparrow$

$$1 - \frac{(d-2)\sigma_i^2}{(d-2)\sigma_1^2 + \|\boldsymbol{x}\|^2} \geq 0 \text{ for any } \boldsymbol{x} \text{ and } \boldsymbol{\Sigma}$$

$+$ Ascending order of the shrinkage factor

$$1 - \frac{(d-2)\sigma_1^2}{(d-2)\sigma_1^2 + \|\boldsymbol{x}\|^2} < \cdots < 1 - \frac{(d-2)\sigma_d^2}{(d-2)\sigma_1^2 + \|\boldsymbol{x}\|^2}$$

(recall $\boldsymbol{\Sigma} = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_d^2)$ with $\sigma_1^2 > \sigma_2^2 > \cdots > \sigma_d^2$)

Example of Bayes I

A Bayes satisfier

- the prior, an extension of $\|\boldsymbol{\theta}\|^{2-d}$

$$\boldsymbol{\theta} \,|\, \lambda \sim N_d(\mathbf{0}, (\lambda^{-1}\sigma_1^2 \boldsymbol{I}_d - \boldsymbol{\Sigma})), \;\; \pi(\lambda) \sim \lambda^{-2} I_{(0,1)}(\lambda)$$

- for $\boldsymbol{\Sigma} = \boldsymbol{I}_d$, the prior density is exactly $\|\boldsymbol{\theta}\|^{2-d}$ since

$$\frac{1}{(2\pi)^{d/2}} \int_0^1 \left(\frac{\lambda}{1-\lambda}\right)^{d/2} \exp\left(-\frac{\lambda\|\boldsymbol{\theta}\|^2}{2(1-\lambda)}\right) \lambda^{-2} \mathrm{d}\lambda$$
$$= \frac{1}{(2\pi)^{d/2}} \int_0^\infty g^{d/2-2} \exp\left(-g\|\boldsymbol{\theta}\|^2/2\right) \mathrm{d}g = \frac{\Gamma(d/2-1)2^{d/2-1}}{(2\pi)^{d/2}} \|\boldsymbol{\theta}\|^{2-d}$$

Example of Bayes II

- the generalized Bayes estimator w.r.t. the prior is

$$\boldsymbol{\delta}_* = \left(\boldsymbol{I} - \frac{\boldsymbol{\Sigma}}{\sigma_1^2} \frac{\int_0^1 \lambda^{d/2-1} \exp(-\|\boldsymbol{x}\|^2 \lambda / \{2\sigma_1^2\}) d\lambda}{\int_0^1 \lambda^{d/2-2} \exp(-\|\boldsymbol{x}\|^2 \lambda / \{2\sigma_1^2\}) d\lambda} \right) \boldsymbol{x}$$

$\Uparrow$ by the way of Strawderman (1971)

- ensemble minimax
- ordinary minimax if $2 \left(\sum \sigma_i^4 / \sigma_1^4 - 2 \right) \geq d - 2$
- admissible

$\Uparrow$ we omit the proofs

Numerical experiment I

- $d = 10$
- $\boldsymbol{\Sigma} = \mathrm{diag}(a^9, a^8, \ldots, a, 1)$
- $a = 1.01, 1.05, 1.25, 1.5$
  Approximately a^9 is $1.09, 1.55, 7.45, 38.4$, respectively
- the James-Stein variant and Bayes

$$\boldsymbol{\delta}_{JS} = \left(\boldsymbol{I} - \boldsymbol{\Sigma} \frac{d-2}{(d-2)\sigma_1^2 + \|\boldsymbol{x}\|^2} \right) \boldsymbol{x}$$

$$\boldsymbol{\delta}_* = \left(\boldsymbol{I} - \frac{\boldsymbol{\Sigma}}{\sigma_1^2} \frac{\int_0^1 \lambda^{d/2-1} \exp(-\|\boldsymbol{x}\|^2 \lambda / \{2\sigma_1^2\}) d\lambda}{\int_0^1 \lambda^{d/2-2} \exp(-\|\boldsymbol{x}\|^2 \lambda / \{2\sigma_1^2\}) d\lambda} \right) \boldsymbol{x}$$

Numerical experiment II

- A sufficient condition for both estimators to be ordinary minimax is given by

$$2\left(\sum_{i=1}^{d}\sigma_i^4/\sigma_1^4-2\right)=2\left(\sum_{i=1}^{d}a^{2(i-10)}-2\right)\geq d-2,$$

where the equality is attained by $a \approx 1.066$

- the inequality above $\begin{cases}\text{is satisfied by } a=1.01, 1.05\\ \text{is not satisfied by } a=1.25, 1.5\end{cases}$

Numerical experiment III

Relative ordinary risk improvement given by

$$1 - \frac{R(\boldsymbol{\theta}, \boldsymbol{\delta}_\phi)}{\mathrm{tr}\boldsymbol{\Sigma}} \quad \text{at} \quad \boldsymbol{\theta} = m\{\mathrm{tr}\boldsymbol{\Sigma}\}^{1/2}\frac{\mathbf{1}_{10}}{\sqrt{10}}$$

– sign means $R(\boldsymbol{\theta}, \boldsymbol{\delta}_\phi) > \mathrm{tr}\boldsymbol{\Sigma}$, non-minimaxity

Table:

	$a \backslash m$	0	2	20	40	60	80	100
δ_*	1.01	0.79	0.14	1.7×10^{-3}	4.8×10^{-4}	2.5×10^{-4}	1.7×10^{-4}	1.3×10^{-4}
	1.05	0.75	0.14	1.7×10^{-3}	4.3×10^{-4}	2.0×10^{-4}	1.2×10^{-4}	8.0×10^{-5}
	1.25	0.63	0.19	1.9×10^{-3}	2.5×10^{-4}	-5.6×10^{-5}	-1.7×10^{-4}	-2.2×10^{-4}
	1.5	0.63	0.27	2.7×10^{-3}	1.6×10^{-4}	-3.0×10^{-4}	-4.6×10^{-4}	-5.4×10^{-4}
δ_{JS}	1.01	0.80	0.14	1.7×10^{-3}	4.8×10^{-4}	2.5×10^{-4}	1.7×10^{-4}	1.3×10^{-4}
	1.05	0.79	0.14	1.7×10^{-3}	4.3×10^{-4}	2.0×10^{-4}	1.2×10^{-4}	8.0×10^{-5}
	1.25	0.72	0.19	1.9×10^{-3}	2.5×10^{-4}	-5.6×10^{-5}	-1.7×10^{-4}	-2.2×10^{-4}
	1.5	0.71	0.25	2.7×10^{-3}	1.6×10^{-4}	-3.0×10^{-4}	-4.6×10^{-4}	-5.4×10^{-4}

Numerical experiment IV

Relative Bayes risk improvement given by

$$1 - \frac{\bar{R}(\boldsymbol{\delta}, \tau)}{\mathrm{tr}\boldsymbol{\Sigma}} \quad \text{for } \tau = 1, 5, 20, 40, 60, 80, 100$$

Non-negativeness means $\bar{R}(\boldsymbol{\delta}_\phi, \tau) \leq \mathrm{tr}\boldsymbol{\Sigma}$, ensemble minimaxity

Table: Bayes Risk Difference

	$a \backslash \tau$	1	5	20	40	60	80	100
δ_*	1.01	0.429	0.139	0.039	0.020	0.013	0.010	0.008
	1.05	0.374	0.144	0.042	0.021	0.015	0.011	0.008
	1.25	0.105	0.082	0.038	0.021	0.014	0.011	0.009
	1.5	0.023	0.022	0.019	0.014	0.012	0.010	0.008
δ_{JS}	1.01	0.406	0.137	0.039	0.020	0.014	0.010	0.008
	1.05	0.393	0.143	0.042	0.022	0.015	0.011	0.009
	1.25	0.122	0.079	0.034	0.020	0.014	0.011	0.009
	1.5	0.028	0.025	0.018	0.013	0.010	0.008	0.007

Problem setting

Estimation of a multivariate normal mean $\boldsymbol{\theta}$

- Under homoscedasticity

$$\boldsymbol{X} \sim N_d(\boldsymbol{\theta}, \boldsymbol{I})$$

- loss $L(\boldsymbol{\delta}, \boldsymbol{\theta}) = \|\boldsymbol{\delta} - \boldsymbol{\theta}\|^2 = \sum_{i=1}^{d} (\delta_i - \theta_i)^2$

James-Stein estimator

- the James-Stein estimator $\left(1 - \frac{c}{\|\boldsymbol{X}\|^2}\right)\boldsymbol{X}$
- the risk

$$d + c\{c - 2(d-2)\}\mathrm{E}\left[\frac{1}{\|\boldsymbol{X}\|^2}\right]$$

- minimaxity under $c \in [0, 2(d-2)]$

James-Stein variants I

Unfamiliar James-Stein variants

- Manhattan distance based JS

$$\left(1 - \frac{c}{\{\sum_i |X_i|\}^2}\right)\boldsymbol{X}, \text{ minimax under } c \in [0, 2(d-2)]$$

- max based JS

$$\left(1 - \frac{c}{\{\max_i |X_i|\}^2}\right)\boldsymbol{X}, \text{ minimax under } c \in \left[0, 2\frac{d-2}{d}\right]$$

$\Uparrow$ the ℓ_p norm James-Stein estimator

$$\left(1 - \frac{c}{\|\boldsymbol{X}\|_p^2}\right)\boldsymbol{X} \text{ where } \|\boldsymbol{x}\|_p = \left\{\sum |x_i|^p\right\}^{1/p}$$

James-Stein variants II

- the risk (by Stein's identity and an inequality)

$$d + \mathrm{E}\left[\frac{c}{\|\boldsymbol{X}\|_p^2}\left(c\frac{\|\boldsymbol{X}\|_2^2}{\|\boldsymbol{X}\|_p^2} - 2(d-2)\right)\right]$$
$$\leq d + \mathrm{E}\left[\frac{c}{\|\boldsymbol{X}\|_p^2}\left(c\max(1, d^{1-2/p}) - 2(d-2)\right)\right]$$

- minimaxity under $c \in [0, 2\min(1, d^{2/p-1})(d-2)]$
- Mathematically interesting, but,, „
- Why and how did I arrive at these variants?

Zhou & Hwang (2005) I

- $\hat{\theta}_{\text{ZH}}$: the i-th component

$$\hat{\theta}_{i\text{ZH}} = \left(1 - \frac{c}{\sum_j |x_j|^{2-\alpha}|x_i|^{\alpha}}\right) x_i \text{ for } \alpha > 0$$

- Risk (by Stein's identity and an inequality)

$$d + E\left[c\frac{\sum |X_j|^{2-2\alpha}}{\{\sum |X_j|^{2-\alpha}\}^2}\left(c - 2(1-\alpha)\frac{\sum |X_j|^{2-\alpha}\sum |X_j|^{-\alpha}}{\sum |X_j|^{2-2\alpha}} - 2(\alpha-2)\right)\right]$$

$$\leq d + \mathrm{E}\left[c\frac{\sum |X_j|^{2-2\alpha}}{\{\sum |X_j|^{2-\alpha}\}^2}\left(c - 2(1-\alpha)d - 2(\alpha-2)\right)\right]$$

- minimaxity under $c \in \left[0, 2(d-2)\left(1 - \alpha\frac{d-1}{d-2}\right)\right]$

Zhou & Hwang (2005) II

- ℓ_p norm representation (recall $\|\boldsymbol{x}\|_p = \{\sum |x_i|^p\}^{1/p}$)

$$\hat{\theta}_{i\mathrm{ZH}} = \left(1 - \frac{c}{\|x\|_{2-\alpha}^{2-\alpha} |x_i|^{\alpha}}\right) x_i$$

- My finding $\hat{\theta}_{i\mathrm{LP}} = \left(1 - \frac{c}{\|x\|_p^{2-\alpha} |x_i|^{\alpha}}\right) x_i \ \forall p > 0$

  $\Uparrow$ minimaxity under

$$c \in \left[0, 2(d-2)\min(1, d^{(2-p-\alpha)/p})\left\{1 - \alpha\frac{d-1}{d-2}\right\}\right]$$

- $\alpha = 0 \Leftarrow$ the James-Stein variants in the previous page

Sparsification

- Zhou & Hwang (2005) introduced the case $\alpha > 0$
- Why interesting?
- Sparsification! minimaxity and sparsity simultaneously
- the positive-part estimator dominates the original one

$$\hat{\theta}^{+}_{i\mathrm{LP}} = \max\left(0, 1 - \frac{c}{\|x\|_p^{2-\alpha}|x_i|^{\alpha}}\right) x_i$$

$\Uparrow$ minimaxity of the positive-part estimator

- the i-th component, $\hat{\theta}^{+}_{i\mathrm{LP}} = 0$ if

$$1 - \frac{c}{\|x\|_p^{2-\alpha}|x_i|^{\alpha}} \leq 0 \ \Leftrightarrow \log|x_i| < -\frac{2-\alpha}{\alpha}\log\|x\|_p + \frac{\log c}{\alpha}$$

A PROBLEM OF $\hat{\theta}^{+}_{i\mathrm{LP}}$ I

- the larger c, the more desirable for sparsity
- Ordinary minimaxity is a very conservative criterion and the upper bound, $2(d-2)\gamma_{\mathrm{OM}}(d,p,\alpha)$,

$$\gamma_{\mathrm{OM}}(d,p,\alpha) = \min(1, d^{(2-p-\alpha)/p}) \left\{1 - \alpha\frac{d-1}{d-2}\right\}$$

  is relatively small!

A problem of $\hat{\theta}^{+}_{i\mathrm{LP}}$ II

Relationship

ordinary minimaxity $\Rightarrow$ ensemble minimaxity

$\not\Leftarrow$

$\Downarrow$

- Ensemble minimaxity must be established for larger c
- ensemble minimaxity and sparsification simultaneously

ENSEMBLE MINIMAXITY

Ensemble Bayes risk under $\boldsymbol{\theta} \sim N_d(\mathbf{0}, \tau \boldsymbol{I}_d)$

$$(1+\tau)\left\{\bar{R}(\hat{\boldsymbol{\theta}}_{\mathrm{LP}}, \tau) - d\right\}$$
$$= \frac{c}{d-2}\left(c E_T\left[\frac{\|T\|_{1-\alpha}^{1-\alpha}}{\|T\|_{p/2}^{2-\alpha}}\right] - 2(d-2) E_T\left[\frac{\|T\|_{1-\alpha/2}^{1-\alpha/2}}{\|T\|_{p/2}^{1-\alpha/2}}\right]\right)$$

where $T = (T_1, \ldots, T_d)^{\mathrm{T}} \sim \mathrm{Dirichlet}(1/2, \ldots, 1/2)$

- Ensemble minimaxity under $c \in [0, 2(d-2)\gamma_{\mathrm{EM}}(d, p, \alpha)]$

$$\gamma_{\mathrm{EM}}(d, p, \alpha) = \frac{E_T\left[\|T\|_{1-\alpha/2}^{1-\alpha/2} / \|T\|_{p/2}^{1-\alpha/2}\right]}{E_T\left[\|T\|_{1-\alpha}^{1-\alpha} / \|T\|_{p/2}^{2-\alpha}\right]}$$

COMPARISON γ_{OM} WITH γ_{EM}

Upperbound for

- minimaxity $2(d-2)\gamma_{\text{OM}}(d,p,\alpha)$,

$$\gamma_{\text{OM}}(d,p,\alpha) = \min(1, d^{(2-p-\alpha)/p}) \left\{ 1 - \alpha \frac{d-1}{d-2} \right\}$$

- ensemble minimaxity $2(d-2)\gamma_{\text{EM}}(d,p,\alpha)]$

$$\gamma_{\text{EM}}(d,p,\alpha) = \frac{E_T\left[\|T\|_{1-\alpha/2}^{1-\alpha/2} / \|T\|_{p/2}^{1-\alpha/2} \right]}{E_T\left[\|T\|_{1-\alpha}^{1-\alpha} / \|T\|_{p/2}^{2-\alpha} \right]}$$

Recall

d the dimension of θ, p from ℓ_p, α for sparsification

γ_{OM} and γ_{EM} for some d, p and α

d	p	$\gamma\backslash\alpha$	0.1Δ	0.2Δ	0.3Δ	0.4Δ	0.5Δ	0.6Δ	0.7Δ	0.8Δ	0.9Δ
10	1	γ_{OM}	0.900	0.800	0.700	0.600	0.500	0.400	0.300	0.200	0.100
		γ_{EM}	5.499	4.691	3.975	3.342	2.787	2.301	1.878	1.513	1.200
	2	γ_{OM}	0.812	0.652	0.515	0.398	0.300	0.216	0.147	0.088	0.040
		γ_{EM}	0.926	0.854	0.782	0.713	0.644	0.577	0.512	0.449	0.388
	∞	γ_{OM}	0.090	0.080	0.070	0.060	0.050	0.040	0.030	0.020	0.010
		γ_{EM}	0.313	0.304	0.293	0.281	0.268	0.254	0.238	0.220	0.201
25	1	γ_{OM}	0.900	0.800	0.700	0.600	0.500	0.400	0.300	0.200	0.100
		γ_{EM}	12.349	9.512	7.268	5.502	4.123	3.052	2.227	1.598	1.123
	2	γ_{OM}	0.771	0.588	0.441	0.324	0.231	0.159	0.102	0.058	0.025
		γ_{EM}	0.883	0.775	0.675	0.583	0.498	0.421	0.351	0.288	0.231
	∞	γ_{OM}	0.036	0.032	0.028	0.024	0.020	0.016	0.012	0.008	0.004
		γ_{EM}	0.174	0.166	0.157	0.148	0.138	0.127	0.115	0.103	0.090

where $\Delta = (d-2)/(d-1)$

Better choices for p and α? I

- Initially, I guessed the case $p \to \infty$ is better

$$\hat{\theta}^{+}_{i\text{MAX}} = \max\left(0, 1 - \frac{c}{\{\max_j |x_j|\}^{2-\alpha} |x_i|^{\alpha}}\right) x_i$$

- Larry said no. Smaller p should be better for sparsity
  Too sensitive to $\max |x_j|$
  $\hat{\theta}^{+}_{i\text{MAX}} = 0$ if

$$\log |x_i| < -\frac{2-\alpha}{\alpha} \log(\max_j |x_j|) + \frac{\log c}{\alpha}$$

  Sparsification is hopeless if $\max |x_j|$ is relatively large

BETTER CHOICES FOR p AND α? I

- Initially, I guessed the case $p \to \infty$ is better

$$\hat{\theta}^{+}_{i\mathrm{MAX}} = \max\left(0, 1 - \frac{c}{\{\max_j |x_j|\}^{2-\alpha}|x_i|^{\alpha}}\right) x_i$$

- Larry said no. Smaller p should be better for sparsity
  Too sensitive to $\max |x_j|$

  $\hat{\theta}^{+}_{i\mathrm{MAX}} = 0$ if

$$\log |x_i| < -\frac{2-\alpha}{\alpha} \log(\max_j |x_j|) + \frac{\log c}{\alpha}$$

  Sparsification is hopeless if $\max |x_j|$ is relatively large

better choices for p and α? I

- Initially, I guessed the case $p \to \infty$ is better

$$\hat{\theta}^{+}_{i\mathrm{MAX}} = \max\left(0, 1 - \frac{c}{\{\max_j |x_j|\}^{2-\alpha}|x_i|^{\alpha}}\right) x_i$$

- Larry said no. Smaller p should be better for sparsity
  Too sensitive to $\max|x_j|$

$\hat{\theta}^{+}_{i\mathrm{MAX}} = 0$ if

$$\log|x_i| < -\frac{2-\alpha}{\alpha}\log(\max_j |x_j|) + \frac{\log c}{\alpha}$$

Sparsification is hopeless if $\max|x_j|$ is relatively large

better choices for p and α? II

- At that time I just suggested the case $p \to 0$ to him
- Actually, I got some theoretical properties of the estimator with $p \to 0$ several months ago
- I really wanted to explain them to Larry,,,,,
- Today I would like to share the results with you

better choices for p and α? II

- At that time I just suggested the case $p \to 0$ to him
- Actually, I got some theoretical properties of the estimator with $p \to 0$ several months ago
- I really wanted to explain them to Larry,,,,,
- Today I would like to share the results with you

better choices for p and α? II

- At that time I just suggested the case $p \to 0$ to him
- Actually, I got some theoretical properties of the estimator with $p \to 0$ several months ago
- I really wanted to explain them to Larry,, ,, ,
- Today I would like to share the results with you

better choices for p and α? II

- At that time I just suggested the case $p \to 0$ to him
- Actually, I got some theoretical properties of the estimator with $p \to 0$ several months ago
- I really wanted to explain them to Larry,, ,, ,
- Today I would like to share the results with you

geometric-mean-based James-Stein I

- the relationship among $\|x\|_p$, the generalized mean of $|x_1|, \dots, |x_d|$ and the geometric mean ($p \to 0$)

$$\left(\frac{1}{d}\sum_{i=1}^{d}|x_i|^p\right)^{1/p} = d^{-1/p}\|x\|_p \to \prod |x_j|^{1/d} \text{ as } p \to 0$$

- geometric-mean-based James-Stein estimator

$$\hat{\theta}^{+}_{i\mathrm{GM}} = \max\left(0, 1 - \frac{c}{(\{\prod |x_j|\}^{1/d})^{2-\alpha}|x_i|^{\alpha}}\right)x_i$$

geometric-mean-based James-Stein II

- geometric-mean-based James-Stein estimator

$$\hat{\theta}^{+}_{i\mathrm{GM}} = \max\left(0, 1 - \frac{c}{(\{\prod |x_j|\}^{1/d})^{2-\alpha}|x_i|^{\alpha}}\right) x_i$$

- not ordinary minimax for any $c > 0$
- ensemble minimax, $\sup_\tau \bar{R}(\hat{\boldsymbol{\theta}}_{\mathrm{GM}}, \tau) \leq d$, if $d \geq 4$ and

$$c \in \left[0, 4\frac{\Gamma\left(\frac{3}{2} - \frac{\alpha}{2} + \frac{\alpha - 2}{2d}\right)}{\Gamma\left(\frac{3}{2} - \alpha + \frac{\alpha - 2}{d}\right)}\left\{\frac{\Gamma\left(\frac{1}{2} + \frac{\alpha - 2}{2d}\right)}{\Gamma\left(\frac{1}{2} + \frac{\alpha - 2}{d}\right)}\right\}^{d-1}\right]$$

Geometric-mean-based James-Stein III

- $(p \to 0)$ the region of sparsification, $\hat{\theta}^{+}_{i\mathrm{GM}} = 0$ if

$$1 - \frac{c}{(\{\prod |x_j|\}^{1/d})^{2-\alpha} |x_i|^{\alpha}} < 0$$

$$\Leftrightarrow \log |x_i| < \frac{d}{d\alpha + 2 - \alpha} \log c - \frac{2-\alpha}{d\alpha + 2 - \alpha} \sum_{j \neq i} \log |x_j|$$

not sensitive to $\max_j |x_j| \Uparrow$

- $(p \to \infty)$ the region of sparsification, $\hat{\theta}^{+}_{i\mathrm{MAX}} = 0$ if

$$\log |x_i| < \frac{\log c}{\alpha} - \frac{2-\alpha}{\alpha} \log(\max_j |x_j|)$$

geometric-mean-based James-Stein IV

- Our recommendation

$$\hat{\theta}^{+}_{i\mathrm{GM}} = \max\left(0, 1 - \frac{c_d(\alpha)}{(\{\prod |x_j|\}^{1/d})^{2-\alpha}|x_i|^{\alpha}}\right) x_i$$

with $c_d(\alpha)$ the upper bound for ensemble minimaxity,

$$c_d(\alpha) = 4\frac{\Gamma\left(\frac{3}{2} - \frac{\alpha}{2} + \frac{\alpha-2}{2d}\right)}{\Gamma\left(\frac{3}{2} - \alpha + \frac{\alpha-2}{d}\right)}\left\{\frac{\Gamma\left(\frac{1}{2} + \frac{\alpha-2}{2d}\right)}{\Gamma\left(\frac{1}{2} + \frac{\alpha-2}{d}\right)}\right\}^{d-1}$$

- The better choice of α is still an open problem

SUMMARY

Estimation of a mean vector $\boldsymbol{X} \sim N_d(\boldsymbol{\theta}, \boldsymbol{\Sigma})$

1. $\boldsymbol{\Sigma} = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_d^2)$, with $\sigma_1^2 > \sigma_2^2 > \cdots > \sigma_d^2$
2. $\boldsymbol{\Sigma} = \boldsymbol{I}_d$

- Ensemble minimaxity of some James-Stein variants under loss $L(\boldsymbol{\delta}, \boldsymbol{\theta}) = \|\boldsymbol{\delta} - \boldsymbol{\theta}\|^2 = \sum_{i=1}^{d} (\delta_i - \theta_i)^2$

$$\left(\boldsymbol{I} - \boldsymbol{\Sigma}\frac{d-2}{(d-2)\sigma_1^2 + \|\boldsymbol{x}\|^2}\right)\boldsymbol{x} \text{ for case 1}$$

$$\hat{\theta}_{i\mathrm{GM}}^{+} = \max\left(0, 1 - \frac{c_d(\alpha)}{(\{\prod |x_j|\}^{1/d})^{2-\alpha}|x_i|^{\alpha}}\right) x_i \text{ for case 2}$$

Thank you!