Matrix superharmonic priors for Bayes estimation under matrix quadratic loss

Takeru Matsuda (RIKEN Center for Brain Science)

William E. Strawderman (Rutgers University)

Abstract

Stein (1974)

When $X \sim \mathrm{N}_n(\mu, I_n)$ $(n \geq 3)$, Bayes estimator with a superharmonic prior $\pi(\mu)$ is minimax under quadratic loss:

$$\Delta\pi := \sum_{a=1}^{n} \frac{\partial^2 \pi}{\partial \mu_a^2} \leq 0 \quad \Rightarrow \quad \mathrm{E}\|\hat{\mu}^{\pi}(x) - \mu\|^2 \leq n$$

This study (M. and Strawderman, *Biometrika* 2021+)

When $X \sim \mathrm{N}_{n,p}(M, I_n, I_p)$ $(n \geq p+2)$, Bayes estimator with a matrix superharmonic prior is minimax under matrix quadratic loss:

$$\widetilde{\Delta}\pi := \left(\sum_{a=1}^{n} \frac{\partial^2 \pi}{\partial M_{ai} \partial M_{aj}} \right)_{ij} \preceq O$$

$$\Rightarrow \quad \mathrm{E}(\hat{M}^{\pi}(X) - M)^{\top}(\hat{M}^{\pi}(X) - M) \preceq n I_p$$

Stein's paradox

$$X \sim \mathrm{N}_n(\mu, I_n)$$

- estimate μ based on X under quadratic loss $\|\hat{\mu} - \mu\|^2$
- Maximum likelihood estimator $\hat{\mu}_{\mathrm{MLE}}(x) = x$ is minimax.

Theorem (Stein, 1956)

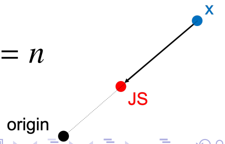
When $n \geq 3$, $\hat{\mu}_{\mathrm{MLE}}(x) = x$ is inadmissible.

- Shrinkage estimators dominate $\hat{\mu}_{\mathrm{MLE}}$.
- e.g. James–Stein estimator (James and Stein, 1961)

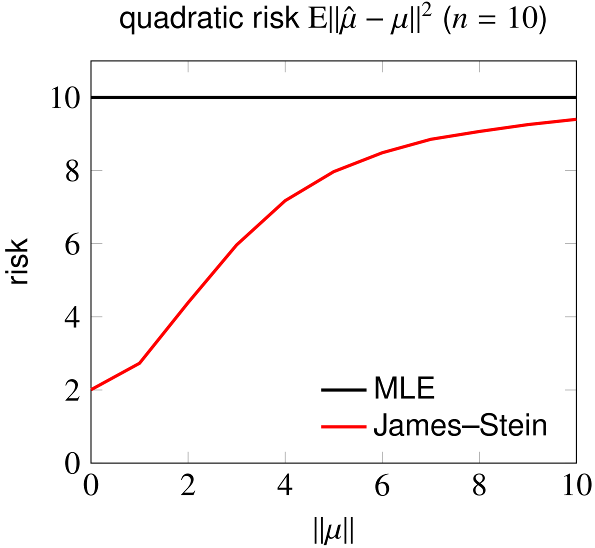
$$\hat{\mu}_{\mathrm{JS}}(x) = \left(1 - \frac{n-2}{\|x\|^2}\right) x$$

$$\mathrm{E}\|\hat{\mu}_{\mathrm{JS}}(x) - \mu\|^2 \leq \mathrm{E}\|\hat{\mu}_{\mathrm{MLE}}(x) - \mu\|^2 = n$$

- JS shrinks x toward the origin.



Risk comparison



- JS attains large risk reduction when μ is close to the origin

superharmonic prior ⇒ minimax

- Bayes estimator of μ with prior $\pi(\mu)$ (posterior mean)

$$\hat{\mu}^{\pi}(x) = \mathrm{E}_{\pi}[\mu \mid x] = \int \mu \pi(\mu \mid x) \mathrm{d}\mu$$

- superharmonic prior

$$\Delta \pi(\mu) = \sum_{a=1}^{n} \frac{\partial^2}{\partial \mu_a^2} \pi(\mu) \leq 0$$

Theorem (Stein, 1974)

The Bayes estimator with a superharmonic prior is minimax.

- e.g. Stein's prior ($n \geq 3$)

$$\pi_{\mathrm{S}}(\mu) = \|\mu\|^{2-n}$$

- Bayes estimator with π_{S} shrinks toward the origin like JS.

Shrinkage estimation of normal mean matrix

$$X \sim \mathrm{N}_{n,p}(M, I_n, I_p) \quad (X_{ai} \sim \mathrm{N}(M_{ai}, 1))$$

- estimate M based on X under Frobenius loss

$$\|\hat{M} - M\|_{\mathrm{F}}^2 = \sum_{a=1}^{n} \sum_{i=1}^{p} (\hat{M}_{ai} - M_{ai})^2$$

- Efron–Morris estimator (= James–Stein estimator when $p = 1$)

$$\hat{M}_{\mathrm{EM}}(X) = X \left(I_p - (n-p-1)(X^\top X)^{-1} \right)$$

Theorem (Efron and Morris, 1972)

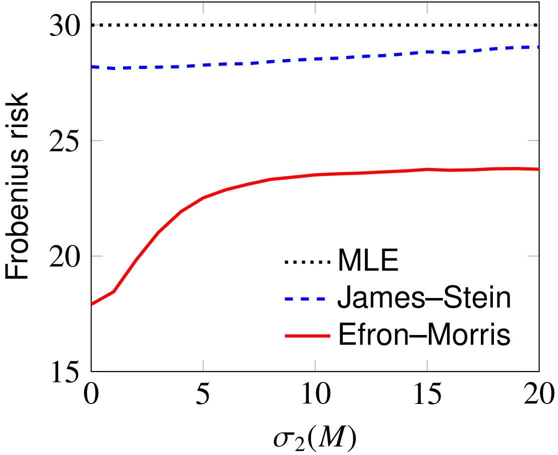
When $n \geq p+2$, $\hat{M}_{\mathrm{EM}}$ is minimax and dominates $\hat{M}_{\mathrm{MLE}}(X) = X$.

- Stein (1974): $\hat{M}_{\mathrm{EM}}$ shrinks singular values separately.

$$\sigma_i(\hat{M}_{\mathrm{EM}}) = \left(1 - \frac{n-p-1}{\sigma_i(X)^2} \right) \sigma_i(X)$$

Risk function (rank 2)

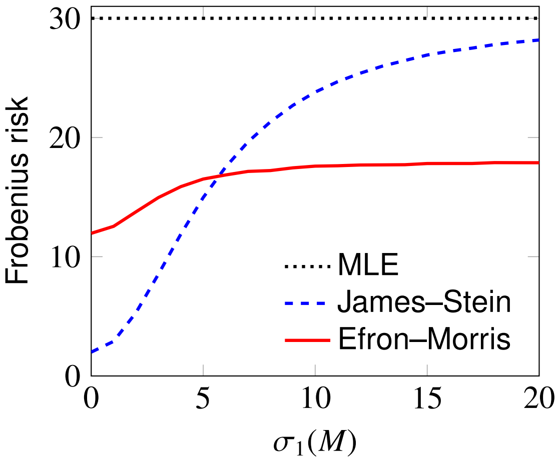
- $n = 10$, $p = 3$, $\sigma_1(M) = 20$, $\sigma_3(M) = 0$



- $\hat{M}_{\mathrm{EM}}$ works well when $\sigma_2(M)$ is small, even if $\sigma_1(M)$ is large.
  - $\hat{M}_{\mathrm{JS}}$ works well if $\|M\|_{\mathrm{F}}^2 = \sigma_1(M)^2 + \sigma_2(M)^2 + \sigma_3(M)^2$ is small.

Risk function (rank 1)

- $n = 10$, $p = 3$, $\sigma_2(M) = \sigma_3(M) = 0$



- $\hat{M}_{\rm EM}$ has constant risk reduction even if $\sigma_1(M)$ is large.
- Therefore, $\hat{M}_{\rm EM}$ works well when M has low rank.

Singular value shrinkage prior (M. and Komaki, 2015)

$$\pi_{\mathrm{SVS}}(M) = \det(M^\top M)^{-(n-p-1)/2} = \prod_{i=1}^{p} \sigma_i(M)^{-(n-p-1)}$$

- puts more weight on matrices with smaller singular values
  $\rightarrow$ shrinks singular values separately
- When $p = 1$, π_{SVS} coincides with Stein's prior $\pi_{\mathrm{S}}(\mu) = \|\mu\|^{2-n}$.

Theorem (M. and Komaki, *Biometrika* 2015)

When $n \geq p + 2$, π_{SVS} is superharmonic:

$$\Delta \pi_{\mathrm{SVS}} = \sum_{a=1}^{n} \sum_{i=1}^{p} \frac{\partial^2 \pi_{\mathrm{SVS}}}{\partial M_{ai}^2} \leq 0.$$

- Bayes estimator with π_{SVS} is minimax under Frobenius loss.
  - similar behavior to EM
  - works well when M has (approximately) low rank

Summary (so far)

vector	matrix
James–Stein estimator $\hat{\mu}_{\mathrm{JS}} = \left(1 - \frac{n-2}{\|x\|^2}\right) x$	Efron–Morris estimator $\hat{M}_{\mathrm{EM}} = X\left(I_p - (n-p-1)(X^\top X)^{-1}\right)$
Stein's prior $\pi_{\mathrm{S}}(\mu) = \|\mu\|^{-(n-2)}$	singular value shrinkage prior $\pi_{\mathrm{SVS}}(M) = \det(M^\top M)^{-(n-p-1)/2}$

- note: JS and EM are not (generalized) Bayes estimators.

Estimation under matrix quadratic loss

$$X \sim \mathrm{N}_{n,p}(M, I_n, I_p) \quad (X_{ai} \sim \mathrm{N}(M_{ai}, 1))$$

- estimate M based on X under matrix quadratic loss

$$L(M, \hat{M}) = (\hat{M} - M)^\top (\hat{M} - M) \in \mathbb{R}^{p \times p}$$

- risk function

$$R(M, \hat{M}) = \mathrm{E}_M[L(M, \hat{M}(X))] \in \mathbb{R}^{p \times p}$$

- We compare $R(M, \hat{M})$ in the Löwner order $\preceq$
  - $A \preceq B \Leftrightarrow B - A$ is positive semidefinite
  - If $R(M, \hat{M}_1) \preceq R(M, \hat{M}_2)$, then $\mathrm{E}_M \|\hat{M}_1 c - Mc\|^2 \leq \mathrm{E}_M \|\hat{M}_2 c - Mc\|^2$ for every c
- cf. multivariate linear regression

Unbiased risk estimate & minimaxity of EM

- matrix divergence

$$(\widetilde{\mathrm{div}}\ g(X))_{ij} = \sum_{a=1}^{n} \frac{\partial}{\partial X_{ai}} g_{aj}(X)$$

Theorem

The matrix quadratic risk of $\hat{M} = X + g(X)$ is given by

$$R(M, \hat{M}) = nI_p + \mathrm{E}_M[\widetilde{\mathrm{div}}\ g(X) + (\widetilde{\mathrm{div}}\ g(X))^\top + g(X)^\top g(X)]$$

Theorem

When $n - p - 1 > 0$, the Efron–Morris estimator is minimax under the matrix quadratic loss:

$$R(M, \hat{M}_{\mathrm{EM}}) = nI_p - (n - p - 1)^2 \mathrm{E}_M[(X^\top X)^{-1}] \leq nI_p$$

Matrix superharmonicity

- “sphere” with center $X \in \mathbb{R}^{n \times p}$ and “radius” $\rho \in \mathbb{R}^p$

$$S_{X,\rho} = \{X + e\rho^\top \mid e \in \mathbb{R}^n, \|e\| = 1\}$$

- average value of f on $S_{X,\rho}$

$$L(f : X, \rho) = \frac{1}{\Omega_n} \int_{S_{0,1}} f(X + e\rho^\top) \mathrm{d}s(e)$$

Definition

An extended real-valued function $f : \mathbb{R}^{n \times p} \to \mathbb{R} \cup \{\infty\}$ is matrix superharmonic if

1. f is lower semicontinuous
2. $f \not\equiv \infty$
3. $L(f : X, \rho) \leq f(X)$ for every $X \in \mathbb{R}^{n \times p}$ and $\rho \in \mathbb{R}^p$

Matrix superharmonic ⇒ superharmonic

Lemma

If a function $f : \mathbb{R}^{n\times p} \to \mathbb{R} \cup \{\infty\}$ is matrix superharmonic, then $f \circ \mathrm{vec}^{-1}$ is superharmonic.

- Proof: For every $X \in \mathbb{R}^{n\times p}$ and $r > 0$,

$$L(f \circ \mathrm{vec}^{-1} : \mathrm{vec}(X), r) = \frac{1}{\Omega_p r^{p-1}} \int_{S_{0,r}} L(f : X, \rho) \mathrm{d}s(\rho) \leq f(X)$$

- The converse does not hold when $p \geq 2$.
  - e.g. $f(X) = \|X\|_{\mathrm{F}}^{2-np}$

Characterization by matrix Laplacian

- Matrix superharmonicity is characterized by a matrix version of the Laplacian.

Definition

For a C^2 function $f : \mathbb{R}^{n\times p} \to \mathbb{R}$, its matrix Laplacian $\widetilde{\Delta} f : \mathbb{R}^{n\times p} \to \mathbb{R}^{p\times p}$ is defined as

$$(\widetilde{\Delta} f(X))_{ij} = \sum_{a=1}^{n} \frac{\partial^2}{\partial X_{ai} \partial X_{aj}} f(X)$$

Theorem

A C^2 function $f : \mathbb{R}^{n\times p} \to \mathbb{R}$ is matrix superharmonic if and only if its matrix Laplacian is negative semidefinite $\widetilde{\Delta} f(X) \preceq O$ for every X.

- Proof: Green's theorem

matrix superharmonic prior ⇒ minimax

$$\hat{M}^{\pi}(X) = \mathrm{E}_{\pi}[M \mid X] = X + \widetilde{\nabla} \log m_{\pi}(X)$$

Theorem

If $\sqrt{m_{\pi}(X)}$ is matrix superharmonic, then $\hat{M}^{\pi}$ is minimax under the matrix quadratic loss.

- Proof: by using the unbiased estimate of risk,

$$R(M, \hat{M}^{\pi}) = nI_p + 4\mathrm{E}_M \left[\frac{\widetilde{\Delta} \sqrt{m_{\pi}(X)}}{\sqrt{m_{\pi}(X)}} \right]$$

Theorem

If $\pi(M)$ is matrix superharmonic, then $\sqrt{m_{\pi}(X)}$ is also matrix superharmonic and $\hat{M}^{\pi}$ is minimax under the matrix quadratic loss.

- When $p = 1$, it reduces to the classical result by Stein (1974).

A class of matrix superharmonic priors

- improper matrix t-prior

$$\pi_{\alpha,\beta}(M) = \det(M^\top M + \beta I_p)^{-(\alpha+n+p-1)/2}$$

Theorem

If $-n - p + 1 \leq \alpha \leq -2p$ and $\beta \geq 0$, then $\pi_{\alpha,\beta}(M)$ is matrix superharmonic and the generalized Bayes estimator with respect to $\pi_{\alpha,\beta}(M)$ is minimax under the matrix quadratic loss.

- When $p = 1$, it reduces to the result by Faith (1993) on (improper) multivariate t-priors.

Matrix superharmonicity of π_{SVS}

- When $\alpha = -2p$ and $\beta = 0$, the prior $\pi_{\alpha,\beta}(M)$ coincides with the singular value shrinkage prior

$$\pi_{\mathrm{SVS}}(M) = \det(M^\top M)^{-(n-p-1)/2}$$

Corollary

When $n - p - 1 > 0$, $\pi_{\mathrm{SVS}}(M)$ is matrix superharmonic and the generalized Bayes estimator with respect to π_{SVS} is minimax under the matrix quadratic loss.

- The matrix superharmonicity of π_{SVS} is strongly concentrated on the space of low rank matrices.

Corollary

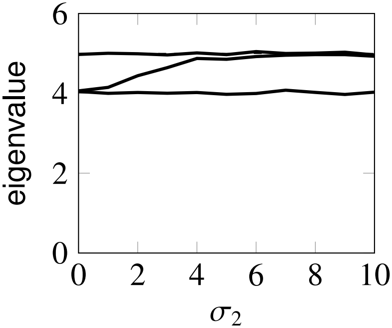
If M has full-rank, then $\widetilde{\Delta}\pi_{\mathrm{SVS}}(M) = O$.

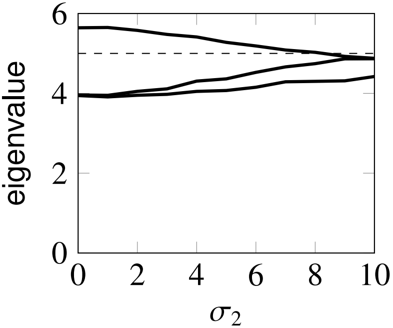
Simulation setting

- We denote the i-th singular value of M by σ_i.
  - $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_p$
- We focus on the eigenvalues $\lambda_1 \geq \cdots \geq \lambda_p$ of the matrix quadratic risk $R(M, \hat{M})$.
  - Since $R(M, \hat{M}) = nI_p$ for MLE $\hat{M} = X$, an estimator is minimax if and only if $\lambda_1 \leq n$ for every M.
- Bayes estimator with $\pi_{\mathrm{SVS}}(M) = \det(M^\top M)^{-(n-p-1)/2}$
- Bayes estimator with Stein's prior $\pi_{\mathrm{S}}(M) = \|M\|_{\mathrm{F}}^{2-np}$
- Efron–Morris estimator $\hat{M}_{\mathrm{EM}} = X(I - (n-p-1)(X^\top X)^{-1})$
  - almost the same risk with Bayes estimators with π_{SVS}
- James–Stein estimator $\hat{M}_{\mathrm{JS}} = (1 - (np-2)/\|X\|_{\mathrm{F}}^2)X$
  - almost the same risk with Bayes estimators with π_{S}

Simulation results (Figure 1)

- eigenvalues of Bayes estimators ($n = 5$, $p = 3$, $\sigma_1 = 10$, $\sigma_3 = 0$)
- left: π_{SVS}, right: π_{S}

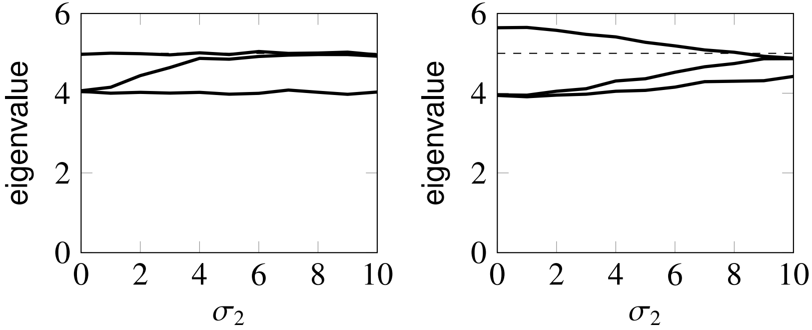




- For π_{SVS}, all eigenvalues do not exceed $n = 5$, which indicates the minimaxity.
  - λ_1 and λ_3 are almost constant with values $\lambda_1 \approx 5$ and $\lambda_3 \approx 4$.
  - λ_2 increases from 4 to 5 with σ_2.
  - These are understood from the fact that π_{SVS} shrinks each singular value separately.

Simulation results (Figure 1)

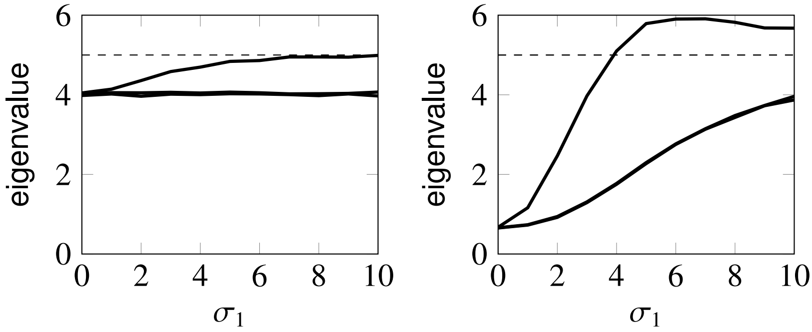
- eigenvalues of Bayes estimators ($n = 5$, $p = 3$, $\sigma_1 = 10$, $\sigma_3 = 0$)
- left: π_{SVS}, right: π_{S}



- For π_{S}, $\lambda_1 \geq n = 5$ when $\sigma_2 \leq 8 \rightarrow$ not minimax.
  - However, since this estimatot is minimax under the Frobenius loss, $\lambda_1 + \lambda_2 + \lambda_3 \leq np = 15$.
  - cf. James–Stein estimator is not minimax componentwise, even though it is minimax under the quadratic loss for the whole vector (Lehmann and Casella, 2006).

Simulation results (Figure 2)

- eigenvalues ($n = 5$, $p = 3$, $\sigma_2 = \sigma_3 = 0$)
- left: π_{SVS}, right: π_{S}

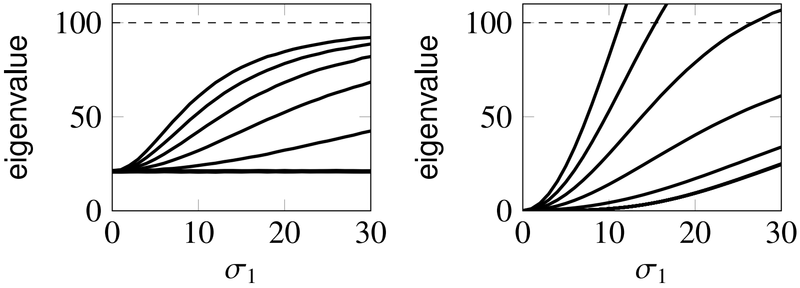


- For π_{SVS}, both λ_2 and λ_3 are almost constant around 4

  $\rightarrow \pi_{\mathrm{SVS}}$ works particularly well when M has low rank

Simulation results (Figure 3)

- eigenvalues ($n = 100$, $p = 20$, $\sigma_i = (6-i)/5 \cdot \sigma_1$ ($i = 2, \ldots, 5$), $\sigma_6 = \cdots = \sigma_{20} = 0$)
- left: $\hat{M}_{\mathrm{EM}}$. right: $\hat{M}_{\mathrm{JS}}$



- The advantage of $\hat{M}_{\mathrm{EM}}$ to the low-rank setting is more pronounced in higher dimensions.
  - $\lambda_6 \approx \cdots \approx \lambda_{20} \approx 20$

Summary

$$X \sim \mathrm{N}_{n,p}(M, I_n, I_p)$$

- The Bayes estimator with a matrix superharmonic prior is minimax under matrix quadratic loss:

$$\widetilde{\Delta}\pi := \left(\sum_{a=1}^{n} \frac{\partial^2 \pi}{\partial M_{ai} \partial M_{aj}} \right)_{ij} \preceq O$$

$$\Rightarrow \quad \mathrm{E}(\hat{M}^{\pi}(X) - M)^{\top}(\hat{M}^{\pi}(X) - M) \preceq nI_p$$

- The matrix t-prior

$$\pi_{\alpha,\beta}(M) = \det(M^{\top} M + \beta I_p)^{-(\alpha+n+p-1)/2}$$

  is matrix superharmonic when $-n - p + 1 \leq \alpha \leq -2p$ and $\beta \geq 0$.
- Matrix superharmonic priors work well for low-rank matrices.
- paper: M. and Strawderman, *Biometrika* 2021+