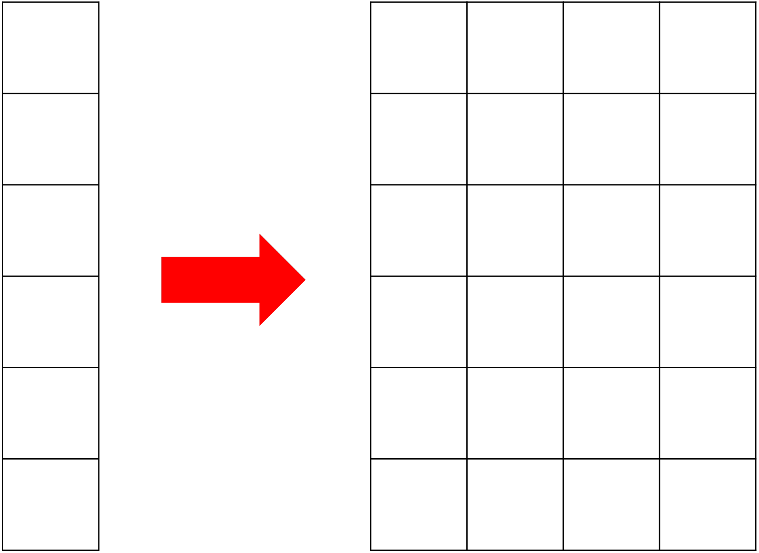
# Matrix estimation via singular value shrinkage

Takeru Matsuda

University of Tokyo & RIKEN Center for Brain Science

June 2, 2025 @ SNAB

# Shrinkage estimation: from vector to matrix



# Stein's paradox

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

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

**Theorem (Stein, 1956)**

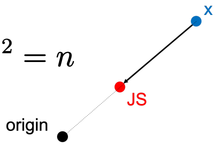
When $n \geq 3$, $\hat{\mu}_{\mathrm{MLE}}(x) = x$ is inadmissible.

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

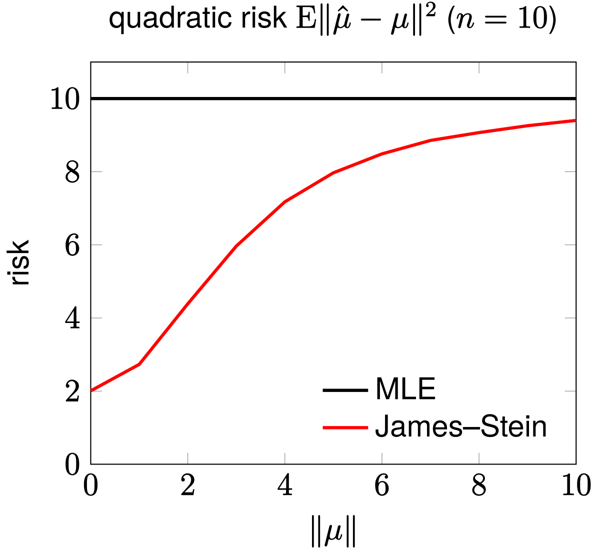
$$\hat{\mu}_{\mathrm{JS}}(x) = \left(1 - \frac{n-2}{\|x\|^2}\right) x$$

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

- JS shrinks $x$ toward the origin.



# Risk function ($n = 10$)




- JS attains large risk reduction when $\mu$ is close to the origin

## superharmonic prior $\Rightarrow$ minimax

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

$$\hat{\mu}^{\pi}(x) = \int \mu \pi(\mu \mid x) \mathrm{d}\mu = \frac{\int \mu p(x \mid \mu) \pi(\mu) \mathrm{d}\mu}{\int p(x \mid \mu) \pi(\mu) \mathrm{d}\mu}$$

- 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 $\pi_{\mathrm{S}}$ shrinks toward the origin like JS.

# Estimation of normal mean matrix

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

- estimate $M$ based on $X$ under Frobenius loss

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

- 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)**

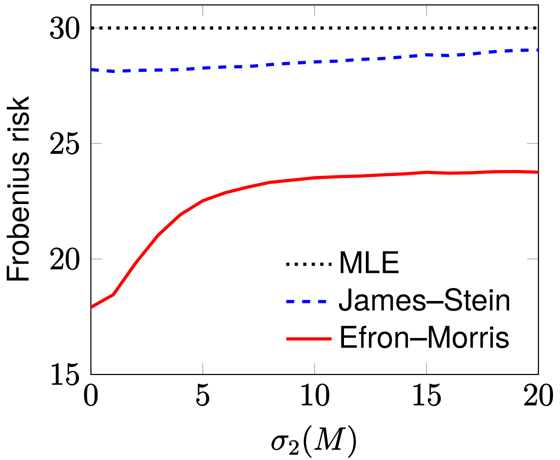
When $n \geq p+2$, $\hat{M}_{\mathrm{EM}}$ is minimax and dominates $\hat{M}_{\mathrm{MLE}}(X) = X$.

- 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)

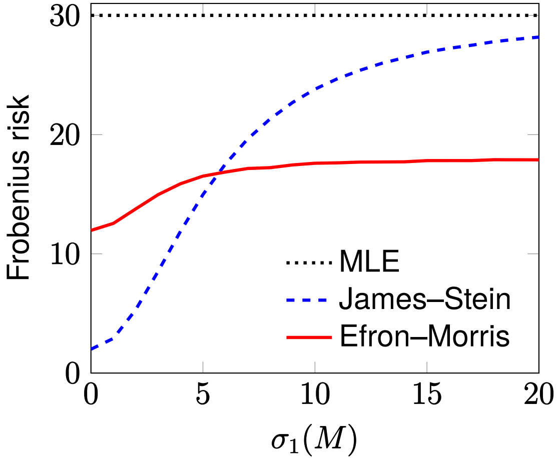
- $n = 10$, $p = 3$, $\sigma_1(M) = 20$, $\sigma_3(M) = 0$




- $\hat{M}_{\mathrm{EM}}$ 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)

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




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

# Bayesian counterpart of Efron–Morris?

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

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

# Singular value shrinkage prior

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

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

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

When $n \geq p + 2$, $\pi_{\mathrm{SVS}}$ 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 $\pi_{\mathrm{SVS}}$ is minimax under Frobenius loss.
  - similar behavior to EM
  - works well when $M$ has (approximately) low rank

# Superharmonicity at low rank matrices

- Previously proposed superharmonic priors mainly shrink to simple subsets (e.g. point, linear subspace).
- In contrast, $\pi_{\mathrm{SVS}}$ shrinks to the set of low rank matrices, which is nonlinear and nonconvex.

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

$\Delta \pi_{\mathrm{SVS}}(M) = 0$ if $M$ has full rank.

- Therefore, superharmonicity of $\pi_{\mathrm{SVS}}$ is strongly concentrated in the same way as the Laplacian of Stein's prior becomes a Dirac delta function.

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

# 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

# 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:

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

# Matrix superharmonic prior ⇒ minimax

## 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:

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

## M. and Strawderman (*Biometrika* 2022)

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:

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

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

# 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.

- By taking $\alpha = -2p$ and $\beta = 0$,

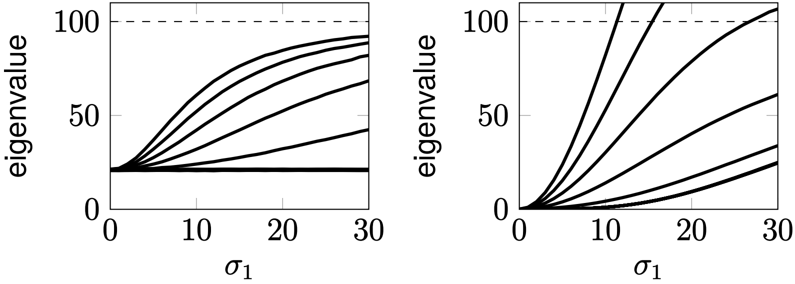
$$\pi_{\mathrm{SVS}}(M) = \det(M^\top M)^{-(n-p-1)/2}$$

**Corollary**

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

# Simulation

- 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: $\hat{M}_{\mathrm{JS}}$



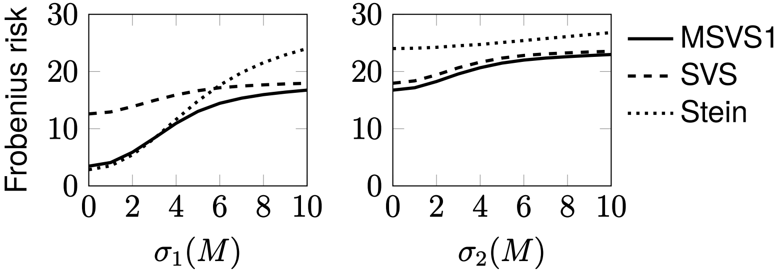

- The advantage of $\hat{M}_{\mathrm{EM}}$ to the low-rank setting is more pronounced in higher dimensions.
  - $\lambda_6 \approx \cdots \approx \lambda_{20} \approx 20$

## Double shrinkage prior (M., Komaki and Strawderman, 2024)

- Bayes estimator with $\pi_{\mathrm{SVS}}$ is inadmissible
- improved by additional scalar / column-wise shrinkage

$$\pi_{\mathrm{MSVS1}}(M) = \pi_{\mathrm{SVS}}(M) \|M\|_{\mathrm{F}}^{-\gamma}$$

$$\pi_{\mathrm{MSVS2}}(M) = \pi_{\mathrm{SVS}}(M) \prod_{i=1}^{p} \|M_{\cdot i}\|^{-\gamma_i}$$



# Empirical Bayes matrix completion (M. and Komaki, 2019)

| | movie 1 | movie 2 | movie 3 | movie 4 |
|---|---|---|---|---|
| user 1 | 4 | 7 | ? | 2 |
| user 2 | 6 | ? | 3 | 8 |
| user 3 | ? | 1 | 9 | ? |
| user 4 | 4 | 5 | ? | 3 |

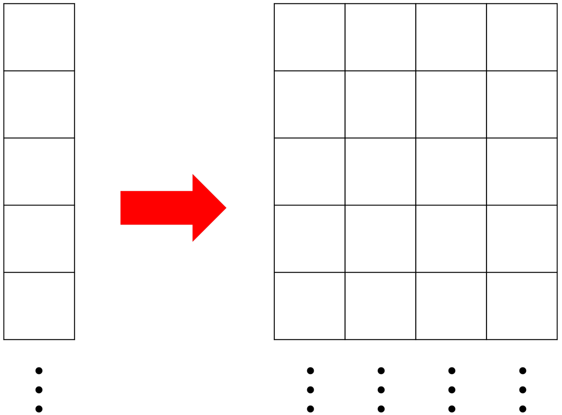
- extension of Efron–Morris estimator to missing data
  - estimation of $M$ from partially observed $X$

$$M \sim \mathrm{N}_{n,p}(0, I_n, \Sigma)$$

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

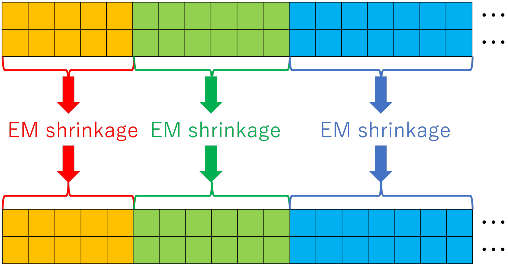
- Application: protein data, movie data

# Application to nonparametric estimation (M., 2024)



- The blockwise Efron–Morris estimator is adaptive minimax over the multivariate Sobolev ellipsoids.
  - extension of Efromovich and Pinsker (1984)
  - adaptation to smoothness, scale and arbitrary quadratic loss

# Application to nonparametric estimation (M., 2024)




**Theorem (M., *IEEE IT* 2024)**

The blockwise Efron–Morris estimator $\hat{\theta}_{\mathrm{BEM}}$ is adaptive minimax over the multivariate Sobolev ellipsoids:

$$\sup_{\theta \in \Theta(\beta, Q)} R_Q(\theta, \hat{\theta}_{\mathrm{BEM}}) \sim \inf_{\hat{\theta}} \sup_{\theta \in \Theta(\beta, Q)} R_Q(\theta, \hat{\theta}) \sim P(\beta, Q) \varepsilon^{4\beta/(2\beta+1)}$$

for every $\beta$ and $Q$.

# Future work

- How about tensors ??

$$X = (X_{ijk})$$

- For tensors, even the definition of rank or singular values is not clear..

- Hopefully, some empirical Bayes method provides a natural shrinkage for tensors.