Singular value shrinkage prior: a matrix version of Stein's prior

Takeru Matsuda

The University of Tokyo

June 19, 2019
Symposium in memory of Charles Stein

Motivation

vector	James–Stein estimator (1961)	Stein's prior (1974)
matrix	Efron–Morris estimator (1972)	?

Stein's 1974 paper

- "Estimation of the mean of a multivariate normal distribution"

- 1. Introduction
- 2. Computation of the risk of an arbitrary estimate of the mean
- 3. The spherically symmetric case
- 4. The risk of an estimate of a matrix of means
- 5. Choice of an estimate in the $p \times p$ case
- 6. Directions in which this work ought to be extended

Abstract

Efron–Morris estimator (Efron and Morris, 1972)

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

minimax estimator of a normal mean matrix
natural extension of the James–Stein estimator

$\downarrow$

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

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

superharmonic ($\Delta \pi_{\mathrm{SVS}} \leq 0$), natural generalization of the Stein prior
works well for low-rank matrices $\rightarrow$ reduced-rank regression

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

estimate unobserved entries of a matrix by exploting low-rankness

Efron–Morris estimator (Efron and Morris, 1972)

Note: singular values of matrices

- Singular value decomposition of $p \times q$ matrix M $(p \geq q)$

$$M = U \Lambda V^\top$$

$$U : p \times q, \quad V : q \times q, \quad U^\top U = V^\top V = I_q$$

$$\Lambda = \mathrm{diag}(\sigma_1(M), \ldots, \sigma_q(M))$$

- $\sigma_1(M) \geq \cdots \geq \sigma_q(M) \geq 0$: singular values of M
- $\mathrm{rank}(M) = \#\{i \mid \sigma_i(M) > 0\}$

Estimation of normal mean matrix

$$X_{ij} \sim \mathrm{N}(M_{ij}, 1) \quad (i = 1, \cdots, p; j = 1, \cdots, q)$$

- estimate M based on X under Frobenius loss $\|\hat{M} - M\|_{\mathrm{F}}^2$
- Efron–Morris estimator (= James–Stein estimator when $q = 1$)

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

Theorem (Efron and Morris, 1972)

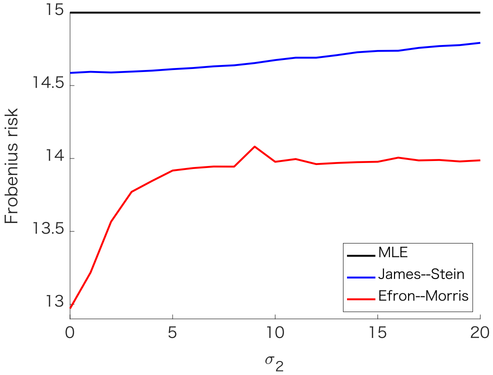
When $p \geq q + 2$, $\hat{M}_{\mathrm{EM}}$ is minimax and dominates $\hat{M}_{\mathrm{MLE}}(X) = X$.

- Stein (1974) noticed that it shrinks the singular values of the observation to zero.

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

Numerical results

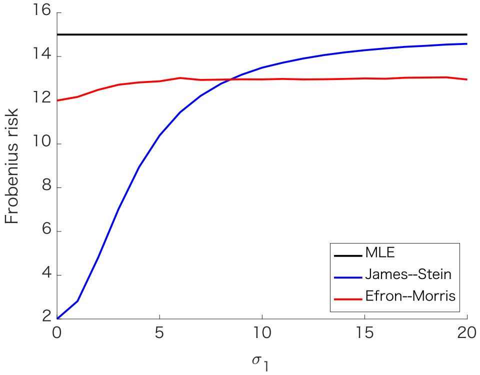
- Risk functions for $p = 5$, $q = 3$, $\sigma_1 = 20$, $\sigma_3 = 0$ (rank 2)
- black: MLE, blue: JS, red: EM



- $\hat{M}_{\mathrm{EM}}$ works well when σ_2 is small, even if σ_1 is large.
  - $\hat{M}_{\mathrm{JS}}$ works well when $\|M\|_F^2 = \sigma_1^2 + \sigma_2^2 + \sigma_3^2$ is small.

Numerical results

- Risk functions for $p = 5$, $q = 3$, $\sigma_2 = \sigma_3 = 0$ (rank 1)
- black: MLE, blue: JS, red: EM



- $\hat{M}_{\mathrm{EM}}$ has constant risk reduction as long as $\sigma_2 = \sigma_3 = 0$, because it shrinks singular values for each.
- Therefore, it works well when M has low rank.

Remark: SURE for matrix mean

- orthogonally invariant estimator

$$X = U\Sigma V^\top, \quad \hat{M} = U\Sigma(I_q - \Phi(\Sigma))V^\top$$

- Stein (1974) derived an unbiased estimate of risk (SURE):

$$pq + \sum_{i=1}^{q} \left\{ \sigma_i^2 \phi_i^2 - 2(p-q+1)\phi_i - 2\sigma_i \frac{\partial \phi_i}{\partial \sigma_i} \right\} - 4 \sum_{i<j} \frac{\sigma_i^2 \phi_i - \sigma_j^2 \phi_j}{\sigma_i^2 - \sigma_j^2}$$

  - regularity conditions → M. and Strawderman (2018)

- SURE is also improved by singular value shrinkage (M. and Strawderman, 2018)
  - extension of Johnstone (1988)

Singular value shrinkage prior (Matsuda and Komaki, 2015)

Superharmonic prior for estimation

$$X \sim \mathrm{N}_p(\mu, I_p)$$

- estimate μ based on X under the quadratic loss
- superharmonic prior

$$\Delta\pi(\mu) = \sum_{i=1}^{p} \frac{\partial^2}{\partial \mu_i^2}\pi(\mu) \leq 0$$

- the Stein prior ($p \geq 3$) is superharmonic:

$$\pi(\mu) = \|\mu\|^{2-p}$$

- Bayes estimator with the Stein prior shrinks to the origin.

Theorem (Stein, 1974)

Bayes estimators with superharmonic priors dominate MLE.

Superharmonic prior for prediction

$$X \sim \mathrm{N}_p(\mu, \Sigma), \qquad Y \sim \mathrm{N}_p(\mu, \widetilde{\Sigma})$$

- We predict Y from the observation X (Σ, $\widetilde{\Sigma}$: known)
- Bayesian predictive density with prior $\pi(\mu)$

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

- Kullback-Leibler loss

$$D(p(y \mid \mu), \hat{p}(y \mid x)) = \int p(y \mid \mu) \log \frac{p(y \mid \mu)}{\hat{p}(y \mid x)} \mathrm{d}y$$

- Bayesian predictive density with the uniform prior is minimax

Superharmonic prior for prediction

$$X \sim \mathrm{N}_p(\mu, \Sigma), \quad Y \sim \mathrm{N}_p(\mu, \widetilde{\Sigma})$$

Theorem (Komaki, 2001)

When $\Sigma \propto \widetilde{\Sigma}$, the Stein prior dominates the uniform prior.

Theorem (George, Liang and Xu, 2006)

When $\Sigma \propto \widetilde{\Sigma}$, superharmonic priors dominate the uniform prior.

Theorem (Kobayashi and Komaki, 2008; George and Xu, 2008)

For general Σ and $\widetilde{\Sigma}$, superharmonic priors dominate the uniform prior.

Motivation

vector	James–Stein estimator $\hat{\mu}_{\mathrm{JS}} = \left(1 - \frac{p-2}{\|x\|^2}\right) x$	Stein's prior $\pi_{\mathrm{S}}(\mu) = \|\mu\|^{-(p-2)}$
matrix	Efron–Morris estimator $\hat{M}_{\mathrm{EM}} = X\left(I_q - (p-q-1)(X^\top X)^{-1}\right)$	?

- note: JS and EM are not generalized Bayes.

Singular value shrinkage prior

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

- We assume $p \geq q + 2$.
- π_{SVS} puts more weight on matrices with smaller singular values, so it shrinks singular values for each.
- When $q = 1$, π_{SVS} coincides with the Stein prior.

Theorem (M. and Komaki, 2015)

π_{SVS} is superharmonic: $\Delta\pi_{\mathrm{SVS}} \leq 0$.

- Therefore, the Bayes estimator and Bayesian predictive density with respect to π_{SVS} are minimax.

Comparison to other superharmonic priors

- Previously proposed superharmonic priors mainly shrink to simple subsets (e.g. point, linear subspace).

- In contrast, our priors shrink 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 π_{SVS} is strongly concentrated in the same way as the Laplacian of the Stein prior becomes a Dirac delta function.

An observation

- James-Stein estimator

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

- Stein's prior

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

- Efron–Morris estimator

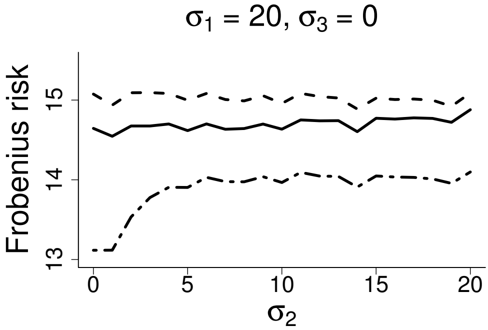
$$\hat{\sigma}_i = \left(1 - \frac{p-q-1}{\sigma_i^2}\right) \sigma_i$$

- Singular value shrinkage prior

$$\pi_{\mathrm{SVS}}(M) = \prod_{i=1}^{q} \sigma_i(M)^{-(p-q-1)}$$

Numerical results

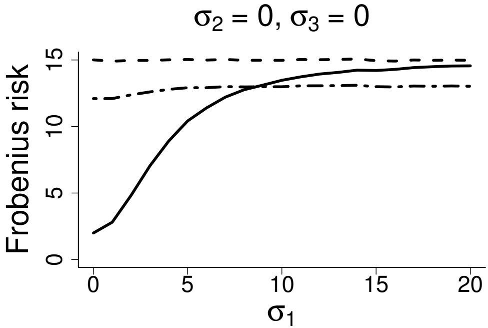
- Risk functions of Bayes estimators
  - $p = 5, q = 3$
  - dashed: uniform prior, solid: Stein's prior, dash-dot: our prior



- π_{SVS} works well when σ_2 is small, even if σ_1 is large.
  - Stein's prior works well when $\|M\|_F^2 = \sigma_1^2 + \sigma_2^2 + \sigma_3^2$ is small.

Numerical results

- Risk functions of Bayes estimators
  - $p = 5, q = 3$
  - dashed: uniform prior, solid: Stein's prior, dash-dot: our prior



- π_{SVS} has constant risk reduction as long as $\sigma_2 = \sigma_3 = 0$, because it shrinks singular values for each.
- Therefore, it works well when M has low rank.

Remark: integral representation

- When $p > 2q$, an integral representation of π_{SVS} is obtained.
  - $\mathrm{d}\Sigma$: Lebesgue measure on the space of positive semidefinite matrices

$$\pi_{\mathrm{SVS}}(M) \propto \int \mathrm{N}_{p,q}(0, I_p \otimes \Sigma)\mathrm{d}\Sigma$$

- cf. Stein's prior

$$\pi_{\mathrm{S}}(\mu) = ||\mu||^{2-p} \propto \int_0^\infty \mathrm{N}_p(0, tI_p)\mathrm{d}t$$

Additional shrinkage

- Efron and Morris (1976) proposed an estimator that further dominates $\hat{M}_{\mathrm{EM}}$ by additional shrinkage to the origin

$$\hat{M}_{\mathrm{MEM}} = X \left\{ I_q - (p-q-1)(X^\top X)^{-1} - \frac{q^2+q-2}{\mathrm{tr}(X^\top X)} I_q \right\}$$

- Motivated from this estimator, we propose another shrinkage prior

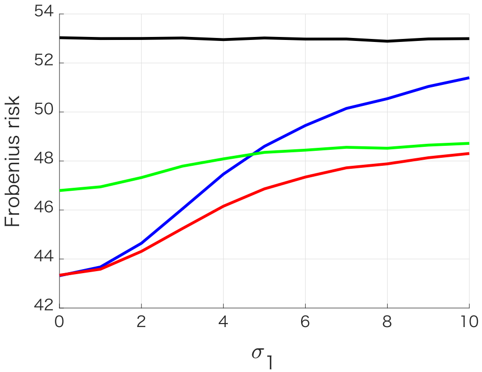
$$\pi_{\mathrm{MSVS}}(M) = \pi_{\mathrm{SVS}}(M) \|M\|_{\mathrm{F}}^{-(q^2+q-2)}$$

Theorem (M. and Komaki, 2017)

The prior π_{MSVS} asymptotically dominates π_{SVS} in both estimation and prediction.

Numerical results

- $p = 10$, $q = 3$, $\sigma_2 = \sigma_3 = 0$ (rank 1)
- black: π_{I}, blue: π_{S}, green: π_{SVS}, red: π_{MSVS}



- Additional shrinkage improves risk when $\|M\|_{\mathrm{F}}$ is small.

Admissibility results

Theorem (M. and Strawderman)

The Bayes estimator with respect to π_{SVS} is inadmissible.
The Bayes estimator with respect to π_{MSVS} is admissible.

- Proof: use Brown's condition

Addition of column-wise shrinkage

$$\pi_{\mathrm{MSVS}}(M) = \pi_{\mathrm{SVS}}(M) \prod_{j=1}^{q} \|M_{\cdot j}\|^{-q+1}$$

- $M_{\cdot j}$: j-th column vector of M

Theorem (M. and Komaki, 2017)

The prior π_{MSVS} asymptotically dominates π_{SVS} in both estimation and prediction.

- This prior can be used for sparse reduced rank regression.

$$Y = XB + E, \quad E \sim \mathrm{N}_{n,q}(0, I_n \otimes \Sigma)$$

$$\to \ \hat{B} = (X^\top X)^{-1} X^\top Y \sim \mathrm{N}_{p,q}(B, (X^\top X)^{-1} \otimes \Sigma)$$

Stein's recommendation

- Efron–Morris estimator

$$\hat{\sigma}_i = \left(1 - \frac{p-q-1}{\sigma_i^2}\right)\sigma_i$$

- Singular value shrinkage prior

$$\pi_{\mathrm{SVS}}(M) = \prod_{i=1}^{q} \sigma_i(M)^{-(p-q-1)}$$

- Stein (1974, Section 5) recommends stronger shrinkage

$$\hat{\sigma}_i = \left(1 - \frac{p+q-2i-1}{\sigma_i^2}\right)\sigma_i$$

  and says it dominates the Efron–Morris estimator.
- Corresponding prior ?

$$\pi(M) = \prod_{i=1}^{q} \sigma_i(M)^{-(p+q-2i-1)}$$

Empirical Bayes matrix completion (Matsuda and Komaki, 2019)

Empirical Bayes viewpoint

- Efron–Morris estimator was derived as an empirical Bayes estimator.

$$M \sim \mathrm{N}_{p,q}(0, I_p \otimes \Sigma) \quad \Leftrightarrow \quad M_{i\cdot} \sim \mathrm{N}_q(0, \Sigma)$$

$$Y \mid M \sim \mathrm{N}_{p,q}(M, I_p \otimes I_q) \quad \Leftrightarrow \quad Y_{ij} \sim \mathrm{N}(M_{ij}, 1)$$

- Bayes estimator (posterior mean)

$$\hat{M}^{\pi}(Y) = Y\left(I_q - (I_q + \Sigma)^{-1}\right)$$

- Since $Y^\top Y \sim W_q(I_q + \Sigma, p)$ marginally,

$$\mathrm{E}[(Y^\top Y)^{-1}] = \frac{1}{p-q-1}(I_q + \Sigma)^{-1}$$

$\rightarrow$ Replace $(I_q + \Sigma)^{-1}$ in $\hat{M}^{\pi}(Y)$ by $(p-q-1)(Y^\top Y)^{-1}$
$\rightarrow$ Efron–Morris estimator

Matrix completion

- Netflix problem
  - matrix of movie ratings by users

	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

- We want to estimate unobserved entries for recommendation.
  $\rightarrow$ matrix completion
- Many studies investigated its theory and algorithm.

Matrix completion

- Low-rankness of the underlying matrix is crucial in matrix completion.
- Existing algorithms employ low rank property.
  - SVT, SOFT-IMPUTE, OPTSPACE, Manopt, ...
- e.g. SVT algorithm
  - $||A||_*$: nuclear norm (sum of singular values)

$$
\begin{aligned}
\underset{\hat{M}}{\text{minimize}} \quad & ||\hat{M}||_* \\
\text{subject to} \quad & |Y_{ij} - \hat{M}_{ij}| \le E_{ij}, \quad (i, j) \in \Omega
\end{aligned}
$$

$\rightarrow$ sparse singular values (low rank)

EB algorithm

- We develop an empirical Bayes (EB) algorithm for matrix completion.
- EB is based on the following hierarchical model
  - Same with the derivation of the Efron–Morris estimator
  - C: scalar or diagonal matrix (unknown)

$$M \sim \mathrm{N}_{p,q}(0, I_p \otimes \Sigma)$$

$$Y \mid M \sim \mathrm{N}_{p,q}(M, I_p \otimes C)$$

- Goal: estimate M from observed entries of Y
  - If Y is fully observed, it reduces to the previous problem.

$\rightarrow$ EM algorithm !!

EB algorithm

EB algorithm

- E step: estimate (Σ, C) from $\hat{M}$ and Y
- M step: estimate M from Y and $(\hat{\Sigma}, \hat{C})$
- Iterate until convergence

- Both steps can be solved analytically.
  - Sherman-Morrison-Woodbery formula
- We obtain two algorithms corresponding to C is scalar or diagonal.
- EB does not require heuristic parameter tuning other than tolerance.

Numerical results

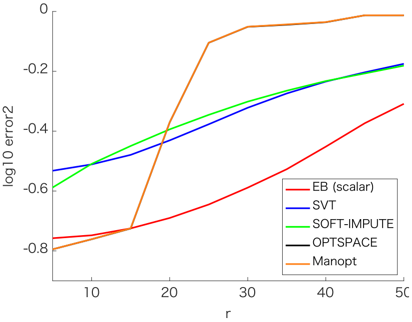
- Results on simulated data
  - 1000 rows, 100 columns, rank = 30, 50 % entries observed
  - observation noise: homogeneous ($R = I_q$)

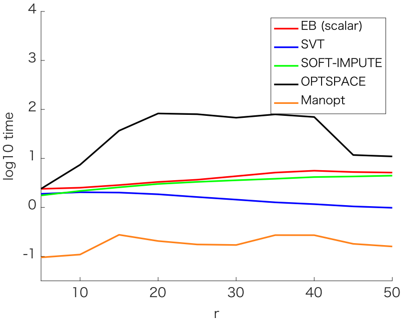
	error	time
EB (scalar)	0.26	4.33
EB (diagonal)	0.26	4.26
SVT	0.48	1.44
SOFT-IMPUTE	0.50	3.58
OPTSPACE	0.89	67.74
Manopt	0.89	0.17

- EB has the best accuracy.

Numerical results: rank

- Performance with respect to rank
  - 1000 rows, 100 columns, 50 % entries observed
  - observation noise: unit variance





- EB has the best accuracy when $r \geq 20$.

Application to real data

- Mice Protein Expression dataset
  - expression levels of 77 proteins measured in the cerebral cortex of 1080 mice
  - from UCI Machine Learning Repository

	error	time
EB (scalar)	0.12	2.90
EB (diagonal)	0.11	3.35
SVT	0.84	0.17
SOFT-IMPUTE	0.29	2.14
OPTSPACE	0.33	12.39
Manopt	0.64	0.19

- EB attains the best accuracy.

Future work (tensor case)

- 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.

Summary

Efron–Morris estimator (Efron and Morris, 1972)

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

minimax estimator of a normal mean matrix
natural extension of the James–Stein estimator

$$\downarrow$$

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

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

superharmonic ($\Delta\pi_{\mathrm{SVS}} \leq 0$), natural generalization of the Stein prior
works well for low-rank matrices → reduced-rank regression

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

estimate unobserved entries of a matrix by exploting low-rankness