Singular value shrinkage priors and empirical Bayes matrix completion

Takeru Matsuda

The University of Tokyo

Apr 11, 2019 @ Banff

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

↓

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

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

superharmonic ($\Delta \pi_{\mathrm{SVS}} \le 0$), 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

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)

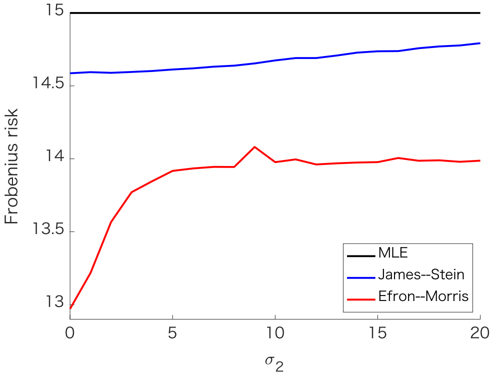
When $p \geq q + 2$, $\hat{M}_{\mathrm{EM}}$ is minimax and dominates $\hat{M}_{\mathrm{MLE}}(X) = X$.

- 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

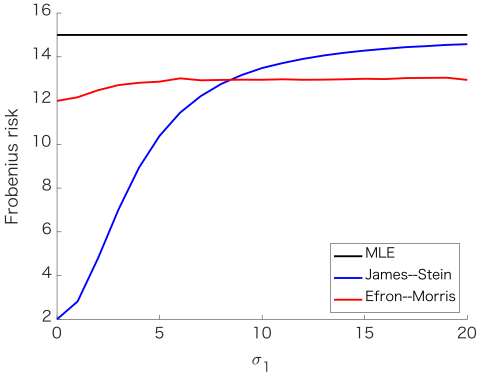
- Risk functions for $p = 5$, $q = 3$, $\sigma_1 = 20$, $\sigma_3 = 0$ (rank 2)
- black: $\hat{M}_{\mathrm{MLE}}$, blue: $\hat{M}_{\mathrm{JS}}$, red: $\hat{M}_{\mathrm{EM}}$



- $\hat{M}_{\mathrm{EM}}$ 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: $\hat{M}_{\mathrm{MLE}}$, blue: $\hat{M}_{\mathrm{JS}}$, red: $\hat{M}_{\mathrm{EM}}$



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

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 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

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

- the Stein prior

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

- Efron–Morris estimator

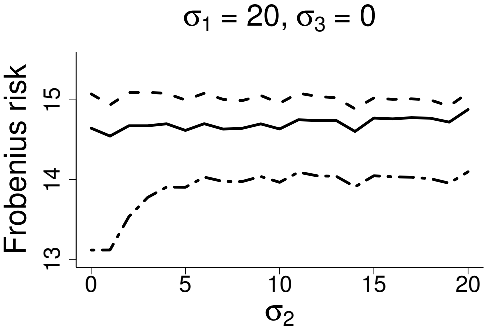
$$\hat{\sigma}_i = \left(1 - \frac{p-q-1}{\sigma_i^2}\right) \sigma_i$$

- Singular value shrinkage prior

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

Numerical results

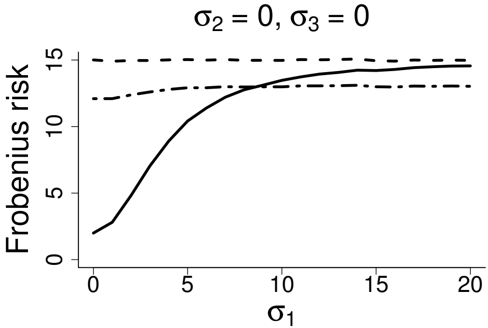
- Risk functions of Bayes estimators
  - $p = 5, q = 3$
  - dashed: uniform prior, solid: Stein's prior, dash-dot: our prior



- 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.

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

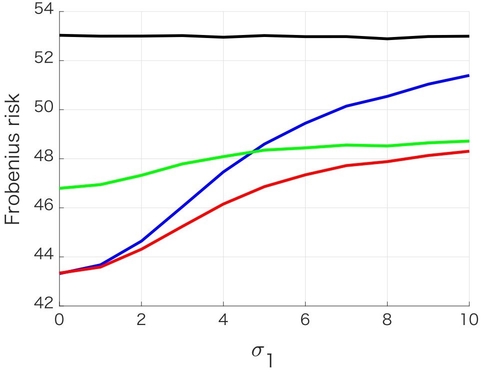
$$\pi_{\mathrm{MSVS}}(M) = \pi_{\mathrm{SVS}}(M)\|M\|_{\mathrm{F}}^{-(q^2+q-2)}$$

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.

$$Y = XB + E, \quad E \sim \mathrm{N}_{n,q}(0, I_n \otimes \Sigma)$$
$$\to \ \hat{B} = (X^\top X)^{-1} X^\top Y \sim \mathrm{N}_{p,q}(B, (X^\top X)^{-1} \otimes \Sigma)$$

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)

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

$\rightarrow$ 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.

Simulation setting

- We compare EB to existing algorithms (SVT, SOFT-IMPUTE, OPTSPACE, Manopt)
- data generation

$$U \sim \mathrm{N}_{p,r}(0, I_p \otimes I_r), \quad V \sim \mathrm{N}_{r,q}(0, I_r \otimes I_q)$$

$$M = UV, \quad Y = M + E, \quad E \sim \mathrm{N}_{p,q}(0, I_p \otimes R)$$

- Observed entries $\Omega \subset \{1, \cdots, p\} \times \{1, \cdots, q\}$: uniformly random
- We evaluate the accuracy by the normalized error for the unobserved entries.

$$\text{error} := \frac{(\sum_{(i,j)\notin\Omega}(\hat{M}_{ij} - M_{ij})^2)^{1/2}}{(\sum_{(i,j)\notin\Omega} M_{ij}^2)^{1/2}}$$

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: heterogeneity

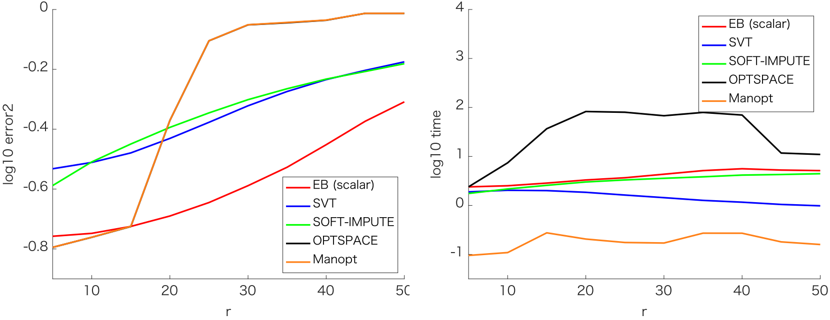
- Results on simulated data with heterogeneity
  - 1000 rows, 100 columns, rank = 30, 50 % entries observed
  - observation noise: heterogeneous ($R = \mathrm{diag}(0.05, 0.1, \cdots, 5)$)

	error	time
EB (scalar)	0.27	3.94
EB (diagonal)	0.24	3.12
SVT	0.43	1.59
SOFT-IMPUTE	0.37	2.10
OPTSPACE	0.28	7.46
Manopt	0.28	0.12

- EB (diagonal) has the best accuracy.
  - accounts for heterogeneity

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.

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, 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