Inadmissibility of the corrected Akaike information criterion

Takeru Matsuda

University of Tokyo, RIKEN Center for Brain Science

Abstract

- multivariate linear regression model

$$Y \sim \mathrm{N}_{n,q}(XB, I_n, \Sigma)$$

- corrected Akaike information criterion
  - minimum variance unbiased estimator of the expected Kullback–Leibler discrepancy

$$\mathrm{AICc} = -2\log p(Y \mid \hat{B}, \hat{\Sigma}) + \frac{2n}{n-p-q-1}\left(pq + \frac{q(q+1)}{2}\right)$$

Theorem (M., *Bernoulli* 2023+)

AICc is inadmissible and dominated by

$$\mathrm{MAICc} = \mathrm{AICc} - c\,\mathrm{tr}(\hat{\Sigma}((X\hat{B})^\top (X\hat{B}))^{-1})$$

as an estimator of the Kullback–Leibler discrepancy.

Contents

- Stein's paradox
- Loss estimation framework
- Inadmissibility of AICc
- Simulation

Stein's paradox

Estimation of normal mean vector

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

- estimate μ based on X by some estimator $\hat{\mu} = \hat{\mu}(x)$
- maximum likelihood estimator (MLE): $\hat{\mu}_{\mathrm{MLE}}(x) = x$
- Is MLE the best estimator ??

  $\rightarrow$ No !! (Stein's paradox, 1956)
- Statistical decision theory provides a framework to compare estimators

Loss and risk

- loss function $L(\mu, \hat{\mu})$: discrepancy between the estimate $\hat{\mu}$ and the true value μ
- e.g. quadratic loss

$$L(\mu, \hat{\mu}) = \|\hat{\mu} - \mu\|^2$$

- risk function $R(\mu, \hat{\mu})$: average loss of an estimator $\hat{\mu} = \hat{\mu}(x)$

$$R(\mu, \hat{\mu}) = \mathrm{E}_\mu[L(\mu, \hat{\mu}(x))] = \int L(\mu, \hat{\mu}(x)) p(x \mid \mu) \mathrm{d}x$$

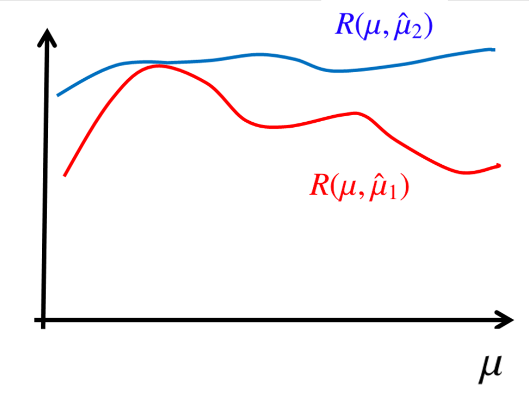
- In statistical decision theory, estimators are compared with the risk functions.
  - smaller risk is preferable

Dominance

Definition

An estimator $\hat{\mu}_1$ is said to dominate another estimator $\hat{\mu}_2$ if

$$R(\mu, \hat{\mu}_1) \leq R(\mu, \hat{\mu}_2) \quad \text{(for every } \mu)$$
$$R(\mu, \hat{\mu}_1) < R(\mu, \hat{\mu}_2) \quad \text{(for some } \mu)$$



Admissibility and minimaxity

Definition

An estimator $\hat{\mu}$ is said to be admissible if no estimator dominates $\hat{\mu}$.

Definition

An estimator $\hat{\mu}$ is said to be inadmissible if there exists an estimator that dominates $\hat{\mu}$.

Definition

An estimator $\hat{\mu}^*$ is said to be minimax if it minimizes the maximum risk:

$$\sup_{\mu} R(\mu, \hat{\mu}^*) = \inf_{\hat{\mu}} \sup_{\mu} R(\mu, \hat{\mu})$$

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)

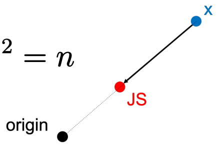
When $n \geq 3$, $\hat{\mu}_{\mathrm{MLE}}(x) = x$ is inadmissible.

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

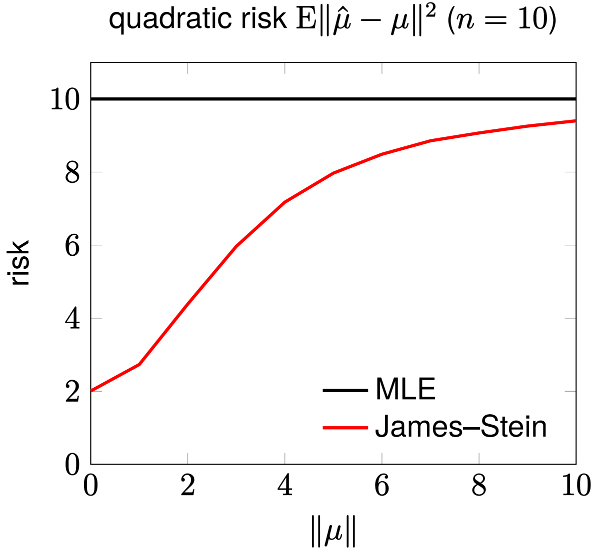
$$\hat{\mu}_{\mathrm{JS}}(x) = \left(1 - \frac{n-2}{\|x\|^2}\right) x$$

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

- JS shrinks x toward the origin.



Risk function ($n = 10$)



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

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)

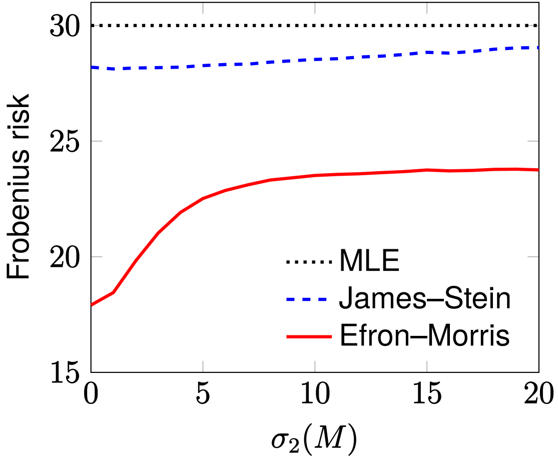
When $n \geq p+2$, $\hat{M}_{\mathrm{EM}}$ is minimax and dominates $\hat{M}_{\mathrm{MLE}}(X) = X$.

- 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)

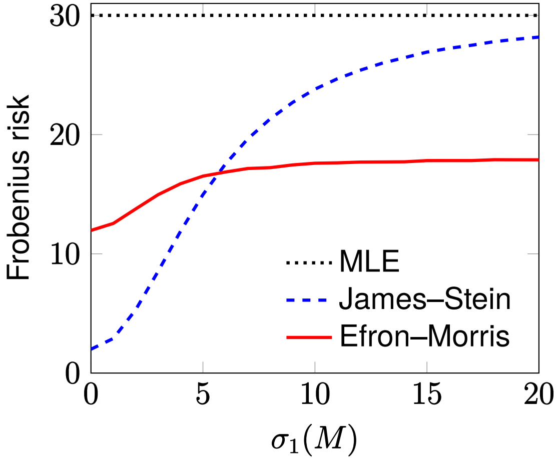
- $n = 10$, $p = 3$, $\sigma_1(M) = 20$, $\sigma_3(M) = 0$



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

Related studies

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

vector	James–Stein estimator (1961)	Stein's prior (1974)
matrix	Efron–Morris estimator (1972)	M. and Komaki (2015)

- Matrix quadratic loss and matrix superharmonicity (M. and Strawderman, 2022)
- Adaptive estimation via singular value shrinkage (M., 2022)
- Empirical Bayes matrix completion (M. and Komaki, 2019)

- レビュー：松田孟留. 縮小推定と優調和性. 応用数理, 2022.

Loss estimation framework

Loss estimation framework

$$Y \sim p(y \mid \theta)$$

- $\hat{\theta}(y)$: estimate of θ
- $\lambda(y)$: estimate of the loss $L(\theta, \hat{\theta}(y))$
  - note: loss depends on both θ and y

Definition

A loss estimator $\lambda_1(y)$ is said to dominate another one $\lambda_2(y)$ if

$$\mathrm{E}_\theta[(\lambda_1(y) - L(\theta, \hat{\theta}(y)))^2] \leq \mathrm{E}_\theta[(\lambda_2(y) - L(\theta, \hat{\theta}(y)))^2] \quad \text{(for every } \theta\text{)}$$
$$\mathrm{E}_\theta[(\lambda_1(y) - L(\theta, \hat{\theta}(y)))^2] < \mathrm{E}_\theta[(\lambda_2(y) - L(\theta, \hat{\theta}(y)))^2] \quad \text{(for some } \theta\text{)}$$

- (In)admissibility of loss estimators are defined accordingly.

Loss estimation for a normal mean vector

$$Y \sim \mathrm{N}_p(\theta, I_p)$$

- quadratic loss

$$L(\theta, \hat{\theta}) = \|\hat{\theta} - \theta\|^2$$

- Stein's unbiased risk estimate (SURE) for $\hat{\theta}(y) = y + g(y)$

$$\lambda^{\mathrm{U}}(y) = p + 2\nabla \cdot g(y) + \|g(y)\|^2$$

$$\mathrm{E}_\theta[\lambda^{\mathrm{U}}(y)] = \mathrm{E}_\theta[L(\theta, \hat{\theta}(y))]$$

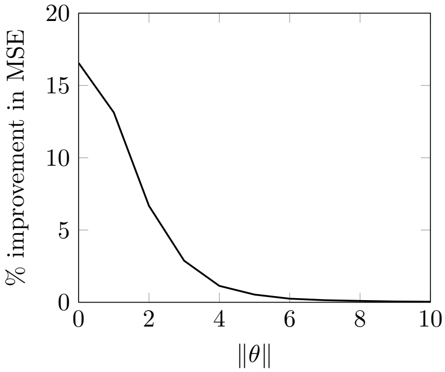
- For MLE $\hat{\theta}(y) = y$, SURE is $\lambda^{\mathrm{U}}(y) = p$

Loss estimation for a normal mean vector

Proposition (Johnstone, 1988)

If $p \geq 5$, then SURE $\lambda^{\mathrm{U}}(y) = p$ for MLE ($\hat{\theta}(y) = y$) is inadmissible and dominated by $\lambda(y) = p - 2(p-4)\|y\|^{-2}$:

$$\mathrm{E}_\theta(\lambda(y) - L(\theta, \hat{\theta}(y)))^2 \leq \mathrm{E}_\theta(\lambda^{\mathrm{U}}(y) - L(\theta, \hat{\theta}(y)))^2$$



Loss estimation for a normal mean matrix

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

- Frobenius loss

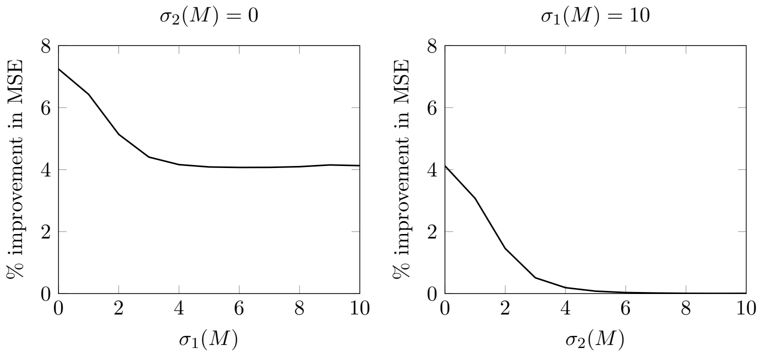
$$L(M, \hat{M}) = \|\hat{M} - M\|_{\mathrm{F}}^2 = \sum_{i,j} (\hat{M}_{ij} - M_{ij})^2$$

Theorem (M., 2023+)

If $p \geq 2q+3$, then SURE $\lambda^{\mathrm{U}}(Y) = pq$ for MLE ($\hat{M}(Y) = Y$) is inadmissible and dominated by

$$\lambda(Y) = pq - \frac{2(p-2q-2)}{q} \mathrm{tr}((Y^\top Y)^{-1}).$$

Loss estimation for a normal mean matrix



- large improvement when some singular values of M are small
- constant reduction of MSE as long as $\sigma_2(M) = 0$

  $\rightarrow$ works well when M is close to low-rank

- (similar to the Efron–Morris estimator)

Inadmissibility of the corrected AIC

Loss estimation for a predictive distribution

$$Y \sim p(y \mid \theta), \quad \widetilde{Y} \sim p(\widetilde{y} \mid \theta)$$

- predict $\widetilde{Y}$ from Y by a predictive distribution $\hat{p}(\widetilde{y} \mid y)$
- loss: Kullback–Leibler discrepancy

$$d(p(\widetilde{y} \mid \theta), \hat{p}(\widetilde{y} \mid y)) = -2 \int p(\widetilde{y} \mid \theta) \log \hat{p}(\widetilde{y} \mid y) \mathrm{d}\widetilde{y}$$

(equivalent to Kullback–Leibler divergence up to constant)

AIC as a loss estimator

- MLE

$$\hat{\theta}(y) = \operatorname*{argmax}_{\theta} \log p(y \mid \theta)$$

- plug-in predictive distribution

$$\hat{p}_{\text{plug-in}}(\tilde{y} \mid y) = p(\tilde{y} \mid \hat{\theta}(y))$$

- AIC is an approximately unbiased loss estimator:

$$\text{AIC} = -2 \log p(y \mid \hat{\theta}(y)) + 2k$$

$$\mathrm{E}_{\theta}[\text{AIC}] \approx \mathrm{E}_{\theta}[d(p(\tilde{y} \mid \theta), \hat{p}_{\text{plug-in}}(\tilde{y} \mid y))]$$

- Question: is AIC admissible ??

Multivariate linear regression model

$$y_i = B^\top x_i + \varepsilon_i, \quad \varepsilon_i \sim \mathrm{N}_q(0, \Sigma), \quad i = 1, \ldots, n$$

$$\downarrow$$

$$Y \sim \mathrm{N}_{n,q}(XB, I_n, \Sigma)$$

- Kullback–Leibler discrepancy

$$d((B, \Sigma), (\hat{B}, \hat{\Sigma})) = -2 \int p(\widetilde{Y} \mid B, \Sigma) \log p(\widetilde{Y} \mid \hat{B}, \hat{\Sigma}) \mathrm{d}\widetilde{Y}$$

Known covariance case

$$Y \sim \mathrm{N}_{n,q}(XB, I_n, \Sigma)$$

$$\hat{B} = (X^\top X)^{-1} X^\top Y$$

$$\mathrm{AIC} = -2 \log p(Y \mid \hat{B}, \Sigma) + 2pq$$

Theorem

If $p \geq 2q+3$, then AIC is inadmissible and dominated by

$$\mathrm{MAIC} = \mathrm{AIC} - \frac{2(p-2q-2)}{q} \mathrm{tr}(\Sigma((X\hat{B})^\top (X\hat{B}))^{-1}).$$

Unknown covariance case

$$Y \sim \mathrm{N}_{n,q}(XB, I_n, \Sigma)$$

$$\hat{B} = (X^\top X)^{-1} X^\top Y, \quad \hat{\Sigma} = \frac{1}{n}(Y - X\hat{B})^\top (Y - X\hat{B})$$

- AIC: approximately unbiased

$$\mathrm{AIC} = -2 \log p(Y \mid \hat{B}, \hat{\Sigma}) + 2 \left(pq + \frac{q(q+1)}{2} \right)$$

$$\mathrm{E}_{B,\Sigma}[\mathrm{AIC}] = \mathrm{E}_{B,\Sigma}[d((B,\Sigma),(\hat{B},\hat{\Sigma}))] + o(1) \quad (n \to \infty)$$

- corrected AIC: exactly unbiased

$$\mathrm{AICc} = -2 \log p(Y \mid \hat{B}, \hat{\Sigma}) + \frac{2n}{n-p-q-1} \left(pq + \frac{q(q+1)}{2} \right)$$

$$\mathrm{E}_{B,\Sigma}[\mathrm{AICc}] = \mathrm{E}_{B,\Sigma}[d((B,\Sigma),(\hat{B},\hat{\Sigma}))]$$

Unknown covariance case

Theorem (M., 2023+)

AIC is inadmissible and dominated by AICc.

- proof: bias-variance decomposition & $\mathrm{AICc} - \mathrm{AIC} = \mathrm{const.}$

Proposition (Davies et al., 2006)

AICc is the minimum variance unbiased estimator of the expected Kullback–Leibler discrepancy.

- proof: use Lehmann–Scheffé theorem
- Is AICc admissible ??

Inadmissibility of the corrected AIC

$$\bar{c} = \frac{4n^2}{(n-p)(q(n-p)+2)}\left(p - 2q - 2 - \frac{q^2+q-2}{n-p-q-1}\right)$$

Theorem (M., 2023+)

If $n - p - q - 1 > 0$ and $\bar{c} > 0$, then for any $c \in (0, \bar{c}]$, AICc is inadmissible and dominated by

$$\mathrm{MAICc} = \mathrm{AICc} - c\,\mathrm{tr}(\hat{\Sigma}((X\hat{B})^\top (X\hat{B}))^{-1}).$$

- In simulation, $c = \bar{c}$ works well.

Single response case

$$y \sim \mathrm{N}_n(X\beta, \sigma^2 I_n)$$

$$\hat{\beta} = (X^\top X)^{-1} X^\top y, \quad \hat{\sigma}^2 = \|y - X\hat{\beta}\|^2 / n$$

$$\bar{c} = \frac{4n^2(p-4)}{(n-p)(n-p+2)}$$

Corollary (M., 2023+)

If $n - p - 2 > 0$ and $\bar{c} > 0$, then for any $c \in (0, \bar{c}]$, AICc is inadmissible and dominated by

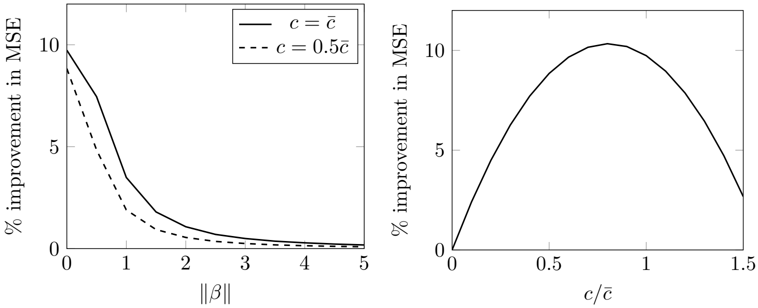
$$\mathrm{MAICc} = \mathrm{AICc} - c\hat{\sigma}^2 \|X\hat{\beta}\|^{-2}.$$

- In simulation, $c = \bar{c}$ works well.

Simulation

Single response

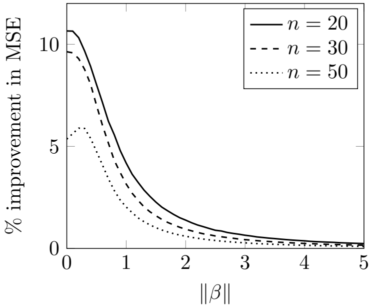
- $X \sim \mathrm{N}_{n,p}(0, I_n, I_p)$, $n = 30$, $p = 10$, $\sigma^2 = 1$

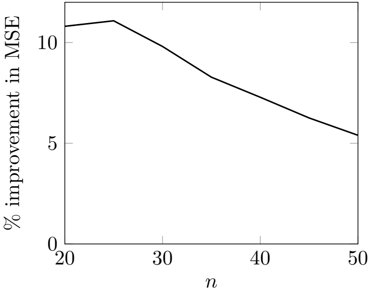


- $c = \bar{c}$ seems to be a reasonable choice
  - We adopt this value in the following experiments

Single response

- $X \sim \mathrm{N}_{n,p}(0, I_n, I_p)$, $p = 10$, $\sigma^2 = 1$

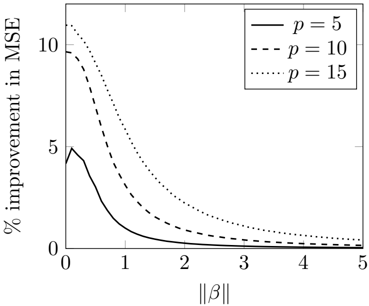


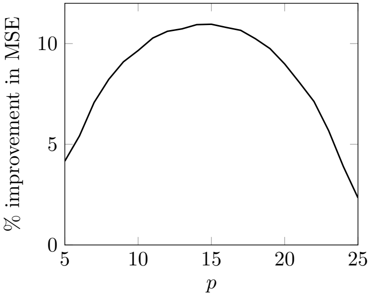


- larger improvement for smaller n

Single response

- $X \sim \mathrm{N}_{n,p}(0, I_n, I_p)$, $n = 30$, $\sigma^2 = 1$

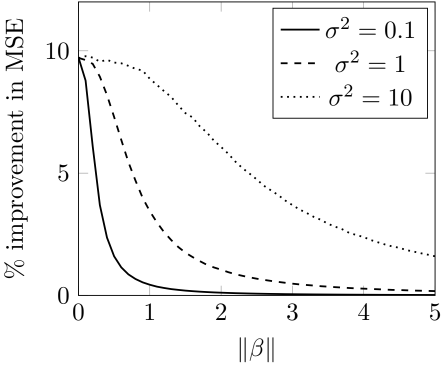




- maximum improvement around $p = 15$

Single response

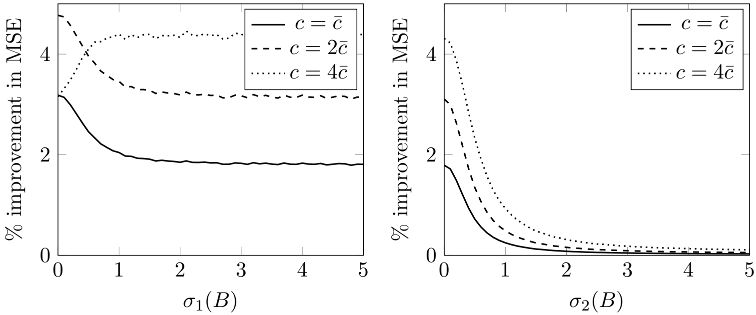
- $X \sim \mathrm{N}_{n,p}(0, I_n, I_p)$, $n = 30$, $p = 10$



- larger improvement for larger σ^2 at $\beta \neq 0$

Multi-response

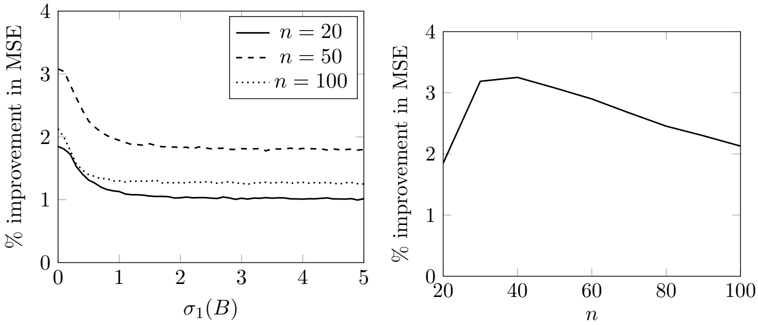
- $X \sim \mathrm{N}_{n,p}(0, I_n, I_p)$, $n = 30$, $p = 10$, $q = 2$



- large improvement when some singular values of M are small
- constant reduction of MSE as long as $\sigma_2(M) = 0$

Multi-response

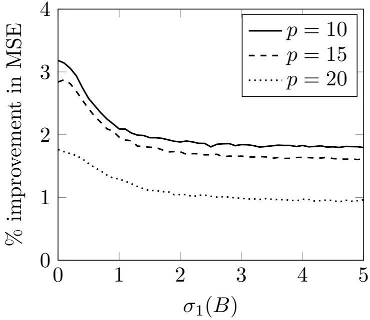
- $X \sim \mathrm{N}_{n,p}(0, I_n, I_p)$, $p = 10$, $q = 2$, $\Sigma = I_2$

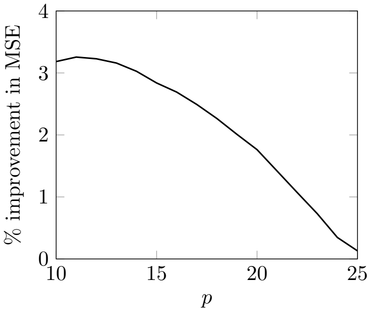


- maximum improvement around $n = 40$

Multi-response

- $X \sim \mathrm{N}_{n,p}(0, I_n, I_p)$, $n = 30$, $q = 2$, $\Sigma = I_2$

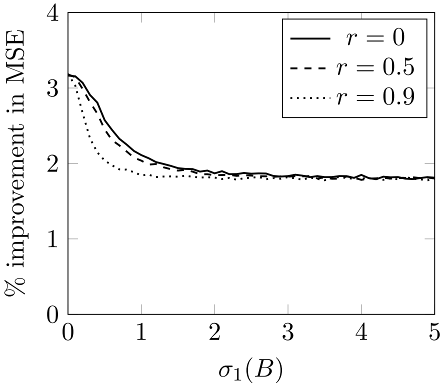




- smaller improvement for larger p

Multi-response

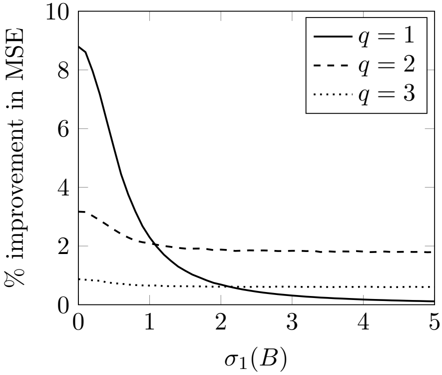
- $X \sim \mathrm{N}_{n,p}(0, I_n, I_p)$, $n = 30$, $p = 10$, $q = 2$, $\Sigma_{11} = \Sigma_{22} = 1$



- largest improvement for $r = 0$ (no correlation)

Multi-response

- $X \sim \mathrm{N}_{n,p}(0, I_n, I_p)$, $n = 30$, $p = 10$, $\Sigma = I_q$



Variable selection

- $X \sim \mathrm{N}_{n,p}(0, I_n, I_p)$, $n = 20$, $p = 10$, $q = 1$, $\sigma^2 = 1$
- $\beta = (0.1, 0.2, 0.3, 0.4, 0.5, 0, 0, 0, 0, 0)^\top$
- k-th submodel: $\beta_{k+1} = \cdots = \beta_p = 0$

	1	2	3	4	5	6	7	8	9	10
AIC	89	8	15	29	352	129	76	76	81	145
AICc	277	147	37	16	460	44	15	4	0	0
MAICc	248	137	34	14	492	54	17	4	0	0

- MAICc selects the true model more frequently than AIC and AICc

Summary & future work

Theorem (M., *Bernoulli* 2023+)

AICc is inadmissible and dominated by

$$\mathrm{MAICc} = \mathrm{AICc} - c\,\mathrm{tr}(\hat{\Sigma}((X\hat{B})^\top (X\hat{B}))^{-1})$$

as an estimator of the Kullback–Leibler discrepancy.

- model generalization by asymptotic arguments ??
- high-dimensional settings ??
  - cf. Bellec and Zhang (2021), Fujikoshi et al. (2014), Yanagihara et al. (2015)
- mis-specified cases ??
  - cf. Fujikoshi and Satoh (1997), Reschenhofer (1999)
- model averaging ?? (e.g. Mallows criterion; Hansen, 2007)
- other information criteria (e.g. TIC, GIC) ??
- Bayesian predictive distribution ?? (cf. Kitagawa, 1997)