Regularized Estimators for High-Dimensional Data Analysis

Youhong Lee (joint with Alex Shkolnik)

Department of Statistics Applied Probability, UC Santa Barbara

April 11, 2023

Outline

1. Stein's Paradox in High Dimensions
2. Regularized Estimators of the Covariance Matrix
3. Main Theorems and Simulations
4. Summary

Section 1

Stein's Paradox in High Dimensions

Charles M. Stein

The "Einstein of the Statistics Department" was also the first Stanford professor arrested for protesting apartheid. Although he rarely published his work, Stein leaves behind a distinctive, intriguing life story.

https://news.stanford.edu/2016/12/01/charles-m-stein-extraordinary-statistician-anti-war-activist-dies-96

Statistical Decision Theory (1930+)

- Decision based on data.
- In estimation, our decision is an estimator for a population quantity.
  - The natural estimator of the population mean is the sample mean.
- Neymann, Pearson, Pitman, and Wald.
  - Wald claims the sample mean is "admissible" (i.e., no estimator has uniformly lower risk) (1939).
  - Peisakoff discovers a mistake in Wald's proof (1950).
  - Stein (1956) proves admissibility fails in dimension 3 and higher.
- Stein's result ended an era of the search for the ideal estimator.

Stein (1956): Model

- Let $Y_1, \cdots, Y_n \in \mathbb{R}^p$ be i.i.d. draws from $N(\theta, \nu^2 \mathrm{I})$, where
  - $\theta = (\theta_1, \cdots, \theta_p) \in \mathbb{R}^p$ is the population mean,
  - ν is the standard deviation.
- Our goal is to estimate the population mean $\theta \in \mathbb{R}^p$.
  - One approach is the maximum likelihood estimator (MLE).
- The maximizer of the Gaussian likelihood function is the sample mean, $\eta \in \mathbb{R}^p$,

$$\eta = \frac{1}{n} \sum_{j=1}^{n} Y_j.$$

Stein (1956): Inadmissibility of the Sample Mean

Let $Y_1, \cdots, Y_n \in \mathbb{R}^p$ be i.i.d. draws from $N(\theta, \nu^2 \mathrm{I})$ with the sample mean $\eta \in \mathbb{R}^p$,

$$\eta = \frac{1}{n} \sum_{j=1}^{n} Y_j.$$

Theorem (Stein 1956)

There exists an estimator η^* such that for $p > 2$,

$$\mathrm{E}\left[|\eta^* - \theta|^2\right] < E\left[|\eta - \theta|^2\right] = \nu^2(p/n),$$

where the loss function is $|v|^2 = \sum_{i=1}^{p} v_j^2$ for $v \in \mathbb{R}^p$ (squared error loss, SEL).

Stein (1956): Inadmissibility of the Sample Mean

Theorem (Stein 1956)

There exists an estimator η^* such that for $p > 2$,

$$\mathrm{E}\left[|\eta^* - \theta|^2\right] < E\left[|\eta - \theta|^2\right] = \nu^2(p/n),$$

under squared error loss (SEL).

- The sample mean is inadmissible in dimension 3 and higher.
  - Admissible for essentially arbitrary symmetric loss $p \leq 2$.
- With a finite sample size n, the risk of the sample mean grows in p.

Bayesian Solution

- A Bayesian derivation is quite elegant (Efron & Morris 1975).
- Let $\eta \sim N(\theta, \nu^2 \mathrm{I})$ with a Gaussian prior on $\theta \in \mathbb{R}^p$.
- The posterior mean (as a Bayes estimator) is given by the conditional expectation

$$\mathrm{E}\left[\theta \,|\, \eta\right] = \mathrm{E}[\theta] + \left(1 - \frac{\nu^2}{\mathrm{Var}(\eta)}\right)(\eta - \mathrm{E}[\theta]).$$

- This motivates the "shrinkage" formula (for some prior mean $m \in \mathbb{R}^p$),

$$\eta(c) = m + c(\eta - m), \quad c \in [0, 1].$$

The James-Stein Estimator

Fix any $m \in \mathbb{R}$ and take $\eta^{\mathrm{JS}} = m + c^{\mathrm{JS}}(\eta - m)$ with the parameter,

$$c^{\mathrm{JS}} = 1 - \frac{\nu^2}{|\eta - m|^2/(p-2)} .$$

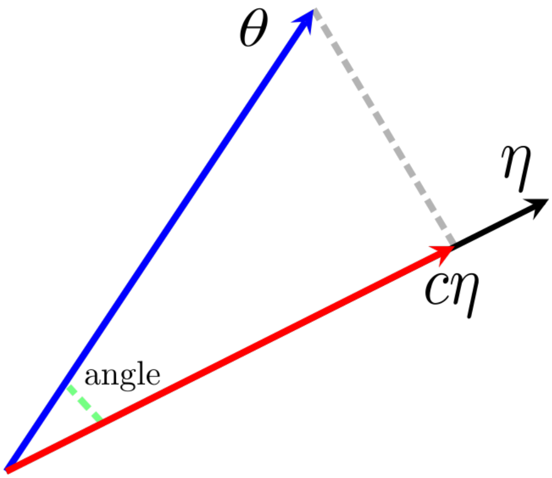
Theorem (James & Stein 1961)

For each $p > 2$,

$$\mathrm{E}\left[|\eta^{\mathrm{JS}} - \theta|^2\right] < \mathrm{E}\left[|\eta - \theta|^2\right] .$$

- James & Stein (1961) considered m at the origin.
- Plug-in estimates of m and ν may be used ($p > 3$).

The James-Stein Estimator in a High Dimension



Section 2

Regularized Estimators of the Covariance Matrix

Estimation of the Population Covariance Matrix

- We consider the problem of estimating a $p \times p$ matrix,

$$\Sigma = \theta\theta^\top + \nu^2 \mathrm{I} \qquad \text{(population covariance)}.$$

- We are no longer interested in the mean (zero w.l.o.g.).

- Our goal is to estimate $\theta \in \mathbb{R}^p$ that correlates the variables.

Estimation of the Population Covariance Matrix

- Let $Y_1, \cdots, Y_n \in \mathbb{R}^p$ be i.i.d. draws from $N(0_p, \Sigma)$ with $\Sigma = \theta\theta^\top + \nu^2 \mathrm{I}$.

$$\mathrm{Y} = \begin{bmatrix} \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots \\ Y_1 & Y_2 & \vdots & Y_n \\ \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots \end{bmatrix} \qquad (p \times n \text{ data matrix}).$$

- Our goal is to estimate $\theta \in \mathbb{R}^p$ that correlates the variables.
  - As for the population mean, we try the MLE.

The Maximum Likelihood Estimator of θ

- Let $Y_1, \cdots, Y_n \in \mathbb{R}^p$ be i.i.d. draws from $N(0_p, \Sigma)$ with $\Sigma = \theta\theta^\top + \nu^2 \mathrm{I}$.

- The Gaussian likelihood $L(\mathrm{Y}|\theta)$ given the parameter θ has

$$L(\mathrm{Y}|\theta) \propto \exp\left(t\langle u, \mathrm{Y}\mathrm{Y}^\top u\rangle\right)$$

  where $t = t(|\theta|) = \frac{|\theta|^2 n/(2\nu^2)}{|\theta|^2+\nu^2}$ and $u = \theta/|\theta|$ is the direction of θ on the unit sphere (Tipping & Bishop (1999)).

- Maximizing the Gaussian likelihood $L(\mathrm{Y}|\theta)$ is equivalent to solving

$$\max_{|v|=1} \langle v, \mathrm{Y}\mathrm{Y}^\top v\rangle.$$

The Maximum Likelihood Estimator of θ

- Maximizing the Gaussian likelihood $L(\mathrm{Y}|\theta)$ is equivalent to solving

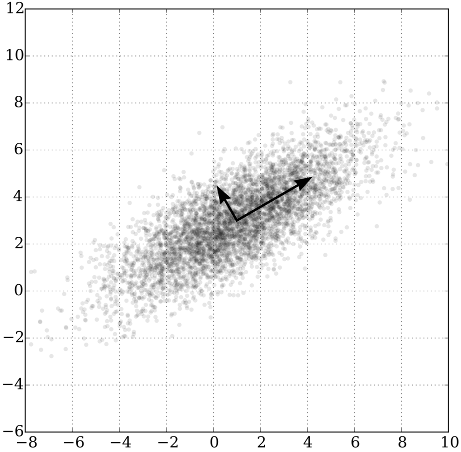
$$s^2 = \max_{|v|=1} \langle v, \mathrm{S}v \rangle.$$

where $\mathrm{S} = \frac{1}{n}\mathrm{Y}\mathrm{Y}^\top$ is the sample covariance matrix.

- We recognize the maximizer as the first principal component.
    - The maximizer v is the direction of maximum variance s^2 in the data.

- Notation:
    - The (scaled) sample principal component is $\eta = sv$.
    - The (scaled) population principal component is θ.

Principal Component Analysis



https://en.wikipedia.org/wiki/Principal_component_analysis

A Bayesian Solution for the First PC

- Following Efron & Morris (1975) we impose a Gaussian prior on the unknown $\theta \in \mathbb{R}^p$ and project it to the unit sphere $\mathbb{S}^{p-1}$.

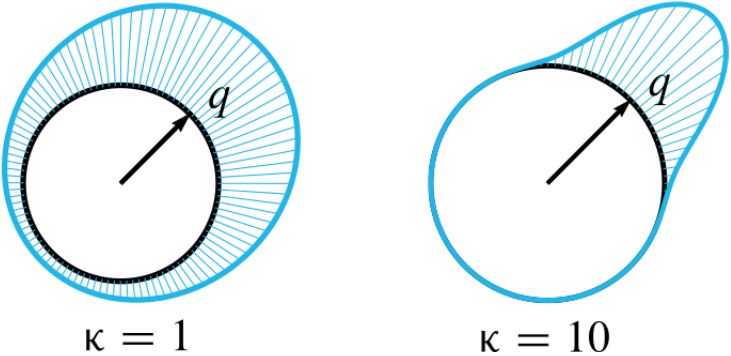
- The prior for the population principal component $u = \theta / |\theta|$ is

$$g(x) \propto \exp(\kappa \langle x, q \rangle), \quad x \in \mathbb{S}^{p-1},$$

  where $\kappa \in (0, \infty)$ is a concentration about the mean $q \in \mathbb{S}^{p-1}$.

  - The von Mises-Fisher distribution (Mardia, Jupp & Mardia 2000).

The von Mises-Fisher Distribution



Straub (2017)

A Bayesian Solution for the First PC

- Following Efron & Morris (1975), we impose a von Mises-Fisher prior on $u \in \mathbb{S}^{p-1}$.

- The posterior for the population principal component $u = \theta/|\theta|$ is

$$f(v|\mathrm{Y}) \propto \exp(t\langle v, \mathrm{S}v\rangle + \kappa\langle v, q\rangle), \quad v \in \mathbb{S}^{p-1},$$

  which is known as the Fisher-Bingham distribution.

- We proceed to maximize this density over v (maximum a posteriori (MAP) estimator).

A Bayesian Solution for the First PC

- Let S be the $p \times p$ sample covariance matrix for i.i.d. draws from $N(0_p, \Sigma)$ with $\Sigma = \theta\theta^\top + \nu^2 \mathrm{I}$.

- Our Bayesian argument motivates solving the penalized problem

$$\max_{|v|=1} \langle v, \mathrm{S}v \rangle + \kappa \langle v, q \rangle \,.$$

- The solution is given in terms of the resolvent of S, i.e.,

$$v(z) \propto (\mathrm{S} - z\mathrm{I})^{-1} q$$

for a parameter $z \in (s^2, \infty)$ that is a function of κ.

A Bayesian Solution for the First PC

- Let S be the $p \times p$ sample covariance matrix for i.i.d. draws from $N(0_p, \Sigma)$ with $\Sigma = \theta\theta^\top + \nu^2\mathrm{I}$.

- For a $q \in \mathbb{S}^{p-1}$, we propose analyzing the Bayesian estimator,

$$v(z) \propto (\mathrm{S} - z\mathrm{I})^{-1} q, \qquad z > s^2.$$

  - As $z \downarrow s^2$ we recover the sample principal component v,

$$s^2 = \max_{|v|=1} \langle v, \mathrm{S}v \rangle .$$

  - As $z \uparrow \infty$ the solution $v(z)$ approaches q.

Assumptions for High-Dimensional Analysis

1. The sample covariance matrix S is composed of n i.i.d. draws from $N(0_p, \Sigma)$ with $\Sigma = \theta\theta^\top + \nu^2\mathrm{I}$.

2. Suppose $p/n \to \infty$ as $n \to \infty$, and the following holds,

$$\langle\theta, \theta\rangle(n/p) \text{ converges in } (0, \infty).$$

3. q is a sequence of nonrandom vectors on $\mathbb{S}^{p-1}$ as p grows.

Inadmissibility of PCA

Theorem (Lee & Shkolnik 2023)

There exists z^* computable from Y such that an estimator $\eta^* = sv(z^*)$ satisfies

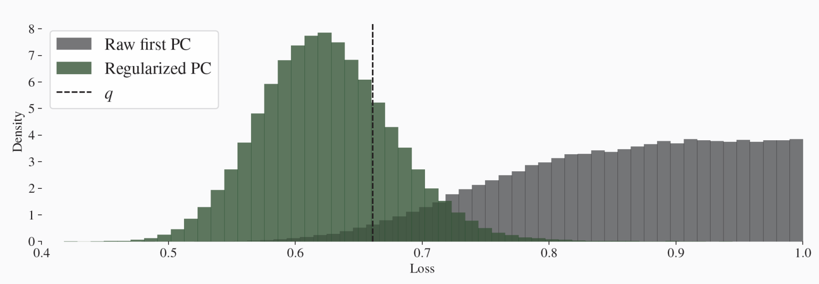
$$|\eta^* - \theta| \sim \sqrt{c}|\eta - \theta|$$

for some random variable $c \in (0, 1)$ almost surely.

- Notation: $f_p \sim g_p$ if $f_p/g_p \to 1$ as $p \to \infty$.

Inadmissibility of PCA

- Our loss function: $\ell(y|u) = |y - u|^2/2 = 1 - \langle y, u \rangle$.



A James-Stein Solution for the First PC

- The statement that PCA is not inadmissible is not new.
  - Goldberg, Papanicalaou & Shkolnik (2022), Gurdogan & Kercheval (2022), Shkolnik (2022), Shkolnik (2023), Goldberg & Kercheval (2023).
  - Thought to be impossibe to show previously (Wang & Fan 2017).

- These works propose variants of the James-Stein estimator,

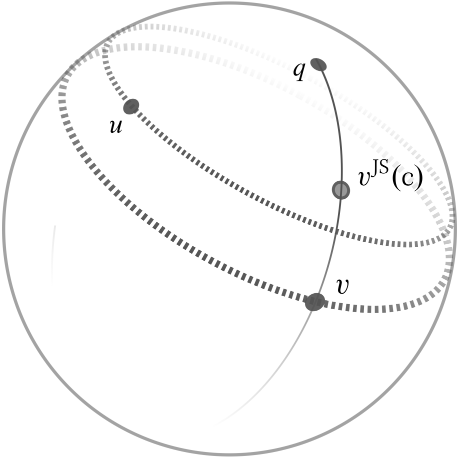
$$v^{\mathrm{JS}} \propto m + c(v - m), \qquad c = 1 - \frac{\nu^2}{|v - m|^2}$$

where $m = \langle v, q \rangle q$ for any vector $q \in \mathbb{S}^{p-1}$.

  - These formulas a borrowed from the original JS estimator.
  - The variance ν^2 may be estimated using the eigenvalues of S.

A James-Stein Solution for the First PC

The JS family occupies the geodesic between v and q.



Comparison between the Bayesian and James-Stein Solutions

- The sample principal component v is the MLE. The estimator

$$v(z) \propto (\mathrm{S} - z\mathrm{I})^{-1} q$$

was derived via Bayesian arguments of Efron & Morris (1975).

  - Is $v(z)$ related to the v^{JS} estimator?
  - What is the interpretation of $v(z)$ geometrically?
  - How do v, v^{JS} and $v(z)$ perform?

Comparison between the Bayesian and James-Stein Solutions

- The sample principal component v is the MLE. The estimator

$$v(z) \propto (\mathrm{S} - z\mathrm{I})^{-1} q$$

was derived via Bayesian arguments of Efron & Morris (1975).

Lemma (Lee & Shkolnik 2023)

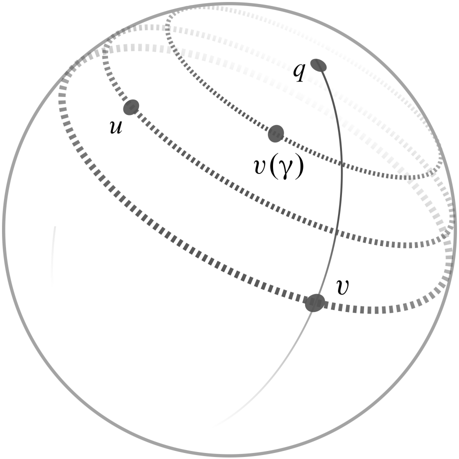
$v(z)$ is the solution $v(\gamma)$ of the regularized PCA problem,

$$\max_{|v|=1} \langle v, \mathrm{S}v \rangle \qquad \text{s.t.} \qquad \langle v, q \rangle \geq \gamma$$

for a parameter $\gamma \in [0, 1]$ that is a function of z.

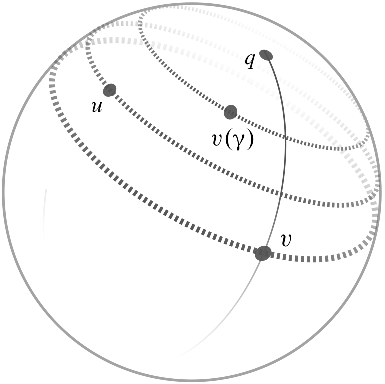
Comparison between the Bayesian and James-Stein Solutions

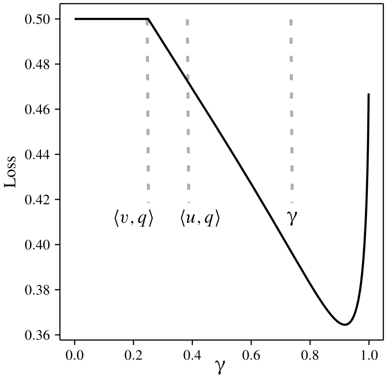
The estimator $v(\gamma)$ has more degrees of freedom than v^{JS}.



Comparison between the Bayesian and James-Stein Solutions

Visualization of the asymptotic loss $|v(\gamma) - u|^2/2$





Comparison between the Bayesian and James-Stein Solutions

- The Neumann expansion of the resolvent $(\mathrm{S} - z\mathrm{I})^{-1}$ yields,

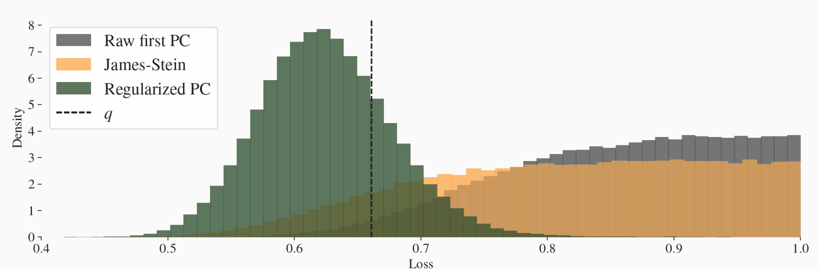
$$v(z) \propto m + c_1(z)(v - m) + \sum_{j \geq 2} c_j(z) v_j$$

  where v_j is the jth sample principal component and $m = \langle v, q \rangle q$.

  - The constant $c_1(z) \neq c$ of v^{JS} at the optimal z.

Comparison between the Bayesian and James-Stein Solutions

- Our loss function: $\ell(y|u) = |y - u|^2/2 = 1 - \langle y, u \rangle$.



Section 3

Main Theorems and Simulations

Assumptions for High-Dimensional Analysis

1. The sample covariance matrix S is composed of n i.i.d. draws from $N(0_p, \Sigma)$ with $\Sigma = \theta\theta^\top + \nu^2\mathrm{I}$.

2. Suppose $p/n \to \infty$ as $n \to \infty$, and the following holds,

$$\langle\theta, \theta\rangle(n/p) \text{ converges in } (0, \infty).$$

3. q is a sequence of nonrandom vectors on $\mathbb{S}^{p-1}$ as p grows.

4. The Gaussian distributional assumption may be removed entirely at the expense of a more abstract set of conditions.

Inadmissibility of PCA

Theorem (Lee & Shkolnik 2023)

There exists z^* computable from Y such that an estimator $\eta^* = sv(z^*)$ satisfies

$$|\eta^* - \theta| \sim \sqrt{c}|\eta - \theta|$$

for some random variable $c \in (0, 1)$ almost surely.

- Notation: $f_p \sim g_p$ if $f_p/g_p \to 1$ as $p \to \infty$
- The random variable c is that of the James-Stein estimator v^{JS}.

Theorem 1: Performance of $v(z)$ over v and q

Theorem 1a (Lee & Shkolnik 2023)

There is a univariate (random) function f on $[0, 1]$ computable from S which has a unique minimum ζ^* that yields z^* ´and $v^* = v(z^*)$ with

$$\sin\left(\alpha(v^*, u)\right) \sim \frac{\sin\left(\alpha(u, q)\right)}{\sin\left(\alpha(v, q)\right)} \sin\left(\alpha(v, u)\right)$$

and $\frac{\sin(\alpha(u,q))}{\sin(\alpha(v,q))} \in (0, 1)$ provided $\langle u, q \rangle \neq 0$ eventually.

- v^* outperforms u asymptotically provided q contains "information".
- Note: $\sin f_p \sim g_p$ if $f_p / g_p \to 1$ as $p \to \infty$, $\sin(\alpha) = \alpha + O(\alpha^3)$.

Theorem 1: Performance of $v(z)$ over v and q

Theorem 1b (Lee & Shkolnik 2023)

There is a univariate (random) function f on $[0,1]$ computable from S which has a unique minimum ζ^* that yields z^* ´and $v^* = v(z^*)$ with

$$\sin\left(\alpha(v^*, u)\right) \sim \frac{\sin\left(\alpha(u,v)\right)}{\sin\left(\alpha(v,q)\right)} \sin\left(\alpha(u,q)\right)$$

and $\frac{\sin(\alpha(u,v))}{\sin(\alpha(v,q))} \in (0,1)$ provided $\langle u, q\rangle \neq 0$ eventually.

- v^* outperforms q asymptotically provided q contains "information".
- Note: sin $f_p \sim g_p$ if $f_p/g_p \to 1$ as $p \to \infty$, $\sin(\alpha) = \alpha + O(\alpha^3)$.

Theorem 2: Performance of $v(z)$ over v^{JS}

Theorem 2 (Lee & Shkolnik 2023)

There is a univariate (random) function f on $[0, 1]$ computable from S which has a unique minimum ζ^* that yields z^* ´and $v^* = v(z^*)$ with

$$\sin\left(\alpha(v^{\mathrm{JS}}, u)\right) - \sin\left(\alpha(v^*, u)\right) \geq \xi_p + o(n/p)$$

for a random variable $\xi_p \geq 0$ w.h.p., and $\xi_p \to 0$ almost surely.

- A key random variable that distinguishes v^* from v^{JS} is

$$|q_N|^2 = \sum_{j=1}^{n} \langle v_j, q \rangle^2$$

which is a proxy for the information $|u_N|^2 = \sum_{j=1}^{n} \langle v_j, u \rangle^2$.

Simulation Study

- We numerically test a Gaussian model parametrized by

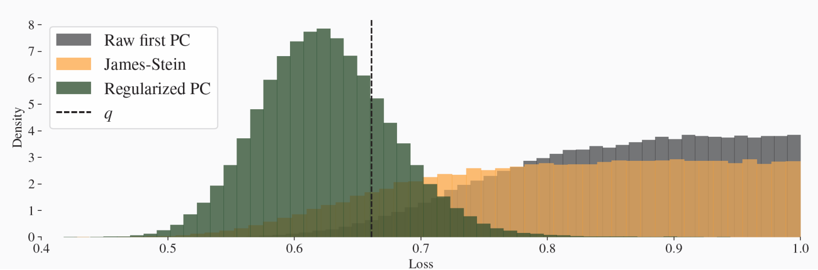
$$\tau^2 = \langle \theta, \theta \rangle (n/p) \qquad \text{(signal strength)}$$

- We test v, v^{JS}, and $v(z^*)$ on the loss function,

$$\ell(y|u) = |y - u|^2/2 = 1 - \langle y, u \rangle \, .$$

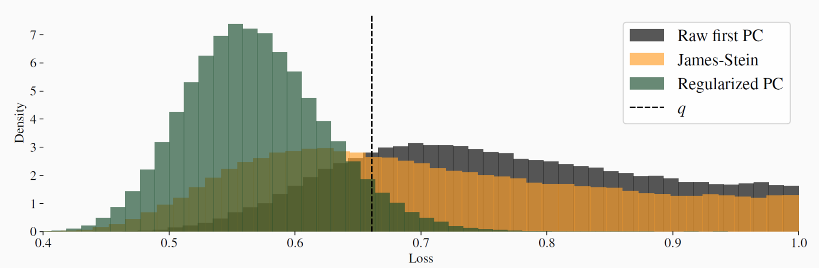
Simulation: Weak Signal

- $\tau^2 \approx 0.15$



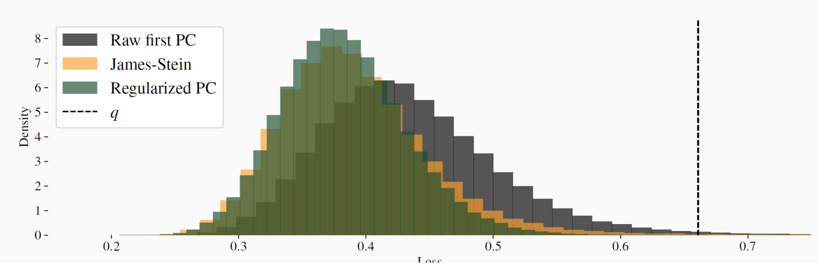
Simulation: Moderate Signal

- $\tau^2 \approx 0.20$



Simulation: Strong Signal

- $\tau^2 \approx 0.60$



Section 4

Summary

Summary

1. The James-Stein estimator shows the inadmissability of the MLE (the sample mean).

2. As the JS estimator has an elegant Bayesian derivation, we propose a regularized estimator for the first population principal component,

$$v(z) \propto (\mathrm{S} - z\mathrm{I})^{-1} q \,.$$

3. We show that the MLE of the population principal component is inadmissible.

4. We explain its relationship to JSE of the first principal component. Theory and numerical results suggest JSE is inadmissible.

Thank You!