High dimensional statistics

Part 2: Shrinkage and Stein's phenomenon

Hugo Richard (research.hugo.richard@gmail.com)

ENPC - February 2022

Introduction

Gaussian sequence model
Observe one sample $\mathbf{y} \in \mathbb{R}^n, \mathbf{y} \sim \mathcal{N}(\boldsymbol{\mu}, \sigma^2 I_n)$.

How to estimate the mean $\boldsymbol{\mu}$?

Introduction

Gaussian sequence model
Observe one sample $\mathbf{y} \in \mathbb{R}^n, \mathbf{y} \sim \mathcal{N}(\boldsymbol{\mu}, \sigma^2 I_n)$.

How to estimate the mean $\boldsymbol{\mu}$?

Example (Practical instance)

Tasks
Proportion of people that vote Trump? (μ_1)

Proportion of people with blue eyes in France? (μ_2)

Proportion of new baby girls in Congo? (μ_3)

Observations
Result of a pole in New York (y_1)

Proportion of people with blue eyes in Paris (y_2)

Proportion of new baby girls in Brazaville (y_3)

Introduction

Gaussian sequence model
Observe one sample $\mathbf{y} \in \mathbb{R}^n, \mathbf{y} \sim \mathcal{N}(\boldsymbol{\mu}, \sigma^2 I_n)$.

How to estimate the mean $\boldsymbol{\mu}$?

Example (Maximum likelihood solution)

Introduction

Gaussian sequence model
Observe one sample $\mathbf{y} \in \mathbb{R}^n, \mathbf{y} \sim \mathcal{N}(\boldsymbol{\mu}, \sigma^2 I_n)$.

How to estimate the mean $\boldsymbol{\mu}$?

Example (Maximum likelihood solution)

Solve

$$\hat{\boldsymbol{\mu}} = \operatorname*{argmax}_{\boldsymbol{\mu}} \mathbb{P}(\mathbf{y}) \tag{1}$$

$$= \operatorname*{argmax}_{\boldsymbol{\mu}} \frac{\exp(-\frac{1}{2}\|\mathbf{y} - \boldsymbol{\mu}\|^2)}{(2\pi)^{\frac{n}{2}}} \tag{2}$$

$$= \mathbf{y} \tag{3}$$

Introduction

Gaussian sequence model
Observe one sample $\mathbf{y} \in \mathbb{R}^n, \mathbf{y} \sim \mathcal{N}(\boldsymbol{\mu}, \sigma^2 I_n)$.

How to estimate the mean $\boldsymbol{\mu}$?

Example (Maximum likelihood solution)
$\hat{\boldsymbol{\mu}} = \mathbf{y}$

How good it is ?
Mean square error $\mathbb{E}_{\mathbf{y}}[\|\hat{\boldsymbol{\mu}}(\mathbf{y}) - \boldsymbol{\mu}\|^2]$

Introduction

Gaussian sequence model
Observe one sample $\mathbf{y} \in \mathbb{R}^n, \mathbf{y} \sim \mathcal{N}(\boldsymbol{\mu}, \sigma^2 I_n)$.

How to estimate the mean $\boldsymbol{\mu}$?

Example (Maximum likelihood solution)
$\hat{\boldsymbol{\mu}} = \mathbf{y}$

How good it is ?
Mean square error $\mathbb{E}_{\mathbf{y}}[\|\hat{\boldsymbol{\mu}}(\mathbf{y}) - \boldsymbol{\mu}\|^2]$

Definition (Admissible estimator)
$\hat{\boldsymbol{\mu}}(\mathbf{y})$ is not admissible if there exists $\hat{\boldsymbol{\mu}}^*(\mathbf{y})$ such that $\forall \boldsymbol{\mu} \in \mathbb{R}^n, \ \mathbb{E}_{\mathbf{y}}[\|\hat{\boldsymbol{\mu}}(\mathbf{y}) - \boldsymbol{\mu}\|^2] \geq \mathbb{E}_{\mathbf{y}}[\|\hat{\boldsymbol{\mu}}^*(\mathbf{y}) - \boldsymbol{\mu}\|^2]$ (inequality strict for at least one value of $\boldsymbol{\mu}$)

Introduction

Introduction

Observe one sample $\mathbf{y} \in \mathbb{R}^n, \mathbf{y} \sim \mathcal{N}(\boldsymbol{\mu}, \sigma^2 I_n)$.

How to estimate the mean $\boldsymbol{\mu}$?

Example (Maximum likelihood solution)

$\hat{\boldsymbol{\mu}} = \mathbf{y}$

Definition (Admissible estimator)

$\hat{\boldsymbol{\mu}}(\mathbf{y})$ is not admissible if there exists $\hat{\boldsymbol{\mu}}^*(\mathbf{y})$ such that $\forall \boldsymbol{\mu} \in \mathbb{R}^n, \; \mathbb{E}_{\mathbf{y}}[\|\hat{\boldsymbol{\mu}}(\mathbf{y}) - \boldsymbol{\mu}\|^2] \geq \mathbb{E}_{\mathbf{y}}[\|\hat{\boldsymbol{\mu}}^*(\mathbf{y}) - \boldsymbol{\mu}\|^2]$ (inequality strict for at least one value of $\boldsymbol{\mu}$)

Main result: The maximum likelihood solution is not admissible

Some estimators of the form $\hat{\mu} = Ay$

$\mathbf{y} \in \mathbb{R}^n, \mathbf{y} \sim \mathcal{N}(\boldsymbol{\mu}, \sigma^2 I_n)$

Shrinkage by a multiplicative constant

Take $\hat{\boldsymbol{\mu}}_s = s\mathbf{y}$ with $s \in [0, 1]$.

Some estimators of the form $\hat{\mu} = Ay$

$\mathbf{y} \in \mathbb{R}^n, \mathbf{y} \sim \mathcal{N}(\boldsymbol{\mu}, \sigma^2 I_n)$

Shrinkage by a multiplicative constant

Take $\hat{\boldsymbol{\mu}}_s = s\mathbf{y}$ with $s \in [0, 1]$.

$$\mathbb{E}_{\mathbf{y}}[\|\hat{\boldsymbol{\mu}}_s - \boldsymbol{\mu}\|^2]$$

Some estimators of the form $\hat{\mu} = Ay$

$\mathbf{y} \in \mathbb{R}^n, \mathbf{y} \sim \mathcal{N}(\boldsymbol{\mu}, \sigma^2 I_n)$

Shrinkage by a multiplicative constant

Take $\hat{\boldsymbol{\mu}}_s = s\mathbf{y}$ with $s \in [0, 1]$.

$$\mathbb{E}_{\mathbf{y}}[\|\hat{\boldsymbol{\mu}}_s - \boldsymbol{\mu}\|^2] = \mathbb{E}_{\mathbf{y}}[\|\hat{\boldsymbol{\mu}}_s - \boldsymbol{\mu}\|^2] \tag{4}$$

Some estimators of the form $\hat{\mu} = Ay$

$\mathbf{y} \in \mathbb{R}^n, \mathbf{y} \sim \mathcal{N}(\boldsymbol{\mu}, \sigma^2 I_n)$

Shrinkage by a multiplicative constant

Take $\hat{\boldsymbol{\mu}}_s = s\mathbf{y}$ with $s \in [0, 1]$.

$$\mathbb{E}_{\mathbf{y}}[\|\hat{\boldsymbol{\mu}}_s - \boldsymbol{\mu}\|^2] = \mathbb{E}_{\mathbf{y}}[\|\hat{\boldsymbol{\mu}}_s - \boldsymbol{\mu}\|^2] \quad (4)$$

$$= \|\mathbb{E}_{\mathbf{y}}[\hat{\boldsymbol{\mu}}_s] - \boldsymbol{\mu}\|^2 + \mathbb{E}_{\mathbf{y}}[\|\hat{\boldsymbol{\mu}}_s - \mathbb{E}_{\mathbf{y}}[\hat{\boldsymbol{\mu}}_s]\|^2] \quad (5)$$

Some estimators of the form $\hat{\mu} = Ay$

$\mathbf{y} \in \mathbb{R}^n, \mathbf{y} \sim \mathcal{N}(\boldsymbol{\mu}, \sigma^2 I_n)$

Shrinkage by a multiplicative constant

Take $\hat{\boldsymbol{\mu}}_s = s\mathbf{y}$ with $s \in [0, 1]$.

$$\begin{aligned}
\mathbb{E}_{\mathbf{y}}[\|\hat{\boldsymbol{\mu}}_s - \boldsymbol{\mu}\|^2] &= \mathbb{E}_{\mathbf{y}}[\|\hat{\boldsymbol{\mu}}_s - \boldsymbol{\mu}\|^2] &(4)\\
&= \|\mathbb{E}_{\mathbf{y}}[\hat{\boldsymbol{\mu}}_s] - \boldsymbol{\mu}\|^2 + \mathbb{E}_{\mathbf{y}}[\|\hat{\boldsymbol{\mu}}_s - \mathbb{E}_{\mathbf{y}}[\hat{\boldsymbol{\mu}}_s]\|^2] &(5)\\
&= (s-1)^2\|\boldsymbol{\mu}\|^2 + s^2 n\sigma^2 &(6)
\end{aligned}$$

Some estimators of the form $\hat{\mu} = Ay$

$\mathbf{y} \in \mathbb{R}^n, \mathbf{y} \sim \mathcal{N}(\boldsymbol{\mu}, \sigma^2 I_n)$

Shrinkage by a multiplicative constant

Take $\hat{\boldsymbol{\mu}}_s = s\mathbf{y}$ with $s \in [0, 1]$.

$$\begin{aligned}
\mathbb{E}_{\mathbf{y}}[\|\hat{\boldsymbol{\mu}}_s - \boldsymbol{\mu}\|^2] &= \mathbb{E}_{\mathbf{y}}[\|\hat{\boldsymbol{\mu}}_s - \boldsymbol{\mu}\|^2] &(4)\\
&= \|\mathbb{E}_{\mathbf{y}}[\hat{\boldsymbol{\mu}}_s] - \boldsymbol{\mu}\|^2 + \mathbb{E}_{\mathbf{y}}[\|\hat{\boldsymbol{\mu}}_s - \mathbb{E}_{\mathbf{y}}[\hat{\boldsymbol{\mu}}_s]\|^2] &(5)\\
&= (s-1)^2\|\boldsymbol{\mu}\|^2 + s^2 n\sigma^2 &(6)
\end{aligned}$$

Taking $s = \operatorname{argmin}_s \mathbb{E}_{\mathbf{y}}[\|\hat{\boldsymbol{\mu}}_s - \boldsymbol{\mu}\|^2]$ we get $s = \frac{1}{1+n\sigma^2/\|\mu\|^2} < 1$

Some estimators of the form $\hat{\mu} = Ay$

$\mathbf{y} \in \mathbb{R}^n, \mathbf{y} \sim \mathcal{N}(\boldsymbol{\mu}, \sigma^2 I_n)$

Shrinkage by a multiplicative constant

Take $\hat{\boldsymbol{\mu}}_s = s\mathbf{y}$ with $s \in [0, 1]$.

$$\mathbb{E}_{\mathbf{y}}[\|\hat{\boldsymbol{\mu}}_s - \boldsymbol{\mu}\|^2] = \mathbb{E}_{\mathbf{y}}[\|\hat{\boldsymbol{\mu}}_s - \boldsymbol{\mu}\|^2] \tag{4}$$

$$= \|\mathbb{E}_{\mathbf{y}}[\hat{\boldsymbol{\mu}}_s] - \boldsymbol{\mu}\|^2 + \mathbb{E}_{\mathbf{y}}[\|\hat{\boldsymbol{\mu}}_s - \mathbb{E}_{\mathbf{y}}[\hat{\boldsymbol{\mu}}_s]\|^2] \tag{5}$$

$$= (s-1)^2\|\boldsymbol{\mu}\|^2 + s^2 n\sigma^2 \tag{6}$$

Taking $s = \mathrm{argmin}_s \, \mathbb{E}_{\mathbf{y}}[\|\hat{\boldsymbol{\mu}}_s - \boldsymbol{\mu}\|^2]$ we get $s = \frac{1}{1+n\sigma^2/\|\mu\|^2} < 1$

Problem We don't know $\boldsymbol{\mu}$

Some estimators of the form $\hat{\mu} = A\mathbf{y}$

$\mathbf{y} \in \mathbb{R}^n, \mathbf{y} \sim \mathcal{N}(\boldsymbol{\mu}, \sigma^2 I_n)$

$n =$ number of samples

We observe $y_1, \dots, y_n,\ y_i \in \mathbb{R}$

$N_k(i) = k$-closest neighbors of i.

Nearest neighbors

Take $\hat{\mu}_i = \frac{1}{k} \sum_{j \in N_k(i)} y_l$

$$\hat{\boldsymbol{\mu}} = A\mathbf{y}$$

where

$$\begin{aligned} A_{il} &= \frac{1}{k} && \text{if } l \in N_k(i) \\ &= 0 && \text{otherwise} \end{aligned}$$

Some estimators of the form $\hat{\mu} = A\mathbf{y}$

Exercice
Ridge solves

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \|X\beta - \mathbf{y}\|^2 + \lambda\|\beta\|^2 \tag{7}$$

1. Prove that $X\hat{\beta} = A_{ridge}(\lambda)\mathbf{y}$.
2. Assume we observe one sample $\mathbf{y} = X\beta^* + \epsilon$ with $\epsilon \sim \mathcal{N}(0, \sigma^2 I_n)$. What is the link with the Gaussian sequence model ?

Bias variance decomposition for linear estimators

$\mathbf{y} \in \mathbb{R}^n, \mathbf{y} = \boldsymbol{\mu} + \boldsymbol{\epsilon}$

$\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2 I_n)$

Take an estimator of the form $\hat{\boldsymbol{\mu}} = A\mathbf{y}$ (A deterministic)

Bias variance decomposition for linear estimators

$\mathbf{y} \in \mathbb{R}^n, \mathbf{y} = \boldsymbol{\mu} + \boldsymbol{\epsilon}$
$\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2 I_n)$
Take an estimator of the form $\hat{\boldsymbol{\mu}} = A\mathbf{y}$ (A deterministic)

$$\mathbb{E}[\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2]$$

Bias variance decomposition for linear estimators

$\mathbf{y} \in \mathbb{R}^n, \mathbf{y} = \boldsymbol{\mu} + \boldsymbol{\epsilon}$
$\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2 I_n)$
Take an estimator of the form $\hat{\boldsymbol{\mu}} = A\mathbf{y}$ (A deterministic)

$$\mathbb{E}[\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2] = \|\mathbb{E}[\hat{\boldsymbol{\mu}}] - \boldsymbol{\mu}\|^2 + \mathbb{E}[\|\hat{\boldsymbol{\mu}} - \mathbb{E}[\hat{\boldsymbol{\mu}}]\|^2] \quad (8)$$

Bias variance decomposition for linear estimators

$\mathbf{y} \in \mathbb{R}^n, \mathbf{y} = \boldsymbol{\mu} + \boldsymbol{\epsilon}$

$\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2 I_n)$

Take an estimator of the form $\hat{\boldsymbol{\mu}} = A\mathbf{y}$ (A deterministic)

$$\mathbb{E}[\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2] = \|\mathbb{E}[\hat{\boldsymbol{\mu}}] - \boldsymbol{\mu}\|^2 + \mathbb{E}[\|\hat{\boldsymbol{\mu}} - \mathbb{E}[\hat{\boldsymbol{\mu}}]\|^2] \quad (8)$$

$$= \|(A - I_n)\boldsymbol{\mu}\|^2 + \mathbb{E}[\|A\boldsymbol{\epsilon}\|^2] \quad (9)$$

Bias variance decomposition for linear estimators

$\mathbf{y} \in \mathbb{R}^n, \mathbf{y} = \boldsymbol{\mu} + \boldsymbol{\epsilon}$

$\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2 I_n)$

Take an estimator of the form $\hat{\boldsymbol{\mu}} = A\mathbf{y}$ (A deterministic)

$$\mathbb{E}[\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2] = \|\mathbb{E}[\hat{\boldsymbol{\mu}}] - \boldsymbol{\mu}\|^2 + \mathbb{E}[\|\hat{\boldsymbol{\mu}} - \mathbb{E}[\hat{\boldsymbol{\mu}}]\|^2] \quad (8)$$

$$= \|(A - I_n)\boldsymbol{\mu}\|^2 + \mathbb{E}[\|A\boldsymbol{\epsilon}\|^2] \quad (9)$$

$$= \|(A - I_n)\boldsymbol{\mu}\|^2 + \sigma^2 \operatorname{tr}(A^\top A) \quad (10)$$

Bias variance decomposition for linear estimators

$\mathbf{y} \in \mathbb{R}^n, \mathbf{y} = \boldsymbol{\mu} + \boldsymbol{\epsilon}$

$\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2 I_n)$

Take an estimator of the form $\hat{\boldsymbol{\mu}} = A\mathbf{y}$ (A deterministic)

$$\mathbb{E}[\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2] = \|\mathbb{E}[\hat{\boldsymbol{\mu}}] - \boldsymbol{\mu}\|^2 + \mathbb{E}[\|\hat{\boldsymbol{\mu}} - \mathbb{E}[\hat{\boldsymbol{\mu}}]\|^2] \tag{8}$$

$$= \|(A - I_n)\boldsymbol{\mu}\|^2 + \mathbb{E}[\|A\boldsymbol{\epsilon}\|^2] \tag{9}$$

$$= \|(A - I_n)\boldsymbol{\mu}\|^2 + \sigma^2 \operatorname{tr}(A^\top A) \tag{10}$$

Problem We don't know $\boldsymbol{\mu}$

Akaike Information Criterion

$\mathbf{y} \in \mathbb{R}^n, \mathbf{y} = \boldsymbol{\mu} + \boldsymbol{\epsilon}$

$\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2 I_n)$

$\hat{\boldsymbol{\mu}} = A\mathbf{y}$

Theorem

$\mathbb{E}[\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2] = \mathbb{E}[\hat{r}(A)]$

with $\hat{r}(A) = \|\hat{\boldsymbol{\mu}} - \mathbf{y}\|^2 + 2\sigma^2 \operatorname{tr}(A) - \sigma^2 n$

Akaike Information Criterion

$\mathbf{y} \in \mathbb{R}^n, \mathbf{y} = \boldsymbol{\mu} + \boldsymbol{\epsilon}$
$\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2 I_n)$
$\hat{\boldsymbol{\mu}} = A\mathbf{y}$

Theorem
$\mathbb{E}[\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2] = \mathbb{E}[\hat{r}(A)]$
with $\hat{r}(A) = \|\hat{\boldsymbol{\mu}} - \mathbf{y}\|^2 + 2\sigma^2 \operatorname{tr}(A) - \sigma^2 n$

Proof.
We have

$$\begin{aligned}\|\hat{\boldsymbol{\mu}} - \mathbf{y}\|^2 &= \|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu} - \boldsymbol{\epsilon}\|^2 \\ &= \|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2 - 2\boldsymbol{\epsilon}^\top(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}) + \|\boldsymbol{\epsilon}\|^2\end{aligned}$$

□

Akaike Information Criterion

$\mathbf{y} \in \mathbb{R}^n, \mathbf{y} = \boldsymbol{\mu} + \boldsymbol{\epsilon}$

$\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2 I_n)$

$\hat{\boldsymbol{\mu}} = A\mathbf{y}$

Theorem

$\mathbb{E}[\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2] = \mathbb{E}[\hat{r}(A)]$
with $\hat{r}(A) = \|\hat{\boldsymbol{\mu}} - \mathbf{y}\|^2 + 2\sigma^2 \operatorname{tr}(A) - \sigma^2 n$

Proof.
We have

$$\|\hat{\boldsymbol{\mu}} - \mathbf{y}\|^2 = \|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu} - \boldsymbol{\epsilon}\|^2$$
$$= \|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2 - 2\boldsymbol{\epsilon}^\top(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}) + \|\boldsymbol{\epsilon}\|^2$$

$$\mathbb{E}[\|\boldsymbol{\epsilon}\|^2] = \sigma^2 n$$

$\mathbb{E}[\boldsymbol{\epsilon}^\top(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})] = \mathbb{E}[\boldsymbol{\epsilon}^\top A \boldsymbol{\epsilon}] = \operatorname{tr}(A\mathbb{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top]) = \sigma^2 \operatorname{tr}(A)$ □

Applications of Akaike information criterion

$\mathbf{y} \in \mathbb{R}^n, \mathbf{y} = \boldsymbol{\mu} + \boldsymbol{\epsilon}$
$\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2 I_n)$
$\hat{\boldsymbol{\mu}}_s = A_{ridge}(\lambda)\mathbf{y}$

How should we choose λ ?

Take $\lambda = \hat{\lambda} = \mathrm{argmin}_\lambda\, \hat{r}(A_{ridge}(\lambda))$

Applications of Akaike information criterion

$\mathbf{y} \in \mathbb{R}^n, \mathbf{y} = \boldsymbol{\mu} + \boldsymbol{\epsilon}$
$\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2 I_n)$
$\hat{\boldsymbol{\mu}}_s = s\mathbf{y}$

How should we choose s ?

Take $s = \hat{s} = \operatorname{argmin}_s \hat{r}(sI)$

Exercice
Find $\hat{s}$

Applications of Akaike information criterion

$\mathbf{y} \in \mathbb{R}^n, \mathbf{y} = \boldsymbol{\mu} + \boldsymbol{\epsilon}$
$\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2 I_n)$
$\hat{\boldsymbol{\mu}}_s = s\mathbf{y}$
Take $s = \hat{s} = (1 - \frac{\sigma^2 n}{\|\mathbf{y}\|^2})$.

Applications of Akaike information criterion

$\mathbf{y} \in \mathbb{R}^n, \mathbf{y} = \boldsymbol{\mu} + \boldsymbol{\epsilon}$

$\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2 I_n)$

$\hat{\boldsymbol{\mu}}_s = s\mathbf{y}$

Take $s = \hat{s} = (1 - \frac{\sigma^2 n}{\|\mathbf{y}\|^2})$.

What comes next ?

- Show that $\hat{\boldsymbol{\mu}} = (1 - \frac{\sigma^2 n}{\|\mathbf{y}\|^2})$ is a better estimator than MLE when $n \geq 5$
- Show that $\hat{\boldsymbol{\mu}} = (1 - \frac{\sigma^2 (n-2)}{\|\mathbf{y}\|^2})$ is a better estimator than MLE when $n \geq 3$

Take home message

What we have done today

- Inadmissible estimator = there exists an estimator with a lower $\mathbb{E}[\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2$ for all $\boldsymbol{\mu}$
- Take $\hat{\boldsymbol{\mu}} = A\mathbf{y}$
- Show $\mathbb{E}[\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2 = \mathbb{E}[\hat{r}(A)]$ for some computable $\hat{r}$
- Take $A = sI$ and compute $\hat{s} = \mathrm{argmin}_s\, \hat{r}(sI)$