

# Lecture 3. Inadmissibility of Maximum Likelihood Estimate and James-Stein Estimator

Yuan Yao

Hong Kong University of Science and Technology

# Outline

Recall: PCA in Noise

Maximum Likelihood Estimate

Example: Multivariate Normal Distribution

James-Stein Estimator

Risk and Bias-Variance Decomposition

Inadmissability

James-Stein Estimators

Stein's Unbiased Risk Estimates (SURE)

Proof of SURE Lemma

## PCA in Noise

- ▶ Data:  $x_i \in \mathbb{R}^p$ ,  $i = 1, \dots, n$
- ▶ PCA looks for Eigen-Value Decomposition (EVD) of sample covariance matrix:

$$\hat{\Sigma}_n = \frac{1}{n} \sum_{i=1}^n (x_i - \hat{\mu}_n)(x_i - \hat{\mu}_n)^T$$

where

$$\hat{\mu}_n = \frac{1}{n} \sum_{i=1}^n x_i$$

- ▶ Geometric view as the best affine space approximation of data
- ▶ What about statistical view when  $x_i = \mu + \varepsilon_i$ ?

## Recall: Phase Transitions of PCA

For rank-1 signal-noise model

$$X = \alpha u + \varepsilon, \quad \alpha \sim \mathcal{N}(0, \sigma_X^2), \quad \varepsilon \sim \mathcal{N}(0, I_p)$$

PCA undergoes a phase transition if  $p/n \rightarrow \gamma$ :

- ▶ The primary eigenvalue of sample covariance matrix satisfies

$$\lambda_{\max}(\hat{\Sigma}_n) \rightarrow \begin{cases} (1 + \sqrt{\gamma})^2 = b, & \sigma_X^2 \leq \sqrt{\gamma} \\ (1 + \sigma_X^2)(1 + \frac{\gamma}{\sigma_X^2}), & \sigma_X^2 > \sqrt{\gamma} \end{cases} \quad (1)$$

- ▶ The primary eigenvector converges to

$$|\langle u, v_{\max} \rangle|^2 \rightarrow \begin{cases} 0 & \sigma_X^2 \leq \sqrt{\gamma} \\ \frac{1 - \frac{\gamma}{\sigma_X^2}}{1 + \frac{\gamma}{\sigma_X^2}}, & \sigma_X^2 > \sqrt{\gamma} \end{cases} \quad (2)$$

## Recall: Phase Transitions of PCA

- ▶ Here the threshold

$$\gamma = \lim_{n,p \rightarrow \infty} \frac{p}{n}$$

- ▶ The **law of large numbers** in traditional statistics assumes  $p$  fixed and  $n \rightarrow \infty$ :

$$\gamma = \lim_{n \rightarrow \infty} p/n = 0.$$

where PCA always works without phase transitions.

- ▶ In **high dimensional statistics**, we allow both  $p$  and  $n$  grow:  $p, n \rightarrow \infty$ , not law of large numbers.
- ▶ What might go wrong? Even the sample mean  $\hat{\mu}_n$ !

## In this lecture

- ▶ Sample mean  $\hat{\mu}_n$  and covariance  $\hat{\Sigma}_n$  are both Maximum Likelihood Estimate (MLE) under Gaussian noise models
- ▶ In high dimensional scenarios (small  $n$ , large  $p$ ), MLE  $\hat{\mu}_n$  is not optimal:
  - Inadmissability: MLE has worse prediction power than [James-Stein Estimator \(JSE\)](#) (Stein, 1956)
  - Many [shrinkage](#) estimates are better than MLE and James-Stein Estimator (JSE)
- ▶ Therefore, penalized likelihood or regularization is necessary in high dimensional statistics

# Outline

Recall: PCA in Noise

## Maximum Likelihood Estimate

Example: Multivariate Normal Distribution

## James-Stein Estimator

Risk and Bias-Variance Decomposition

Inadmissability

James-Stein Estimators

Stein's Unbiased Risk Estimates (SURE)

Proof of SURE Lemma

## Maximum Likelihood Estimate

- ▶ Statistical model  $f(X|\theta)$  as a conditional probability function on  $\mathbb{R}^p$  with parameter space  $\theta \in \Theta$
- ▶ The likelihood function is defined as the probability of observing the given data  $x_i \sim f(X|\theta)$  as a function of  $\theta$ ,

$$\mathcal{L}(\theta) = \prod_{i=1}^n f(x_i|\theta)$$

- ▶ A Maximum Likelihood Estimator is defined as

$$\begin{aligned} \hat{\theta}_n^{MLE} &\in \arg \max_{\theta \in \Theta} \mathcal{L}(\theta) = \arg \max_{\theta \in \Theta} \prod_{i=1}^n f(x_i|\theta) \\ &= \arg \max_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^n \log f(x_i|\theta). \end{aligned} \quad (3)$$



## Maximum Likelihood Estimate

- ▶ For example, consider the normal distribution  $\mathcal{N}(\mu, \Sigma)$ ,

$$f(X|\mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^p |\Sigma|}} \exp \left[ -\frac{1}{2} (X - \mu)^T \Sigma^{-1} (X - \mu) \right],$$

where  $|\Sigma|$  is the determinant of covariance matrix  $\Sigma$ .

- ▶ Take independent and identically distributed (i.i.d.) samples  $x_i \sim \mathcal{N}(\mu, \Sigma)$  ( $i = 1, \dots, n$ )

## Maximum Likelihood Estimate (continued)

- ▶ To get the MLE given  $x_i \sim \mathcal{N}(\mu, \Sigma)$  ( $i = 1, \dots, n$ ), solve

$$\max_{\mu, \Sigma} \prod_{i=1}^n f(x_i | \mu, \Sigma) = \max_{\mu, \Sigma} \prod_{i=1}^n \frac{1}{\sqrt{2\pi|\Sigma|}} \exp[-(X_i - \mu)^T \Sigma^{-1} (X_i - \mu)]$$

- ▶ Equivalently, consider the logarithmic likelihood

$$\begin{aligned} J(\mu, \Sigma) &= \log \prod_{i=1}^n f(x_i | \mu, \Sigma) \\ &= -\frac{1}{2} \sum_{i=1}^n (X_i - \mu)^T \Sigma^{-1} (X_i - \mu) - \frac{n}{2} \log |\Sigma| + C \end{aligned}$$

where  $C$  is a constant independent to parameters

## MLE: sample mean $\hat{\mu}_n$

- ▶ To solve  $\mu$ , the log-likelihood is a quadratic function of  $\mu$ ,

$$0 = \left. \frac{\partial J}{\partial \mu} \right|_{\mu=\mu^*} = - \sum_{i=1}^n \Sigma^{-1} (x_i - \mu^*)$$

$$\Rightarrow \mu^* = \frac{1}{n} \sum_{i=1}^n x_i =: \hat{\mu}_n$$

## MLE: sample covariance $\hat{\Sigma}_n$

- ▶ To solve  $\Sigma$ , the first term in (4)

$$\begin{aligned} & -\frac{1}{2} \sum_{i=1}^n \mathbf{Tr}(x_i - \mu)^T \Sigma^{-1} (x_i - \mu) \\ = & -\frac{1}{2} \sum_{i=1}^n \mathbf{Tr}[\Sigma^{-1} (x_i - \mu)(x_i - \mu)^T], \quad \mathbf{Tr}(ABC) = \mathbf{Tr}(BCA) \\ = & -\frac{n}{2} (\mathbf{Tr} \Sigma^{-1} \hat{\Sigma}_n), \quad \hat{\Sigma}_n := \frac{1}{n} \sum_{i=1}^n (x_i - \hat{\mu}_n)(x_i - \hat{\mu}_n)^T, \\ = & -\frac{n}{2} \mathbf{Tr}(\Sigma^{-1} \hat{\Sigma}_n^{\frac{1}{2}} \hat{\Sigma}_n^{\frac{1}{2}}) \\ = & -\frac{n}{2} \mathbf{Tr}(\hat{\Sigma}_n^{\frac{1}{2}} \Sigma^{-1} \hat{\Sigma}_n^{\frac{1}{2}}), \quad \mathbf{Tr}(ABC) = \mathbf{Tr}(BCA) \\ = & -\frac{n}{2} \mathbf{Tr}(S), \quad S := \hat{\Sigma}_n^{\frac{1}{2}} \Sigma^{-1} \hat{\Sigma}_n^{\frac{1}{2}} \end{aligned}$$

## MLE: sample covariance $\hat{\Sigma}_n$

Use  $S$  to represent  $\Sigma$ :

- ▶ Notice that

$$\Sigma = \hat{\Sigma}_n^{\frac{1}{2}} S^{-1} \hat{\Sigma}_n^{\frac{1}{2}}$$
$$\Rightarrow -\frac{n}{2} \log |\Sigma| = \frac{n}{2} \log |S| - \frac{n}{2} \log |\hat{\Sigma}_n| = f(\hat{\Sigma}_n)$$

where we use for determinant of squared matrices of equal size,  
 $\det(AB) = |AB| = \det(A) \det(B) = |A| \cdot |B|$ .

- ▶ Therefore,

$$\max_{\Sigma} J(\Sigma) \Leftrightarrow \min_S \frac{n}{2} \mathbf{Tr}(S) - \frac{n}{2} \log |S| + \text{Const}(\hat{\Sigma}_n, 1)$$

## MLE: sample covariance $\hat{\Sigma}_n$

- ▶ Since  $S = \hat{\Sigma}_n^{\frac{1}{2}} \Sigma^{-1} \hat{\Sigma}_n^{\frac{1}{2}}$  is symmetric and positive semidefinite, let  $S = U \Lambda U^T$  be its eigenvalue decomposition,  $\Lambda = \mathbf{diag}(\lambda_i)$  with  $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p \geq 0$ . Then we have

$$J(\lambda_i) = \frac{n}{2} \sum_{i=1}^p \lambda_i - \frac{n}{2} \sum_{i=1}^p \log(\lambda_i) + Const$$

$$\Rightarrow 0 = \left. \frac{\partial J}{\partial \lambda_i} \right|_{\lambda_i^*} = \frac{n}{2} - \frac{n}{2} \frac{1}{\lambda_i^*} \Rightarrow \lambda_i^* = 1$$

$$\Rightarrow S^* = I_p$$

- ▶ Hence the MLE solution

$$\Sigma^* = \hat{\Sigma}_n = \frac{1}{n} \sum_{i=1}^n (X_i - \hat{\mu}_n)(X_i - \hat{\mu}_n)^T,$$

## Note

- ▶ In statistics, it is often defined

$$\hat{\Sigma}_n = \frac{1}{n-1} \sum_{i=1}^n (X_i - \hat{\mu}_n)(X_i - \hat{\mu}_n)^T,$$

where the denominator is  $(n - 1)$  instead of  $n$ . This is because that for sample covariance matrix, a single sample  $n = 1$  leads to no variance at all.

## Consistency of MLE

Under some regularity conditions, the maximum likelihood estimator  $\hat{\theta}_n^{MLE}$  has the following nice *limit* properties for fixed  $p$  and  $n \rightarrow \infty$ :

A. (Consistency)  $\hat{\theta}_n^{MLE} \rightarrow \theta_0$ , in probability and almost surely.

B. (Asymptotic Normality)  $\sqrt{n}(\hat{\theta}_n^{MLE} - \theta_0) \rightarrow \mathcal{N}(0, I_0^{-1})$  in distribution, where  $I_0$  is the Fisher Information matrix

$$I(\theta_0) := \mathbf{E}\left[\left(\frac{\partial}{\partial \theta} \log f(X|\theta_0)\right)^2\right] = -\mathbf{E}\left[\frac{\partial^2}{\partial \theta^2} \log f(X|\theta_0)\right].$$

C. (Asymptotic Efficiency)  $\lim_{n \rightarrow \infty} \text{cov}(\hat{\theta}_n^{MLE}) = I^{-1}(\theta_0)$ . Hence  $\hat{\theta}_n^{MLE}$  is the **Uniformly Minimum-Variance Unbiased Estimator**, i.e. the estimator with the least variance among the class of unbiased estimators, for any unbiased estimator  $\hat{\theta}_n$ ,  
 $\lim_{n \rightarrow \infty} \text{var}(\hat{\theta}_n^{MLE}) \leq \lim_{n \rightarrow \infty} \text{var}(\hat{\theta}_n)$ .



## However, large $p$ small $n$ ?

- ▶ The asymptotic results all hold under the assumption by **fixing  $p$  and taking  $n \rightarrow \infty$** , where MLE satisfies  $\hat{\mu}_n \rightarrow \mu$  and  $\hat{\Sigma}_n \rightarrow \Sigma$ .
- ▶ However, when  **$p$  becomes large compared to finite  $n$** ,  $\hat{\mu}_n$  is not the best estimator for *prediction* measured by expected mean squared error from the truth, to be shown below.

# Outline

Recall: PCA in Noise

Maximum Likelihood Estimate

Example: Multivariate Normal Distribution

**James-Stein Estimator**

Risk and Bias-Variance Decomposition

Inadmissability

James-Stein Estimators

Stein's Unbiased Risk Estimates (SURE)

Proof of SURE Lemma

## Prediction Error and Risk

- ▶ To measure the *prediction* performance of an estimator  $\hat{\mu}_n$ , it is natural to consider the expected squared loss in regression, i.e. given a response  $y = \mu + \epsilon$  with zero mean noise  $\mathbf{E}[\epsilon] = 0$ ,

$$\mathbf{E} \|y - \hat{\mu}_n\|^2 = \mathbf{E} \|\mu - \hat{\mu} + \epsilon\|^2 = \mathbf{E} \|\mu - \hat{\mu}\|^2 + \mathbf{Var}(\epsilon), \quad \mathbf{Var}(\epsilon) = \mathbf{E}(\epsilon^T \epsilon).$$

- ▶ Since  $\mathbf{Var}(\epsilon)$  is a constant for all estimators  $\hat{\mu}$ , one may simply look at the first part which is often called as *risk* in literature,

$$\mathcal{R}(\hat{\mu}, \mu) = \mathbf{E} \|\mu - \hat{\mu}\|^2$$

It is the *mean square error* (MSE) between  $\mu$  and its estimator  $\hat{\mu}$ , that measures the expected prediction error.

## Bias-Variance Decomposition

- ▶ The risk or MSE enjoy the following important *bias-variance decomposition*, as a result of the Pythagorean theorem.

$$\begin{aligned}\mathcal{R}(\hat{\mu}_n, \mu) &= \mathbf{E} \|\hat{\mu}_n - \mathbf{E}[\hat{\mu}_n] + \mathbf{E}[\hat{\mu}_n] - \mu\|^2 \\ &= \mathbf{E} \|\hat{\mu}_n - \mathbf{E}[\hat{\mu}_n]\|^2 + \|\mathbf{E}[\hat{\mu}_n] - \mu\|^2 \\ &=: \mathbf{Var}(\hat{\mu}_n) + \mathbf{Bias}(\hat{\mu}_n)^2\end{aligned}$$

- ▶ Consider multivariate Gaussian model,  $x_1, \dots, x_n \sim \mathcal{N}(\mu, \sigma^2 I_p)$  ( $i = 1, \dots, n$ ), and the maximum likelihood estimators (MLE) of the parameters ( $\mu$  and  $\Sigma = \sigma^2 I_p$ )

## Example: Bias-Variance Decomposition of MLE

- ▶ Consider multivariate Gaussian model,  $Y_1, \dots, Y_n \sim \mathcal{N}(\mu, \sigma^2 I_p)$  ( $i = 1, \dots, n$ ), and the maximum likelihood estimators (MLE) of the parameters ( $\mu$  and  $\Sigma = \sigma^2 I_p$ )
- ▶ The MLE estimator satisfies

$$\mathbf{Bias}(\hat{\mu}_n^{MLE}) = 0$$

and

$$\mathbf{Var}(\hat{\mu}_n^{MLE}) = \frac{p}{n} \sigma^2$$

In particular for  $n = 1$ ,  $\mathbf{Var}(\hat{\mu}^{MLE}) = \sigma^2 p$  for  $\hat{\mu}^{MLE} = Y$ .

## Example: Bias-Variance Decomposition of Linear Estimators

- ▶ Consider  $Y \sim \mathcal{N}(\mu, \sigma^2 I_p)$  and linear estimator  $\hat{\mu}_C = CY$

- ▶ Then we have

$$\mathbf{Bias}(\hat{\mu}_C) = \|(I - C)\mu\|^2$$

and

$$\begin{aligned}\mathbf{Var}(\hat{\mu}_C) &= \mathbf{E}[(CY - C\mu)^T(CY - C\mu)] \\ &= \mathbf{E}[\text{tr}((Y - \mu)^T C^T C(Y - \mu))] \\ &= \sigma^2 \text{tr}(C^T C).\end{aligned}$$

- ▶ Linear estimator includes an important case, the *Ridge regression* (a.k.a. Tikhonov regularization) with  $C = X(X^T X + \lambda I)^{-1} X^T$ ,

$$\min_{\beta} \frac{1}{2} \|Y - X\beta\|^2 + \frac{\lambda}{2} \|\beta\|^2, \quad \lambda > 0.$$

## Example: Bias-Variance Decomposition of Diagonal Estimators

- ▶ For simplicity, one may restrict our discussions on the diagonal linear estimators  $C = \mathbf{diag}(c_i)$  (up to an change of orthonormal basis for Ridge regression), whose risk is

$$\mathcal{R}(\hat{\mu}_C, \mu) = \sigma^2 \sum_{i=1}^p c_i^2 + \sum_{i=1}^p (1 - c_i)^2 \mu_i^2.$$

- ▶ For hyper-rectangular model class  $|\mu_i| \leq \tau_i$ , the minimax risk is

$$\inf_{c_i} \sup_{|\mu_i| \leq \tau_i} \mathcal{R}(\hat{\mu}_C, \mu) = \sum_{i=1}^p \frac{\sigma^2 \tau_i^2}{\sigma^2 + \tau_i^2}.$$

For sparse models such that  $\#\{i : \tau_i = O(\sigma)\} = k \ll p$ , it is possible to trade bias with variance toward a **smaller risk using linear estimators than MLE!**

## Note

$$\mathcal{R}(\hat{\mu}_C, \mu) = \sigma^2 \sum_{i=1}^p c_i^2 + \sum_{i=1}^p (1 - c_i)^2 \mu_i^2.$$

- ▶ For the supreme over  $|\mu_i| \leq \tau_i$ ,

$$\Rightarrow \sup_{|\mu_i| \leq \tau_i} \mathcal{R}(\hat{\mu}_C, \mu) = \sigma^2 \sum_{i=1}^p c_i^2 + \sum_{i=1}^p (1 - c_i)^2 \tau_i^2 =: J(c).$$

- ▶ To see the infimum over  $c_i$ ,

$$0 = \frac{\partial J(c)}{\partial c_i} = 2\sigma^2 c_i - 2\tau_i^2(1 - c_i) \Rightarrow c_i = \frac{\tau_i^2}{\sigma^2 + \tau_i^2}$$

- ▶ The minimax risk is thus

$$\inf_{c_i} \sup_{|\mu_i| \leq \tau_i} \mathcal{R}(\hat{\mu}_C, \mu) = \sum_{i=1}^p \frac{\sigma^2 \tau_i^2}{\sigma^2 + \tau_i^2}.$$



## Formality: Inadmissibility

### Definition (Inadmissible, Charles Stein (1956))

An estimator  $\hat{\mu}_n$  of the parameter  $\mu$  is called **inadmissible** on  $\mathbb{R}^p$  with respect to the squared risk if there exists another estimator  $\mu_n^*$  such that

$$\mathbf{E} \|\mu_n^* - \mu\|^2 \leq \mathbf{E} \|\hat{\mu}_n - \mu\|^2 \quad \text{for all } \mu \in \mathbb{R}^p,$$

and there exist  $\mu_0 \in \mathbb{R}^p$  such that

$$\mathbf{E} \|\mu_n^* - \mu_0\|^2 < \mathbf{E} \|\hat{\mu}_n - \mu_0\|^2.$$

In this case, we also call that  $\mu_n^*$  **dominates**  $\hat{\mu}_n$ . Otherwise, the estimator  $\hat{\mu}_n$  is called **admissible**.

## Stein's Phenomenon

- ▶ (Charles Stein (1956)) For  $p \geq 3$ , there exists  $\hat{\mu}$  such that  $\forall \mu \in \mathbb{R}^p$ ,

$$\mathcal{R}(\hat{\mu}, \mu) < \mathcal{R}(\hat{\mu}^{\text{MLE}}, \mu)$$

which makes MLE inadmissible.

- ▶ What are such estimators?

## James-Stein Estimator

### Example (James-Stein Estimator)

$$\hat{\mu}^{JS} = \left(1 - \frac{\sigma^2(p-2)}{\|\hat{\mu}^{MLE}\|^2}\right) \hat{\mu}^{MLE}. \quad (4)$$

Such an estimator shrinks each component of  $\hat{\mu}^{MLE}$  toward 0.

- ▶ Charles Stein shows in 1956 that MLE is inadmissible, while the following original form of James-Stein estimator is demonstrated by his student Willard James in 1961.
- ▶ Bradley Efron summarizes the history and gives a simple derivation of these estimators from an *Empirical Bayes* point of view, while we shall give a *ridge regression* derivation.

## James-Stein Estimator with Shrinkage toward Mean

- ▶ A varied form of James-Stein estimator can shrink MLE toward other points such as the component mean of  $\hat{\mu}^{MLE}$ :

$$\hat{\mu}_i^{JS_1} = \bar{X} + \left(1 - \frac{\sigma^2(p-3)}{S(\hat{\mu}^{MLE})}\right) (\hat{\mu}_i^{MLE} - \bar{X}), \quad (5)$$

where  $\bar{X} = \sum_{i=1}^p X_i/p$  and  $S(X) = \sum_i (X_i - \bar{X})^2$ ,

- ▶ *Positive part James-Stein estimator:*

$$\tilde{\mu}^{JS_{1+}} := \bar{X} + \left(1 - \frac{\sigma^2(p-3)}{S(\hat{\mu}^{MLE})}\right)_+ (\hat{\mu}_i^{MLE} - \bar{X}), \quad (x)_+ := \min(0, x)$$

- ▶ Both dominate MLE if  $p > 3$  and can be derived from ridge regression.

## James-Stein Estimator as Multi-task Ridge Regression

James-Stein estimators can be written as a *multi-task ridge regression*:

$$(\hat{\mu}_i, \hat{\mu}) := \arg \min_{\mu_i, \mu} \sum_{i=1}^p [(\mu_i - X_i)^2 + \lambda(\mu_i - \mu)^2]. \quad (6)$$

- ▶ Denote  $\bar{X} = \sum_{i=1}^p X_i/p$  and  $S(X) = \sum_i (X_i - \bar{X})^2$
- ▶ Taking  $\lambda = \sigma^2(p-3)/(S - \sigma^2(p-3))$ ,  $\hat{\mu}_i$  gives  $\hat{\mu}^{JS_1}$ ;
- ▶ Taking  $\lambda = \min(S, \sigma^2(p-3))/(S - \min(S, \sigma^2(p-3)))$  with  $1/0 = \infty$ , it gives  $\hat{\mu}^{JS_{1+}}$ .

## Proof sketch

Consider

$$J(\mu_i, \mu) = \sum_{i=1}^p [(\mu_i - X_i)^2 + \lambda(\mu_i - \mu)^2]$$

▶  $\partial J / \partial \mu = 0$  we get

$$2\lambda \sum_{i=1}^p (\hat{\mu} - \hat{\mu}_i) = 0 \Rightarrow \hat{\mu} = \frac{1}{p} \sum_{i=1}^p \hat{\mu}_i$$

▶  $0 = \partial J / \partial \mu_i = 2(\hat{\mu}_i - X_i) + 2\lambda(\hat{\mu}_i - \hat{\mu}) = 0$

$$\Rightarrow \hat{\mu}_i = \frac{1}{1 + \lambda} (X_i + \lambda \hat{\mu}) = \hat{\mu} + \frac{1}{1 + \lambda} (X_i - \hat{\mu})$$

whose average over  $i$  gives

$$\frac{1}{p} \sum_{i=1}^p \hat{\mu}_i = \hat{\mu} = \hat{\mu} + \frac{1}{(1 + \lambda)} \left( \frac{1}{p} \sum_{i=1}^p X_i - \hat{\mu} \right) \Rightarrow \hat{\mu} = \frac{1}{p} \sum_{i=1}^p X_i = \bar{X}$$

## Proof sketch

Now

$$\hat{\mu}_i = \bar{X} + \frac{1}{1+\lambda} (X_i - \bar{X}) = \bar{X} + \left(1 - \frac{\lambda}{1+\lambda}\right) (X_i - \bar{X})$$

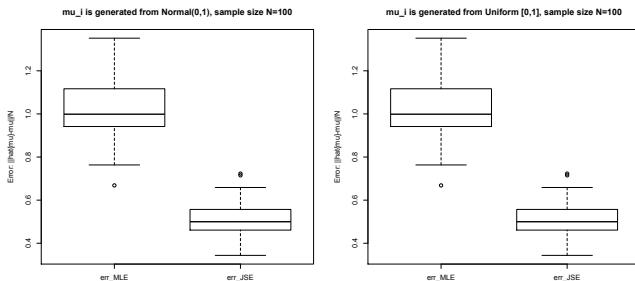
- ▶ Taking  $\lambda = \sigma^2(p-3)/(S - \sigma^2(p-3))$ ,  $\frac{\lambda}{1+\lambda} = \frac{\sigma^2(p-3)}{S}$  ( $\hat{\mu}_{JS_1}$ )
- ▶ Taking  $\lambda = \min(S, \sigma^2(p-3))/(S - \min(S, \sigma^2(p-3)))$ ,
  - if  $S > \sigma^2(p-3)$ , the same as above
  - if  $S \leq \sigma^2(p-3)$ ,  $\lambda = S/(S - S)$  and  $\frac{\lambda}{1+\lambda} = \frac{S}{S-S+S} = 1$  ( $\hat{\mu}_{JS_{1+}}$ )
- ▶ Note: taking  $\hat{\mu} = \bar{X} = 0$  and  $\lambda = \sigma^2(p-2)/(S - \sigma^2(p-2))$ , it gives James-Stein shrinkage toward 0 ( $\hat{\mu}_{JS_0}$ ). □

## Example

- ▶ Let's look at an example of James-Stein Estimator
  - R: [https://github.com/yuany-pku/2017\\_CSIC5011/blob/master/slides/JSE.R](https://github.com/yuany-pku/2017_CSIC5011/blob/master/slides/JSE.R)



## Illustration that JSE dominates MLE



**Figure:** Comparison of risks between Maximum Likelihood Estimators and James-Stein Estimators with  $X_i \sim \mathcal{N}(0, I_p)$  (left) and  $X_{ij} \sim \mathcal{U}[0, 1]$  (right) for  $i = 1, \dots, N$  and  $j = 1, \dots, p$  where  $p = N = 100$ .

## Efron's Batting Example in 1970

Table: Efron's Batting example.  $\hat{\mu}^{MLE}$  is obtained from the mean hits in these early games, while  $\mu$  is obtained by averages over the remainder of the season.

Players	hits/AB	$\hat{\mu}_i^{(MLE)}$	$\mu_i$	$\hat{\mu}_i^{(JS)}$	$\hat{\mu}_i^{(JS_1)}$
Clemente	18/45	0.4	<b>0.346</b>	0.378	0.294
F.Robinson	17/45	0.378	<b>0.298</b>	0.357	0.289
F.Howard	16/45	0.356	<b>0.276</b>	0.336	0.285
Johnstone	15/45	0.333	<b>0.222</b>	0.315	0.28
Berry	14/45	0.311	<b>0.273</b>	0.294	0.275
Spencer	14/45	0.311	<b>0.27</b>	0.294	0.275
Kessinger	13/45	0.289	<b>0.263</b>	0.273	0.27
L.Alvarado	12/45	0.267	<b>0.21</b>	0.252	0.266
Santo	11/45	0.244	<b>0.269</b>	0.231	0.261
Swoboda	11/45	0.244	<b>0.23</b>	0.231	0.261
Unser	10/45	0.222	<b>0.264</b>	0.21	0.256
Williams	10/45	0.222	<b>0.256</b>	0.21	0.256
Scott	10/45	0.222	<b>0.303</b>	0.21	0.256
Petrocelli	10/45	0.222	<b>0.264</b>	0.21	0.256
E.Rodriguez	10/45	0.222	<b>0.226</b>	0.21	0.256
Campaneris	9/45	0.2	<b>0.286</b>	0.189	0.252
Munson	8/45	0.178	<b>0.316</b>	0.168	0.247
Alvis	7/45	0.156	<b>0.2</b>	0.147	0.242
Mean Square Error	-	0.075545	-	0.072055	0.021387

## James-Stein Estimator Dominates MLE

### Theorem (James-Stein (1956, 1961))

Suppose  $Y \sim \mathcal{N}_p(\mu, I)$ . Then  $\hat{\mu}^{\text{MLE}} = Y$ .  $\mathcal{R}(\hat{\mu}, \mu) = \mathbf{E}_\mu \|\hat{\mu} - \mu\|^2$ , and define

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

Then if  $p \geq 3$  and for all  $\mu \in \mathbb{R}^p$

$$\mathcal{R}(\hat{\mu}^{JS}, \mu) < \mathcal{R}(\hat{\mu}^{\text{MLE}}, \mu)$$

## More Estimators Dominates MLE

- ▶ *Stein estimator.*  $a = 0, b = \sigma^2 p,$

$$\tilde{\mu}_S := \left(1 - \frac{\sigma^2 p}{\|y\|^2}\right) y$$

- ▶ *James-Stein estimator.*  $c \in (0, 2(p - 2))$

$$\tilde{\mu}_{JS}^c := \left(1 - \frac{\sigma^2 c}{\|y\|^2}\right) y$$

- ▶ *Positive part James-Stein estimator.*

$$\tilde{\mu}_{JS+} := \left(1 - \frac{\sigma^2(p - 2)}{\|y\|^2}\right)_+ y, \quad (x)_+ := \min(0, x)$$

- ▶ *Positive part Stein estimator.*

$$\tilde{\mu}_{S+} := \left(1 - \frac{\sigma^2 p}{\|y\|^2}\right)_+ y$$

$$\mathcal{R}(\tilde{\mu}_{JS+}) < \mathcal{R}(\tilde{\mu}_{JS}) < \mathcal{R}(\hat{\mu}_n), \quad \mathcal{R}(\tilde{\mu}_{S+}) < \mathcal{R}(\tilde{\mu}_S) < \mathcal{R}(\hat{\mu}_n)$$

## Stein's Unbiased Risk Estimates

### Lemma (Stein's Unbiased Risk Estimates (SURE))

Suppose  $Y \sim \mathcal{N}_p(\mu, I)$  and  $\hat{\mu} = Y + g(Y)$ . If  $g$  satisfies

1.  $g$  is weakly differentiable.
2.  $\sum_{i=1}^p \int |\partial_i g_i(x)| dx < \infty$

Denote

$$U(Y) := p + 2\nabla^T g(Y) + \|g(Y)\|^2 \quad (7)$$

Then

$$\mathcal{R}(\hat{\mu}, \mu) = \mathbf{E}U(Y) = \mathbf{E}(p + 2\nabla^T g(Y) + \|g(Y)\|^2) \quad (8)$$

where  $\nabla^T g(Y) := \sum_{i=1}^p \frac{\partial}{\partial y_i} g_i(Y)$ .

## Examples of weakly differentiable $g$

- ▶ For linear estimator  $\hat{\mu} = CY$ ,

$$g(Y) = (C - I)Y$$

- ▶ For James-Stein estimator

$$g(Y) = -\frac{p-2}{\|Y\|^2}Y$$

## Soft-Thresholding

- ▶ Soft-Thresholding solves LASSO ( $\ell_1$ -regularized MLE)

$$\hat{\mu} = \arg \min_{\mu} J_1(\mu) = \arg \min_{\mu} \frac{1}{2} \|Y - \mu\|^2 + \lambda \|\mu\|_1$$

- ▶ Subgradients of objective function leads to

$$\begin{aligned} 0 \in \partial_{\mu_j} J_1(\mu) &= (\mu_j - y_j) + \lambda \mathbf{sign}(\mu_j) \\ \Rightarrow \hat{\mu}_j(y_j) &= \mathbf{sign}(y_j)(|y_j| - \lambda)_+ \end{aligned}$$

where the set-valued map  $\mathbf{sign}(x) = 1$  if  $x > 0$ ,  $\mathbf{sign}(x) = -1$  if  $x < 0$ , and  $\mathbf{sign}(x) = [-1, 1]$  if  $x = 0$ , is the subgradient of absolute function  $|x|$ .

- ▶ Then

$$g_i(x) = \begin{cases} -\lambda & x_i > \lambda \\ -x_i & |x_i| \leq \lambda \\ \lambda & x_i < -\lambda \end{cases}$$

which is weakly differentiable



## Hard-Thresholding, a Counter Example

- ▶ Hard-Thresholding solves the  $\ell_0$ -regularized MLE where  $\|x\|_0 := \#\{x_i \neq 0\}$

$$\hat{\mu} = \arg \min_{\mu} J_0(\mu) = \arg \min_{\mu} \frac{1}{2} \|Y - \mu\|^2 + \lambda \|\mu\|_0$$

that is NP-hard

- ▶ Closed-form solution

$$\hat{\mu}_i(y_i) = \begin{cases} y_i & y_i > \lambda \\ 0 & |y_i| \leq \lambda \\ y_i & y_i < -\lambda \end{cases}$$

- ▶ Then

$$g_i(x) = \begin{cases} 0 & |x_i| > \lambda \\ -x_i & |x_i| \leq \lambda \end{cases}$$

which is **NOT** weakly differentiable!



## Sketchy Proof of SURE Lemma

### Proof.

Assume that  $\sigma = 1$ . Let  $\phi(y)$  be the density function of Gaussian distribution  $\mathcal{N}_p(0, I)$ .

$$\begin{aligned}\mathcal{R}(\hat{\mu}, \mu) &= \mathbf{E}_{\mu} \|Y + g(Y) - \mu\|^2 \\ &= \mathbf{E} (p + 2(Y - \mu)^T g(Y) + \|g(Y)\|^2)\end{aligned}$$

$$\begin{aligned}\mathbf{E}(Y - \mu)^T g(Y) &= \sum_{i=1}^p \int_{-\infty}^{\infty} (y_i - \mu_i) g_i(Y) \phi(Y - \mu) dY \\ &= \sum_{i=1}^p \int_{-\infty}^{\infty} -g_i(Y) \frac{\partial}{\partial y_i} \phi(Y - \mu_i) dY, \quad \text{derivative of } \phi \\ &= \sum_{i=1}^p \int_{-\infty}^{\infty} \frac{\partial}{\partial y_i} g_i(Y) \phi(Y - \mu_i) dY, \quad \text{Integration by parts} \\ &= \mathbf{E} \nabla^T g(Y)\end{aligned}$$



## Risk of Linear Estimator

Suppose  $Y \sim \mathcal{N}(\mu, I_p)$

$$\hat{\mu}_C(Y) = Cy$$

$$\Rightarrow g(Y) = (C - I)Y$$

$$\Rightarrow \nabla^T g(Y) = - \sum_i \frac{\partial}{\partial y_i} ((C - I)Y) = \text{tr}(C) - p$$

$$\begin{aligned}\Rightarrow U(Y) &= p + 2\nabla^T g(Y) + \|g(Y)\|^2 \\ &= p + 2(\text{tr}(C) - p) + \|(I - C)Y\|^2 \\ &= -p + 2\text{tr}(C) + \|(I - C)Y\|^2\end{aligned}$$

In general, if  $Y \sim \mathcal{N}(\mu, \sigma^2 I)$ ,

$$\mathcal{R}(\hat{\mu}_C, \mu) = \|(I - C(\lambda))Y\|^2 - p\sigma^2 + 2\sigma^2 \text{tr}(C(\lambda)).$$

## When Linear Estimator is Admissible?

Theorem (Lemma 2.8 in Johnstone's book (GE))

$Y \sim N(\mu, I)$ ,  $\forall \hat{\mu} = CY$ ,  $\hat{\mu}$  is admissible iff

1.  $C$  is symmetric.
2.  $0 \leq \rho_i(C) \leq 1$  (eigenvalue).
3.  $\rho_i(C) = 1$  for at most two  $i$ .

## Risk of James-Stein Estimator

- ▶ Suppose  $Y \sim \mathcal{N}(\mu, I_p)$  and for  $p \geq 3$ ,

$$\hat{\mu}^{JS} = \left(1 - \frac{p-2}{\|Y\|^2}\right) Y \Rightarrow g(Y) = -\frac{p-2}{\|Y\|^2} Y$$

- ▶ Now

$$U(Y) = p + 2\nabla^T g(Y) + \|g(Y)\|^2$$

$$\|g(Y)\|^2 = \frac{(p-2)^2}{\|Y\|^2}$$

$$\nabla^T g(Y) = -\sum_i \frac{\partial}{\partial y_i} \left( \frac{p-2}{\|Y\|^2} Y \right) = -\frac{(p-2)^2}{\|Y\|^2}$$

- ▶ Finally

$$\Rightarrow \mathcal{R}(\hat{\mu}^{JS}, \mu) = \mathbf{E} U(Y) = p - \mathbf{E} \frac{(p-2)^2}{\|Y\|^2} < p = \mathcal{R}(\hat{\mu}^{MLE}, \mu)$$

## Further Analysis of the Risk of James-Stein Estimator

- ▶ To find an upper bound of the risk of James-Stein estimator, notice that  $\|Y\|^2 \sim \chi^2(\|\mu\|^2, p)$  and <sup>1</sup>

$$\chi^2(\|\mu\|^2, p) \stackrel{d}{=} \chi^2(0, p + 2N), \quad N \sim \text{Poisson}\left(\frac{\|\mu\|^2}{2}\right)$$

we have

$$\begin{aligned} \mathbf{E}\left(\frac{1}{\|Y\|^2}\right) &= \mathbf{E}_N \mathbf{E}_Y \left[ \frac{1}{\|Y\|^2} \middle| N \right] \\ &= \mathbf{E} \frac{1}{p + 2N - 2} \\ &\geq \frac{1}{p + 2 \mathbf{E} N - 2} \quad (\text{Jensen's Inequality}) \\ &= \frac{1}{p + \|\mu\|^2 - 2} \end{aligned}$$

---

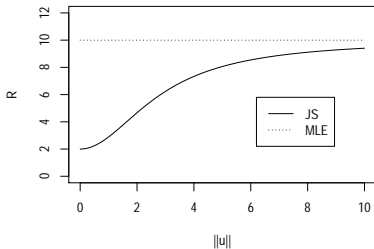
<sup>1</sup>This is a homework.

## Upper Bound for JSE

### Proposition (Upper bound of MSE for JSE)

Let  $Y \sim \mathcal{N}(\mu, I_p)$  for  $p \geq 3$ ,

$$\mathcal{R}(\hat{\mu}^{\text{JS}}, \mu) \leq p - \frac{(p-2)^2}{p-2 + \|\mu\|^2} = 2 + \frac{(p-2)\|\mu\|^2}{p-2 + \|\mu\|^2}$$



## Risk of Soft-Thresholding

► Recall

$$g_i(x) = \begin{cases} -\lambda & x_i > \lambda \\ -x_i & |x_i| \leq \lambda \\ \lambda & x_i < -\lambda \end{cases} \Rightarrow \frac{\partial}{\partial x_i} g_i(x) = -I(|x_i| \leq \lambda)$$

► Then

$$\begin{aligned} \mathcal{R}(\hat{\mu}_\lambda, \mu) &= \mathbf{E}(p + 2\nabla^T g(Y) + \|g(Y)\|^2) \\ &= \mathbf{E} \left( p - 2 \sum_{i=1}^p I(|y_i| \leq \lambda) + \sum_{i=1}^p y_i^2 \wedge \lambda^2 \right) \\ &\leq 1 + (2 \log p + 1) \sum_{i=1}^p \mu_i^2 \wedge 1 \quad \text{if we take } \lambda = \sqrt{2 \log p} \end{aligned}$$

## Risk of Soft-Thresholding (continued)

- ▶ Consider the risk upper bound

$$1 + (2 \log p + 1) \sum_{i=1}^p \mu_i^2 \wedge 1$$

- ▶ The risk of soft-thresholding for each  $\mu_i$  is bounded by 1: so if  $\mu$  is sparse ( $s = \#\{i : \mu_i \neq 0\}$ ) but large in magnitudes (s.t.  $\|\mu\|_2^2 \geq p$ ), we may expect the risk of soft-thresholding  $\leq O(s \log p) \ll O(p)$ , the risk of MLE.



## Summary

The following results are about mean estimation under noise:

- ▶ Sample mean as the maximum likelihood estimator is consistent as  $n \rightarrow \infty$  with fixed  $p < \infty$ , and the minimum variance unbiased estimator.
- ▶ For high dimensional statistics, there are many estimators (shrinkage) that dominate MLE in terms of prediction power, e.g.
  - Linear estimator may dominate MLE for sparse targets
  - James-Stein estimator uniformly dominates MLE if  $p \geq 3$
  - Soft-thresholding (Lasso) estimator may dominate MLE and even JSE for sparse targets
- ▶ Therefore, regularization lies in the core of high dimensional statistics against the noise