A First Look Into Optimality of Estimation

Statistical Consultary, 2020 FALL

Shenduo Zhang
School of mathematics and statistics
Xi'an Jiaotong University
zhangshenduo@gmail.com

November 24, 2020

Abstract

We give a short introduction to the concept of minimaxity and admissible estimator, Jame-Stein estimator as well as numerical experiements to illustrate its trade-off.

***K*eywords** First keyword · Second keyword · More

1 Introduction

The study of optimality of statistical estimation was one of the most active area of statistics since 1980s. There has been enormous number of research done and there're still many challenging problems open, particularly in non-parametric and high-dimensional models. The pioneer work was done by Cencov, using van Tree's inequality. The breakthrough was achieved by two Russian mathematician Ibragimov and Has'minskii.

A complete introduction to minimaxity would need at least a book of more than 200 pages. This is only an introduction of it. Part of this book was based on a lecture the author attended, parts of it come from the other two books in the reference. The idea of the numerical experienments come from an online R tutorial website.

2 Review of linear model

Linear model is the most basic statistical model. Although as simple as it is, it's also the most important model in statistics. The idea of linear model first can be traced back to Gauss, which contributed to the methodology later known to be least square method. Many attention has been given to it's descedent, generalized linear model. We want to start with the most basic linear model,

Definition 1 (Linear model)**.**

$$Y = X\beta + \xi, \xi \sim N(0, \sigma^2 I_V)$$

W, V are two linear spaces, $\beta \in W$ is the parameter to estimate, $X : W \mapsto V$ is a fixed design matrix, $\xi \in V$ is noise, σ is known.

This problem can be solved by the least square estimator.

Definition 2 (Least square estimator)**.**

$$\hat{\beta} := \arg\min_{\hat{\beta} \in W} \left\| Y - X\hat{\beta} \right\|^2$$

And the model can be generalized geometrically to random shift model,

Definition 3 (Random shift model)**.**

$$Y = \mu + \xi, \xi \sim N(0, \sigma^2 I_V)$$

V is a linear space, $\mu \in L \subset V$, L is called the regression manifold, σ is known.

The problem of random shift model can be solve by projection the response Y to the regression manifold L. And in the special case when the regression manifold is actually a linear subspace, which is spaned by the design matrix $\mathrm{Im}X$, projection estimator is equivalent to LS-estimator. The connection between those two models are characterised by normalized equation,

$$\mu = X\hat{\beta} \tag{1}$$

$$X^*X\hat{\beta} = X^*Y \tag{2}$$

$$\tag{3}$$

If one introduce the concept of MP inverse, the estimator can be written as

$$\hat{\beta} \in X^+Y + \ker(X)$$

where X^+ is the MP inverse.

This reduction of geometrical random shift model to linear model is known to be the coordinate free approach. We will use the coordinate free approach but focus on this specific basic linear model rather than assume L to be a general regression manifold.

2.1 Gauss-Markov theorem

It can be shown that the above projection estimator has many properties that makes it a "good" estimator, for instance a very important one being the Gauss-Markov theorem, which guarantee projection estimator is the best linear unbiased estimator(BLUE).

Theorem 1 (Gauss-Markov)**.** *Suppose $< Y, d >$ for some $d \in V$ is a linear unbiased for linear functinoal $< \mu, c >$, $\mu \in L$. Then,*

$$\mathbf{Var}(< Y, d >) \geq \mathbf{Var}(< \hat{\mu}, c >), \mu \in L \tag{4}$$

and more over, $< \hat{\mu}, c >$ is the unique linear unbiased estimator with the smallest possible variance.

However, is projection estimator really the "best" estimator for linear model out there? The answer apparently depends on how one define an estimator being "good" and the "best". Here, it's called the best because the we are restricted to unbiased linear estimators, and we adopt the criteria of variance to judge the estimator.

3 Minimaxity of projection estimator

First, we claim projection estimator is the "best" estimator. To explain why, we need to first specify the criteria we adopted here.

General semi-parameter settings:

1. $(\Omega, \mathcal{F}, \mathbb{P}_\theta)$ where the data $\mathcal{X}$ (or $(\mathbf{X}, \mathbf{Y})$ in supervised situation) was drawn. $\mathbb{P}$ is index by θ.
2. $\theta \in \Theta$ where θ is the estimation target, Θ is the candidate family. In the parameter setting, θ is the parameter to estimate, for instance, the median μ. In the non-parametric setting, θ is the target function to estimate. For instance, regression function $f(\cdot)$ or density function $p(\cdot)$.
3. $d(\cdot, \cdot) : \mathbb{R}^d \times \mathbb{R}^d \mapsto \mathbb{R}$ is a semi-distance to judge the performance of estimation. In parametric setting, d can be taken as the inner product educed from l_2 norm. In non-parameter setting, d can be taken to be inner product educed from L^2 norm, L_1 norm or L^∞ norm. I haven't found any examples of application of it which is neat enough to fit into the scope of this introduction.

Definition 4 (Risk)**.**

$$R(\theta; \hat{\theta}) = \mathbb{E}_\theta d\Big(\theta - \hat{\theta}\Big).$$

In the linear model we are interested, the risk has the following form,

$$R(\mu; \hat{\mu}) = \mathbb{E}_\mu \|\mu - \hat{\mu}\|^2.$$

Risk is the error that an estimator $\hat{\theta}$ made when estimating the real parameter θ. To judge if a estimator is good or bad, we are interested in the risk upper bound when estimating any "parameters" within one's interests. So we define the following maximum risk,

Definition 5 (Maximum Risk)**.**

$$r(\theta;\hat{\theta}) = \sup_{\theta\in\Theta} R(\theta;\hat{\theta}) = \mathbb{E}_\theta \left\|\theta - \hat{\theta}\right\|^2.$$

In the linear model we are interested, the maximum risk has the following form,

$$r(\mu;\hat{\mu}) = \sup_{\mu\in L} R(\mu;\hat{\mu}) = \mathbb{E}_\mu \|\mu - \hat{\mu}\|^2.$$

Remark 1. *It's important to clarify that generally speaking, we won't know what's the value of risk of our estimation unless we have help from some oracle. But we do know the maximum risk once we are given enough regularity condition. However, in the linear model, we are interested in the case $\theta = \mu \in L = \Theta$, L is a linear subspace of a finite dimensional linear space V. And by computing the risk for any specific μ, we found the risk is only dependent to L rather than specific μ. So, we happened to know the risk without any help from any oracle. And it's non-asymptotic. This is not generally true even in the parametric sense. For instance, consider the same model but the dimension of V is extremly high and μ is sparse. If we know where are the non zero elements located, the risk will be significantly smaller than the maximum risk, more specifically it will reduce from $O(\dim L)$ to the rate of degree of sparsity of μ.*

$$R(\mu;\hat{\mu}) = r(\mu;\hat{\mu}) = \sigma^2 \dim(L) \tag{5}$$

In the non-parametric senerio, where the supremum is taken over a function class Θ, those two terms can be significantly different. And usually the risk can be only guarantee in asymptotic sense, because we are always somehow discretilizing the smooth target function, for instance, cut off its fourier spectrum at some frequency or approximate the integral with step function. But linear model is enough to illustrate the idea.

Now we claim that the projection estimator is indeed the "best" estimator for linear model. We introduce the following definition of minimaxity.

Definition 6 (Minimax estimator)**.** *An estimator T^* is called minimax estimator if*

$$\sup_{\theta\in\Theta} R(\theta, T^*) = \inf_T \sup_{\theta\in\Theta} R(\theta, T) \tag{6}$$

The infinimum is taken over all the estimator. This standard of optimality is saying that, no matter what methodology one adopt when making estimation, there is certain error that one can not get ride of when estimating μ.

And without any suprise, projection estimator $\hat{\mu}$ is a minimax estimator for linear models.

Theorem 2 (Minimaxity of projection estimator)**.** $\forall$ *estimator $T(Y)$ of μ,*

$$\sup_{\mu\in L} \mathbb{E}_\mu \|T(Y) - \mu\|^2 \geq \sup_{\mu\in L} \mathbb{E}_\mu \|\hat{\mu} - \mu\|^2 = \sigma^2 \dim(L) \tag{7}$$

Definition 7 (minimax risk)**.** $\inf_T \sup_{\theta\in\Theta} R(\mu;T)$ *is called minimax risk.*

This error lower bound $\sigma^2 \dim L$ is known to be the minimax risk of linear model,

If the risk upper bound of an estimator reaches minimax lower bound of the model, we call it a minimax estimator. In the asymptotic senerio, we are usually interested in rate minimaxity, which means the upper bound on rate matches the lower bound on risk. Before we give the proof, we want to introduce another concept of optimality.

3.0.1 Bayes estimator

We adopt the general setting here first, consider such model. $X \sim \mathbb{P}_\theta; \theta \in \Theta \subset V$, where V is an inner product space, specifically l_2 or L_2. $\Pi(\mathrm{d}\theta) = \pi(\theta)\mathrm{d}\theta$ is a probability distribution on Θ called prior distribution and $\pi(\theta)$ is called the prior density. We reduce it to computing the average risk over a continuously parameterized subfamily of Θ rather than maximum risk over all candidates. Here in linear model (or more generally speaking, parametric settings), $\Theta = V = \mathbb{R}^d$, however in the non-parametric senerio, V is finite dimensional, so we can even construct prior not on a subset but on the whole candidate set. However, in non-parametric senerio, we usually need to find a subfamily of the candidate set, because the candidate set is too large(infinite dimensional). Even construct a prior on such huge space is not an easy thing to do(for me).

I believe this approach will carry through when V is infinite dimensional, but we restricted ourself to the finite dimensional case.

To compute the average, we need to define Bayes risk and Bayes estimator.

Definition 8 (Bayes estimator)**.** *Let T be an estimator of θ, $R(\theta;T)$ be the risk of T,*

$$R_\Pi(T) = \int_\Theta R(\theta;T)\Pi(\mathrm{d}\theta) = \int_\Theta R(\theta;T)\pi(\theta)\mathrm{d}\theta \tag{8}$$

is called the average risk with respect to prior Π. The estimator T_Π is Bayes estimator with respect to Π if and only if $\forall T$, estimator of θ, one has

$$R_\Pi(T) \geq R_\Pi(T_\Pi) \tag{9}$$

For this minimaxity of linear model, we introduce the following theorem

Theorem 3. *Suppose exists an estimator $T(\mathbf{X})$ and a sequence of prior distribution $\{\Pi_k\}$ such that*

$$R_{\Pi_k}(T_{\Pi_k}) \to \sup_{\theta\in\Theta} R(\theta;T) \quad (k\to\infty) \tag{10}$$

Then T is a minimax estimator.

Proof. $\forall$ esimator $\tilde{T}$,

$$\sup_{\theta\in\Theta} R(\theta;\tilde{T}) \geq R_{\Pi_k}(\tilde{T}) \geq R_{\Pi_k}(T_{\Pi_k}) \to \sup_{\theta\in Theta} R(\theta;T) \tag{11}$$

□

Now if we can find such a sequence of Bayes estimator, we are done with the proof. Without any suprise, the Bayes estimator is the mean of posterior distribution.

Proposition 1.

$$T_\Pi(x) := \int \theta p(\theta|x)\mathrm{d}\theta \tag{12}$$

which is the mean of the posterior density. $T_\Pi(X)$ is Bayes estimator of θ with respect to Π.

Proof. Let $\tilde{\theta}$ be a random variable in Θ, $\tilde{\theta}\sim\Pi$, $X\sim p(\cdot|\theta)$.

$$T_\Pi(x) = \int_\Theta \theta p(\theta|x)\mathrm{d}\theta = \mathbb{E}(\tilde{\theta}|X=x) \tag{13}$$

For any other estimator T, we have the following tower property,

$$\begin{aligned}
R_\Pi(T) &= \int_\Theta \mathbb{E}_\theta\|T(X)-\theta\|^2\pi(\theta)\mathrm{d}\theta && (14)\\
&= \int_\Theta \mathbb{E}\Big(\|T(X)-\theta\|\Big|\tilde{\theta}\Big)\pi(\theta)\mathrm{d}\theta && (15)\\
&= \mathbb{E}\mathbb{E}\left(\left\|T(X)-\tilde{\theta}\right\|^2\middle|\tilde{\theta}\right) && (16)\\
&= \mathbb{E}\left\|T(X)-\tilde{\theta}\right\|^2 && (17)\\
&= \mathbb{E}\|T(X)-T_\Pi(X)\|^2 + \mathbb{E}\left\|T_\Pi(X)-\tilde{\theta}\right\|^2 + 2\mathbb{E}\Big\langle T(X)-T_\Pi(X), T_\Pi(X)-\tilde{\theta}\Big\rangle && (18)
\end{aligned}$$

Using conditional expectation to rule out randomness of $\tilde{\theta}$ and using linearty of inner product and expectation,

$$\mathbb{E}\Big\langle T(X)-T_\Pi(X), T_\Pi(X)-\tilde{\theta}\Big\rangle = \mathbb{E}\Big\langle T(X)-T_\Pi(X), \mathbb{E}\Big(T_\Pi(X)-\tilde{\theta}\Big|X\Big)\Big\rangle = 0 \tag{19}$$

Hence, for any estimator T,

$$R_\Pi(T) = R_\Pi(T_\Pi) + \mathbb{E}\|T(X)-T_\Pi(X)\|^2. \tag{20}$$

T_Π is Bayes optimal. □

Now to give lower bound on minimaxity of linear model, it only suffices to find an sequence of explicit prior such that the Bayes risk of its posterior mean converges to the corresponding minimax risk.

Proof. The prior on L is constructed as follow,

$$\Pi_k \sim N(0, \tau_k^2 I_L), \tau_k \to +\infty. \tag{21}$$

The density of Y conditioning on a fixed μ is

$$p(y|\mu) = \frac{1}{(2\pi)^{n/2}\sigma^n} \exp\left\{ -\frac{\|y-\mu\|^2}{2\sigma^2} \right\}. \tag{22}$$

Compute posterior, up to a constant we have the following estimate holds

$$\begin{aligned} p(\mu|y) = \frac{p(y|\mu)\pi(\mu)}{\int_L p(y|\mu)\pi(\mu)\mathrm{d}\mu} &\asymp \exp\left\{ -\frac{\|\mu\|}{\tau_k^2} - \frac{\|y-\mu\|^2}{2\sigma^2} \right\} \\ &\asymp \exp\left\{ -\frac{\|\mu\|^2}{\tau_k^2} + \frac{\langle y, \mu \rangle}{\sigma^2} \right\} \end{aligned}$$

It's clear that $\mu|Y = y$ will follows a normal distribution with independent coordinates, we just need to work out its mean and corresponding variance. Set the bias to be a and covariance operator to be $\sigma^2 I_L$, compare its density function, we obtain the following equation,

$$\exp\left\{ -(\frac{1}{\tau_k^2} + \frac{1}{2\sigma^2})\|\mu\|^2 + \frac{1}{\sigma^2}\langle y, \mu \rangle \right\} = \exp\left\{ -\frac{1}{2b^2}\|\mu\|^2 + \frac{1}{b^2}\langle \mu, a \rangle \right\} \tag{23}$$

Notice that we can use the orthogonal decompesition to y and obtain

$$-(\frac{1}{\tau_k^2} + \frac{1}{2\sigma^2})\|\mu\|^2 + \frac{1}{\sigma^2}\langle P_L y, \mu \rangle = -\frac{1}{2b^2}\|\mu\|^2 + \frac{1}{b^2}\langle \mu, a \rangle \tag{24}$$

We obtain the following equation

$$\begin{cases} \frac{1}{\tau_k^2} + \frac{1}{2\sigma^2} = \frac{1}{2b^2} \\ \frac{1}{\sigma^2} P_L y = \frac{1}{b^2} a \end{cases} \tag{25}$$

Solve the equation would give

$$\begin{cases} a = \frac{\tau_K^2}{2\sigma^2 + \tau_k^2} P_L y \\ b = \frac{\sigma^2 \tau_k^2}{2\sigma^2 + \tau_k^2} \end{cases} \tag{26}$$

Hence we have the posterior $\mu|Y \sim N(\frac{\tau_K^2}{2\sigma^2+\tau_k^2} P_L Y, \frac{\sigma^2\tau_k^2}{2\sigma^2+\tau_k^2} I_L)$. Compute its mean would give the Bayes estimator $T_{\Pi_k} = \frac{\tau_k^2}{2\sigma^2+\tau_k^2}\hat{\mu}$. Then it only suffice to prove when $\tau_k \to \infty$, $R_\Pi(T_\Pi) \to \sigma^2 \dim L$. Indeed,

$$\begin{aligned} R(T_{\Pi_k}; \mu) &= \mathbb{E}_\mu \left\| \frac{\tau_k^2}{2\sigma^2 + \tau_k^2}\hat{\mu} - \mu \right\|^2 \\ &= \mathbb{E}_\mu \left\| \frac{\tau_k^2}{2\sigma^2 + \tau_k^2}(\hat{\mu} - \mu) - \frac{2\sigma^2}{2\sigma^2 + \tau_k^2}\mu \right\|^2 \\ &= \mathbb{E}_\mu \left\| \frac{\tau_k^2}{2\sigma^2 + \tau_k^2}(\hat{\mu} - \mu) \right\|^2 + \mathbb{E}_\mu \left\| \frac{2\sigma^2}{2\sigma^2 + \tau_k^2}\mu \right\|^2 - \mathbb{E}_\mu C\langle \hat{\mu} - \mu, \mu \rangle \end{aligned}$$

Here note that $\langle \hat{\mu} - \mu, \mu \rangle = \langle P_L(\mu + \xi) - \mu, \mu \rangle = \langle P_L(\xi), \mu \rangle$. Take the expectation, this term is zero. Hence we have the following equality

$$R(T_{\Pi_k}; \mu) = \left(\frac{\tau_k^2}{2\sigma^2 + \tau_k^2} \right)^2 \sigma^2 \dim L + \left(\frac{2\sigma^2}{2\sigma^2 + \tau_k^2}\|\mu\| \right)^2 \tag{27}$$

Now we integrate it with respect to the prior to compute its Bayes risk,

$$\begin{aligned} R_{\Pi_k}(T_{\Pi_k}) &= \left(\frac{\tau_k^2}{2\sigma^2 + \tau_k^2} \right)^2 \sigma^2 \dim L + \left(\frac{2\sigma^2}{2\sigma^2 + \tau_k^2} \right)^2 \int_L \|\mu\|^2 \Pi_k(\mathrm{d}\mu) \\ &= \left(\frac{\tau_k^2}{2\sigma^2 + \tau_k^2} \right)^2 \sigma^2 \dim L + \left(\frac{2\sigma^2}{2\sigma^2 + \tau_k^2} \right)^2 \tau_k^2 \dim L \end{aligned}$$

The last equality follows from the definition of second moment of μ with respect to the prior or expand it explicitly by coordinates. As one can see that as $\tau_k \to \infty$, the first term converges to $\sigma^2 \dim L$ which is desire minimax rate and the second term vanish. □

3.1 More on minimaxity

We are just one step to another commonly used approach to minimaxity, it's convenient to introduce it without giving a proof. It follows a similar idea of the reduction to bayes risk. It's an information inequality called **van Trees inequality**. This approach is only useful when dealing with semi-parametric model with quadratic loss, because its proof is based on repeatedly usage of Cauchy-Schwartz inequality. The parameter space must be a Hilbert space or at least equiped with Cauchy-Swartz inequality for the proof to carry out. The overall idea is similar, bounding the maximum risk below by the average risk with respect to some Bayes prior on a candidate set. Rather than finding an explicit sequence of prior whose Bayes risk converge to the minimax risk, we work with a information lower bound. This approach is significantly easier to handle when dealing with optimality under non-parametric estimation comparing to other approach requiring heuristic construction. Now we introduce the following notation and basic assumption,

1. $\{P_{\theta_t}, t \in T\}$ be a family of probability measures on $(\mathcal{X}, \mathcal{A})$, where $\{\theta_t, t \in T\}$ is the subset of our candidate family Θ.
2. There exists a σ-finite measure ν on $(\mathcal{X}, \mathcal{A})$ such that $\forall t \in T, P_t \ll \nu$. Denote the probability density of P_t with respect to ν to be $p(\cdot, t)$.
3. $\mu(\cdot)$ is a probability distribution on T with absolutely continuous with respect to Lebesgue measure.

We would only state the theorem without proving it. This theorem is named after Harry L. Van Trees, who is 90 years old by the time this is written.

Theorem 4 (van Trees inequality)**.** *Under the above semi-parametric setting, we assume the following holds*

1. *$p(x, t)$ is meansurable and absolutely continuous in t for almost all x with respect to the measure ν.*
2. *The Fisher information*
$$\mathcal{I}(t) = \int \left(\frac{\frac{\partial}{\partial t} p(x,t)}{p(x,t)} \right)^2 p(x,t)\nu(\mathrm{d}x) \tag{28}$$
is finite and belongs to $L^1(T)$.
3. *The prior density μ is absolutely continuous on T and vanish on end point $\mu(t_1) = \mu(t_2) = 0$, and has finite Fisher information*
$$\mathcal{J}(\mu) = \int_T \frac{\mu'(t)^2}{\mu(t)} \mathrm{d}t \tag{29}$$

Then, for any estimator $\hat{t}(\mathbf{X})$, the Bayes risk is bounded as follows:

$$\int_T \mathbb{E}_t \left[\left(\hat{t}(\mathbf{X}) - t \right)^2 \right] \mu(t) \mathrm{d}t \geq \frac{1}{\int \mathcal{I}(t)\mu(t)\mathrm{d}t + \mathcal{J}(\mu)}. \tag{30}$$

If we can find such prior, whose bayes risk has the desiring rate of convergence, and compute its fisher information getting a non-asymptotic upper bound on information, the minimaxity follows immediately. Usually the choice of μ can be solved by a variation equation maximizing $\mathcal{J}$ with respect to the regularity condition. The computation is tedious requiring variation and solving several differential equation after integration by parts and an application of Green formula. I will just post its form without giving a proof, since me myself has only prove half of it without working on the constant. See [2] section 2 for a more detailed discussion.

4 Admissibility

We now ask again the same question, is projection estimator the best estimator for linear model out there? We have known that first, it's the best linear unbiased estimator, and second, it's a minimax estimator. Does this mean there're no other estimators better than projection?

Now, we make another completely reversed claim, the projection estimator is not the "best" estimator. To specify this, we need to first introduce the concept of admissible estimator.

Definition 9 (Admissible estimator)**.** *An estimator $\hat{\theta}$ is called inadmissible if there $\exists T(Y)$ of θ such that*

$$\mathbb{E}_\theta \|T(Y) - \theta\|^2 \leq \mathbb{E}_\mu \left\|\hat{\theta} - \theta\right\|^2 \tag{31}$$

with the strict inequality for some $\theta \in \Theta$. Otherwise, it's called admissible.

4.0.1 Stein

Theorem 5 (James-Stein)**.** *When $\dim L \geq 3$, there $\exists$ an estimator $T(Y)$ of μ such that $\forall \mu \in L$,*

$$\mathbb{E}_\mu \|T(Y) - \mu\|^2 < \mathbb{E}_\mu \|\hat{\mu} - \mu\|^2 = \sigma^2 \dim L \tag{32}$$

in another word, projection estimator $\hat{\mu}$ is an inadmissible estimator of linear model for $\dim L \geq 3$.

By consider a shrinkage approach,

$$T(Y) = \hat{\mu} + \sigma^2 g(\hat{\mu}) \tag{33}$$

where $g : \mathbb{R}^d \mapsto \mathbb{R}^d$ is a smooth function.

First if we try to plug in the shrinkage estimator T, check its risk,

$$\begin{aligned}
\mathbb{E}_\mu \|T(Y) - \mu\|^2 &= \mathbb{E}_\mu \left\|\hat{\mu} + \sigma^2 g(\hat{\mu}) - \mu\right\|^2 \\
&= \mathbb{E}_\mu \left(\|\mu - \hat{\mu}\|^2 + \sigma^4 \|g(\hat{\mu})\|^2 + 2\sigma^2 \langle \mu - \hat{\mu}, g(\hat{\mu})\rangle\right) \\
&= \mathbb{E}_\mu \|\mu - \hat{\mu}\|^2 + \sigma^4 \mathbb{E}_\mu \|g(\hat{\mu})\|^2 + 2\sigma^2 \mathbb{E}_\mu \langle \mu - \hat{\mu}, g(\hat{\mu})\rangle
\end{aligned}$$

Theorem 6 (Stein identity)**.** *Let $X \sim N(\theta, \sigma^2 I_d)$, let g be a differentiable fucntion satisfying $\mathbb{E}|g'(X)| < \infty$. Then*

$$\mathbb{E}\langle g(X), X - \theta\rangle = \sigma^2 \mathbb{E} \nabla \cdot g'(X) \tag{34}$$

Proof. In the case when $d = 1$,

$$\mathbb{E} g(X)(X - \theta) = \frac{1}{\sqrt{2\pi}\sigma} \int_\infty^\infty g(x)(x - \theta) e^{-(x-\theta)^2/(2\sigma^2)} \mathrm{d}x$$

Integration by parts,

$$\mathbb{E} g(X)(X - \theta) = \frac{1}{\sqrt{2\pi}\sigma^2} \left[-\sigma^2 g(x) e^{-(x-\theta)^2/(2\sigma^2)}\right]\Big|_\infty^\infty + \sigma^2 \int_\infty^\infty g'(x) e^{-(x-\theta)^2/(2\sigma^2)} \mathrm{d}x$$

The first term is 0 and the second term is $\sigma^2 \mathbb{E} g'(X)$. □

Use the identity with $\hat{\mu} \sim N(\mu, \sigma^2 I_d)$ to see that $\mathbb{E}\langle \mu - \hat{\mu}, g(\hat{\mu})\rangle = \sigma^2 \mathbb{E} \nabla \cdot g'(X)$. Plug back to the risk to see that

$$\mathbb{E}_\mu \|T(Y) - \mu\|^2 = \mathbb{E}_\mu \|\mu - \hat{\mu}\|^2 + \sigma^4 \mathbb{E}_\mu \|g(\hat{\mu})\|^2 + 2\sigma^4 \mathbb{E} \nabla \cdot g'(\hat{\mu}) \tag{35}$$

The first term is the risk of projection estimator, the second term is non-negative. If we can construct a function g such that the second term is 0 and the last term is negative, we end up with an estimator whose risk is strictly less than projection estimator.

Consider the following choice of g,

$$g = \nabla \log \psi \tag{36}$$

ψ is a smooth, non-constant, positive function $\psi : \mathbb{R}^d \mapsto \mathbb{R}$. We have the following equation

$$g(x) = \frac{\nabla \psi(x)}{\psi(x)} \tag{37}$$

$$\nabla \cdot g(x) = \frac{\Delta \psi(x) \psi(x) - \|\nabla \psi\|^2}{\psi^2(x)} = \frac{\Delta \psi(x)}{\psi(x)} - \|g(x)\|^2 \tag{38}$$

If we plug it back to the risk 35, we found that

$$\mathbb{E}_\mu \|T(Y) - \mu\|^2 = \mathbb{E}_\mu \|\mu - \hat{\mu}\|^2 - \sigma^4 \mathbb{E}_\mu \|g(\hat{\mu})\|^2 + 2\sigma^4 \mathbb{E}_\mu \frac{\Delta \psi(\hat{\mu})}{\psi(\hat{\mu})} \tag{39}$$

Then observe that the second term is non-negative which will definitely help us to reduce the bias if we can make the third term vanish. So the problem is reduced to choose a $\psi : \mathbb{R}^d \mapsto \mathbb{R}$ that is smooth, positive, and $\Delta\psi = 0$. In another words, we are interested to find a smooth harmonic function $\psi \neq$constant and is positive. Such function exists if $d \geq 3$. However the maximum principle guarantee that ψ does not exists in 1 and 2 dimension. Now it only suffice to show that for some $\mu \in \mathbb{R}^d$,

$$\mathbb{E}_\mu \|g(\hat{\mu}) > 0\| \tag{40}$$

This is obvious since $g \neq 0$ and it's continuous, so there exists a small domain on which $\|g(x)\| \geq c > 0$. Now apply Markov theorem on to this set since $\hat{\mu}$ has finite first moment to obtain the claim.

Now we gave one explicit construction for ψ.

$$\psi(x) = \|x\|^{-(d-2)}. \tag{41}$$

Corresponding Jame-Stein estimator has the following from

$$T_{JS}(Y) = \hat{\mu} - (d-2)\frac{\sigma^2\hat{\mu}}{\|\hat{\mu}\|} = \hat{\mu}(1 - \frac{\sigma^2(d-2)}{\|\hat{\mu}\|^2}) \tag{42}$$

for $d > 2$.

What JS estimator does is to shrinkage the projection estimator towards the origin. This means that JS-estimator is a biased estimator. What we did was to relax the candidate available in Gauss-Markov theorem to allow us to do this shrinkage operation. The risk of JS estimator is complicated to compute, we will just give the result of it.

$$\mathbb{E}_\mu \|T_{JS} - \mu\|^2 = \sigma^2 d - \sigma^2(d-2)^2 \mathbb{E}\frac{1}{\chi^2_{d,\frac{\|\mu\|}{\sigma}}} \tag{43}$$

where $\chi^2_{d,\frac{\|\mu\|}{\sigma}}$ is the biased Chi-square distribution.

However, JS estimator is still not admissible. The construction of admissible estimator is beyond the scope of this introduction report. But by comparing the risk of MLE and JS-estimator, it's actually not hard to see, if $\|\mu\| = 0$, then $\mathbb{E}\frac{1}{\chi^2_{d,\frac{\|\mu\|}{\sigma}}} = 1/(d-2)$, hence the improvements on risk is characterized by the ratio $2/d$, which is independent of n.

However, JS-estimator is still not admissible estimator, and there're other type of shrinkage estimator avaliable. See [3] and [1] for a more detailed discussion.

4.1 Numerical results

As we illustrated earlier, JS estimator shrinks projection estimator by introducing a delibrated bias to reduce risk. However this can lead to biased estimation of parameters as we will see soon.

We use a numerical experiement to illustrate this. We generate data from a linear model

$$Y = X\beta + \xi \tag{44}$$

where $X \in \mathbb{R}^{20} \times \mathbb{R}^7$, $\beta = (-3, -2, -1, 0, 1, 2, 3)^T$ and $\xi \sim N(0, I_7)$. We run the MLE and JS estimator respectively and plot then in the same figure 1.

Here we can clearly see although JS estimator commits a smaller risk, but it's not unbiased. This would result in extremly inaccurate estimation for parameters when sample size is large. This means the estimation is more robust, in the sense that the expected error is smaller, but the estimation is actually worse comparing to MLE.

If we cann't obtain an unbiased estimator, like in the Gaussian sequential model, i.e. infintie dimensional linear model, then JS estimator can be considered since there are no unbiased estimator whatsoever. But in the case where we have an unbiased estimator, using JS estimator doesn't make much sense. JS estimator can be generalized to all linear estimator, for instance N-W estimator or projection estimator in non-parametric senerio. The discussion for the relationship between JS-type shrinkage estimator and linear estimator in infinite dimensional case, or non-parametric case is beyond the scope of this introduction level report and is much more complicated. See [2] in Chapter 3 for more detailed discussions.

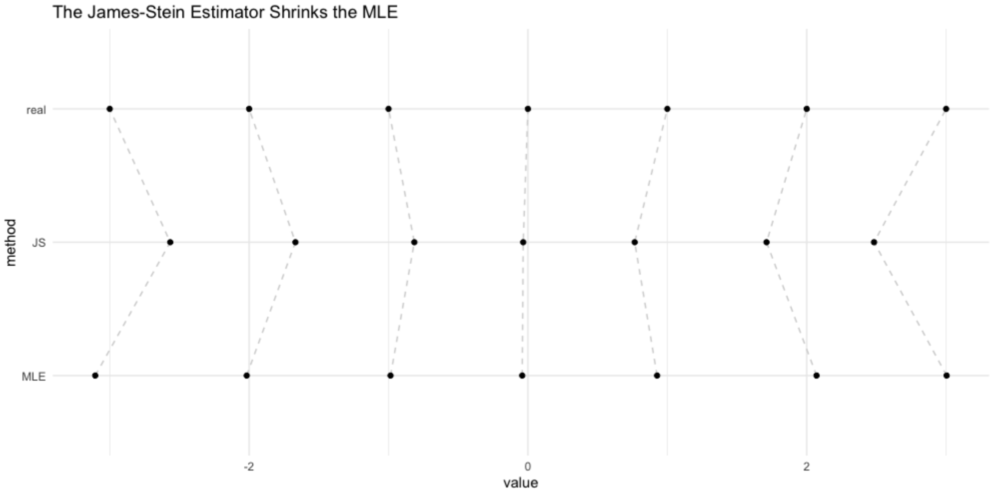

Figure 1: Comparison of true value, MLE and JS estimators