Foundations of Statistical Inference

J. Berestycki & D. Sejdinovic

Department of Statistics
University of Oxford

MT 2019

Chapter 9: Stein's Paradox

Stein's paradox and the James-Stein Estimator

Stein's paradox has been described as *"the most striking theorem of post-war mathematical statistics"* (Efron, 1992).

Setup

Let $X_i \sim N(\mu_i, 1),\ i = 1, 2, ..., p$ be jointly independent so we have one data point for each of the p μ_i-parameters.

Let $X = (X_1, ..., X_p)$ and $\mu = (\mu_1, ..., \mu_p)$. The goal is to estimate μ. Consider $\hat{\mu}_{MLE} = X$

- MLE
- MVUE
- Is it admissible? (for a quadratic loss function, say)

Recall that $\hat{\mu}$ is inadmissible if we can find $\tilde{\mu}$ such that

$$R(\mu, \hat{\mu}) \geq R(\mu, \tilde{\mu}),\ \forall \mu$$

with strict inequality for some μ.

Stein's paradox and the James-Stein Estimator

Stein's paradox has been described as *"the most striking theorem of post-war mathematical statistics"* (Efron, 1992).

Setup

Let $X_i \sim N(\mu_i, 1),\ i = 1, 2, ..., p$ be jointly independent so we have one data point for each of the p μ_i-parameters.

Let $X = (X_1, ..., X_p)$ and $\mu = (\mu_1, ..., \mu_p)$. The goal is to estimate μ. Consider $\hat{\mu}_{MLE} = X$

- MLE
- MVUE
- Is it admissible? (for a quadratic loss function, say)

Recall that $\hat{\mu}$ is inadmissible if we can find $\tilde{\mu}$ such that

$$R(\mu, \hat{\mu}) \geq R(\mu, \tilde{\mu}),\ \forall \mu$$

with strict inequality for some μ.

Stein's paradox and the James-Stein Estimator

Stein's paradox has been described as *"the most striking theorem of post-war mathematical statistics"* (Efron, 1992).

Setup

Let $X_i \sim N(\mu_i, 1),\ i = 1, 2, ..., p$ be jointly independent so we have one data point for each of the p μ_i-parameters.

Let $X = (X_1, ..., X_p)$ and $\mu = (\mu_1, ..., \mu_p)$. The goal is to estimate μ.
Consider $\hat{\mu}_{MLE} = X$

- MLE
- MVUE
- Is it admissible? (for a quadratic loss function, say)

Recall that $\hat{\mu}$ is inadmissible if we can find $\tilde{\mu}$ such that

$$R(\mu, \hat{\mu}) \geq R(\mu, \tilde{\mu}), \ \forall \mu$$

with strict inequality for some μ.

Stein's paradox and the James-Stein Estimator

Stein's paradox has been described as *"the most striking theorem of post-war mathematical statistics"* (Efron, 1992).

Setup

Let $X_i \sim N(\mu_i, 1),\ i = 1, 2, ..., p$ be jointly independent so we have one data point for each of the p μ_i-parameters.

Let $X = (X_1, ..., X_p)$ and $\mu = (\mu_1, ..., \mu_p)$. The goal is to estimate μ.
Consider $\hat{\mu}_{MLE} = X$

- MLE
- MVUE
- Is it admissible? (for a quadratic loss function, say)

Recall that $\hat{\mu}$ is inadmissible if we can find $\tilde{\mu}$ such that

$$R(\mu, \hat{\mu}) \geq R(\mu, \tilde{\mu}),\ \forall \mu$$

with strict inequality for some μ.

Stein's paradox and the James-Stein Estimator

Answer:

If $p \geq 3$, $\hat{\mu}$ is inadmissible for quadratic loss!

Theorem

An estimator with lower risk is given by the James-Stein estimator

$$\hat{\mu}_{JSE} = \left(1 - \frac{p-2}{\sum_i X_i^2}\right) X$$

Stein's paradox and the James-Stein Estimator

Answer:

If $p \geq 3$, $\hat{\mu}$ is inadmissible for quadratic loss!

Theorem

An estimator with lower risk is given by the James-Stein estimator

$$\hat{\mu}_{JSE} = \left(1 - \frac{p-2}{\sum_i X_i^2}\right) X$$

Implications of Stein's Paradox

Suppose we are interested in estimating

1. the weight of a randomly chosen loaf of bread from a supermarket.
2. the height of a random chosen blade of grass from a garden.
3. the speed of a randomly chosen car as it passes a speed camera.

These are totally unrelated quantities. It seem implausible that by combining information across the data points that we might end up with a better way of estimating the vector of three parameters.

The James-Stein estimator tells us that we can get a better estimate (on average) for the vector of three parameters by simultaneously using the three unrelated measurements.

Implications of Stein's Paradox

Suppose we are interested in estimating

1. the weight of a randomly chosen loaf of bread from a supermarket.
2. the height of a random chosen blade of grass from a garden.
3. the speed of a randomly chosen car as it passes a speed camera.

These are totally unrelated quantities. It seem implausible that by combining information across the data points that we might end up with a better way of estimating the vector of three parameters.

The James-Stein estimator tells us that we can get a better estimate (on average) for the vector of three parameters by simultaneously using the three unrelated measurements.

Implications of Stein's Paradox

Suppose we are interested in estimating

1. the weight of a randomly chosen loaf of bread from a supermarket.
2. the height of a random chosen blade of grass from a garden.
3. the speed of a randomly chosen car as it passes a speed camera.

These are totally unrelated quantities. It seem implausible that by combining information across the data points that we might end up with a better way of estimating the vector of three parameters.

The James-Stein estimator tells us that we can get a better estimate (on average) for the vector of three parameters by simultaneously using the three unrelated measurements.

Proof

Consider the alternative estimator

$$\hat{\mu}_{JSE} = \left(1 - \frac{a}{\sum_i X_i^2}\right) X \quad \text{(the James-Stein estimator)}$$

Note This estimator 'shrinks' X towards 0 (when $\sum_i X_i^2 > a$).

We will show that if $a = p - 2$ then $R(\mu, \hat{\mu}_{JSE}) < R(\mu, \hat{\mu}_{MLE})$ for every $\mu \in \mathbb{R}^n$, so that the MLE is inadmissible in this case.

First, the risk for $\hat{\mu}_{MLE}$ is

$$R(\mu, \hat{\mu}_{MLE}) = \sum_{i=1}^{p} \mathbb{E}(|\mu_i - \hat{\mu}_{MLE,i}|^2) = \sum_{i=1}^{p} \mathbb{E}(|\mu_i - X_i|^2) = p$$

recognizing $\mathrm{Var}(X_i) = 1$.

Proof

Consider the alternative estimator

$$\hat{\mu}_{JSE} = \left(1 - \frac{a}{\sum_i X_i^2}\right) X \quad \text{(the James-Stein estimator)}$$

Note This estimator 'shrinks' X towards 0 (when $\sum_i X_i^2 > a$).

We will show that if $a = p - 2$ then $R(\mu, \hat{\mu}_{JSE}) < R(\mu, \hat{\mu}_{MLE})$ for every $\mu \in \mathbb{R}^n$, so that the MLE is inadmissible in this case.

First, the risk for $\hat{\mu}_{MLE}$ is

$$R(\mu, \hat{\mu}_{MLE}) = \sum_{i=1}^{p} \mathbb{E}(|\mu_i - \hat{\mu}_{MLE,i}|^2) = \sum_{i=1}^{p} \mathbb{E}(|\mu_i - X_i|^2) = p$$

recognizing $\mathrm{Var}(X_i) = 1$.

Proof

Consider the alternative estimator

$$\hat{\mu}_{JSE} = \left(1 - \frac{a}{\sum_i X_i^2}\right) X \quad \text{(the James-Stein estimator)}$$

Note This estimator 'shrinks' X towards 0 (when $\sum_i X_i^2 > a$).

We will show that if $a = p - 2$ then $R(\mu, \hat{\mu}_{JSE}) < R(\mu, \hat{\mu}_{MLE})$ for every $\mu \in \mathbb{R}^n$, so that the MLE is inadmissible in this case.

First, the risk for $\hat{\mu}_{MLE}$ is

$$R(\mu, \hat{\mu}_{MLE}) = \sum_{i=1}^{p} \mathbb{E}(|\mu_i - \hat{\mu}_{MLE,i}|^2) = \sum_{i=1}^{p} \mathbb{E}(|\mu_i - X_i|^2) = p$$

recognizing $\mathrm{Var}(X_i) = 1$.

Proof

Consider the alternative estimator

$$\hat{\mu}_{JSE} = \left(1 - \frac{a}{\sum_i X_i^2}\right) X \quad \text{(the James-Stein estimator)}$$

Note This estimator 'shrinks' X towards 0 (when $\sum_i X_i^2 > a$).

We will show that if $a = p - 2$ then $R(\mu, \hat{\mu}_{JSE}) < R(\mu, \hat{\mu}_{MLE})$ for every $\mu \in \mathbb{R}^n$, so that the MLE is inadmissible in this case.

First, the risk for $\hat{\mu}_{MLE}$ is

$$R(\mu, \hat{\mu}_{MLE}) = \sum_{i=1}^{p} \mathbb{E}(|\mu_i - \hat{\mu}_{MLE,i}|^2) = \sum_{i=1}^{p} \mathbb{E}(|\mu_i - X_i|^2) = p$$

recognizing $\text{Var}(X_i) = 1$.

Proof

Consider the alternative estimator

$$\hat{\mu}_{JSE} = \left(1 - \frac{a}{\sum_i X_i^2}\right) X \quad \text{(the James-Stein estimator)}$$

Note This estimator 'shrinks' X towards 0 (when $\sum_i X_i^2 > a$).

We will show that if $a = p - 2$ then $R(\mu, \hat{\mu}_{JSE}) < R(\mu, \hat{\mu}_{MLE})$ for every $\mu \in \mathbb{R}^n$, so that the MLE is inadmissible in this case.

First, the risk for $\hat{\mu}_{MLE}$ is

$$R(\mu, \hat{\mu}_{MLE}) = \sum_{i=1}^{p} \mathbb{E}(|\mu_i - \hat{\mu}_{MLE,i}|^2) = \sum_{i=1}^{p} \mathbb{E}(|\mu_i - X_i|^2) = p$$

recognizing $\mathrm{Var}(X_i) = 1$.

Proof

Consider the alternative estimator

$$\hat{\mu}_{JSE} = \left(1 - \frac{a}{\sum_i X_i^2}\right) X \quad \text{(the James-Stein estimator)}$$

Note This estimator 'shrinks' X towards 0 (when $\sum_i X_i^2 > a$).

We will show that if $a = p - 2$ then $R(\mu, \hat{\mu}_{JSE}) < R(\mu, \hat{\mu}_{MLE})$ for every $\mu \in \mathbb{R}^n$, so that the MLE is inadmissible in this case.

First, the risk for $\hat{\mu}_{MLE}$ is

$$R(\mu, \hat{\mu}_{MLE}) = \sum_{i=1}^{p} \mathbb{E}(|\mu_i - \hat{\mu}_{MLE,i}|^2) = \sum_{i=1}^{p} \mathbb{E}(|\mu_i - X_i|^2) = p$$

recognizing $\mathrm{Var}(X_i) = 1$.

Stein's Lemma

Lemma (Stein's Lemma)

For independent Normal RV $X = (X_1, \ldots, X_p)$; $X_i \overset{ind}{\sim} \mathcal{N}(\mu_i, 1)$

$$\mathbb{E}((X_i - \mu_i)h(X)) = \mathbb{E}\left(\frac{\partial h(X)}{\partial X_i}\right).$$

This can be shown by integrating by parts. Noting if $f_i(x) = -e^{-(x-\mu_i)^2/2}$ then $f_i'(x) = (x-\mu_i)e^{-(x-\mu_i)^2/2}$

$$\int (x_i - \mu_i)e^{-(x_i-\mu_i)^2/2}dx = -e^{-(x_i-\mu_i)^2/2}$$

Stein's Lemma

Lemma (Stein's Lemma)

For independent Normal RV $X = (X_1, \ldots, X_p)$; $X_i \stackrel{ind}{\sim} \mathcal{N}(\mu_i, 1)$

$$\mathbb{E}((X_i - \mu_i)h(X)) = \mathbb{E}\left(\frac{\partial h(X)}{\partial X_i}\right).$$

This can be shown by integrating by parts. Noting if $f_i(x) = -e^{-(x-\mu_i)^2/2}$ then $f_i'(x) = (x - \mu_i)e^{-(x-\mu_i)^2/2}$

$$\int (x_i - \mu_i)e^{-(x_i-\mu_i)^2/2} dx = -e^{-(x_i-\mu_i)^2/2}$$

Stein's Lemma

Lemma (Stein's Lemma)

For independent Normal RV $X = (X_1, \dots, X_p)$; $X_i \overset{ind}{\sim} \mathcal{N}(\mu_i, 1)$

$$\mathbb{E}((X_i - \mu_i)h(X)) = \mathbb{E}\left(\frac{\partial h(X)}{\partial X_i}\right).$$

This can be shown by integrating by parts. Noting if $f_i(x) = -e^{-(x-\mu_i)^2/2}$ then $f_i'(x) = (x - \mu_i)e^{-(x-\mu_i)^2/2}$

$$\int (x_i - \mu_i)e^{-(x_i-\mu_i)^2/2}dx = -e^{-(x_i-\mu_i)^2/2}$$

Stein's Lemma

$$\int (x_i - \mu_i) e^{-(x_i-\mu_i)^2/2} dx = -e^{-(x_i-\mu_i)^2/2}$$

and thus we have

$$\begin{aligned}\int_{-\infty}^{\infty} (x_i - \mu_i) h(x) e^{-(x_i-\mu_i)^2/2} dx_i &= -h(x) e^{-(x_i-\mu_i)^2/2} \Big|_{x_i=-\infty}^{x_i=\infty} \\ &\quad + \int_{-\infty}^{\infty} \frac{\partial h(x)}{\partial x_i} e^{-(x_i-\mu_i)^2/2} dx_i\end{aligned}$$

The first term is zero if $h(x)$ (for eg) is bounded, giving the lemma.

Stein's Lemma

$$\int (x_i - \mu_i) e^{-(x_i-\mu_i)^2/2} dx = -e^{-(x_i-\mu_i)^2/2}$$

and thus we have

$$\begin{aligned}\int_{-\infty}^{\infty} (x_i - \mu_i) h(x) e^{-(x_i-\mu_i)^2/2} dx_i &= -h(x) e^{-(x_i-\mu_i)^2/2} \Big|_{x_i=-\infty}^{x_i=\infty} \\ &\quad + \int_{-\infty}^{\infty} \frac{\partial h(x)}{\partial x_i} e^{-(x_i-\mu_i)^2/2} dx_i\end{aligned}$$

The first term is zero if $h(x)$ (for eg) is bounded, giving the lemma.

Proof (continued)

$$R(\mu, \hat{\mu}_{JSE}) = \sum_{i=1}^{p} \mathbb{E}(|\mu_i - \hat{\mu}_i|^2) \quad \text{with} \quad \hat{\mu}_i = \left(1 - \frac{a}{\sum_i X_i^2}\right) X_i$$

$$\begin{aligned}
\mathbb{E}(|\mu_i - \hat{\mu}_i|^2) &= \mathbb{E}(|\mu_i - X_i|^2) - 2a\mathbb{E}\left(\frac{(X_i - \mu_i)X_i}{\sum_j X_j^2}\right) \\
&\quad + a^2\mathbb{E}\left(\frac{X_i^2}{(\sum_j X_j^2)^2}\right)
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}\left(\frac{(X_i - \mu_i)X_i}{\sum_j X_j^2}\right) &= \mathbb{E}\left(\frac{\partial}{\partial X_i}\frac{X_i}{\sum_j X_j^2}\right) \quad \text{Stein's lemma} \\
&= \mathbb{E}\left(\frac{\sum_j X_j^2 - 2X_i^2}{(\sum_j X_j^2)^2}\right) = \mathbb{E}\left(\frac{1}{\sum_j X_j^2} - 2\frac{X_i^2}{(\sum_j X_j^2)^2}\right)
\end{aligned}$$

Proof (continued)

$$R(\mu, \hat{\mu}_{JSE}) = \sum_{i=1}^{p} \mathbb{E}(|\mu_i - \hat{\mu}_i|^2) \quad \text{with} \quad \hat{\mu}_i = \left(1 - \frac{a}{\sum_i X_i^2}\right) X_i$$

$$\mathbb{E}(|\mu_i - \hat{\mu}_i|^2) = \mathbb{E}(|\mu_i - X_i|^2) - 2a\mathbb{E}\left(\frac{(X_i - \mu_i)X_i}{\sum_j X_j^2}\right) + a^2\mathbb{E}\left(\frac{X_i^2}{(\sum_j X_j^2)^2}\right)$$

$$\mathbb{E}\left(\frac{(X_i - \mu_i)X_i}{\sum_j X_j^2}\right) = \mathbb{E}\left(\frac{\partial}{\partial X_i}\frac{X_i}{\sum_j X_j^2}\right) \quad \text{Stein's lemma}$$

$$= \mathbb{E}\left(\frac{\sum_j X_j^2 - 2X_i^2}{(\sum_j X_j^2)^2}\right) = \mathbb{E}\left(\frac{1}{\sum_j X_j^2} - 2\frac{X_i^2}{(\sum_j X_j^2)^2}\right)$$

Proof (continued)

$$R(\mu, \hat{\mu}_{JSE}) = \sum_{i=1}^{p} \mathbb{E}(|\mu_i - \hat{\mu}_i|^2) \quad \text{with} \quad \hat{\mu}_i = \left(1 - \frac{a}{\sum_i X_i^2}\right) X_i$$

$$\mathbb{E}(|\mu_i - \hat{\mu}_i|^2) = \mathbb{E}(|\mu_i - X_i|^2) - 2a\mathbb{E}\left(\frac{(X_i - \mu_i)X_i}{\sum_j X_j^2}\right) + a^2\mathbb{E}\left(\frac{X_i^2}{(\sum_j X_j^2)^2}\right)$$

$$\mathbb{E}\left(\frac{(X_i - \mu_i)X_i}{\sum_j X_j^2}\right) = \mathbb{E}\left(\frac{\partial}{\partial X_i}\frac{X_i}{\sum_j X_j^2}\right) \quad \text{Stein's lemma}$$

$$= \mathbb{E}\left(\frac{\sum_j X_j^2 - 2X_i^2}{(\sum_j X_j^2)^2}\right) = \mathbb{E}\left(\frac{1}{\sum_j X_j^2} - 2\frac{X_i^2}{(\sum_j X_j^2)^2}\right)$$

Proof (continued)

$$
\begin{aligned}
R(\mu, \hat{\mu}_{JSE}) &= \sum_{i=1}^{p} \mathbb{E}(|\mu_i - \hat{\mu}_i|^2) \quad \text{with} \quad \hat{\mu}_i = \left(1 - \frac{a}{\sum_i X_i^2}\right) X_i \\
\mathbb{E}(|\mu_i - \hat{\mu}_i|^2) &= \mathbb{E}(|\mu_i - X_i|^2) - 2a\mathbb{E}\left(\frac{(X_i - \mu_i)X_i}{\sum_j X_j^2}\right) \\
&\qquad + a^2 \mathbb{E}\left(\frac{X_i^2}{(\sum_j X_j^2)^2}\right) \\
\mathbb{E}\left(\frac{(X_i - \mu_i)X_i}{\sum_j X_j^2}\right) &= \mathbb{E}\left(\frac{\partial}{\partial X_i} \frac{X_i}{\sum_j X_j^2}\right) \quad \text{Stein's lemma} \\
&= \mathbb{E}\left(\frac{\sum_j X_j^2 - 2X_i^2}{(\sum_j X_j^2)^2}\right) = \mathbb{E}\left(\frac{1}{\sum_j X_j^2} - 2\frac{X_i^2}{(\sum_j X_j^2)^2}\right)
\end{aligned}
$$

Proof (continued)

Putting the pieces together,

$$\begin{aligned}\sum_{i=1}^{p} \mathbb{E}(|\mu_i - \hat{\mu}_i|^2) &= R(\mu, \hat{\mu}_{MLE}) - (2ap - 4a)\mathbb{E}\left(\frac{1}{\sum_j X_j^2}\right) \\ &\quad + a^2 \mathbb{E}\left(\frac{1}{\sum_j X_j^2}\right) \\ &= p - (2a(p-2) - a^2)\mathbb{E}\left(\frac{1}{\sum_j X_j^2}\right)\end{aligned}$$

and this is less than p if $2ap - 4a - a^2 > 0$ and in particular at $a = p - 2$, which minimizes the risk over $a \in \mathbb{R}$.

Note We have not shown that the James Stein estimator is itself admissible.

Proof (continued)

Putting the pieces together,

$$\begin{aligned}\sum_{i=1}^{p} \mathbb{E}(|\mu_i - \hat{\mu}_i|^2) &= R(\mu, \hat{\mu}_{MLE}) - (2ap - 4a)\mathbb{E}\left(\frac{1}{\sum_j X_j^2}\right) \\ &\quad + a^2 \mathbb{E}\left(\frac{1}{\sum_j X_j^2}\right) \\ &= p - (2a(p-2) - a^2)\mathbb{E}\left(\frac{1}{\sum_j X_j^2}\right)\end{aligned}$$

and this is less than p if $2ap - 4a - a^2 > 0$ and in particular at $a = p - 2$, which minimizes the risk over $a \in \mathbb{R}$.

Note We have not shown that the James Stein estimator is itself admissible.

Proof (continued)

Putting the pieces together,

$$\begin{aligned}\sum_{i=1}^{p}\mathbb{E}(|\mu_i-\hat{\mu}_i|^2) &= R(\mu,\hat{\mu}_{MLE})-(2ap-4a)\mathbb{E}\left(\frac{1}{\sum_j X_j^2}\right)\\ &\quad +a^2\mathbb{E}\left(\frac{1}{\sum_j X_j^2}\right)\\ &= p-(2a(p-2)-a^2)\mathbb{E}\left(\frac{1}{\sum_j X_j^2}\right)\end{aligned}$$

and this is less than p if $2ap-4a-a^2>0$ and in particular at $a=p-2$, which minimizes the risk over $a\in\mathbb{R}$.

Note We have not shown that the James Stein estimator is itself admissible.

Proof (continued)

Putting the pieces together,

$$\begin{aligned}\sum_{i=1}^{p} \mathbb{E}(|\mu_i - \hat{\mu}_i|^2) &= R(\mu, \hat{\mu}_{MLE}) - (2ap - 4a)\mathbb{E}\left(\frac{1}{\sum_j X_j^2}\right) \\ &\quad + a^2 \mathbb{E}\left(\frac{1}{\sum_j X_j^2}\right) \\ &= p - (2a(p-2) - a^2)\mathbb{E}\left(\frac{1}{\sum_j X_j^2}\right)\end{aligned}$$

and this is less than p if $2ap - 4a - a^2 > 0$ and in particular at $a = p - 2$, which minimizes the risk over $a \in \mathbb{R}$.

Note We have not shown that the James Stein estimator is itself admissible.

The risk of the James-Stein estimator

Remember $R(\mu, \hat{\mu}_{MLE}) = p$. When $a = p - 2$

If $\mu_i = 0 \Rightarrow X_i \sim \mathsf{N}(0,1) \Rightarrow \sum_j X_j^2 \sim \chi_p^2 \Rightarrow \mathbb{E}\left(\frac{1}{\sum_j X_j^2}\right) = 1/(p-2)$

$\Rightarrow R(\mu, \hat{\mu}_{JSE}) = 2$.

If $\mu_i = \lambda \Rightarrow X_i = \lambda + Z_i$ where $Z_i \sim \mathsf{N}(0,1)$ and $\sum_j X_j^2 \sim \lambda^2 + \chi_p^2$

$\Rightarrow \mathbb{E}\left(\frac{1}{\sum_j X_j^2}\right) \to 0$ as $\lambda \to \infty$ so $R(\mu, \hat{\mu}_{JSE}) \to p$.

So we get a smaller difference between $R(\mu, \hat{\mu}_{MLE})$ and $R(\mu, \hat{\mu}_{JSE})$ as $\mathbb{E}\left(\frac{1}{\sum_j X_j^2}\right)$ gets smaller.

The risk of the James-Stein estimator

Remember $R(\mu, \hat{\mu}_{MLE}) = p$. When $a = p - 2$

If $\mu_i = 0 \Rightarrow X_i \sim \mathsf{N}(0, 1) \Rightarrow \sum_j X_j^2 \sim \chi_p^2 \Rightarrow \mathbb{E}\left(\frac{1}{\sum_j X_j^2}\right) = 1/(p-2)$

$\Rightarrow R(\mu, \hat{\mu}_{JSE}) = 2$.

If $\mu_i = \lambda \Rightarrow X_i = \lambda + Z_i$ where $Z_i \sim \mathsf{N}(0, 1)$ and $\sum_j X_j^2 \sim \lambda^2 + \chi_p^2$

$\Rightarrow \mathbb{E}\left(\frac{1}{\sum_j X_j^2}\right) \to 0$ as $\lambda \to \infty$ so $R(\mu, \hat{\mu}_{JSE}) \to p$.

So we get a smaller difference between $R(\mu, \hat{\mu}_{MLE})$ and $R(\mu, \hat{\mu}_{JSE})$ as $\mathbb{E}\left(\frac{1}{\sum_j X_j^2}\right)$ gets smaller.

The risk of the James-Stein estimator

Remember $R(\mu, \hat{\mu}_{MLE}) = p$. When $a = p - 2$

If $\mu_i = 0 \Rightarrow X_i \sim \mathsf{N}(0,1) \Rightarrow \sum_j X_j^2 \sim \chi_p^2 \Rightarrow \mathbb{E}\left(\frac{1}{\sum_j X_j^2}\right) = 1/(p-2)$

$\Rightarrow R(\mu, \hat{\mu}_{JSE}) = 2$.

If $\mu_i = \lambda \Rightarrow X_i = \lambda + Z_i$ where $Z_i \sim \mathsf{N}(0,1)$ and $\sum_j X_j^2 \sim \lambda^2 + \chi_p^2$

$\Rightarrow \mathbb{E}\left(\frac{1}{\sum_j X_j^2}\right) \to 0$ as $\lambda \to \infty$ so $R(\mu, \hat{\mu}_{JSE}) \to p$.

So we get a smaller difference between $R(\mu, \hat{\mu}_{MLE})$ and $R(\mu, \hat{\mu}_{JSE})$ as $\mathbb{E}\left(\frac{1}{\sum_j X_j^2}\right)$ gets smaller.

The risk of the James-Stein estimator

Remember $R(\mu, \hat{\mu}_{MLE}) = p$. When $a = p - 2$

If $\mu_i = 0 \Rightarrow X_i \sim \mathsf{N}(0,1) \Rightarrow \sum_j X_j^2 \sim \chi_p^2 \Rightarrow \mathbb{E}\left(\frac{1}{\sum_j X_j^2}\right) = 1/(p-2)$

$\Rightarrow R(\mu, \hat{\mu}_{JSE}) = 2$.

If $\mu_i = \lambda \Rightarrow X_i = \lambda + Z_i$ where $Z_i \sim \mathsf{N}(0,1)$ and $\sum_j X_j^2 \sim \lambda^2 + \chi_p^2$

$\Rightarrow \mathbb{E}\left(\frac{1}{\sum_j X_j^2}\right) \to 0$ as $\lambda \to \infty$ so $R(\mu, \hat{\mu}_{JSE}) \to p$.

So we get a smaller difference between $R(\mu, \hat{\mu}_{MLE})$ and $R(\mu, \hat{\mu}_{JSE})$ as $\mathbb{E}\left(\frac{1}{\sum_j X_j^2}\right)$ gets smaller.

The risk of the James-Stein estimator

Remember $R(\mu, \hat{\mu}_{MLE}) = p$. When $a = p - 2$

If $\mu_i = 0 \Rightarrow X_i \sim \mathsf{N}(0,1) \Rightarrow \sum_j X_j^2 \sim \chi_p^2 \Rightarrow \mathbb{E}\left(\frac{1}{\sum_j X_j^2}\right) = 1/(p-2)$
$\Rightarrow R(\mu, \hat{\mu}_{JSE}) = 2$.

If $\mu_i = \lambda \Rightarrow X_i = \lambda + Z_i$ where $Z_i \sim \mathsf{N}(0,1)$ and $\sum_j X_j^2 \sim \lambda^2 + \chi_p^2$
$\Rightarrow \mathbb{E}\left(\frac{1}{\sum_j X_j^2}\right) \to 0$ as $\lambda \to \infty$ so $R(\mu, \hat{\mu}_{JSE}) \to p$.

So we get a smaller difference between $R(\mu, \hat{\mu}_{MLE})$ and $R(\mu, \hat{\mu}_{JSE})$ as $\mathbb{E}\left(\frac{1}{\sum_j X_j^2}\right)$ gets smaller.

The risk of the James-Stein estimator

Geometrically, the James-Stein estimator shrinks each component of X towards the origin shrinkage estimator).

There is nothing special about the origin. Fix $\mu_0 \in \mathbb{R}^p$ and define

$$\hat{\mu}_{JSE}^{(\mu_0)} = \mu_0 + \left(1 - \frac{p-2}{\|X - \mu_0\|^2}\right)(X - \mu_0).$$

As $R(\hat{\mu}_{JSE}^{(\mu_0)}, \mu + \mu_0) = R(\hat{\mu}_{JSE}, \mu)$, it is also strictly better than X.

Exercise A better estimator is $\bar{X}\mathbf{1}_p + \left(1 - \frac{a}{V}\right)(X - \bar{X}\mathbf{1}_p)$ where $V = \sum_{j-1}^{p}(X_j - \bar{X})^2$ and $\mathbf{1}_p$ is p-dimensional vector of 1's.

The risk of the James-Stein estimator

Geometrically, the James-Stein estimator shrinks each component of X towards the origin shrinkage estimator).

There is nothing special about the origin. Fix $\mu_0 \in \mathbb{R}^p$ and define

$$\hat{\mu}_{JSE}^{(\mu_0)} = \mu_0 + \left(1 - \frac{p-2}{\|X - \mu_0\|^2}\right)(X - \mu_0).$$

As $R(\hat{\mu}_{JSE}^{(\mu_0)}, \mu + \mu_0) = R(\hat{\mu}_{JSE}, \mu)$, it is also strictly better than X.

Exercise A better estimator is $\bar{X}\mathbf{1}_p + \left(1 - \frac{a}{V}\right)(X - \bar{X}\mathbf{1}_p)$ where $V = \sum_{j-1}^{p}(X_j - \bar{X})^2$ and $\mathbf{1}_p$ is p-dimensional vector of 1's.

The risk of the James-Stein estimator

Geometrically, the James-Stein estimator shrinks each component of X towards the origin shrinkage estimator).

There is nothing special about the origin. Fix $\mu_0 \in \mathbb{R}^p$ and define

$$\hat{\mu}_{JSE}^{(\mu_0)} = \mu_0 + \left(1 - \frac{p-2}{\|X - \mu_0\|^2}\right)(X - \mu_0).$$

As $R(\hat{\mu}_{JSE}^{(\mu_0)}, \mu + \mu_0) = R(\hat{\mu}_{JSE}, \mu)$, it is also strictly better than X.

Exercise A better estimator is $\bar{X}\mathbf{1}_p + \left(1 - \frac{a}{V}\right)(X - \bar{X}\mathbf{1}_p)$ where $V = \sum_{j-1}^{p}(X_j - \bar{X})^2$ and $\mathbf{1}_p$ is p-dimensional vector of 1's.

Proof that $R(\hat{\mu}_{JSE}^{(\mu_0)}, \mu + \mu_0) = R(\hat{\mu}_{JSE}, \mu)$

Let us write $Y_i = X_i - \mu_0$. If the parameter value is $\mu + \mu_0$ then $Y_i \sim N(\mu, 1)$. Since

$$
\begin{aligned}
R(\hat{\mu}_{JSE}^{(\mu_0)}, \mu + \mu_0) &= \mathbb{E}_{\mu+\mu_0}[\left(\mu - \left(1 - \frac{p-2}{\|X - \mu_0\|^2}\right)(X - \mu_0)\right)^2] \\
&= \mathbb{E}_{\mu}[\left(\mu - \left(1 - \frac{p-2}{\|Y\|^2}\right)Y\right)^2] \\
&= R(\hat{\mu}_{JSE}, \mu)
\end{aligned}
$$

The risk of the James-Stein estimator

Note that the shrinkage factor becomes negative when $\|X - \mu_0\|^2 < p - 2$. It can be shown that

$$\hat{\mu}_{JSE+}^{(\mu_0)} = \mu_0 + \left(1 - \frac{p-2}{\|X - \mu_0\|^2}\right)_+ (X - \mu_0).$$

dominates strictly $\hat{\mu}_{JSE}^{(\mu_0)}$.

James Stein's estimator is not admissible. Neither is $\hat{\mu}_{JSE+}^{(\mu_0)}$

The risk of the James-Stein estimator

Note that the shrinkage factor becomes negative when $\|X - \mu_0\|^2 < p - 2$. It can be shown that

$$\hat{\mu}_{JSE+}^{(\mu_0)} = \mu_0 + \left(1 - \frac{p-2}{\|X - \mu_0\|^2}\right)_+ (X - \mu_0).$$

dominates strictly $\hat{\mu}_{JSE}^{(\mu_0)}$.

James Stein's estimator is not admissible. Neither is $\hat{\mu}_{JSE+}^{(\mu_0)}$

The risk of the James-Stein estimator

Note that the shrinkage factor becomes negative when $\|X - \mu_0\|^2 < p - 2$. It can be shown that

$$\hat{\mu}_{JSE+}^{(\mu_0)} = \mu_0 + \left(1 - \frac{p-2}{\|X - \mu_0\|^2}\right)_+ (X - \mu_0).$$

dominates strictly $\hat{\mu}_{JSE}^{(\mu_0)}$.
James Stein's estimator is not admissible. Neither is $\hat{\mu}_{JSE+}^{(\mu_0)}$

Generalisation of James-Stein estimator

How crucial are the normality and square error loss assumptions?

1. Normality can be relaxed. Similar but more involved results hold for a wide range of distributions.
2. Can be generalized to different loss functions but ...
3. Does not apply to losses such as $L(\hat{\theta}, \theta) = (\hat{\theta}_1 - \theta)^2$. (then we cant improve on $\hat{\mu} = X$)

The core principle is to employ shrinkage to reduce variance (at the expense of introducing bias).

Generalisation of James-Stein estimator

How crucial are the normality and square error loss assumptions?

1. Normality can be relaxed. Similar but more involved results hold for a wide range of distributions.
2. Can be generalized to different loss functions but ...
3. Does not apply to losses such as $L(\hat{\theta}, \theta) = (\hat{\theta}_1 - \theta)^2$. (then we cant improve on $\hat{\mu} = X$)

The core principle is to employ shrinkage to reduce variance (at the expense of introducing bias).

The baseball example

Player	n_i	Z_i	π_i
Baines	415	0.284	0.289
Barfield	476	0.246	0.256
Bell	583	0.254	0.265
Biggio	555	0.276	0.287
Bonds	519	0.301	0.297
Bonilla	625	0.280	0.279
Brett	544	0.329	0.305
Brooks Jr.	568	0.266	0.269
Browne	513	0.267	0.271

$n_i =$ number of times at bat, $Z_i =$ batting average during 1990 season, $\pi_i =$ true batting average (overall career average).

Model : $Z_i = n_i^{-1} Bin(n_i, \pi_i)$.

transform

$X_i = \sqrt{n_i} \sin^{-1}(2Z_i - 1) \simeq N(\theta_i, 1)$

with $\theta_i = \sqrt{n_i} \sin^{-1}(2\pi_i - 1)$.

To see this

$$\begin{aligned} X_i - \theta_i &= g(Z_i) - g(\pi_i) \simeq g'(\pi_i)(Z_i - \pi_i) \\ &= = \frac{\sqrt{n_i}(Z_i - \pi_i)}{\sqrt{\pi_i(1-\pi_i)}}. \end{aligned}$$

The baseball example

Player	n_i	Z_i	π_i
Baines	415	0.284	0.289
Barfield	476	0.246	0.256
Bell	583	0.254	0.265
Biggio	555	0.276	0.287
Bonds	519	0.301	0.297
Bonilla	625	0.280	0.279
Brett	544	0.329	0.305
Brooks Jr.	568	0.266	0.269
Browne	513	0.267	0.271

$n_i =$ number of times at bat, $Z_i =$ batting average during 1990 season, $\pi_i =$ true batting average (overall career average).

Model : $Z_i = n_i^{-1} Bin(n_i, \pi_i)$.

transform

$X_i = \sqrt{n_i} \sin^{-1}(2Z_i - 1) \simeq N(\theta_i, 1)$

with $\theta_i = \sqrt{n_i} \sin^{-1}(2\pi_i - 1)$.

To see this

$$\begin{aligned} X_i - \theta_i &= g(Z_i) - g(\pi_i) \simeq g'(\pi_i)(Z_i - \pi_i) \\ &= = \frac{\sqrt{n_i}(Z_i - \pi_i)}{\sqrt{\pi_i(1-\pi_i)}}. \end{aligned}$$

The baseball example

Player	n_i	Z_i	π_i
Baines	415	0.284	0.289
Barfield	476	0.246	0.256
Bell	583	0.254	0.265
Biggio	555	0.276	0.287
Bonds	519	0.301	0.297
Bonilla	625	0.280	0.279
Brett	544	0.329	0.305
Brooks Jr.	568	0.266	0.269
Browne	513	0.267	0.271

$n_i =$ number of times at bat, $Z_i =$ batting average during 1990 season, $\pi_i =$ true batting average (overall career average).

Model : $Z_i = n_i^{-1} Bin(n_i, \pi_i)$.

transform

$X_i = \sqrt{n_i}\sin^{-1}(2Z_i - 1) \simeq N(\theta_i, 1)$

with $\theta_i = \sqrt{n_i}\sin^{-1}(2\pi_i - 1)$.

To see this

$$\begin{aligned} X_i - \theta_i &= g(Z_i) - g(\pi_i) \simeq g'(\pi_i)(Z_i - \pi_i) \\ &= = \frac{\sqrt{n_i}(Z_i - \pi_i)}{\sqrt{\pi_i(1-\pi_i)}}. \end{aligned}$$

The baseball example

Player	n_i	Z_i	π_i
Baines	415	0.284	0.289
Barfield	476	0.246	0.256
Bell	583	0.254	0.265
Biggio	555	0.276	0.287
Bonds	519	0.301	0.297
Bonilla	625	0.280	0.279
Brett	544	0.329	0.305
Brooks Jr.	568	0.266	0.269
Browne	513	0.267	0.271

$n_i =$ number of times at bat, $Z_i =$ batting average during 1990 season, $\pi_i =$ true batting average (overall career average).

Model : $Z_i = n_i^{-1} Bin(n_i, \pi_i)$.

transform

$X_i = \sqrt{n_i} \sin^{-1}(2Z_i - 1) \simeq N(\theta_i, 1)$

with $\theta_i = \sqrt{n_i} \sin^{-1}(2\pi_i - 1)$.

To see this

$$\begin{aligned} X_i - \theta_i &= g(Z_i) - g(\pi_i) \simeq g'(\pi_i)(Z_i - \pi_i) \\ &= = \frac{\sqrt{n_i}(Z_i - \pi_i)}{\sqrt{\pi_i(1-\pi_i)}}. \end{aligned}$$

The baseball example

So $\|X - \theta\|^2 = 2.56 < 9$.
Using $\theta_0 = \sqrt{n}\sin^{-1}(2\pi_0 - 1)$ with $\pi_0 = 0.275$ we get

$$\|\hat{\theta}^{(\theta_0)}_{JSE+} - \theta\|^2 = 1.50.$$

The baseball example

So $\|X - \theta\|^2 = 2.56 < 9$.
Using $\theta_0 = \sqrt{\bar{n}} \sin^{-1}(2\pi_0 - 1)$ with $\pi_0 = 0.275$ we get

$$\|\hat{\theta}^{(\theta_0)}_{JSE+} - \theta\|^2 = 1.50.$$

The baseball example 2

	Y_i	n_i	p_i	AB	X_i	JS_i	μ_i	HR	$\hat{HR}$	$\hat{HR}_s$
McGwire	7	58	0.138	509	−6.56	−7.12	−6.18	70	61	50
Sosa	9	59	0.103	643	−5.90	−6.71	−7.06	66	98	75
Griffey	4	74	0.089	633	−9.48	−8.95	−8.32	56	34	43
Castilla	7	84	0.071	645	−9.03	−8.67	−9.44	46	54	61
Gonzalez	3	69	0.074	606	−9.56	−9.01	−8.46	45	26	35
Galaragga	6	63	0.079	555	−7.49	−7.71	−7.94	44	53	48
Palmeiro	2	60	0.070	619	−9.32	−8.86	−8.04	43	21	28
Vaughn	10	54	0.066	609	−5.01	−6.15	−7.73	40	113	78
Bonds	2	53	0.067	552	−8.59	−8.40	−7.62	37	21	24
Bagwell	2	60	0.063	540	−9.32	−8.86	−8.23	34	18	24
Piazza	4	66	0.057	561	−8.72	−8.48	−8.84	32	34	38
Thome	3	66	0.068	440	−9.27	−8.83	−8.47	30	20	25
Thomas	2	72	0.050	585	−10.49	−9.59	−9.52	29	16	28
T. Martinez	5	64	0.053	531	−8.03	−8.05	−8.86	28	41	41
Walker	3	42	0.051	454	−6.67	−7.19	−7.24	23	32	24
Burks	2	38	0.042	504	−6.83	−7.29	−7.15	21	27	19
Buhner	6	58	0.062	244	−6.98	−7.38	−8.15	15	25	21

$Y_i = \#$ home runs in pre-season, $n_i = \#$ times at bat, $p_i =$ true full-season strike rate.

Naive estimator is $\hat{p}_i = Y_i/n_i$.

The baseball example 2

	Y_i	n_i	p_i	AB	X_i	JS_i	μ_i	HR	$\hat{HR}$	$\hat{HR}_s$
McGwire	7	58	0.138	509	−6.56	−7.12	−6.18	70	61	50
Sosa	9	59	0.103	643	−5.90	−6.71	−7.06	66	98	75
Griffey	4	74	0.089	633	−9.48	−8.95	−8.32	56	34	43
Castilla	7	84	0.071	645	−9.03	−8.67	−9.44	46	54	61
Gonzalez	3	69	0.074	606	−9.56	−9.01	−8.46	45	26	35
Galaragga	6	63	0.079	555	−7.49	−7.71	−7.94	44	53	48
Palmeiro	2	60	0.070	619	−9.32	−8.86	−8.04	43	21	28
Vaughn	10	54	0.066	609	−5.01	−6.15	−7.73	40	113	78
Bonds	2	53	0.067	552	−8.59	−8.40	−7.62	37	21	24
Bagwell	2	60	0.063	540	−9.32	−8.86	−8.23	34	18	24
Piazza	4	66	0.057	561	−8.72	−8.48	−8.84	32	34	38
Thome	3	66	0.068	440	−9.27	−8.83	−8.47	30	20	25
Thomas	2	72	0.050	585	−10.49	−9.59	−9.52	29	16	28
T. Martinez	5	64	0.053	531	−8.03	−8.05	−8.86	28	41	41
Walker	3	42	0.051	454	−6.67	−7.19	−7.24	23	32	24
Burks	2	38	0.042	504	−6.83	−7.29	−7.15	21	27	19
Buhner	6	58	0.062	244	−6.98	−7.38	−8.15	15	25	21

$Y_i = \#$ home runs in pre-season, $n_i = \#$ times at bat, $p_i =$ true full-season strike rate.
Naive estimator is $\hat{p}_i = Y_i/n_i$.

The baseball example 2

	Y_i	n_i	p_i	AB	X_i	JS_i	μ_i	HR	$\hat{HR}$	$\hat{HR}_s$
McGwire	7	58	0.138	509	−6.56	−7.12	−6.18	70	61	50
Sosa	9	59	0.103	643	−5.90	−6.71	−7.06	66	98	75
Griffey	4	74	0.089	633	−9.48	−8.95	−8.32	56	34	43
Castilla	7	84	0.071	645	−9.03	−8.67	−9.44	46	54	61
Gonzalez	3	69	0.074	606	−9.56	−9.01	−8.46	45	26	35
Galaragga	6	63	0.079	555	−7.49	−7.71	−7.94	44	53	48
Palmeiro	2	60	0.070	619	−9.32	−8.86	−8.04	43	21	28
Vaughn	10	54	0.066	609	−5.01	−6.15	−7.73	40	113	78
Bonds	2	53	0.067	552	−8.59	−8.40	−7.62	37	21	24
Bagwell	2	60	0.063	540	−9.32	−8.86	−8.23	34	18	24
Piazza	4	66	0.057	561	−8.72	−8.48	−8.84	32	34	38
Thome	3	66	0.068	440	−9.27	−8.83	−8.47	30	20	25
Thomas	2	72	0.050	585	−10.49	−9.59	−9.52	29	16	28
T. Martinez	5	64	0.053	531	−8.03	−8.05	−8.86	28	41	41
Walker	3	42	0.051	454	−6.67	−7.19	−7.24	23	32	24
Burks	2	38	0.042	504	−6.83	−7.29	−7.15	21	27	19
Buhner	6	58	0.062	244	−6.98	−7.38	−8.15	15	25	21

As before define $f_n(y) = n^{1/2} \sin^{-1}(2y - 1)$ and $X_i = f_{n_i}(Y_i/n_i), \theta_i = f_{n_i}(p_i)$. so that $X_i \sim N(\theta_i, 1)$.

The baseball example 2

Use the estimator

$$JS_i = \bar{X} + (1 - (p-3)/V)(X_i - \bar{X})$$

where $V = \|X - \bar{X}\|^2 = \sum(X_i - \bar{X})^2$ and $\bar{X} = \frac{1}{p}\sum X_i$. The true θ_i must be clustered more closely around their mean than the X_i.
$\sum(X_i - \theta_i)^2 = 19.68$ compared with $\sum(JS_i - \theta_i)^2 = 8.07$.

The baseball example 2

Use the estimator

$$JS_i = \bar{X} + (1 - (p-3)/V)(X_i - \bar{X})$$

where $V = \|X - \bar{X}\|^2 = \sum(X_i - \bar{X})^2$ and $\bar{X} = \frac{1}{p}\sum X_i$. The true θ_i must be clustered more closely around their mean than the X_i.
$\sum(X_i - \theta_i)^2 = 19.68$ compared with $\sum(JS_i - \theta_i)^2 = 8.07$.

The baseball example 2

Use the estimator

$$JS_i = \bar{X} + (1 - (p-3)/V)(X_i - \bar{X})$$

where $V = \|X - \bar{X}\|^2 = \sum(X_i - \bar{X})^2$ and $\bar{X} = \frac{1}{p}\sum X_i$. The true θ_i must be clustered more closely around their mean than the X_i.
$\sum(X_i - \theta_i)^2 = 19.68$ compared with $\sum(JS_i - \theta_i)^2 = 8.07$.

The baseball example 2

	Y_i	n_i	p_i	AB	X_i	JS_i	μ_i	HR	$\hat{HR}$	$\hat{HR}_s$
McGwire	7	58	0.138	509	−6.56	−7.12	−6.18	70	61	50
Sosa	9	59	0.103	643	−5.90	−6.71	−7.06	66	98	75
Griffey	4	74	0.089	633	−9.48	−8.95	−8.32	56	34	43
Castilla	7	84	0.071	645	−9.03	−8.67	−9.44	46	54	61
Gonzalez	3	69	0.074	606	−9.56	−9.01	−8.46	45	26	35
Galaragga	6	63	0.079	555	−7.49	−7.71	−7.94	44	53	48
Palmeiro	2	60	0.070	619	−9.32	−8.86	−8.04	43	21	28
Vaughn	10	54	0.066	609	−5.01	−6.15	−7.73	40	113	78
Bonds	2	53	0.067	552	−8.59	−8.40	−7.62	37	21	24
Bagwell	2	60	0.063	540	−9.32	−8.86	−8.23	34	18	24
Piazza	4	66	0.057	561	−8.72	−8.48	−8.84	32	34	38
Thome	3	66	0.068	440	−9.27	−8.83	−8.47	30	20	25
Thomas	2	72	0.050	585	−10.49	−9.59	−9.52	29	16	28
T. Martinez	5	64	0.053	531	−8.03	−8.05	−8.86	28	41	41
Walker	3	42	0.051	454	−6.67	−7.19	−7.24	23	32	24
Burks	2	38	0.042	504	−6.83	−7.29	−7.15	21	27	19
Buhner	6	58	0.062	244	−6.98	−7.38	−8.15	15	25	21

HR is actual $\#$ of home runs in the whole season, $\hat{HR}$ is just the extrapolation from the pre-season, $\hat{HR}_s$ is the prediction based on the JS estimator. It does better on aggregate.