Estimating the multivariate mean vector

March 7, 2024

Stein Paradox

- In 1956, it was shown that for a simple example the regular Maximal likelihood estimator is not optimal.

References

- C. Stein (1956) *Inadmissibility of the usual estimator of the mean of a multivariate normal distribution*, Proc. Third Berkeley Symposium, 1, 197–206, Univ. California Press
- W. James and C. Stein (1961), *Estimation with quadratic loss*, Proc. Fourth Berkeley Symposium, 1, 361–380.

Stein Paradox

- In 1956, it was shown that for a simple example the regular Maximal likelihood estimator is not optimal.
- We will look into a strictly better shrinkage estimator from 1961.

References

- C. Stein (1956) *Inadmissibility of the usual estimator of the mean of a multivariate normal distribution*, Proc. Third Berkeley Symposium, 1, 197–206, Univ. California Press
- W. James and C. Stein (1961), *Estimation with quadratic loss*, Proc. Fourth Berkeley Symposium, 1, 361–380.

Loss function

- Suppose that

$$\mathbf{Y} \sim \mathcal{N}_{\mathrm{p}}\left(\boldsymbol{\mu}, \mathbf{I}\right)$$

Loss function

- Suppose that

$$\mathbf{Y} \sim \mathcal{N}_{\mathsf{p}}\left(\boldsymbol{\mu}, \mathbf{I}\right)$$

- Goal find an estimator of $\boldsymbol{\mu}$ given we observed a single observation, $\mathbf{Y} = \mathbf{y}$.

Loss function

- Suppose that

$$\mathbf{Y} \sim \mathcal{N}_p(\boldsymbol{\mu}, \mathbf{I})$$

- Goal find an estimator of $\boldsymbol{\mu}$ given we observed a single observation, $\mathbf{Y} = \mathbf{y}$.
- What is the best estimator in terms of squared error

$$L(\hat{\boldsymbol{\mu}}, \boldsymbol{\mu}) = ||\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}||^2 = \sum_{i=1}^{p} (\hat{\mu}_i - \mu_i)^2$$

- The Maximum likelihood is the sample mean $\hat{\boldsymbol{\mu}}^{mle} = \mathbf{y}$ (recall $n = 1$).

MLE estimator

- For $\hat{\mu}^{\text{mle}}(\mathbf{Y}) = \mathbf{Y}$ we can analyse the expected loss

$$\mathbb{E}_{\mathbf{Y}}\left[||\hat{\boldsymbol{\mu}}^{\text{mle}}(\mathbf{Y}) - \boldsymbol{\mu}||^2\right] = \mathbb{E}_{\mathbf{Y}}\left[||\mathbf{Y} - \boldsymbol{\mu}||^2\right]$$

MLE estimator

- For $\hat{\mu}^{\text{mle}}(\mathbf{Y}) = \mathbf{Y}$ we can analyse the expected loss

$$\mathbb{E}_{\mathbf{Y}}\left[||\hat{\mu}^{\text{mle}}(\mathbf{Y}) - \boldsymbol{\mu}||^2\right] = \mathbb{E}_{\mathbf{Y}}\left[||\mathbf{Y} - \boldsymbol{\mu}||^2\right]$$

Using $\mathbf{Y} = \boldsymbol{\mu} + \mathbf{Z}$ where $\mathbf{Z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ we get

$$\mathbb{E}_{\mathbf{Y}}\left[||\mathbf{Y} - \boldsymbol{\mu}||^2\right] = \mathbb{E}_{\mathbf{Z}}\left[||Z||^2\right] = p.$$

MLE estimator

- For $\hat{\mu}^{\text{mle}}(\mathbf{Y}) = \mathbf{Y}$ we can analyse the expected loss

$$\mathbb{E}_{\mathbf{Y}}\left[||\hat{\boldsymbol{\mu}}^{\text{mle}}(\mathbf{Y}) - \boldsymbol{\mu}||^2\right] = \mathbb{E}_{\mathbf{Y}}\left[||\mathbf{Y} - \boldsymbol{\mu}||^2\right]$$

Using $\mathbf{Y} = \boldsymbol{\mu} + \mathbf{Z}$ where $\mathbf{Z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ we get

$$\mathbb{E}_{\mathbf{Y}}\left[||\mathbf{Y} - \boldsymbol{\mu}||^2\right] = \mathbb{E}_{\mathbf{Z}}\left[||Z||^2\right] = p.$$

- For $p = 1, 2$ this is the optimal estimator, however for $p \geq 3$ it is not the case!

Bias-Variance Tradeoff

$$\mathrm{MSE}(\hat{\mu}_i) = \mathrm{E}(\hat{\mu}_i - \mu_i)^2 = \mathrm{B}_i^2 + \mathrm{Var}_i,$$

where $\mathrm{B}_i = \mathrm{E}\hat{\mu}_i - \mu_i$ is the bias of $\hat{\mu}_i$

and $\mathrm{Var}_i = \mathrm{E}(\hat{\mu}_i - \mathrm{E}(\hat{\mu}_i))^2$ is the variance of $\hat{\mu}_i$.

Bias-Variance Tradeoff

$$\mathrm{MSE}(\hat{\mu}_i) = \mathrm{E}(\hat{\mu}_i - \mu_i)^2 = \mathrm{B}_i^2 + \mathrm{Var}_i,$$

where $\mathrm{B}_i = \mathrm{E}\hat{\mu}_i - \mu_i$ is the bias of $\hat{\mu}_i$

and $\mathrm{Var}_i = \mathrm{E}(\hat{\mu}_i - \mathrm{E}(\hat{\mu}_i))^2$ is the variance of $\hat{\mu}_i$.

In our problem $\mathrm{E}(\hat{\mu}_{\mathrm{MLE}}) = \mu$ and $\mathrm{MSE}(\hat{\mu}_{\mathrm{MLE}}) = \sum_{i=1}^{p} \mathrm{Var}_i$

Bias-Variance Tradeoff

$$\mathrm{MSE}(\hat{\mu}_i) = E(\hat{\mu}_i - \mu_i)^2 = B_i^2 + \mathrm{Var}_i,$$

where $B_i = E\hat{\mu}_i - \mu_i$ is the bias of $\hat{\mu}_i$

and $\mathrm{Var}_i = E(\hat{\mu}_i - E(\hat{\mu}_i))^2$ is the variance of $\hat{\mu}_i$.

In our problem $E(\hat{\mu}_{MLE}) = \mu$ and $\mathrm{MSE}(\hat{\mu}_{MLE}) = \sum_{i=1}^{p} \mathrm{Var}_i$

Can we improve MSE by introducing some bias and reducing the variance ?

Shrinking towards zero

Consider the estimate $\hat{\mu}_{\mathsf{c}} = \mathsf{c}\hat{\mu}_{\mathsf{MLE}}$

Shrinking towards zero

Consider the estimate $\hat{\mu}_c = c\hat{\mu}_{MLE}$

$$B_i(c) = c\mu_i - \mu_i = (c-1)\mu_i \text{ and } Var_i(c) = c^2\sigma^2$$

Shrinking towards zero

Consider the estimate $\hat{\mu}_c = c\hat{\mu}_{MLE}$

$$B_i(c) = c\mu_i - \mu_i = (c-1)\mu_i \text{ and } Var_i(c) = c^2\sigma^2$$

$$MSE_i(c) = (c-1)^2\mu_i^2 + c^2\sigma^2$$

Shrinking towards zero

Consider the estimate $\hat{\mu}_c = c\hat{\mu}_{MLE}$

$$B_i(c) = c\mu_i - \mu_i = (c-1)\mu_i \text{ and } Var_i(c) = c^2\sigma^2$$

$$MSE_i(c) = (c-1)^2\mu_i^2 + c^2\sigma^2$$

$$MSE(c) = E||\hat{\mu}_c - \mu||^2 = (c-1)^2||\mu||^2 + c^2 p\sigma^2$$

Shrinking towards zero

Consider the estimate $\hat{\mu}_c = c\hat{\mu}_{MLE}$

$$B_i(c) = c\mu_i - \mu_i = (c-1)\mu_i \text{ and } Var_i(c) = c^2\sigma^2$$

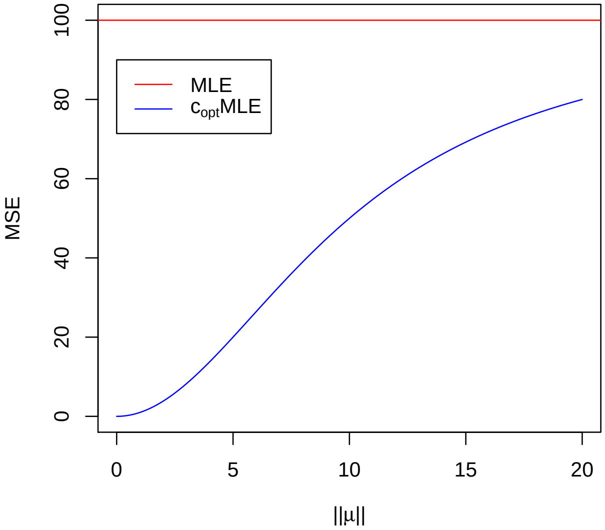
$$MSE_i(c) = (c-1)^2\mu_i^2 + c^2\sigma^2$$

$$MSE(c) = E||\hat{\mu}_c - \mu||^2 = (c-1)^2||\mu||^2 + c^2 p\sigma^2$$

Using elementary calculus we can show that the optimal value of c is equal to

$$c_{opt} = \text{argmin}_{c\in R} MSE(c) = \frac{||\mu||^2}{||\mu||^2 + p\sigma^2} \in [0, 1) \quad .$$

Improvement in MSE, $p = 100, \sigma = 1$



Shrinking towards the common mean

Consider an estimator

$$\hat{\mu}_{\mathsf{d}} = (1 - \mathsf{d})\hat{\mu}_{\mathsf{MLE}} + \mathsf{d}\bar{\mathsf{X}}$$

Shrinking towards the common mean

Consider an estimator

$$\hat{\mu}_d = (1 - d)\hat{\mu}_{MLE} + d\bar{X}$$

$$d_{opt} = \frac{\sigma^2}{\sigma^2 + Var(\mu)} \in (0, 1], \quad \text{with} \quad Var(\mu) = \frac{1}{p-1}\sum(\mu_i - \bar{\mu})^2.$$

Shrinking towards the common mean

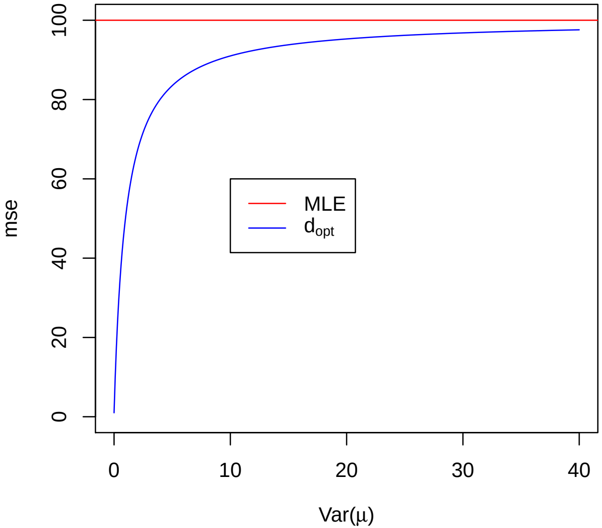
Consider an estimator

$$\hat{\mu}_d = (1 - d)\hat{\mu}_{MLE} + d\bar{X}$$

$$d_{opt} = \frac{\sigma^2}{\sigma^2 + \mathrm{Var}(\mu)} \in (0, 1], \quad \text{with} \quad \mathrm{Var}(\mu) = \frac{1}{p-1}\sum(\mu_i - \bar{\mu})^2.$$

$$d = 1 \quad \text{if and only if} \quad \mu_1 = \ldots = \mu_p$$

Improvement in MSE, p = 100, $\sigma = 1$



James-Stein estimators (1961)

$$c_{opt} = \frac{||\mu||^2}{||\mu||^2 + p\sigma^2} = \left(1 - \frac{p\sigma^2}{||\mu||^2 + p\sigma^2} = \right) = \left(1 - \frac{p\sigma^2}{E||X||^2}\right)$$

James-Stein estimators (1961)

$$c_{opt} = \frac{||\mu||^2}{||\mu||^2 + p\sigma^2} = \left(1 - \frac{p\sigma^2}{||\mu||^2 + p\sigma^2} =\right) = \left(1 - \frac{p\sigma^2}{E||X||^2}\right)$$

$$c_{JS} = \left(1 - \frac{(p-2)\sigma^2}{||X||^2}\right)$$

James-Stein estimators (1961)

$$c_{opt} = \frac{||\mu||^2}{||\mu||^2 + p\sigma^2} = \left(1 - \frac{p\sigma^2}{||\mu||^2 + p\sigma^2} = \right) = \left(1 - \frac{p\sigma^2}{E||X||^2}\right)$$

$$c_{JS} = \left(1 - \frac{(p-2)\sigma^2}{||X||^2}\right)$$

$$d_{opt} = \frac{\sigma^2}{\sigma^2 + Var(\mu)}$$

James-Stein estimators (1961)

$$c_{opt} = \frac{||\mu||^2}{||\mu||^2 + p\sigma^2} = \left(1 - \frac{p\sigma^2}{||\mu||^2 + p\sigma^2} = \right) = \left(1 - \frac{p\sigma^2}{E||X||^2}\right)$$

$$c_{JS} = \left(1 - \frac{(p-2)\sigma^2}{||X||^2}\right)$$

$$d_{opt} = \frac{\sigma^2}{\sigma^2 + \mathrm{Var}(\mu)}$$

$$d_{JS} = \frac{p-3}{p-1} \frac{\sigma^2}{\mathrm{Var}(X)}$$

James-Stein estimators (1961)

$$c_{opt} = \frac{||\mu||^2}{||\mu||^2 + p\sigma^2} = \left(1 - \frac{p\sigma^2}{||\mu||^2 + p\sigma^2} = \right) = \left(1 - \frac{p\sigma^2}{E||X||^2}\right)$$

$$c_{JS} = \left(1 - \frac{(p-2)\sigma^2}{||X||^2}\right)$$

$$d_{opt} = \frac{\sigma^2}{\sigma^2 + \mathrm{Var}(\mu)}$$

$$d_{JS} = \frac{p-3}{p-1} \frac{\sigma^2}{\mathrm{Var}(X)}$$

Stein's identity

$$X \sim N(\mu, \sigma^2 I_{p \times p})$$

Stein's identity

$$X \sim N(\mu, \sigma^2 I_{p \times p})$$

$$\hat{\mu} = X + g(X)$$

Stein's identity

$$X \sim N(\mu, \sigma^2 I_{p \times p})$$

$$\hat{\mu} = X + g(X)$$

Under weak regularity conditions on $g(\cdot)$

Stein's identity

$$X \sim N(\mu, \sigma^2 I_{p\times p})$$

$$\hat{\mu} = X + g(X)$$

Under weak regularity conditions on $g(\cdot)$

$$E||\hat{\mu} - \mu||^2 = E\left(||g(X)||^2 + 2\sigma^2 \sum_{i=1}^{p} \frac{\partial g_i(X)}{\partial X_i} \right) + p\sigma^2$$

James Stein estimator

Theorem (James and Stein (1961))

Let $\mathbf{Y} \sim \mathcal{N}_p\left(\boldsymbol{\mu}, \sigma^2 \mathbf{I}\right)$, *and* $L(\hat{\boldsymbol{\mu}}, \boldsymbol{\mu}) = \mathbb{E}_Y\left[||\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}||^2\right]$ *then for* $p \geq 3$

$$L(\hat{\boldsymbol{\mu}}^{JS}, \boldsymbol{\mu}) \leq L(\hat{\boldsymbol{\mu}}^{MLE}, \boldsymbol{\mu}).$$

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

[1]

[1] Samworth, Richard J., and Statslab Cambridge. "Stein's paradox." eureka 62 (2012): 38-41.

James Stein estimator

Theorem (James and Stein (1961))

Let $\mathbf{Y} \sim \mathcal{N}_p\left(\boldsymbol{\mu}, \sigma^2 \mathbf{I}\right)$, *and* $L(\hat{\boldsymbol{\mu}}, \boldsymbol{\mu}) = \mathbb{E}_Y\left[||\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}||^2\right]$ *then for* $p \geq 3$

$$L(\hat{\boldsymbol{\mu}}^{JS}, \boldsymbol{\mu}) \leq L(\hat{\boldsymbol{\mu}}^{MLE}, \boldsymbol{\mu}).$$

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

1

- One can further prove that $\hat{\boldsymbol{\mu}}^{JS+} = \left(1 - \sigma^2 \frac{p-2}{||\mathbf{Y}||^2}\right)_+ \mathbf{Y}$ is even better.

[1] Samworth, Richard J., and Statslab Cambridge. "Stein's paradox." eureka 62 (2012): 38-41.

Baseball data

- We want to predict the batting average of eighteen baseball players the season 1970. We will use the betting average of the players for each players first 45 bats.

```r
library(Rgbp)
data(baseball)
p <- baseball$Hits/baseball$At.Bats
p.true <- baseball$RemainingAverage
p.MLE <- p
```

Baseball data

- We want to predict the batting average of eighteen baseball players the season 1970. We will use the betting average of the players for each players first 45 bats.
- Number of hits $H_i \sim \text{Bin}(n = 45, p_i)$.

```
library(Rgbp)
data(baseball)
p <- baseball$Hits/baseball$At.Bats
p.true <- baseball$RemainingAverage
p.MLE <- p
```

Baseball data

- We want to predict the batting average of eighteen baseball players the season 1970. We will use the betting average of the players for each players first 45 bats.
- Number of hits $H_i \sim \text{Bin}(n = 45, p_i)$.
- The MLE estimator is $\hat{p}_i = \frac{h_i}{n}$.

```
library(Rgbp)
data(baseball)
p <- baseball$Hits/baseball$At.Bats
p.true <- baseball$RemainingAverage
p.MLE <- p
```

simple James Stein estimator I

- To use the James Stein estimator we need to know the standard deviation which we estimate from the data.

```
pbar  <- mean(p)
sigma2 <- pbar * (1-pbar)/baseball$At.Bats
p.JS <- (1 -sigma2/(length(p)-2) ) * p
c.JS <- 1 -sigma2/(length(p)-2) =0.0.9997292
```

simple James Stein estimator I

- To use the James Stein estimator we need to know the standard deviation which we estimate from the data.
- Using $\mathbb{V}\left[\frac{H_i}{n}\right] = \frac{1}{n}p_i(1-p_i)$ if $H_i \sim \mathrm{Bin}(n, p_i)$

```
pbar  <- mean(p)
sigma2 <- pbar * (1-pbar)/baseball$At.Bats
p.JS <- (1 -sigma2/(length(p)-2) ) * p
c.JS <- 1 -sigma2/(length(p)-2) =0.0.9997292
```

- To use the James Stein estimator we need to know the standard deviation which we estimate from the data.
- Using $\mathbb{V}\left[\frac{H_i}{n}\right] = \frac{1}{n}p_i(1-p_i)$ if $H_i \sim \mathrm{Bin}(n, p_i)$
- Pool the estimate.

```
pbar  <- mean(p)
sigma2 <- pbar * (1-pbar)/baseball$At.Bats
p.JS <- (1 -sigma2/(length(p)-2) ) * p
c.JS <- 1 -sigma2/(length(p)-2) =0.0.9997292
```

simple James Stein estimator II

If we now compare square root of the mean square error:

```r
Loss.MLE  = sqrt(mean((p.MLE - p.true)^2))
Loss.JS = sqrt(mean((p.JS - p.true)^2))
cat('Loss.MLE = ',  round(Loss.MLE,digits = 4),'\n'
```

```
## Loss.MLE =  0.069
```

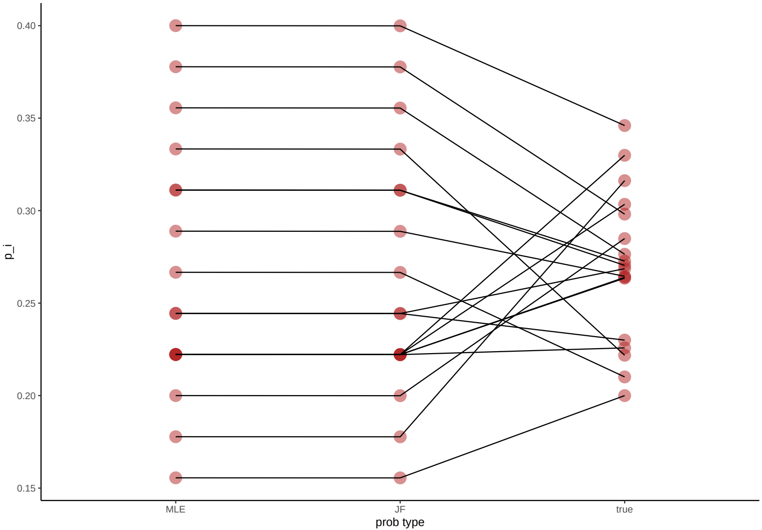
```r
cat('Loss.JS = ', round(Loss.JS,digits=4),'\n')
```

```
## Loss.JS =  0.069
```

```r
cat('RATIO = ',round(Loss.JS/Loss.MLE,6),'\n')
```

```
## RATIO =  0.999837
```

simple James Stein estimator III



simple James Stein estimator IV

```
dJS <- (length(p)-3)sigma2/(((length(p)-1) Var(pMLE))
dJS =0.7883
pbar <- mean(p)
p.JS <- (1-d) pMLE + d pbar
```

```
cat('Loss.JS = ', round(Loss.JS,digits=4),'\n')
```

```
## Loss.JS =  0.0384
```

```
cat('RATIO = ',round(Loss.JS/Loss.MLE,6),'\n')
```

```
## RATIO =  0.555981 (compared to  0.999837)
```

Empirical Bayes interpretation

$$\mu_1, \ldots, \mu_p - \text{ iid } N(0, \tau^2)$$

Empirical Bayes interpretation

$$\mu_1, \ldots, \mu_p - \quad \text{iid} \quad N(0, \tau^2)$$

$$E(\mu_i | X_i) = \left(1 - \frac{\sigma^2}{\sigma^2 + \tau^2}\right) X_i$$

Empirical Bayes interpretation

$$\mu_1, \ldots, \mu_p - \quad \text{iid} \quad N(0, \tau^2)$$

$$E(\mu_i | X_i) = \left(1 - \frac{\sigma^2}{\sigma^2 + \tau^2}\right) X_i$$

$$\text{Var } X_i = \tau^2 + \sigma^2 = E\frac{||X||^2}{p}$$

Empirical Bayes interpretation

$$\mu_1, \ldots, \mu_p - \quad \text{iid} \quad N(0, \tau^2)$$

$$E(\mu_i | X_i) = \left(1 - \frac{\sigma^2}{\sigma^2 + \tau^2}\right) X_i$$

$$\text{Var } X_i = \tau^2 + \sigma^2 = E \frac{||X||^2}{p}$$

$$E \frac{(p-2)}{||X||^2} = \frac{1}{\sigma^2 + \tau^2}$$

Empirical Bayes interpretation

$$\mu_1, \ldots, \mu_p - \quad \text{iid} \quad N(0, \tau^2)$$

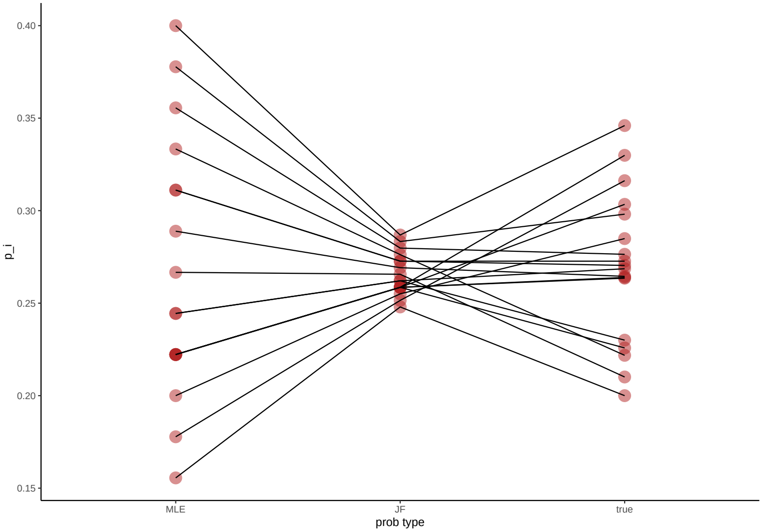
$$E(\mu_i | X_i) = \left(1 - \frac{\sigma^2}{\sigma^2 + \tau^2}\right) X_i$$

$$\text{Var } X_i = \tau^2 + \sigma^2 = E \frac{||X||^2}{p}$$

$$E \frac{(p-2)}{||X||^2} = \frac{1}{\sigma^2 + \tau^2}$$

$$\hat{\mu}_{EB} = \left(1 - \frac{\sigma^2 (p-2)}{||X||^2}\right) X = \mu_{JS}$$

simple James Stein estimator VI



Hard thresholding

When signal is sparse even better results can be obtained by hard thresholding

$$\hat{\mu}_i = \begin{cases} X_i & \text{when} \quad H_{0i} \text{ is rejected} \\ 0 & \text{when} \quad H_{0i} \text{ is not rejected} \end{cases}, \tag{1}$$

where the decisions are made by Bonferroni or BH multiple testing procedures. Bonferroni is optimal for very sparse signals while BH ”adapts” to the unknown sparsity (see Abramovich, Benjamini, Donoho and Johnstone, Ann.Statist. 2006)

Optimality with respect to Bayes risk

γ_0 - loss for type I error, γ_A - loss for type II error

Optimality with respect to Bayes risk

γ_0 - loss for type I error, γ_A - loss for type II error

$\mu_i \sim (1 - \epsilon)\delta_0 + \epsilon N(0, \tau^2)$

Optimality with respect to Bayes risk

γ_0 - loss for type I error, γ_A - loss for type II error

$\mu_i \sim (1-\epsilon)\delta_0 + \epsilon N(0, \tau^2)$

Bayes oracle $\rightarrow$ Bayes classifier:

$$\text{Reject } H_{0i} \quad \text{when} \quad \frac{\phi(X_i; \sigma^2+\tau^2)}{\phi(X_i; \sigma^2)} > \frac{1-\epsilon}{\epsilon}\frac{\gamma_0}{\gamma_A}$$

Optimality with respect to Bayes risk

γ_0 - loss for type I error, γ_A - loss for type II error

$\mu_i \sim (1-\epsilon)\delta_0 + \epsilon N(0, \tau^2)$

Bayes oracle $\rightarrow$ Bayes classifier:

$$\text{Reject } H_{0i} \quad \text{when} \quad \frac{\phi(X_i; \sigma^2 + \tau^2)}{\phi(X_i; \sigma^2)} > \frac{1-\epsilon}{\epsilon} \frac{\gamma_0}{\gamma_A}$$

[B., Ghosh, Tokdar, *IMS Collections* 2008] and [B., Ghosh, Ochman, Tokdar *QREI*, 2007]: empirical comparison of BH with several Bayesian multiple testing procedures with respect to minimizing the Bayes classification risk.

Optimality with respect to Bayes risk

γ_0 - loss for type I error, γ_A - loss for type II error

$\mu_i \sim (1-\epsilon)\delta_0 + \epsilon N(0,\tau^2)$

Bayes oracle $\rightarrow$ Bayes classifier:

$$\text{Reject } H_{0i} \quad \text{when} \quad \frac{\phi(X_i;\sigma^2+\tau^2)}{\phi(X_i;\sigma^2)} > \frac{1-\epsilon}{\epsilon}\frac{\gamma_0}{\gamma_A}$$

[B., Ghosh, Tokdar, *IMS Collections* 2008] and [B., Ghosh, Ochman, Tokdar *QREI*, 2007]: empirical comparison of BH with several Bayesian multiple testing procedures with respect to minimizing the Bayes classification risk.

M.B, A.Chakrabarti, F.Frommlet, J.K.Ghosh, Ann.Statist. 2011: proof of the asymptotic Bayes optimality of BH

Facts to remember

In high dimensional problems unbiased estimators can often be improved by biased estimators with reduced variance.

Facts to remember

In high dimensional problems unbiased estimators can often be improved by biased estimators with reduced variance.

When $p > 2$ then the maximum likelihood estimator of the vector of means for the multivariate normal distribution with independent covariates is not admissible. It can be improved by James-Stein estimator.

Facts to remember

In high dimensional problems unbiased estimators can often be improved by biased estimators with reduced variance.

When $p > 2$ then the maximum likelihood estimator of the vector of means for the multivariate normal distribution with independent covariates is not admissible. It can be improved by James-Stein estimator.

In case when the signal is sparse this can be further improved by thresholding rules.

Facts to remember

In high dimensional problems unbiased estimators can often be improved by biased estimators with reduced variance.

When $p > 2$ then the maximum likelihood estimator of the vector of means for the multivariate normal distribution with independent covariates is not admissible. It can be improved by James-Stein estimator.

In case when the signal is sparse this can be further improved by thresholding rules.

Hard thresholded estimator of μ using BH multiple testing rule adapts to the unknown sparsity and is asymptotically optimal in the sense discussed in (Abramovich, Benjamini, Donoho and Johnstone, Ann.Statist. 2006) or (B., Chakrabarti, Frommlet, Ghosh, Ann.Statist. 2011