James-Stein estimation

Statistical Learning
CLAMSES - University of Milano-Bicocca

Aldo Solari

References

- Samworth (2012). Stein's paradox. eureka, 62:38–41
- Candés (2022) Lecture notes (Stats 300C - Theory of Statistics)

- A very surprising result arises in a remarkably simple estimation problem.
- Let $X_1, \ldots, X_p$ be independent random variables, with $X_i \sim N(\mu_i, 1)$ for $i = 1, \ldots, p$. Writing $X = (X_1, \ldots, X_p)^t$, suppose we want to find a good estimator $\hat{\mu} = \hat{\mu}(X)$ of $\mu = (\mu_1, \ldots, \mu_p)^t$
- Squared error loss function:

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

  where $\|\cdot\|$ denotes the Euclidean norm
- Risk function: $R(\hat{\mu}, \mu) = \mathbb{E}[L(\hat{\mu}, \mu)]$

Inadmissible estimators

- If $\hat{\mu}$ and $\tilde{\mu}$ are both estimators of μ, we say that $\hat{\mu}$ strictly dominates $\tilde{\mu}$ if $R(\hat{\mu}, \mu) \leq R(\tilde{\mu}, \mu)$ for all μ, with strict inequality for some value of μ. In this case we say that $\tilde{\mu}$ is *inadmissible*.
- If $\hat{\mu}$ is not strictly dominated by any estimator of μ, it is said to be admissible. Note that admissible estimators are not necessarily sensible: for $p = 1$, the estimator $\hat{\mu} = 37$ (which ignores the data!) is *admissible*.
- On the other hand decision theory dictates that inadmissible estimators can be discarded
- $\hat{\mu} = X$ is a very obvious estimator of μ: it is the maximum likelihood estimator and the uniform minimum variance unbiased estimator with

$$R(\hat{\mu}, \mu) = p \quad \forall\ \mu \in \mathbb{R}^p$$

since $\|X - \mu\|^2 \sim \chi^2_p$

James-Stein estimator

- It has been proved that $\hat{\mu} = X$ is admissible for $p = 1, 2$
- James and Stein (1961) showed that the estimator

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

strictly dominates $\hat{\mu} = X$ for $p \geq 3$:

$$R(\hat{\mu}_{JS}, \mu) = p - (p-2)^2 \mathbb{E}\left(\frac{1}{\|X\|^2}\right) < p \quad \forall\ \mu \in \mathbb{R}^p$$

$\|X\|^2 = \sum_{i=1}^p X_i^2$ follows a noncentral χ^2 distribution with p degrees of freedom and noncentrality parameter $\|\mu\|^2$. Using a result about noncentral χ^2 variables, we can write

$$\|X\|^2 \sim \chi^2_{p+2K}$$

where $K \sim \text{Poisson}(\|\mu\|^2/2)$.

$$\begin{aligned}
\mathbb{E}\left(\frac{1}{\|X\|^2}\right) &= \mathbb{E}\left(\frac{1}{\chi^2_{p+2K}}\right) = \mathbb{E}\left\{\mathbb{E}\left(\frac{1}{\chi^2_{p+2K}}\right)|K\right\} \\
&= \mathbb{E}\left\{\frac{1}{(p-2)+2K}\right\} \geq \frac{1}{(p-2)+\|\mu\|^2}
\end{aligned}$$

with equality if $\mu = 0$, where we used $\mathbb{E}(1/\chi^2_p) = 1/(p-2)$ for $p > 2$ and Jensen's inequality. Then

$$R(\hat{\mu}_{\mathcal{JS}}, \mu) \leq p - \frac{p-2}{1+\|\mu\|^2/(p-2)}$$

Oracle linear estimator

- A linear estimator of the form

$$\tilde{\mu} = bX = (bX_1, \ldots, bX_p)^t$$

  with $0 \leq b \leq 1$ shrinks X towards the origin
- The risk of a linear estimator is

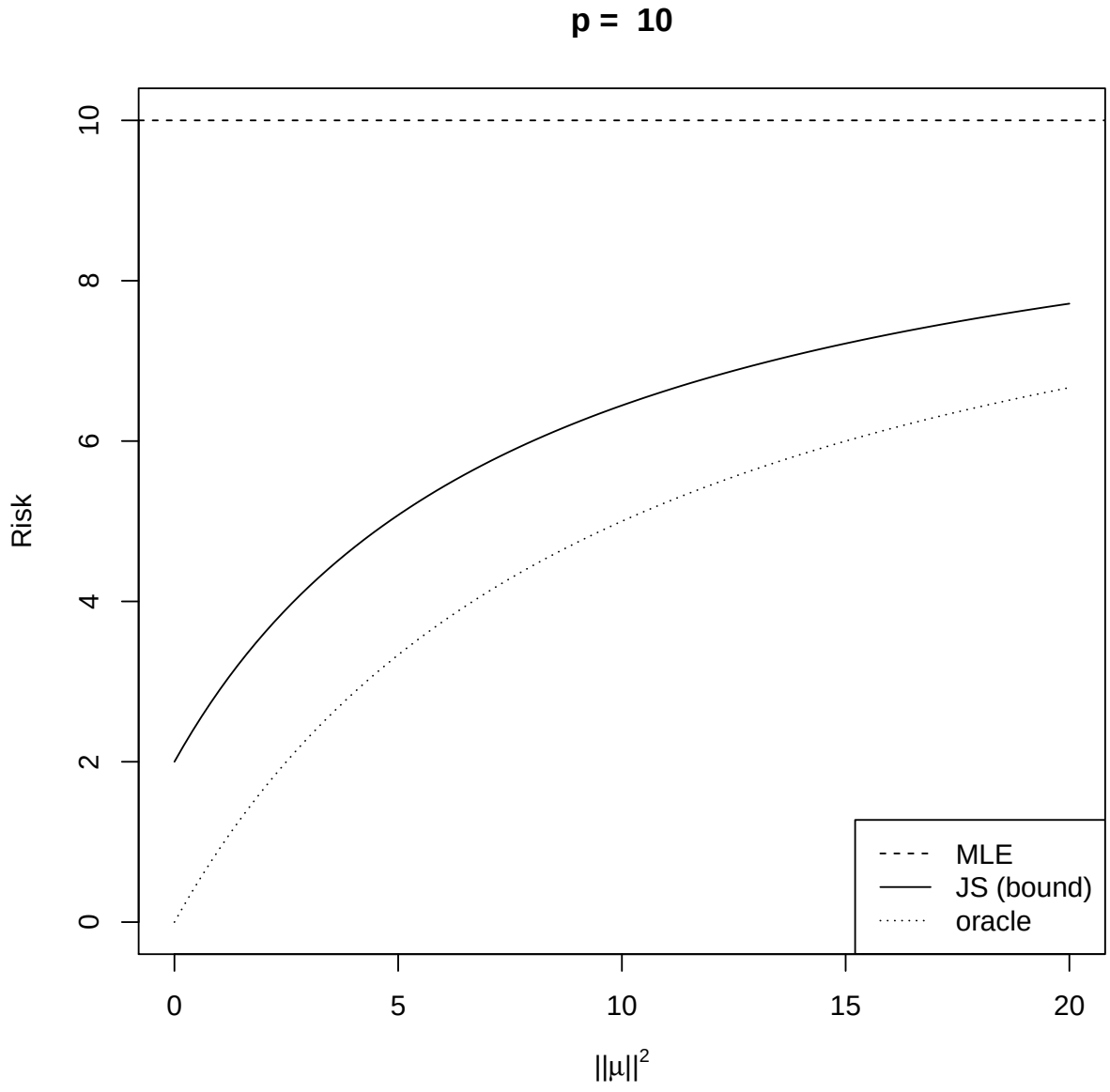
$$R(\tilde{\mu}, \mu) = (1-b)^2 \|\mu\|^2 + b^2 p$$

  minimized by

$$b^* = \frac{\|\mu\|^2}{p + \|\mu\|^2}$$

- The risk of the oracle linear estimator $\tilde{\mu}^* = b^* X$ is

$$R(\tilde{\mu}^*, \mu) = p\|\mu\|^2 / (p + \|\mu\|^2)$$


p = 10
Risk
0
2
4
6
8
10
0
5
10
15
20
||μ||2
MLE
JS (bound)
oracle

- Geometrically, the James-Stein estimator shrinks each component of X towards the origin, and the biggest improvement comes when μ is close to zero
- For $\mu = 0$ we have $R(\hat{\mu}_{JS}, 0) = 2$ for all $p \geq 2$
- As $\|\mu\|^2 \to \infty$, $R(\hat{\mu}_{JS}, \mu) \to p$

Stein's heuristic argument (1956)

- Stein argued that a good estimate should obey $\hat{\mu}_i \approx \mu_i$ for every i. Thus we should also have $\hat{\mu}_i^2 \approx \mu_i^2$, which further implies $\sum_i \hat{\mu}_i^2 \approx \sum_i \mu_i^2$
- Consider the estimator $\hat{\mu} = X$. For this estimator we have

$$\mathbb{E}\|X\|^2 = \mathbb{E}\sum_i X_i^2 = \mathbb{E}\sum_i (\mu_i + Z_i)^2 = \|\mu\|^2 + p$$

  where $Z_i \sim N(0, 1)$
- This suggests that for large p, $\|X\|^2$ is likely to be considerably larger than $\|\mu\|^2$, and hence we may be able to obtain a better estimator by shrinking the estimator $\hat{\mu} = X$ towards o.

Positive James-Stein estimator

- If the shrinkage in $\hat{\mu}_{JS}$ is too large, it is possible that the estimator switches to the other sign when $\|X\|^2 < p - 2$
- By precluding the possibility of a sign reversal, the positive JS estimator

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

  where $(a)_+ = \max(a, 0)$ denotes the positive part
- $\hat{\mu}_{JS}^{+}$ further improves upon the $\hat{\mu}_{JS}$ estimate, i.e., $R(\hat{\mu}_{JS}^{+}, \mu) < R(\hat{\mu}_{JS}, \mu)$ for all μ
- However, this estimator is not admissible either.

Shrinking toward an arbitrary point

- In terms of choosing a point to shrink towards, though, there is nothing special about the origin, and we could equally well shrink towards any pre-chosen $m \in \mathbb{R}^p$ using the estimator

$$\hat{\mu}^m_{JS} = m + \left(1 - \frac{p-2}{\|X-m\|^2}\right)(X-m)$$

- In this case, we have $R(\hat{\mu}^m_{JS}, \mu - m) = R(\hat{\mu}_{JS}, \mu)$, so $\hat{\mu}^m_{JS}$ still strictly dominates $\hat{\mu} = X$

Correlated data

- Assume that $X \sim N_p(\mu, \Sigma)$ where Σ is a known covariance matrix
- A a generalization of James-Stein estimator

$$\hat{\mu}_{JS}^{\Sigma} = \left(1 - \frac{c(\tilde{p} - 2)}{X^t \Sigma^{-1} X}\right) X$$

  with $0 < c < 2$ and $\tilde{p} = \text{tr}(\Sigma)/\lambda_{\text{max}}(\Sigma)$ is the effective dimension of the problem, where $\lambda_{\text{max}}(\Sigma)$ is the maximum eigenvalue of Σ
- If $\tilde{p} > 2$, then the generalization of the JS estimator $\hat{\mu}_{JS}^{\Sigma}$ dominates the MLE $\hat{\mu} = X$

Linear model

- We can apply the previous result to the case of linear regression $y \sim N_n(X\beta, \sigma^2 I_n)$, where the MLE is the OLS estimator $\hat{\beta} = (X^t X)^{-1} X^t y \sim N_p(\beta, \sigma^2 (X^t X)^{-1})$, so with $\mu = X\beta$ and $\hat{\mu} = X\hat{\beta}$ we have $R(\hat{\mu}, \mu) = \sigma^2 p$
- James-Stein estimator becomes

$$\hat{\beta}_{JS} = \left(1 - \frac{(p-2)\sigma^2}{\hat{\beta}^t X^t X \hat{\beta}}\right) \hat{\beta}$$

- Letting $\hat{\mu}_{JS} = X\hat{\beta}_{JS}$ and $\mu = X\beta$, the James–Stein Theorem guarantees that

$$R(\hat{\mu}_{JS}, \mu) \leq \sigma^2 p$$

no matter what β is, as long as $p \geq 3$

- It is natural to ask how crucial the normality and squared error loss assumptions are to the Stein phenomenon
- The normality assumption is not critical at all;
- The original result can also be generalised to different loss functions, but there is an important caveat here: the Stein phenomenon only holds when we are interested in simultaneous estimation of all components of μ. If our loss function were $L(\hat{\mu}, \mu) = (\hat{\mu}_1 - \mu_1)^2$ then we could not improve on $\hat{\mu} = X$

$p = 5$, $\mu = (\sqrt{p/2}, \sqrt{p/2}, 0, 0, 0)^t$, $\|\mu\|^2 = p$

10^4 repetitions

	Risk	Risk1	Risk2	Risk3	Risk4	Risk5
MLE	5.00	1.01	1.01	1.00	0.98	0.99
JS	3.65	1.08	1.07	0.50	0.49	0.50

$R(\hat{\mu}_{JS}, \mu) \leq p - (p-2)/(1 + p/(p-2)) = 3.875$

An Empirical Bayes interpretation

Bayesian setup

- Consider the Bayesian setup

$$\mu_i \sim N(0, \tau^2) \quad X|\mu \sim N(\mu, I_p) \tag{1}$$

- Given the data X, the posterior of μ is

$$\mu|X \sim N(\lambda X, \lambda I_p)$$

  where $\lambda = \tau^2/(1+\tau^2)$

- The Bayes estimator is simply the mean of the posterior

$$\hat{\mu}_B = \lambda X = \left(1 - \frac{1}{1+\tau^2}\right) X$$

- Assuming (1), the Bayes risk is $R(\hat{\mu}_B, \mu) = \lambda p$

Connection to James-Stein

- We cannot directly compute the shrinkage factor $\lambda = \tau^2/(1+\tau^2)$, but perhaps we can estimate it using the data
- Since $X_i = \mu_i + Z_i \sim N(0, 1+\tau^2)$, where $Z_i \sim N(0,1)$. This implies $\|X\|^2 \sim (1+\tau^2)\chi^2_p$
- Combining this result with $\mathbb{E}[(p-2)/\chi^2_p] = 1$, we arrive at an unbiased estimate for λ

$$\hat{\lambda} = \left(1 - \frac{(p-2)}{\|X\|^2}\right)$$

- Assuming (1), the Bayes risk is $R(\hat{\mu}_{JS}, \mu) = \left(1 + \frac{2}{p\tau^2}\right) R(\hat{\mu}_B, \mu)$

$p = 5$, $\tau^2 = 2$, $\mu_i \sim N(0, \tau^2)$

10^4 repetitions

	Bayes Risk	B.Risk1	B.Risk2	B.Risk3	B.Risk4	B.Risk5
MLE	5.01	1.01	1.01	1.00	0.99	1.00
BAYES	3.34	0.67	0.68	0.67	0.67	0.66
JS	4.02	0.81	0.82	0.80	0.80	0.79

$R(\hat{\mu}, \mu) = 5$, $R(\hat{\mu}_B, \mu) = 3.33$, $R(\hat{\mu}_{JS}, \mu) = 4$,

Shrinking Toward the Group Mean

- In practice, instead of arbitrarily picking some point, it might instead make sense to chose $m = \bar{X}$ as so as to adapt to the true center of μ_i
- Consider the Bayesian setup

$$\mu_i \sim N(m, \tau^2) \quad X|\mu \sim N(\mu, I_p) \tag{2}$$

  with m and τ^2 unknown
- The marginal distribution of our data is

$$X_i \overset{i.i.d.}{\sim} N(m, 1 + \tau^2)$$

  and the posterior mean is

$$\mu|X \sim N(m + \lambda(X - m), \lambda I_p)$$

- $\hat{\mu}_B = m + \lambda(X - m)$ but m is unknown. Taking the empirical Bayes approach, we can use the unbiased estimator $\bar{X}$ in its place
- Similarly, we can use the sample variance $S = \sum_i (X_i - \bar{X})^2 \sim (1 + \tau^2)\chi^2_{p-1}$ to estimate λ. Now we have $\mathbb{E}[(p-3)/\chi^2_{p-1}] = 1$
- This gives us the estimator

$$\hat{\mu}^{\bar{X}}_{JS} = \bar{X} + \left(1 - \frac{p-3}{S}\right)(X - \bar{X})$$

If $p > 3$, this estimator dominates the MLE everywhere

A baseball data example

Player	MLE	TRUTH
1	0.34	0.30
2	0.33	0.35
3	0.32	0.22
4	0.31	0.28
5	0.29	0.26
6	0.29	0.27
7	0.28	0.30
8	0.26	0.27
9	0.24	0.23
10	0.23	0.26
11	0.23	0.26
12	0.22	0.21
13	0.22	0.26
14	0.22	0.27
15	0.21	0.32
16	0.21	0.23
17	0.20	0.28
18	0.14	0.20

The column labelled MLE is the batting average for 18 players in the 1970 season, using the first 90 at bats.

The column labelled TRUTH is the batting average for the remainder of the 1970 season.

- Each player Batting average =(# hits / # at bats) value is a binomial proportion

$$Y_i \sim \text{Binomial}(n, \pi_i)/n$$

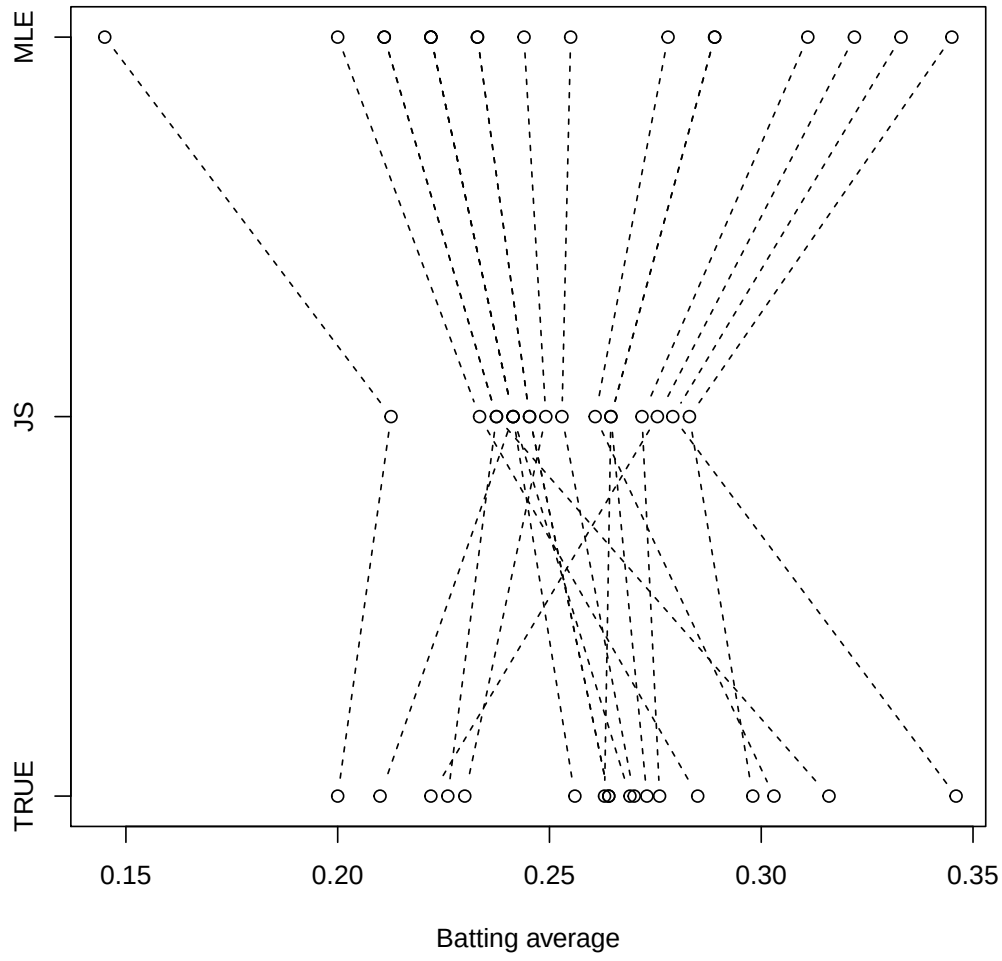
  where π_i is the true average and $n = 90$
- Since batting averages are binomial, we can use the normal approximation

$$Y_i \approx N\left(\pi_i, \frac{\pi_i(1-\pi_i)}{n}\right)$$

  but the variance depends on the mean
- One solution is to make a variance stabilizing transformation

$$X_i = 2\sqrt{n+0.5}\,\arcsin\left(\sqrt{\frac{nY_i + 3/8}{n+3/4}}\right) \approx N(\mu_i, 1)$$

  where $\mu_i = 2\sqrt{n+0.5}\,\arcsin(\sqrt{\frac{n\pi_i+3/8}{90+3/4}})$
- Inverted back $y_i^{JS} = \frac{1}{n}\left[(n+0.75)(\sin(\frac{\hat{\mu}_i^{JS}}{2\sqrt{n+0.5}}))^2 - 0.375\right]$



$$\sum_i (y_i - y_i^{\text{TRUE}})^2 = 0.0425 \quad \sum_i (y_i^{JS} - y_i^{\text{TRUE}})^2 = 0.0205$$