

Empirical Bayes and the James-Stein estimator

Bradley Efron (2010) Large-Scale Inference: Empirical Bayes
Methods for Estimation, Testing, and Prediction

Jesse C. Chen
jessekelighine.com

2023-12-12

- 1 Motivation: Bayesian inference and empirical Bayes
 - Bayesian Inference
 - Bayesian Estimator
 - Empirical Bayes
 - James-Stein Estimator
- 2 James-Stein Theorem
- 3 MLE or JS estimator?
 - Example: 1970s Major league players
 - Remarks on JS estimator
- 4 Conclusion

Bayesian Inference

The concept of Bayesian inference is different from that of the Frequentists'. The main difference is whether there is a **prior belief** on the parameters of interest.

Consider a parameter vector $\mu \sim g(\cdot)$ where g is some density (μ is N dimensional), and $g(\mu)$ in turn give rise to an observable data vector z where

$$z | \mu \sim f_{\mu}(z).$$

In Bayesian statistics, $g(\mu)$ is called a **prior** distribution, which represents our prior knowledge of the parameters μ . A statistician's job is to inference μ given the observations z .

In Frequentists' world, no prior knowledge on the parameters μ are assumed. In Bayesian statistics, we obtain the following from Bayes' formula for the inference:

$$g(\mu | z) = g(\mu) \frac{f_{\mu}(z)}{f(z)}$$

where $g(\mu | z)$ is the **posterior** distribution, i.e., the **updated** distribution after observing z . If we have a reasonable prior belief on μ , it is often the case that Bayesian statistics performs better than Frequentists' arguments.

The James-Stein estimator is a Frequentists' method that harvests the power of Bayesian statistics through empirical estimation.

Motivation from Bayesian estimators

Consider the special case where

$$\boldsymbol{\mu} \sim \mathcal{N}_N(0, AI) \quad \text{and} \quad \mathbf{z} | \boldsymbol{\mu} \sim \mathcal{N}(\boldsymbol{\mu}, I).$$

We can calculate that the posterior distribution of $\boldsymbol{\mu}$ as

$$\boldsymbol{\mu} | \mathbf{z} \sim \mathcal{N}_N(B\mathbf{z}, BI) \quad \text{where} \quad B = \frac{A}{A+1}.$$

- The “obvious” MLE estimator for μ is simply

$$\hat{\mu}^{(\text{MLE})} = (z_1, \dots, z_N)' = \mathbf{z},$$

which, proven by Fisher, is a MVUE in the one dimensional case.

- On the other hand, the Bayesian estimator is based on the posterior distribution, thus, we have

$$\hat{\mu}^{(\text{Bayes})} = B\mathbf{z} = \left(\frac{A}{A+1} \right) \mathbf{z} = \left(1 - \frac{1}{A+1} \right) \mathbf{z}$$

Note that while $\hat{\mu}^{(\text{MLE})}$ is an unbiased estimator, $\hat{\mu}^{(\text{Bayes})}$ is not, since it is **shrunk** towards $\mathbf{0}$.

Then what is the advantage of $\hat{\mu}^{(\text{Bayes})}$?

Consider the mean-square error of the two estimators:

- For MLE, we have

$$\mathbb{E} \left[\|\hat{\boldsymbol{\mu}}^{(\text{MLE})} - \boldsymbol{\mu}\|^2 \right] := \mathbb{E} \left[\sum_{i=1}^N (\hat{\mu}_i^{(\text{MLE})} - \mu_i)^2 \right] = N$$

- For Bayesian estimator, we have

$$\mathbb{E} \left[\|\hat{\boldsymbol{\mu}}^{(\text{Bayes})} - \boldsymbol{\mu}\|^2 \right] = BN = \left(\frac{A}{1+A} \right) N < N$$

That is, the Bayesian estimator offers some saving in mean-square error. If $A = 1$, then the mean-square error of the Bayesian estimator is only half of that of the MLE.

Empirical Bayes

- Unfortunately, without knowing A , it is not possible to construct the Bayesian estimator.
- However, we can use the *estimator* of A to construct the Bayesian estimator. (aka “plug-in principle”)
- Consider the marginal distribution of z , we have

$$z \sim \mathcal{N}_N(0, (A + 1)I).$$

From the marginal distribution we can construct an estimator

$$\hat{B} := 1 - (N - 2)/S$$

where $S := \|z\|^2$ such that $\mathbb{E}[\hat{B}] = B = A/(1 + A)$.

James-Stein Estimator

Therefore, the James-Stein estimator is defined as

$$\hat{\boldsymbol{\mu}}^{(\text{JS})} := \hat{B}\mathbf{z} = \left(1 - \frac{N-2}{S}\right)\mathbf{z}$$

If we allow the mean of the prior distribution of $\boldsymbol{\mu}$ be $\mathbf{M} = (M, \dots, M)'$ instead of 0, we have a similar definition of James-Stein estimator:

$$\hat{\boldsymbol{\mu}}^{(\text{JS})} := \hat{\mathbf{M}} + \hat{B}(\mathbf{z} - \hat{\mathbf{M}}) \quad \text{where} \quad \begin{cases} \hat{\mathbf{M}} = (\bar{z}, \dots, \bar{z})' \\ \hat{B} = 1 - \frac{N-3}{\|\mathbf{z} - \bar{z}\|^2} \end{cases}$$

It remains to show that the James-Stein estimator is still better than MLE (the plug-in principle introduces more variance).

Surprisingly, the JS estimator performs nearly as well as the Bayesian estimator. If we consider mean-square error the ratio between the Bayesian and the JS estimator, we have

$$\frac{\mathbb{E} \left[\|\hat{\boldsymbol{\mu}}^{(\text{JS})} - \boldsymbol{\mu}\|^2 \right]}{\mathbb{E} \left[\|\hat{\boldsymbol{\mu}}^{(\text{Bayes})} - \boldsymbol{\mu}\|^2 \right]} = 1 + \frac{2}{NA}$$

where the ratio tends to one on the order of $O(1/N)$.

However, the biggest shock JS estimator brought was not that the estimator performs better, since biasing for the some choices of $\boldsymbol{\mu}$ is implied in the derivation. The true shock came from the fact that MLE is **dominated** in higher dimensions by the JS estimator.

Domination over MLE

James-Stein Theorem, aka Stein's Paradox:

Theorem (James-Stein (1961))

For $N \geq 4$, the James-Stein estimator **everywhere** dominates the MLE $\hat{\mu}^{(MLE)}$ in terms of expected total square, i.e.,

$$\mathbb{E}_{\mu} \left[\|\hat{\mu}^{(JS)} - \mu\|^2 \right] < \mathbb{E}_{\mu} \left[\|\hat{\mu}^{(MLE)} - \mu\|^2 \right]$$

for every choice of μ . MLE is said to be inadmissible.

That is, JS estimator not only performs well when μ are near the estimated mean, it performs well **no matter one's prior belief** (no matter the realisation of μ).

Sketch of Proof

1 Show

$$\mathbb{E}_{\boldsymbol{\mu}} [\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2] = \mathbb{E}_{\boldsymbol{\mu}} [\|\mathbf{z} - \hat{\boldsymbol{\mu}}\|^2] - N + \sum_{i=1}^N \text{Cov}_{\boldsymbol{\mu}}(\hat{\mu}_i, z_i).$$

2 Show the following given z_i is normal:

$$\text{Cov}_{\boldsymbol{\mu}}(\hat{\mu}_i, z_i) = \mathbb{E}_{\boldsymbol{\mu}}[\partial \hat{\mu}_i / \partial z_i].$$

3 Obtain from (1) and (2) that

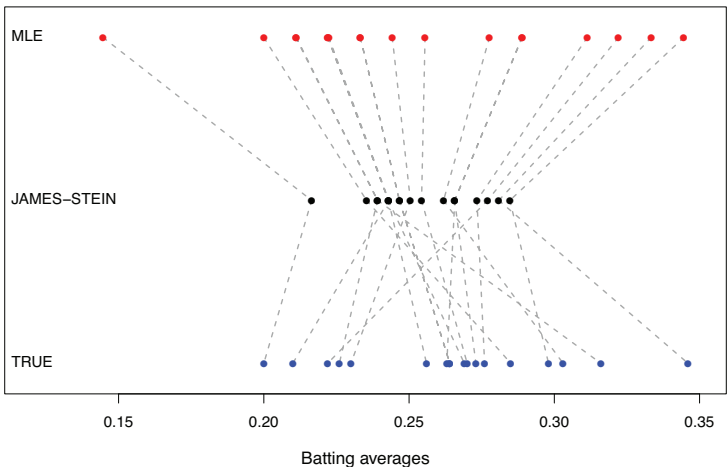
$$\mathbb{E}_{\boldsymbol{\mu}} [\|\hat{\boldsymbol{\mu}}^{(\text{JS})} - \boldsymbol{\mu}\|^2] = N - \mathbb{E}_{\boldsymbol{\mu}} [(N-3)^2/S].$$

4 Since MLE's mean-square error is N , given that $N \geq 3$, it is guaranteed JS estimator dominates MLE. \square

MLE or JS estimator? Example: Major league

Name	hits/AB	$\hat{\mu}_i^{(\text{MLE})}$	μ_i	$\hat{\mu}_i^{(\text{JS})}$
Clemente	18/45	.400	.346	.294
F Robinson	17/45	.378	.298	.289
F Howard	16/45	.356	.276	.285
Johnstone	15/45	.333	.222	.280
Berry	14/45	.311	.273	.275
Spencer	14/45	.311	.270	.275
Kessinger	13/45	.289	.263	.270
L Alvarado	12/45	.267	.210	.266
Santo	11/45	.244	.269	.261
Swoboda	11/45	.244	.230	.261
Unser	10/45	.222	.264	.256
Williams	10/45	.222	.256	.256
Scott	10/45	.222	.303	.256
Petrocelli	10/45	.222	.264	.256
E Rodriguez	10/45	.222	.226	.256
Campaneris	9/45	.200	.286	.252
Munson	8/45	.178	.316	.247
Alvis	7/45	.156	.200	.242
Grand Average		.265	.265	.265

- Obviously, the MLE estimator is the early game average.
- JS estimator improves the estimation by a lot. The mean-square error ratio of JS estimation and MLE is about 0.28, a significant reduction of total error.
- But what is the problem of JS estimator?



JS estimator over shrink the estimates, and it is a poor estimator for extreme values.


	μ_i	$\text{MSE}_i^{(\text{MLE})}$	$\text{MSE}_i^{(\text{JS})}$
1	-.81	.95	.61
2	-.39	1.04	.62
3	-.39	1.03	.62
4	-.08	.99	.58
5	.69	1.06	.67
6	.71	.98	.63
7	1.28	.95	.71
8	1.32	1.04	.77
9	1.89	1.00	.88
10	4.00	1.08	2.04!!
Total Sqerr		10.12	8.13

- Left chart is created through simulations with $z \mid \mu \sim \mathcal{N}_{10}(\mu, I)$.
- Notice that although the total error improves under JS estimator, but individual estimation suffers from the shrinkage.
- Traditional statistical methods are conservative in protecting individual effects from the tyranny of the majority, while JS estimator does not.

- However, sacrificing individual performance does imply a better performance over all. In cases of large scale inferences, a better overall performance is favoured.
- If we insist on preserving some correctness of individual inferences, compromising methods are available. One such compromising method is to choose MLE (adjusted with overall s.d.) whenever the JS estimator and MLE disagree too much. This method is called *Limited translation estimates*, developed by Efron & Morris in the 1970s.
- JS estimator are also applied to adjust regression results and construct confidence intervals. More methods, throughout the years, are introduced to “generalized” the result of JS estimator.
- In fact, Stein proved that MLE is inadmissible before proposing the JS estimator. He also proved that JS estimator itself is also inadmissible, but no explicit form has been found.

Conclusion

James and Stein not only introduced an estimator that dominates MLE in high dimension inference, the JS estimator also inspired other modern statistical methods through the following two ways:

- 1 Learning from the experience of others:** This is a borrowing from Bayesian statistics into Frequentists' arsenal: the idea that estimation can be improved using "indirect evidence." Later employed into *large scale hypothesis testing* and *false discovery rate control*.
- 2 Shrinkage:** The *shrinkage principle* and *bias-variance trade-off* are perhaps the two of the most influential ideas in modern day statistics and machine learning. Lasso, Ridge, and other regulation methods in regression are now common place. 

References

- 1 Efron, B. (2010). *Empirical Bayes and the James—Stein Estimator*. Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction (Institute of Mathematical Statistics Monographs, pp. 1-14). Cambridge: Cambridge University Press. doi:
<https://doi.org/10.1017/CB09780511761362.002>
- 2 Efron, B., & Morris, C. (1977). *Stein's Paradox in Statistics*. Scientific American, 236(5), 119-127. Retrieved from
<http://www.jstor.org/stable/24954030>