**Stat210A: Theoretical Statistics** **Lecture Date: October 28, 2014**

# The James-Stein Phenomenon

*Lecturer: Michael I. Jordan* *Scribe: Simon Walter*

## 1 Background

We wish to estimate $\theta = (\theta_1, \theta_2, \ldots, \theta_p)$ and we observe $X = (X_1, X_2, \ldots, X_p)$ where

$$X_i = \theta_i + \epsilon_i \quad \epsilon_i \overset{iid}{\sim} N(0,1)$$

This is often called a denoising problem.

Since we are simultaneously interested in more than one parameter, we need a composite loss; that is a loss function that combines the error for each component of our parameter vector into a single real number. We use the squared error composite loss. For a decision rule $\delta = (\delta_1(X), \delta_2(X), \ldots, \delta_p(X))$ we have:

$$L(\theta, \delta) = \sum_{i=1}^{p} (\theta_i - \delta_i(X))^2 = ||\theta - \delta(X)||_2^2$$

This leads to risk:

$$R(\theta, \delta) = \mathbb{E}_\theta ||\delta(X) - \theta||_2^2$$

Intuitively because of the independence of the $X_i$ we would expect a good $\delta_i$ is a function of $X_i$ alone; the natural choice is $\delta_i = X_i$. This estimator is the MLE, UMVUE, (and we may see later minimax).

The risk of the MLE is

$$R(\hat{\theta}_{\text{MLE}}, \theta) = \mathbb{E}||X - \theta||^2 = p\text{Var}(X_1) = p$$

So $R(\hat{\theta}_{\text{MLE}}, \theta)$ grows linearly in $p$ and we might suspect that there are improvements over the MLE in high dimensions. The standard strategy to seek such an improvement is regularization. For example if we are confident that $\theta$ is near the zero vector; then shrinkage towards zero is likely beneficial. This is exactly the strategy of the James-Stein (**J-S**) estimator defined[1] for $p \geq 3$:

$$\theta_{JS}(X) = \left(1 - \frac{p-2}{||X||_2^2}\right) X$$

Notice that, unlike the MLE, each $\delta_i$ is a function of the entire data vector $X$. We should not be surprised that this estimator has superior risk when $\theta$ is near zero, but it would be surprising if it had superior risk were $\theta$ a long way from zero. Figure 1 below shows the risk of the J-S estimator and the MLE for several choices of $p$. Note that for each $p$ shown in the Figure this risk of the J-S estimator at the origin is 2, and, strikingly, we will show that this holds for all $p$. The figure also suggests it can be shown that

$$\lim_{\theta \to \pm\infty} R(\hat{\theta}_{JS}, \theta) = p$$

[1] We could define the J-S estimator for $p = 2$ but it would reduce to the MLE.

These facts lend credibility to the view that the MLE is dominated by the J-S estimator. However, we should note that the J-S estimator only dominates the MLE if we are interested in the $\theta_i$ simultaneously; if we are only interested in the $\theta_i$ sequentially then the MLE is superior.

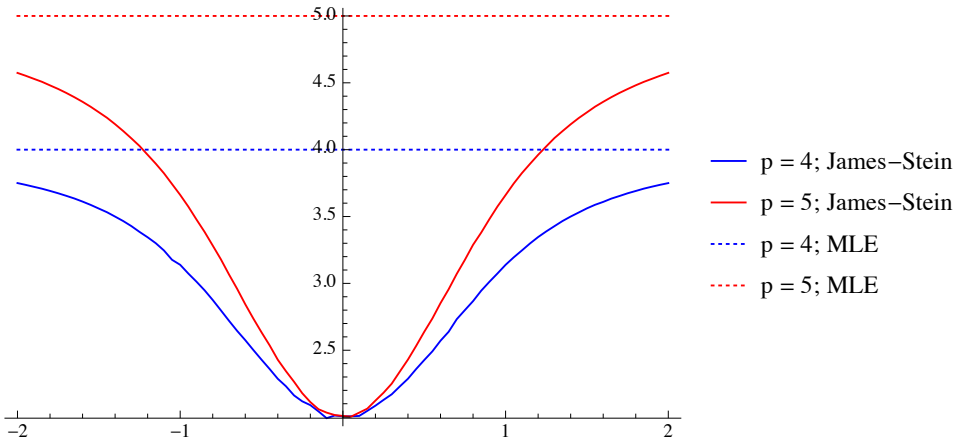


Figure 1: The risk of the James-Stein estimator.

In the remainder of this lecture we will establish results and develop intuition for the domination of the James-Stein estimator. Intuition is particularly valuable because an analogue of the James-Stein Phenomenon arises in some situations but not others.

## 2 Empirical Bayes view of the J-S estimator

In this section we derive the James-Stein estimator using an empirical Bayes point of view. This treatment is developed in detail by Efron and Morris (1972).

The J-S estimator shrinks $X$ towards zero but it is not hard to see that 0 is not special and dominance of the MLE can be acheived for an analogous estimator that shrinks towards any point. (To make this more precise notice that if we make the transformation $\theta \mapsto \theta + a$ then shrinkage towards 0 in the untransformed problem amounts to shrinkage towards $a$ in the transformed problem.) This feels a little Bayesian and it turns out that an empirical Bayesian interpretation of the James-Stein phenomenon is available.[2]

We choose the prior

$$\Theta_i \overset{iid}{\sim} N(0, \tau^2)$$

A little algebra reveals that the posterior distribution is

$$\Theta_i | X \overset{ind}{\sim} N\left(\frac{X_i}{1 + 1/\tau^2}, \frac{1}{1 + 1/\tau^2}\right)$$

The algebra is available in Keener §11.1 p. 206. Alternatively we can use the formulas for conditional distributions form normal distribution theory. These formulas are available on Wikipedia, for example.[3]

[2]Empirical Bayes methods were detailed in Tamara's recitation. Briefly the idea is to impose a prior on $\theta$ with particular (hyper)parameters that are then estimated from the data by, for example, maximum likelihood estimation or matching moments of the posterior distribution of with the empirical moments. This approach is not philosophically Bayesian, although it is inspired and informed by Bayesian methods

[3]`http://en.wikipedia.org/wiki/Multivariate_normal_distribution#Conditional_distributions`

It then follows (Keener p. 207) that the Bayes decision rule is the posterior mean

$$\delta_i(X) = \frac{X_i}{1 + 1/\tau^2} = \left(1 - \frac{1}{1+\tau^2}\right) X_i$$

Notice that $\delta_i$ shrinks the MLE, $X_i$ towards zero (and we can obtain shrinkage towards any point by modifying the mean of the prior distribution). However $\delta_i$ still only depends on $X_i$, so we do not yet have a coupling estimator.

To obtain a concrete estimator we must set $\tau^2$. Were we philosophically Bayesian, we could do so subjectively by determining our prior knowledge of the problem; or we could be objective and choose a prior that is non-informative (for example the Jeffreys prior). We will do neither, instead we will adopt an empirical Bayes strategy and estimate $\tau$ from the data. One justification for this is that we were never really Bayesian in the first place but were simply borrowing some Bayesian ideas. Another view is that we might have liked to impose a hyper-prior on $\tau$ and hence obtain a hierarchical Bayesian model but this is analytically (and numerically) difficult so we obtain an approximate empirical solution.

Since our estimator $\delta_i$ depends on $\tau$ only through $\frac{1}{1+\tau^2}$ this is the quantity we wish to estimate. We will adopt a slightly heuristic approach here and invert an unbiased estimator of $1 + \tau^2$. We notice

$$\begin{aligned}\mathbb{E}(X_i^2) &= \mathrm{Var}(\Theta_i + \epsilon_i) \\ &= \tau^2 + 1\end{aligned}$$

So we conclude $\frac{1}{p}\sum_{i=1}^{p} X_i^2$ is unbiased for $1 + \tau^2$ (in fact it is UMVU by the Lehmann-Scheffé Theorem). So an empirical Bayes estimator of $\theta_i$ is

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

This is indeed a coupling estimator (i.e. it uses the entire vector $X$ to for each $\theta_i$). The constant multiple of $1/||X||^2$ is slightly different to that ultimately chosen for the J-S estimator. In fact if we use an unbiased estimator of $\tau^2/(1+\tau^2)$ we obtain the J-S estimator exactly (see Lehmann and Casella (1998) Problem 4.7.1 on p. 298); proving this was left as an exercise in lecture.

There are, however, other ways of estimating $\tau$ if we use maximum likelihood estimation we obtain the Efron-Morris estimator

$$\hat{\theta}_{EM} = \bar{X} + \left(1 - \frac{p-3}{||X - \bar{X}||^2}\right)(X - \bar{X})$$

This seems more natural than the J-S estimator because it imposes shrinkage towards $\bar{X}$ which is less arbitrary than the shrinkage towards 0.

At this point in the lecture a question was asked whether there was a connection between the James-Stein phenomenon and a result that seems sometimes to be called Polya's theorem: a random walk will return to the origin (or any visited point) almost surely in one or two dimensions but will almost surely not return to any visited point in higher dimensions. This result is sometimes communicated with the aphorism, attributed to Shizuo Kakatuni, that a drunk person will find their way home eventually but a drunk bird may be lost forever. We were directed to Brown (1971), there is also a question and answer on `stats.stackexchange` that is a helpful summary of the result of Brown (1971): `http://stats.stackexchange.com/a/13647/24370`.

There are further variations on the J-S estimator, for example the positive part J-S estimator:

$$\delta_{JS}^{+} = \left(1 - \frac{p-2}{||X||^2}\right)^{+} X_i$$

where $t^+ = t\mathbb{I}(t > 0)$. It can be shown that this dominates the J-S estimator but it too is inadmissible We did not discuss how one might prove the positive part estimator is inadmissible in lecture but the standard proof of this fact shows that because of the lack of smoothness of the estimator when $\left(1 - \frac{p-2}{||X||^2}\right)$ is near zero it is not Bayes or the limit of Bayes and hence does not belong to the complete class of admissible estimator. There are however examples of explicit estimators that dominate the positive part estimator; see, for example, Shao and Strawderman (1994). By now we will not be surprised to learn that this dominating estimator is not admissible either. However, the modern view is that that although the positive part estimator is not admissible it is 'almost' admissible and only limited improvement is possible.

It may appear the J-S and Efron-Morris estimators are simply theoretical oddities but they have demonstrated their value in practice as well. Perhaps the most prominent application to actual data is Efron and Morris (1975). Moreover the insight that shrinkage is important is the precursor of many modern methods, like ridge regression, the lasso and the Dantzig selector.

# 3 Computing the risk of the J-S estimator

In this section we explicitly compute the risk of the J-S estimator. We require Stein's lemma for Gaussian random variables (Keener 11.1) this lemma was also proved (in greater generality) in Homework Problem 2.5.

**Lemma 1** (Stein's identity)**.** *Suppose* $X \sim N(\mu, \sigma^2)$ *and* $h : \mathbb{R} \to \mathbb{R}$ *is a differentiable (or absolutely continuous) function with*

$$\mathbb{E}|h'(X)| < \infty$$

*Then*

$$\mathbb{E}(X - \mu)h(X) = \sigma^2 \mathbb{E}h'(X)$$

This result follows immediately from an application of integration by parts to the expectation on the LHS of the identity.

There is also a matrix version of this identity (Keener Lemma 11.2). It was remarked that in class we often skip the matrix version of results but it is important that we review them in Keener in our own time. We will use $Dh(x)$ to denote the $p \times p$ matrix of partial derivatives:

$$[Dh(x)]_{ij} = \frac{\partial h_i(x)}{\partial x_j}$$

**Lemma 2** (Stein's matrix identity)**.** *Suppose* $X = (X_1, X_2, \dots X_p)$ *has components* $X_i \overset{iid}{\sim} N(\theta_i, 1)$*. If*

$$\mathbb{E}||Dh(X)|| < \infty$$

*Then*

$$\mathbb{E}(X - \theta)^\top h(X) = \mathbb{E}[\mathrm{tr}(Dh(X))]$$

This result is proved using Stein's identity and the tower property of expectation, see Keener p. 209. The importance of this lemma is that it converts inner products into derivatives and often derivatives are substantially easier to control. To make this concrete, suppose we have $h = \int h'$ where $h'$ is easier to control; Stein's identity permits us to focus on $h'$. Chaining, a method in empirical process theory, is a similar idea. It was also suggested that large deviation inequalities and Chernoff and Hoeffding bounds were conceptually similar to methods informed by Stein's identity.

Notice that the risk of an decision rule is a function of an unknown parameter and therefore is not known *a priori* hence it is amenable to estimation.

**Theorem 3.** *Suppose* $X_i \sim N(\theta_i, 1)$ *and* $\delta(X) = X - h(X)$ *Define* $\hat{R} = p + ||h(X)||^2 - 2\text{tr}(Dh(X))$

*Then*

$$R(\theta, \delta) = \mathbb{E}_\theta \hat{R}$$

*Proof.*

$$\begin{aligned}
R(\theta, \delta) &= \mathbb{E}_\theta \left[ \sum_{i=1}^{p} (X_i - \theta_i - h_i(X))^2 \right] \\
&= \mathbb{E}_\theta \left[ \sum_{i=1}^{p} (X_i - \theta_i)^2 + \sum_{i=1}^{p} h_i^2(X) - 2 \sum_{i=1}^{p} (X_i - \theta_i) h_i(X) \right] \\
&= \mathbb{E}_\theta \left[ \sum_{i=1}^{p} (X_i - \theta_i)^2 \right] + \mathbb{E}_\theta \left[ \sum_{i=1}^{p} h_i^2(X) \right] - 2\mathbb{E}_\theta \left[ \sum_{i=1}^{p} (X_i - \theta_i) h_i(X) \right] \\
&= p + \mathbb{E}_\theta ||h(X)||^2 - 2\mathbb{E}_\theta [\text{tr}(Dh(X))] \\
&= \mathbb{E}_\theta \hat{R}
\end{aligned}$$

The penultimate equality applies Lemma 2. □

Using this strategy to estimate the risk is called Stein's unbiased risk estimation (**SURE**) after Stein (1981). **SURE** is not a term used by Keener but we should know it.

Notice that if we set

$$h_i(X) = \frac{p-2}{||X||^2} X_i$$

the $\delta(X)$ in Theorem 3 becomes the J-S estimator. Then the product rule implies

$$\frac{\partial h_i(X)}{\partial X_i} = \frac{p-2}{||X||^2} - \frac{2(p-2)X_i^2}{||X||^4}$$

It then follows that

$$\text{tr}[Dh(X)] = \frac{(p-2)^2}{||X||^2} \quad \text{and} \quad ||h(X)||^2 = \frac{p-2}{||X||^2}$$

The detail of these computations are given in Keener p. 210 but were left as an exercise in lecture.

So we have

$$\begin{aligned}
\hat{R} &= p + \frac{(p-2)^2}{||X||^2} - 2\frac{(p-2)^2}{||X||^2} \\
&= p - \frac{(p-2)^2}{||X||^2}
\end{aligned}$$

And hence Theorem 3 implies

$$R(\theta, \hat{\theta}_{JS}) = \mathbb{E}_\theta(\hat{R}) = p - \mathbb{E}_\theta \left[ \frac{(p-2)^2}{||X||^2} \right] < p = R(\theta, \hat{\theta}_{MLE})$$

To determine the risk of the J-S estimator exactly for a particular $\theta$ we need $\mathbb{E}_\theta \left( (p-2)/||X||^2 \right)$. When $\theta = 0$, $1/||X||^2$ is an inverse $\chi_p^2$ distribution and it can be shown (Keener p. 211 refers to Keener p. 70 equation (4.10)) that

$$\mathbb{E}_0 \left[ \frac{1}{||X||^2} \right] = \frac{1}{p-2}$$

It then follows from our prior application of Theorem 3 that

$$R(0, \theta_{JS}) = p - \frac{(p-2)^2}{p-2} = 2$$

Note that this result does not hold when $p = 1$; a proof of the admissibility of the MLE in this case (which is surprisingly difficult) is given in Keener p. 215-6. This result was referred to in lecture but not explicitly covered.

# 4 Historical background

The James-Stein phenomenon was discovered in the 1950s. At that time the result caused discord, in part because it challenged the classical view that focus on unbiased estimation was (in many cases) good enough. It was also preceded by several proofs of the admissibility of the MLE for estimating a normal mean in one dimension. However, in the 1970s the empirical Bayes view of the J-S estimator was given by Efron and Morris (1972) and gradually the controversy abated. Nonetheless the paper we will review in the next section was published in 1990. This suggest the J-S phenomenon was a matter of ongoing concern at that time.

# 5 The J-S estimator and linear regression

In this section of the lecture we read through most of Stigler (1990). We will review the key points of the paper, but much of the content of the paper will not be covered.

The setup for Stigler is the same as that given in section 1 of these notes. Now suppose that we have $\theta_i = \alpha + \beta X_i$ for some constants $\alpha$ and $\beta$. In Figure 2 we plot the pairs $(X_i, \theta_i)$. Which line should we choose to estimate the vector $\theta$? There are three obvious choices $\theta_i = X_i$, $\theta_i = \mathbb{E}(\theta_i|X)$ and $\theta_i = \mathbb{E}(X_i|\theta)$. The first corresponds to the MLE, the second corresponds is the solution to the regression of $\theta$ on $X$ and it is a classical result that it is the best unbiased estimator of $\theta$ (the Gauss-Markov theorem). The third regresses $X$ on $\theta$ and would be optimal if we were interested in estimating $\theta$ from $X$, not the reverse. So the most satisfactory choice is then $\theta_i = \mathbb{E}(\theta_i|X)$

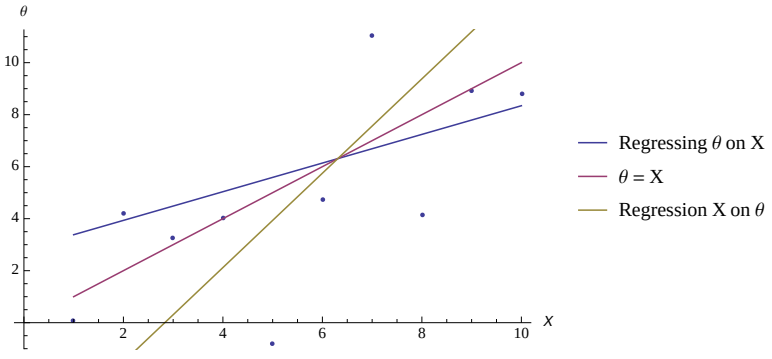


Figure 2: Viewing estimation of the mean of a multivariate normal distribution as a regression problem.

Unfortunately, although we have $X$ we do not have $\theta$ so it is not clear we can find the regression line. Initially, we will suppose that we have $\theta$ and we have a joint distribution for $(\theta, X)$ so we can find both $\mathbb{E}(X|\theta)$ and $\mathbb{E}(\theta|X)$. Then the solution to the regression problem is given by the normal equations:

$$\hat{\theta}_i = \bar{\theta} + \hat{\beta}(X_i - \bar{X})$$
$$\hat{\beta} = \frac{\sum_{i=1}^{p}(X_i - \bar{X})(\theta_i - \bar{\theta})}{\sum_{i=1}^{p}(X_i - \bar{X})^2}$$

Now if we suppose that the $\theta_i$ follow a distribution with finite second moment then we have

$$\text{Cov}(X, \theta_i) = \frac{1}{p-1}\sum_{i-1}^{p}(X_i - \bar{X})(\theta_i - \bar{\theta})$$

Using the relation of $X$ and $\theta$ we obtain

$$\text{Cov}(X, \theta) = \text{Var}(\theta) = \text{Var}(X) - 1$$

So a reasonable estimate of $\text{Cov}(X, \theta)$ is

$$\hat{\text{Var}}(X) - 1 = \frac{1}{p-1}\sum_{i=1}^{p}(X_i - \bar{X})^2 - 1$$

So whatever the distribution on $\theta$ the relationship between $X$ and $\theta$ ensures

$$\mathbb{E}\left(\sum_{i=1}^{p}(X_i - \bar{X})^2 - (p-1)\right) = \mathbb{E}\left(\sum_{i=1}^{p}(X_i - \bar{X})(\theta_i - \bar{\theta})\right)$$

Therefore we deduce that a good estimator of our regression coefficient $\beta$ is

$$\hat{\beta} = \frac{\sum_{i=1}^{p}(X_i - \bar{X})^2 - (p-1)}{\sum_{i=1}^{p}(X_i - \bar{X})^2}$$
$$= 1 - \frac{p-1}{\sum_{i=1}^{p}(X_i - \bar{X})^2}$$

Substituting this into our original equation and using $\bar{X}$ to estimate $\bar{\theta}$ we obtain

$$\hat{\theta}_i^{EM} = \bar{X} + \left(1 - \frac{p-1}{\sum_{i=1}^{p}(X_i - \bar{X})^2}\right)(X_i - \bar{X})$$

Which is a version of the Efron-Morris estimator with $p-1$ in place of $p-3$. Stigler notes that we can obtain a version of the James-Stein estimator if we follow an analogous argument except we force the regression line to go through the origin.

Now we have a heuristic argument that minimizes not just the risk but the loss for every $\theta$. The distillation of this argument into a proof was covered in the next lecture.

As an aside, consider the case $p = 2$. In this case all the regression lines we might choose coincide, therefore the estimator arising from this argument should coincide with the line $\theta_i = X_i$, and hence no improvement would be expected over the MLE.

# References

Brown, L. D. (1971). Admissible estimators, recurrent diffusions, and insoluble boundary value problems. *The Annals of Mathematical Statistics*, 42(3):855–903.

Efron, B. and Morris, C. (1972). Limiting the risk of Bayes and empirical Bayes estimators–Part II: The empirical Bayes case. *Journal of the American Statistical Association*, 67(337):pp. 130–139.

Efron, B. and Morris, C. (1975). Data analysis using Stein's estimator and its generalizations. *Journal of the American Statistical Association*, 70(350):pp. 311–319.

Lehmann, E. and Casella, G. (1998). *Theory of Point Estimation*. Springer Texts in Statistics. Springer.

Shao, P. Y.-S. and Strawderman, W. E. (1994). Improving on the James-Stein positive-part estimator. *The Annals of Statistics*, 22(3):1517–1538.

Stein, C. M. (1981). Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics*, 9(6):1135–1151.

Stigler, S. M. (1990). The 1988 neyman memorial lecture: A galtonian perspective on shrinkage estimators. *Statistical Science*, 5(1):147–155.