Stein's Paradox

Dr Richard J. Samworth, Statslab Cambridge

Perhaps the most surprising result in Statistics arises in a remarkably simple estimation problem. Let $X_1, \ldots, X_p$ be independent random variables, with $X_i \sim N(\theta_i, 1)$ for $i = 1, \ldots, p$. Writing $X = (X_1, \ldots, X_p)^T$, suppose we want to find a good estimator $\hat{\theta} = \hat{\theta}(X)$ of $\theta = (\theta_1, \ldots, \theta_p)^{\mathrm{T}}$. To define more precisely what is meant by a *good* estimator, we use the language of statistical decision theory. We introduce a **loss function** $L(\hat{\theta}, \theta)$, which measures the loss incurred when the true value of our unknown parameter is θ, and we estimate it by $\hat{\theta}$. We will be particularly interested in the squared error loss function $L(\hat{\theta}, \theta) = \|\hat{\theta} - \theta\|^2$, where $\|\cdot\|$ denotes the Euclidean norm, but other choices, such as the absolute error loss $L(\hat{\theta}, \theta) = \sum_{i=1}^{p} |\hat{\theta}_i - \theta_i|$ are of course perfectly possible.

Now $L(\hat{\theta}, \theta)$ is a random quantity, which is not ideal for comparing the overall performance of two different estimators (as opposed to the losses they each incur on a particular data set). We therefore introduce the **risk function**

$$R(\hat{\theta}, \theta) = \mathbb{E}\{L(\hat{\theta}, \theta)\}.$$

If $\hat{\theta}$ and $\tilde{\theta}$ are both estimators of θ, we say $\hat{\theta}$ **strictly dominates** $\tilde{\theta}$ if $R(\hat{\theta},\theta) \leq R(\tilde{\theta},\theta)$ for all θ, with strict inequality for some value of θ. In this case, we say $\tilde{\theta}$ is **inadmissible**. If $\hat{\theta}$ is not strictly dominated by any estimator of θ, it is said to be **admissible**. Notice that admissible estimators are not necessarily sensible: for instance, in our problem above with $p = 1$ and the squared error loss function, the estimator $\hat{\theta} = 37$ (which ignores the data!) is admissible. On the other hand, decision theory dictates that inadmissible estimators can be discarded, and that we should restrict our choice of estimator to the set of admissible ones.

This discussion may seem like overkill in this simple problem, because there is a very obvious estimator of θ: since all the components of X are independent, and $\mathbb{E}(X_i) = \theta_i$ (in other words X_i is an **unbiased** estimator of θ_i), why not just use $\hat{\theta}^0(X) = X$? Indeed, this estimator appears to have several desirable properties (for example, it is the maximum likelihood estimator and the uniform minimum variance unbiased estimator), and by the early 1950's, three proofs had emerged to show that $\hat{\theta}^0$ is admissible for squared error loss when $p = 1$. Nevertheless, Stein (1956) stunned the statistical world when he proved that, although $\hat{\theta}^0$ is admissible for squared error loss when $p = 2$, it is inadmissible when $p \geq 3$. In fact, James and Stein (1961) showed that the estimator

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

strictly dominates $\hat{\theta}^0$. The proof of this remarkable fact is relatively straightforward, and is given in the Appendix.

One of the things that is so surprising about this result is that even though all of the components of X are independent, the ith component of $\hat{\theta}^{JS}$ depends on all of the components of X. To give an unusual example to emphasise the point, suppose that we were interested in estimating the proportion of the US electorate who will vote for Barack Obama, the proportion of babies born in China that are girls and the proportion of Britons with light-coloured eyes. Then our James–Stein estimate of the proportion of democratic voters depends on our hospital and eye colour data! The reader might reasonably complain that in the above examples, the data would be binomially rather than normally distributed. However, one can easily transform binomially distributed data so that it is well approximated by a normal distribution with unit variance (see the baseball example below), and then consider the estimation problem on the transformed scale, before applying the inverse transform.

Geometrically, the James–Stein estimator shrinks each component of X towards the origin, and it is therefore not particularly surprising that the biggest improvement in risk over $\hat{\theta}^0$ comes when $\|\theta\|$ is close to zero; see Figure 1 for plots of the risk functions of $\hat{\theta}^0$ and $\hat{\theta}^{JS}$ when $p = 5$. A simple calculation shows that $R(\hat{\theta}^{JS}, 0) = 2$ for all $p \geq 2$, so the improvement in risk can be substantial when p is moderate or large. 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 $\theta_0 \in \mathbb{R}^p$ using the estimator

$$\hat{\theta}^{JS}_{\theta_0}(X) = \theta_0 + \left(1 - \frac{p-2}{\|X-\theta_0\|^2}\right)(X-\theta_0).$$

In this case, we have $R(\hat{\theta}^{JS}_{\theta_0}, \theta-\theta_0) = R(\hat{\theta}^{JS}, \theta)$, so $\hat{\theta}^{JS}_{\theta_0}$ still strictly dominates $\hat{\theta}^0$ when $p \geq 3$.

Note that the shrinkage factor in $\hat{\theta}^{JS}_{\theta_0}$ becomes negative when $\|X - \theta_0\|^2 < p - 2$, and indeed it can be proved that $\hat{\theta}^{JS}_{\theta_0}$ is strictly dominated by the positive-part James–Stein estimator

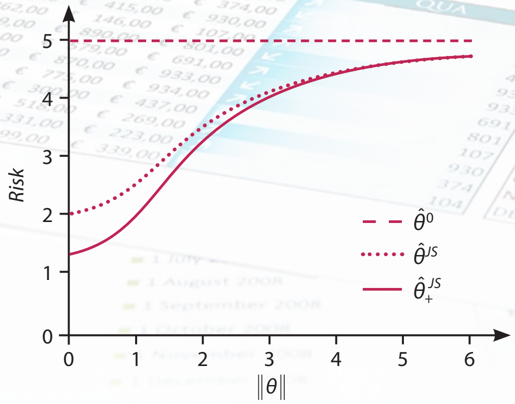
$$\hat{\theta}^{JS}_{+,\theta_0}(X) = \theta_0 + \left(1 - \frac{p-2}{\|X-\theta_0\|^2}\right)_+ (X-\theta_0),$$

where $x_+ = \max(x, 0)$. The risk of the positive-part James–Stein estimator $\hat{\theta}^{JS}_+ = \hat{\theta}^{JS}_{+,0}$ is also included in Figure 1 for comparison. Remarkably, even the positive-part James–Stein estimator is inadmissible, though it cannot be improved by much, and it took until Shao and Strawderman (1994) to find a (still inadmissible!) estimator to strictly dominate it.

Generalisations and Related Problems

It is natural to ask how crucial the normality and squared error loss assumptions are to the Stein phenomenon. As a consequence of many papers written since Stein's original masterpiece, it is now known that the normality assumption is not critical at all; similar (but more complicated) results can be proved for very wide classes of distributions. 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 θ. If our loss function were $L(\hat{\theta}, \theta) = (\hat{\theta}_1 - \theta_1)^2$, for example, then we could not improve on $\hat{\theta}^0$. This explains why it wouldn't make much sense to use the James–Stein estimator in our bizarre example above; it is inconceivable that we would be simultaneously interested in three such different quantities to the extent that we would want to incorporate all three estimation errors into our loss function.

Although Stein's result is very clean to state and prove, it may seem somewhat removed from practical statistical problems. Nevertheless, the idea at the heart of Stein's proposal, namely that of employing shrinkage to reduce variance (at the expense of introducing bias) turns out to be a very powerful one that has had a huge impact on statistical methodology. In particular, many modern statistical models may involve thousands or even millions of parameters (e.g. in microarray experiments in genetics, or fMRI studies in neuroimaging); in such circumstances, we would almost certainly want estimators to set some of the parameters to zero, not only to improve performance but also to ensure the interpretability of the fitted model.



▲ Figure 1: Risks with respect to squared error loss of the usual estimator $\hat\theta^0$, the James–Stein estimator $\hat\theta^{JS}$ and the positive-part James–Stein estimator $\hat\theta_+^{JS}$ when $p = 5$.

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

▲ Table 1: Table showing number of times at bat n_i, batting average Z_i in 1990, and career batting average π_i, of $p = 9$ baseball players.

Another important problem that is closely related to estimation is that of constructing a confidence set for θ, the aim being to give an idea of the uncertainty in our estimate of θ. Given $\alpha \in (0,1)$, an **exact (1 – α)-level confidence set** is a subset $C = C(X)$ of $\mathbb{R}^p$ such that, whatever the true value of θ, the confidence set contains it with probability exactly $1 - \alpha$. The usual, exact $(1 - \alpha)$-level confidence set for θ in our original normal distribution set-up is a sphere centred at X. More precisely, it is

$$C^0(X) = \{\vartheta \in \mathbb{R}^p : \|\vartheta - X\|^2 \le \chi^2_p(\alpha)\},$$

where $\chi^2_p(\alpha)$ denotes the upper α-point of the χ^2_p distribution (in other words, if $Z \sim \chi^2_p$, then $\mathbb{P}\{Z > \chi^2_p(\alpha)\} = \alpha$). But in the light of what we have seen in the estimation problem, it is natural to consider confidence sets that are spheres centred at $\hat\theta_+^{JS}$ (or $\hat\theta^{JS}_{+,\theta_0}$, for some $\theta_0 \in \mathbb{R}^p$). Since the distribution of $\|\hat\theta_+^{JS} - \theta\|^2$ depends on $\|\theta\|$, we can no longer obtain an exact $(1-\alpha)$-level confidence set, but it may be possible to construct much smaller confidence sets – using bootstrap methods to obtain the radius, for example – which still have at least $(1-\alpha)$-level coverage (e.g. Samworth, 2005).

A baseball data example

The following example is adapted from Samworth (2005). The data in Table 1 give the baseball batting averages (number of hits divided by number of times at bat) of $p = 9$ baseball players, all of whom were active in 1990. The source was *www.baseball-reference.com*. For $i = 1, \ldots, p$, let n_i and Z_i respectively denote the number of times at bat and batting average of the ith player during the 1990 season. Further, let π_i denote the player's true batting average, taken to be his career batting average. (Each player had at least 3000 at bats in his career.) We consider the model where $Z_1, \ldots, Z_p$ are independent, with $Z_i \sim n_i^{-1}\,\mathrm{Bin}(n_i, \pi_i)$.

We make the transformation

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

and let $\theta_i = \sqrt{n_i}\,\sin^{-1}(2\pi_i - 1)$, which means that X_i is approximately distributed as $N(\theta_i, 1)$. A heuristic argument (which can be made rigorous) to justify this is that by a Taylor expansion applied to the function $g(x) = \sqrt{n_i}\,\sin^{-1}(2x - 1)$, we have

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

and this latter expression has an approximate $N(0, 1)$ distribution when n_i is large, by the central limit theorem. In fact, since $\min_i n_i \ge 400$, an exact calculation gives that the variance of each X_i is between 1 and 1.005 for $\pi_i \in [0.2, 0.8]$. For our prior guess $\theta_0 = (\theta_{0,1}, \ldots, \theta_{0,p})^{\mathrm{T}}$, we take $\theta_{0,i} = \sqrt{\bar n}\,\sin^{-1}(2\pi_0 - 1)$, with $\pi_0 = 0.275$ and $\bar n = p^{-1}\sum_{i=1}^{p} n_i$. We find that $\|X - \theta\|^2 = 2.56$, somewhat below its expected value of around 9, though since the variance of a χ^2_9 random variable is 18, this observation is only around 1.5 standard deviations away from its mean. On the other hand, $\|\hat\theta^{JS}_{+,\theta_0} - \theta\|^2 = 1.50$, so Stein estimation does provide an improvement in this case.

Letting $\pi = (\pi_1, \ldots, \pi_p)$ and recalling that θ is a function of π, the usual 95% confidence set for π is

$$\{\pi \in [0,1]^p : \|X - \theta\|^2 \le 16.9\}.$$

On the other hand, the 95% confidence set for π constructed using the bootstrap approach is

$$\{\pi \in [0,1]^p : \|\hat{\theta}^{JS}_{+,\theta_0}(X) - \theta\|^2 \le 12.5\}.$$

Numerical integration gives that the volume ratio of the bootstrap confidence set to the usual confidence set in this case is 0.26, so the benefits of having centred the confidence set more appropriately are quite substantial.

References

1. W. James, and C. Stein, *Estimation with quadratic loss*, Proc. Fourth Berkeley Symposium, 1, 361–380, Univ. California Press (1961)
2. R. Samworth, *Small confidence sets for the mean of a spherically symmetric distribution*, J. Roy. Statist. Soc., Ser. B, 67, 343–361 (2005)
3. C. Stein, *Inadmissibility of the usual estimator of the mean of a multivariate normal distribution*, Proc. Third Berkeley Symposium, 1, 197–206, Univ. California Press (1956)
4. P. Y.-S. Shao and W. E. Strawderman, *Improving on the James–Stein positive-part estimator*, Ann. Statist., 22, 1517–1538 (1994)

Appendix

First note that since $\|X-\theta\|^2 \sim \chi^2_p$, we have $R(\hat{\theta}^0, \theta) = p$ for all $\theta \in \mathbb{R}^p$. To compute the risk of the James–Stein estimator, note that we can write

$$\begin{aligned} R(\hat{\theta}^{JS}, \theta) &= \mathbb{E}\left\{\left\| X - \theta - \frac{(p-2)X}{\|X\|^2} \right\|^2\right\} \\ &= p - 2(p-2)\sum_{i=1}^{p} \mathbb{E}\left\{\frac{X_i(X_i - \theta_i)}{\|X\|^2}\right\} + (p-2)^2 \mathbb{E}\left(\frac{1}{\|X\|^2}\right). \end{aligned}$$

Consider the expectation inside the sum when $i = 1$. We can simplify this expectation by writing it out as an n-fold integral, and computing the inner integral by parts:

$$\begin{aligned} \mathbb{E}\left\{\frac{X_1(X_1 - \theta_1)}{\|X\|^2}\right\} &= \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} \frac{x_1}{\|x\|^2} \times \frac{(x_i - \theta_i)}{(2\pi)^{p/2}} e^{-\|x-\theta\|^2/2}\, dx_1 \ldots dx_p \\ &= \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} \frac{\|x\|^2 - 2x_1^2}{\|x\|^4} \times \frac{1}{(2\pi)^{p/2}} e^{-\|x-\theta\|^2/2}\, dx_1 \ldots dx_p, \end{aligned}$$

since the integrated term vanishes. Repeating virtually the same calculation for components $i = 2, \ldots, p$, we obtain

$$\begin{aligned} \sum_{i=1}^{p} \mathbb{E}\left\{\frac{X_i(X_i - \theta_i)}{\|X\|^2}\right\} &= \sum_{i=1}^{p} \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} \left(\frac{\|x\|^2 - 2x_i^2}{\|x\|^4}\right) \frac{1}{(2\pi)^{p/2}} e^{-\|x-\theta\|^2/2}\, dx_1 \ldots dx_p \\ &= \sum_{i=1}^{p} \mathbb{E}\left(\frac{\|X\|^2 - 2X_i^2}{\|X\|^4}\right) = (p-2)\mathbb{E}\left(\frac{1}{\|X\|^2}\right). \end{aligned}$$

We therefore conclude that

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

for all $\theta \in \mathbb{R}^p$, as required.