Explaining the Stein Paradox

Kwong Hiu Yung

1999/06/10

Abstract

This report offers several rationale for the Stein paradox. Sections 1 and 2 defines the multivariate normal mean estimation problem and introduces Stein's paradox. Sections 3–8 advances the Galtonian perspective and explains Stein's paradox using regression analysis. Sections 9 and 10 approaches Stein's paradox through the conventional empirical Bayes approach. The closing sections 11 and 12 compares admissibility to equivariance and to minimaxity as criteria for simultaneous estimation.

1 Estimating Mean of Multivariate Normal

Stein's paradox arises in estimating the mean of a multivariate normal random variable. Let $X_1, X_2, \ldots, X_k$ be independent normal random variables, such that $X_i \sim N(\theta_i, 1)$. All k random variables have a common known variance, but their unknown means differ and vary separately. In other words, $(X_1, X_2, \ldots, X_k) \sim N(\theta, I)$, where $\theta = (\theta_1, \theta_2, \ldots, \theta_k)$ and I is the $k \times k$ identity matrix.

Consider estimating the mean θ under squared-error loss $L(\theta, d) = \sum_{i=1}^{k} (\theta_i - d_i)^2 = |\theta - d|^2$. A natural and intuitive estimate of θ would be X itself. Charles Stein showed that this naïve estimator $\hat{\theta} = X$ is not even admissible. For $k \geq 3$, the obvious estimate X is dominated by

$$\hat{\theta}^{JS} = (1 - \frac{k-2}{S^2})X, \text{where } S^2 = \sum_{i=1}^{k} X_i^2. \quad (1)$$

The James-Stein estimator $\hat{\theta}^{JS}$ shrinks the naïve estimate towards zero by a factor $1 - \frac{k-2}{S^2}$, where $S^2 = \sum X_i^2$ depends on the other random variables. Although the risk is optimal for $k-2$, more generally, for $k \geq 3$ and $0 < c < 2(k-2)$, any estimator of the form

$$\hat{\theta}_i^{JS} = (1 - \frac{c}{S^2})X_i \quad (2)$$

has uniformly smaller risk for all θ. Similarly, for $k \geq 4$ and $0 < c < 2(k-3)$, the family of Efron-Morris estimator

$$\hat{\theta}_i^{EM} = \bar{X} + (1 - \frac{c}{S'^2})(X_i - \bar{X}), \quad (3)$$

where

$$S'^2 = \sum_{i=1}^{k} (X_i - \bar{X})^2 \quad (4)$$

dominates the naïve estimator X. In this case, the optimal $c = k - 3$. The Efron-Morris estimators shrink towards the mean $\bar{X}$ rather than 0.

Because each X_i is independently distributed, estimating each θ_i as separate experiments would seem intuitive. Paradoxically, combining observations actually provides a much better estimate of the means.

So Stein's paradox states not only that the natural estimator X is not admissible, but also that X is dominated by $\hat{\theta}^{JS}$, an estimator which combines independent experiments.

2 Stein's Paradox Violates Intuition

On the surface, Stein's paradox seems unacceptable. Intuitively each independent experiment should not affect the other experiments. For example, consider

X_i as test scores of each student in a class. Suppose that $X_i \sim N(\theta_i, 1)$, where θ_i is the theoretical IQ of student i.

Given that each student takes his test independently, the natural estimate of each student's IQ would be his individual test score. Since each student took the test individually and separately, justice dictates that each student be judged on his personal performance. In particular, the natural estimate of θ_i would intuitively be X_i.

Yet according to Stein's paradox, the performance of the entire class should be factored into the estimate of each individual's IQ. Rather than estimate Tom's IQ by his personal test score, Stein's estimator shrinks Tom's IQ by a factor based on the standard deviation of all test scores. Stein's paradox asserts that making use of test scores of the whole class provides a better estimate of each individual's IQ.

3 Galtonian Perspective on Stein's Paradox

Consider pairs $\{(X_i, \theta_i)|i = 1, 2, \ldots, k\}$, where the X_i's are known but the θ_i's are unknown. Since $X_i \sim N(\theta_i, 1)$, write $X_i = \theta_i + Z_i$, where $Z_i \sim N(0, 1)$. Because X_i are mutually independent, the Z_i are also mutually independent.

Francis Galton over a century ago offered a perspective that can shed light on Stein's paradox. On a graph with X as the horizontal axis and θ as the vertical axis, the point (X_i, θ_i) would lie near the diagonal line $\theta = X$. Because of the error term Z_i, the point (X_i, θ_i) is horizontally shifted from the diagonal line $\theta = X$ by the random distance Z_i.

Since $E\bar{X} = \bar{\theta}$ and $Var\bar{X} = 1/k$, the mean $\bar{X}$ should be centered about $\bar{\theta}$. In other words, $(\bar{X}, \bar{\theta})$ should lie near the center diagonal line $\theta = X$ on the X/θ coordinate system.

4 Regression Analysis Picture

Given an X_i, the corresponding θ_i can be estimated using standard regression $\hat{\theta} = E[\theta|X]$. The naïve estimate $\hat{\theta} = X$, apparently agrees with $\hat{X} = E[X|\theta] = \theta$, the regression line of X on θ. However, the regression line of X on θ is not appropriate in estimating θ on the basis of X. Instead the opposite regression line should be used. Using the regression of θ on X would provide a more meaningful estimate of θ. Since the distribution of θ given X is unknown, the regression line $\hat{\theta} = E[\theta|X]$ must be approximated.

5 Linear Regression Motivates the Efron-Morris Estimator

The optimal estimator $\hat{\theta} = E[\theta|X]$ is unavailable because the distribution of θ given X is unknown. Instead, restrict attention to only linear estimators of the form

$$\hat{\theta}_i = a + bX_i, i = 1, 2, \ldots, k. \tag{5}$$

Minimizing the loss function $L(\theta, \hat{\theta}) = \sum_{i=1}^{k} (\theta_i - \hat{\theta}_i)^2$ is equivalent to finding the least squares fit to $\{(X_i, \theta_i)|i = 1, 2, \ldots, k\}$ in the X/θ coordinate system. The usual least squares fit would be

$$\hat{\theta}_i = \bar{\theta} + \hat{\beta}(X_i - \bar{X}), \text{where } \hat{\beta} = \frac{\sum (X_i - \bar{X})(\theta_i - \hat{\theta})}{\sum (X_j - \hat{X})^2} \tag{6}$$

if the points $\{(X_i, \theta_i)|i = 1, 2, \ldots, k\}$ were actually known.

Since θ_i are unknown, the points $\{(X_i, \theta_i)|i = 1, 2, \ldots, k\}$ are not available and so the least squares coefficients cannot be computed exactly. Instead, consider estimating the coefficients using only the available data $X_1, X_2, \ldots, X_k$.

Since $X \sim N(\theta, I)$ is a full-rank exponential family, X is complete sufficient for θ. So $\bar{X}$ is not only an unbiased estimator of $\bar{\theta}$ but also the UMVU estimator of $\bar{\theta}$.

The numerator of $\hat{\beta}$ can be estimated by $\sum (X_i - \bar{X})^2 - (k - 1)$ since $E[\sum (X_i - \bar{X})^2 - (k - 1)] = E[\sum (X_i - \bar{X})(\theta_i - \bar{\theta})] = \sum (\theta_i - \bar{\theta})^2$. An estimate of $\hat{\beta}$ would be

$$\frac{\sum (X_i - \bar{X})^2 - (k - 1)}{\sum (X_j - \bar{X})^2} = 1 - \frac{k - 1}{\sum (X_j - \bar{X})^2}. \tag{7}$$

The estimate of the least square linear fit becomes

$$\hat{\theta}_i^{EM} = \bar{X} + (1 - \frac{k-1}{\sum (X_j - \bar{X})^2})(X_i - \bar{X}), \quad (8)$$

which is just an Efron-Morris estimator. Although the constant $c = k-1$ is not the optimal choice among the Efron-Morris family, the estimator suggested by the estimated least square linear fit nonetheless dominates the naïve estimator X.

6 Linear Regression Motivates the James-Stein Estimator

The James-Stein estimator can also be found through a similar analysis. In the X/θ coordinate system, consider using the regression of θ on X to get an estimator $\hat{\theta} = E[\theta|X]$. Once again the distribution of θ given X is unavailable. This time, restrict attention to only estimators proportional to X, of the form

$$\hat{\theta}_i = bX_i. \quad (9)$$

Minimizing the loss function $L(\theta, \hat{\theta}) = \sum (\theta_i - \hat{\theta}_i)^2$ is equivalent to finding the least squares fit through the origin. The least squares estimator is

$$\hat{\beta} = \frac{\sum \theta_i X_i}{\sum X_i^2}. \quad (10)$$

Since $E[\sum X_i^2 - k] = E[\sum \theta_i X] = \sum \theta_i^2$ is an estimate of the numerator of $\hat{\beta}$, the least square fit through the original can be estimated by

$$\hat{\theta}_i^{JS} = (1 - \frac{k}{\sum X_i^2})X_i, \quad (11)$$

which has the form of the James-Stein estimator with $c = k$ instead of the optimal $c = k - 2$.

7 Shrinkage Estimators Dominate

The regression analysis picture provides a simple route toward the shrinkage estimators of James-Stein and Efron-Morris. The Galtonian perspective also explains why the shrinkage estimators are preferred over the naïve estimator.

The naïve estimator X corresponds to an estimate of the regression line $\hat{X} = E[X|\theta] = \theta$. To predict θ on the basis of X, however, the proper regression line to use is $\hat{\theta} = E[\theta|X]$. Shrinkage estimators are just estimates of the proper regression line.

When $k = 1, 2$, the two regression lines meet at each point (X_i, θ_i). In this case, the James-Stein shrinkage estimator does not dominate the usual estimator. The superiority of the James-Stein estimator becomes apparently only when the two regression lines differ, for $k \geq 3$.

8 Minimizing the Risk Directly

From the Galtonian viewpoint, Stein's estimator is just an estimate of the least squares fit to the regression of θ on X in the X/θ coordinate system. The least squares fit minimizes the loss function, which in turn minimizes the risk.

A similar strategy follows from minimizing the risk directly. Among the class of estimators $\hat{\theta} = bX$, the risk is function

$$R(\theta, bX) = E_\theta[\sum_{i=1}^{k} (\theta_i - bX_i)^2] \quad (12)$$

$$= kb^2 + (1-b)^2|\theta|^2, \quad (13)$$

achieves its minimum at

$$b^* = \frac{|\theta|^2}{k + |\theta|^2} \quad (14)$$

to give the minimum risk

$$R(\theta, b^*X) = k\frac{|\theta|^2}{k + |\theta|^2} < k. \quad (15)$$

In actuality, θ is the unknown estimand. Since

$$E[\sum_{i=1}^{k} X_i^2] = \sum_{i=1}^{k} (\theta_i^2 + 1) = \sum_{i=1}^{k} \theta_i^2 + k, \quad (16)$$

estimating the numerator and denominator of b^* separately gives

$$\hat{b}^* = \frac{\sum X_i^2 - k}{\sum X_i^2} = 1 - \frac{k}{\sum X_i^2}. \quad (17)$$

As before, the estimating the optimal b^* leads to the shrinking factor.

9 Empirical Bayes Approach to the James-Stein Estimator

The empirical Bayes picture offers an alternative explanation for Stein's paradox. From the Bayesian standpoint, consider

$$\begin{aligned} X_i|\theta_i &\sim N(\theta_i, 1) \quad (18)\\ \theta_i &\sim N(0, \tau^2), \quad (19)\end{aligned}$$

where each i is an independent experiment. Then the posterior density of θ_i given X_i is

$$\theta_i|X_i \sim N((1-B)X_i, 1-B), \text{ where } B = \frac{1}{\tau^2+1}. \quad (20)$$

Under squared-error loss, the Bayes estimator is $\hat{\theta}_i = E[\theta_i|X_i] = (1-B)X_i$. In the context of the original problem, however, the Bayes estimator is unsatisfactory because τ^2 is unknown. Since $X_i|\theta_i \sim N(\theta_i, 1)$, write $X_i = \theta_i + Z_i$, where Z_i is independent of θ_i and $Z_i \sim N(0,1)$. Then $X_i \sim N(0, \tau^2+1)$. The natural estimator of $B = 1/(\tau^2+1)$ is therefore

$$\hat{B} = \frac{c}{S^2}, \text{ where } S^2 = \sum X_i^2. \quad (21)$$

Thus, the empirical Bayes argument gives

$$\hat{\theta} = (1-\hat{B})X = (1 - \frac{c}{S^2})X, \quad (22)$$

which has the form of the James-Stein shrinkage estimator.

10 Empirical Bayes Approach to the Efron-Morris Estimator

The empirical Bayes picture also provides a route towards the Efron-Morris estimator. From the Bayesian standpoint, consider independent experiments

$$\begin{aligned} X_i|\theta_i &\sim N(\theta_i, 1) \quad (23)\\ \theta_i &\sim N(\mu, \tau^2), \quad (24)\end{aligned}$$

which in turn gives

$$\begin{aligned} \theta_i|X_i &\sim N(\mu + (1-B)(X_i - \mu), 1-B) \quad (25)\\ X_i &\sim N(\mu, \frac{1}{B}), \text{ where } B = 1/(\tau^2+1). \quad (26)\end{aligned}$$

Using $\hat{\mu} = \bar{X}$ and $\hat{B} = c/\sum (X_j - \bar{X})^2$ gives the Efron-Morris estimator

$$\hat{\theta}_i = \bar{X} + (1 - \frac{c}{\sum (X_j - \bar{X})^2})(X_i - \bar{X}). \quad (27)$$

11 Equivariance and Minimaxity in Simultaneous Estimation

Consider $X_1 \sim N(\theta_1, 1)$, with unknown estimand $\theta_1 \in \Re$. The estimator $\hat{\theta}_1 = X_1$ is invariant under the translation group $G_1 = \{g_a | g_a X_1 = X_1 + a\}$. The class of all equivariant estimators has form $X_1 + b$. Under the invariant loss function $L_1(\theta_1, d_1) = (\theta_1 - d_1)^2$, $\hat{\theta}_1 = X_1$ is the minimum risk equivariance estimator because the normal distribution is symmetric about its mean.

Now generalize to $X \sim N(\theta, I)$, where unknown estimand $\theta \in \Re_1 \times \Re_2 \times \ldots \times \Re_k$ has components varying separately. Consider the translation group $G = G_1 \times G_2 \times \ldots \times G_k$. The class of all equivariant estimators have form $X + b$, where constant $b \in \Re^k$. Under the invariant loss function $L(\theta, d) = \sum L_i(\theta_i, d_i) = |\theta - d|^2$, the minimum risk equivariant estimator is X, again because the normal distribution is symmetric about its mean.

Now reconsider the one-dimensional estimation problem under the minimax criterion. Under the one-dimensional squared-error loss function $L_1(\theta_1, d_1) = (\theta_1 - d_1)^2$, the Bayes estimator corresponding to the prior $\theta_1 \sim N(\mu, b^2)$ is

$$\hat{\theta}_1 = \frac{X_1 + \mu/b^2}{1 + 1/b^2}, \tag{28}$$

with posterior risk $r_1 = 1/(1 + 1/b^2)$. As $b \to \infty$, $r_1 \to 1$. Since $R(\theta_1, X_1) = E_{\theta_1}(\theta_1 - X_1)^2 = 1$ for all θ_1, the usual least favorable sequence argument shows that X_1 is minimax for estimating θ_1.

For the multivariate estimation problem, take multivariate prior θ $N(\mu, b^2 I)$. Then under the multivariate squared-error loss function $L(\theta, d) = \sum L_i(\theta_i, d_i) = |\theta - d|^2$, the Bayes estimator is $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_k)$, where

$$\hat{\theta}_i = \frac{X_i + \mu/b^2}{1 + 1/b^2}. \tag{29}$$

The posterior risk becomes $r = \sum r_i = k/(1 + 1/b^2)$. As $b \to \infty$, $r \to k$. Since $R(\theta, X) = E_\theta \sum (\theta_i - X_i)^2 = k$ for all θ, the vector X is minimax for θ.

The above arguments for equivariance and for minimaxity apply to many other simultaneous estimation situations. Under quite general assumptions, minimum risk equivariance and minimaxity both extend by components.

12 Conclusion: Paradox of Admissibility in Simultaneous Estimation

Unlike minimum risk equivariance and minimaxity, the admissibility of component estimators does not extend to the admissibility of the simultaneous estimator. Stein's paradox points out clearly that the admissibility of X_1 does not imply the admissibility of the vector X. The Galtonian picture and the Bayesian picture offer insight on why admissibility does not extend by components.

In the Galtonian perspective, the independent samples X_i are coupled through regression of $\{(X_i, \theta_i)\}$ in the X/θ coordinate system. As is typical in regression analysis, the prediction of any single point is influenced by its neighboring points.

In the Bayes perspective, the independent experiments $X_i|\theta_i$ are coupled through the common prior for all the θ_i. In estimating the common variance of θ_i, all the independent samples X_i are pooled together to provide a more accurate empirical Bayes estimate.

Minimum risk equivariance and minimaxity are both strong optimality criteria. Minimum risk equivariance with respect to a transitive group is a concrete property which defines a total ordering on the space of equivariant estimators. Minimaxity is also a concrete property which defines a total ordering on the space of estimators. So imposing minimum risk equivariance or minimaxity on component estimators is usually sufficient to guarantee that the same concrete property will hold over the product space.

Admissibility is *not* a concrete optimality criterion. Since the risk functions of two estimators may cross, comparison of risk functions in their entirety does not define a total ordering on the space of estimators. Because admissibility is a weak criterion representing the absence of optimality, the product of admissible estimators does not guarantee admissibility. On the other hand, inadmissibility is a concrete optimality criterion. So in general, the product of *inadmissible* estimators will remain inadmissible.

13 References

1. Efron, B. and Morris, C. N. (1977). Stein's paradox in statistics. *Scientific American.* **236** (5) 119–127.

2. Lehmann, E. L. (1983). *Theory of Point Estimation.* Wiley, New York.

3. Stigler, S. M. (1990). The 1998 Neyman Memorial Lecture: A Galtonian Perspective on Shrinkage Estimators. *Statistical Science.* **5** (1) 147–156.