Reversing the Stein Effect

Michael Perlman and Sanjay Chaudhuri

June 18, 2019

INTRODUCTION

- We identify the *Reverse Stein Effect*.
- A statistician who shrinks his/her data toward a point chosen without reliable knowledge about the underlying value of the parameter to be estimated but based instead upon the observed data will not be protected by the minimax property of shrinkage estimators such as that of James and Stein.
- But instead, he/she will likely incur a greater error than if shrinkage were not used.

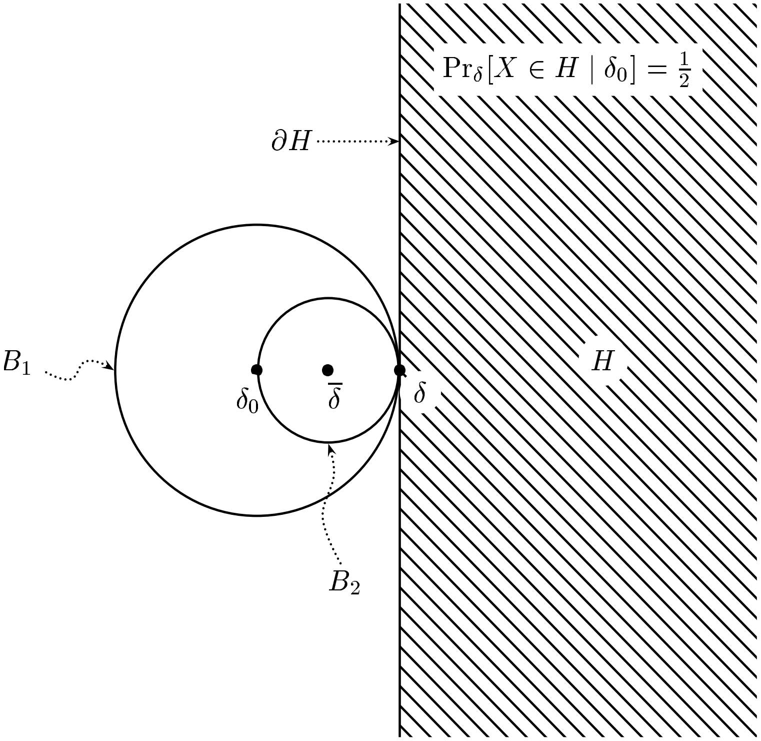
THE CASE FOR SHRINKAGE: THE STEIN EFFECT

- Suppose, $X = Y + \delta \in \mathbb{R}^p$ is an observed random vector.
- δ is an unknown location parameter,
- The unobserved random vector Y is absolutely continuous.
- We further assume that, $Y \equiv X - \delta$ is *directionally symmetric*, that is, $\vec{Y} \overset{d}{=} -\vec{Y}$, where $\vec{Y} := Y/\|Y\|$ is the unit vector in the direction of Y.
- We consider the "shrinkage" estimators for δ of the form:

  $$\widehat{\delta}_\gamma \equiv \widehat{\delta}_\gamma(X; \delta_0) = \gamma(X - \delta_0) \cdot (X - \delta_0) + \delta_0, \tag{1}$$

  where $\gamma \equiv \gamma(X - \delta_0) \in [0, 1)$ and δ_0 is *any* fixed shrinkage target point in $\mathbb{R}^p$.
- The improvement offered by such shrinkage estimators is often referred to as *the Stein Effect.*

The Stein Effect (contd.)

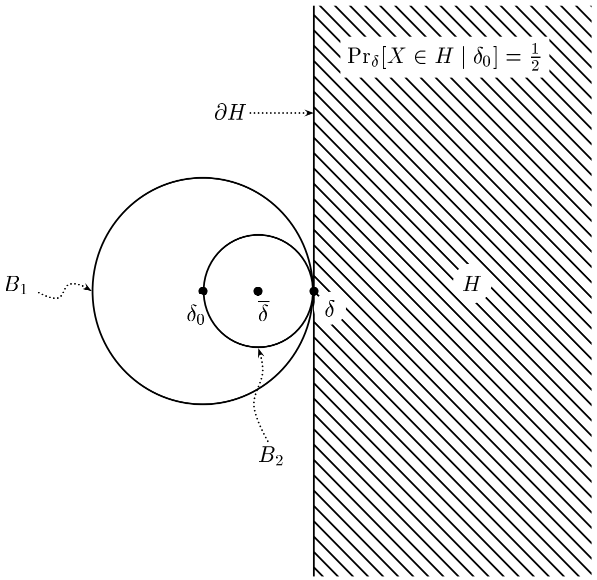


- For fixed δ and δ_0, let $B_1 \equiv B_1(\|\delta - \delta_0\|; \delta_0) \subset \mathbb{R}^p$ denote the ball of radius $\|\delta - \delta_0\|$ centred at δ_0.
- Let H be the halfspace bounded by a hyperplane ∂H tangent to B_1 at δ.
- $$\begin{aligned} & \Pr_\delta\left[\,\|X - \delta_0\| > \|\delta - \delta_0\| \,\middle|\, \delta_0\right] \\ = & \Pr_\delta\left[\, X \in B_1^c \,\middle|\, \delta_0\right] \\ > & \Pr_\delta[X \in H \mid \delta_0] = \frac{1}{2}, \end{aligned} \tag{2}$$
- The inequality in (2) follows from directional symmetry.

- Furthermore, under somewhat stronger but still general assumptions,

$$\lim_{p\to\infty} \Pr_\delta\left[\,\|X - \delta_0\| > \|\delta - \delta_0\|\,\right] \equiv \lim_{p\to\infty} \Pr_\delta\left[\, X \in B_1^c\right] = 1. \tag{3}$$

The Stein Effect (contd.)



- Thus $\|X - \delta_0\|$ is usually an overestimate of $\|\delta - \delta_0\|$, so an estimator of the form $\gamma(X - \delta_0) \cdot (X - \delta_0)$ for $\delta - \delta_0$ should be preferable to $X - \delta_0$ itself.
- Writing δ as $(\delta - \delta_0) + \delta_0$ immediately leads to estimators for δ of the form (1).
- Now, suppose, $\tilde{\gamma} \equiv \tilde{\gamma}(X - \delta_0, \delta - \delta_0)$ is allowed to depend on δ and $B_2 \equiv B_2(\|\delta - \delta_0\|; \bar{\delta})$ is the ball of radius $\frac{1}{2}\|\delta - \delta_0\|$ centred at $\frac{1}{2}(\delta_0 + \delta) \equiv \bar{\delta}$.
- Then $\{\, X \mid \exists\, \tilde{\gamma} \in [0, 1) \ni \|\widehat{\delta}_{\tilde{\gamma}} - \delta\| < \|X - \delta\|\} = B_2^c$.
- Since $B_2^c \supset B_1^c$, $\Pr_\delta[X \in B_2^c \mid \delta_0] > \frac{1}{2}$.

- Furthermore, under the assumptions,

$$\lim_{p \to \infty} \Pr_\delta[X \in B_2^c] = 1. \qquad (4)$$

- This shows that if δ were known then usually *some* shrinkage factor $\tilde{\gamma}$ applied to $X - \delta_0$ will move X closer to δ, again suggesting a search for estimators of the form (1).

THE STEIN PARADOX

- Assume now that $Y \sim N_p(0, \sigma^2 I_p)$, the multivariate normal distribution with mean 0 and covariance matrix $\sigma^2 I_p$, where $\sigma^2 > 0$ is known, so $X \sim N_p(\delta, \sigma^2 I_p)$.

- In this simple case, the James-Stein (JS) estimator for δ is given by

$$\widehat{\delta}_{JS} \equiv \widehat{\delta}_{JS}(X; \delta_0) = \left(1 - \frac{\sigma^2(p-2)}{\|X - \delta_0\|^2}\right)(X - \delta_0) + \delta_0, \tag{5}$$

  where δ_0 is a fixed but *arbitrary* point in $\mathbb{R}^p$.

- The truncated $\equiv$ "plus-rule" JS estimator

$$\widehat{\delta}^+_{JS} \equiv \widehat{\delta}^+_{JS}(X; \delta_0) = \left(1 - \frac{\sigma^2(p-2)}{\|X - \delta_0\|^2}\right)^+ (X - \delta_0) + \delta_0 \tag{6}$$

  is a shrinkage estimator of the form (1).

- These renowned estimators have the property that when $p \geq 3$, they dominate X under both the mean square error (MSE) and Pitman closeness (PC) criteria.

The Stein Paradox (contd.)

- For every fixed $\delta, \delta_0 \in \mathbb{R}^p$,

  $$\mathsf{E}_\delta[\|\widehat{\delta}^+_{JS}(X;\delta_0) - \delta\|^2 \mid \delta_0] < \mathsf{E}_\delta[\|\widehat{\delta}_{JS}(X;\delta_0) - \delta\|^2 \mid \delta_0] < \mathsf{E}_\delta[\|X - \delta\|^2] \equiv p\sigma^2, \tag{7}$$

- We can also show:

  $$\begin{aligned} &\mathsf{Pr}_\delta[\, \|\widehat{\delta}^+_{JS}(X;\delta_0) - \delta\| < \|X - \delta\| \mid \delta_0] > \mathsf{Pr}_\delta[\, \|\widehat{\delta}_{JS}(X;\delta_0) - \delta\| < \|X - \delta\| \mid \delta_0] \\ &\qquad = \mathsf{Pr}\left[\chi^2_p\left(\frac{\|\delta - \delta_0\|^2}{4\sigma^2}\right) \geq \frac{\|\delta - \delta_0\|^2}{4\sigma^2} + \frac{p-2}{2}\right] > \frac{1}{2} \end{aligned} \tag{8}$$

  and approaches 1 as $p \to \infty$ if $\frac{\|\delta - \delta_0\|}{\sigma} = o(p)$.

- Note especially that:

  (A) the improvements offered by the JS estimators can be great, especially when p is large: if $\delta = \delta_0$ then MSE($\widehat{\delta}^+_{JS}$) <MSE($\widehat{\delta}_{JS}$) $= 2\sigma^2 \ll p\sigma^2$, and if $\|\delta - \delta_0\| = o(p)$ with σ^2 fixed then $\mathsf{Pr}_\delta[\, \|\widehat{\delta} - \delta\| < \|X - \delta\|] \to 1$ as $p \to \infty$ for both $\widehat{\delta} = \widehat{\delta}_{JS}$ and $\widehat{\delta}^+_{JS}$;

  (B) the MSE and PC dominances of X by $\widehat{\delta}_{JS}$ and $\widehat{\delta}^+_{JS}$ hold even if the true mean δ is arbitrarily far from the shrinkage target δ_0.

The Stein Paradox (contd.)

- Of the two properties (A) and (B), it is (B) that is most surprising.
- It is not difficult to construct estimators that satisfy (A).
- For example a Bayes estimator w.r. to a normal prior centred at δ_0.
- However, such a Bayes estimator will not satisfy (B), the difference stemming from the fact that the Bayes estimator will have a constant shrinkage factor while the shrinkage factors in (5) and (6) are adaptive.
- When first discovered, the domination of X by the JS estimators was highly surprising, because the estimator X itself is:

  (a) the best unbiased estimator of δ,

  (b) the best translation-invariant estimator of δ,

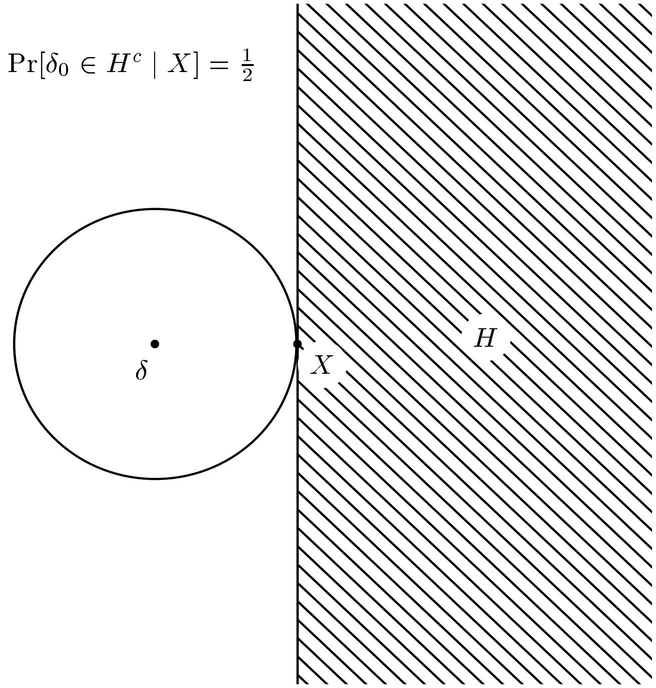
  (c) the maximum likelihood estimator (MLE) of δ,

  (d) a minimax estimator of δ, and

  (e) an admissible estimator of δ when $p = 1$ or 2.

  So compelling were these properties of X that its domination by the JS estimators came to be known as *the Stein Paradox*.

The Reverse Stein Effect



- Suppose that the position $X \sim N_p(\delta, \sigma^2 I_p)$ is known and suppose $p \geq 3$.
- If we use the truncated James-Stein estimator $\widehat{\delta}^{+}_{JS}(X; \delta_0)$ to estimate δ by shrinking X toward a fixed point δ_0, then by (7) and (8), $\widehat{\delta}^{+}_{JS}(X; \delta_0)$ is more likely to be closer to δ than our present location X is, no matter where δ is! And what's more, we can shrink X toward any δ_0 that we like!
- Consider the set,

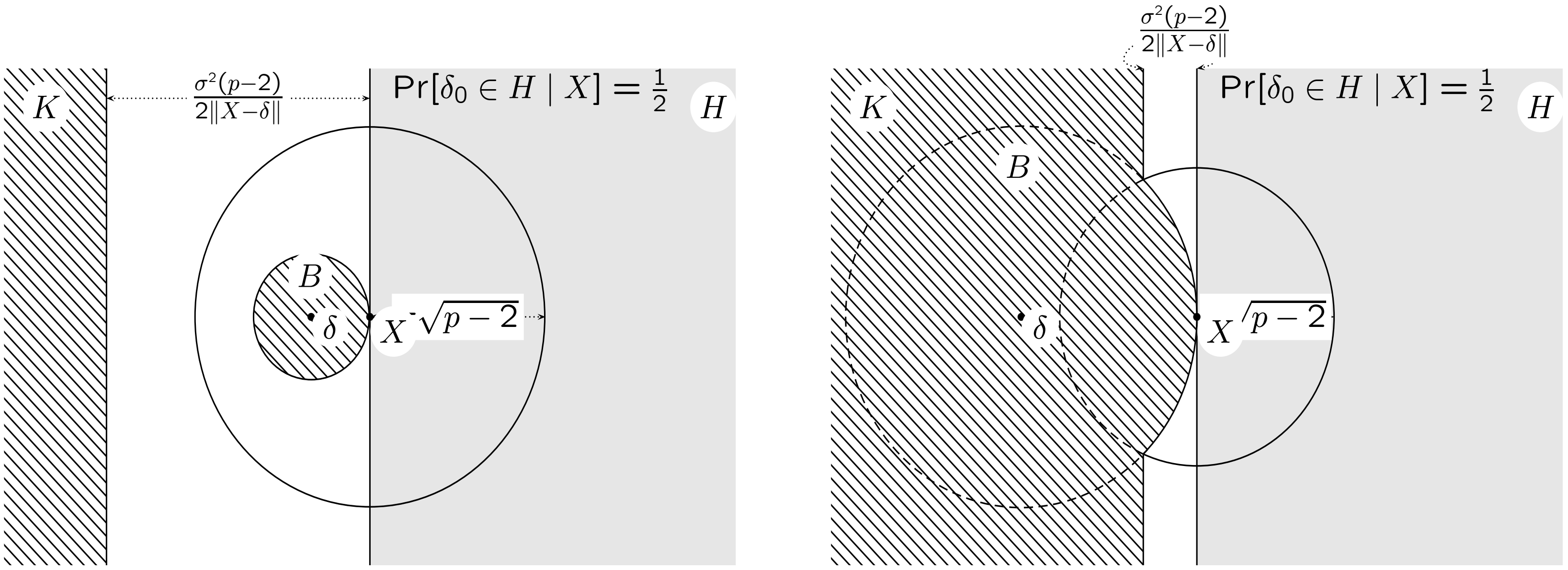
$$\{\, \delta_0 \mid \exists \breve{\gamma} \in [0, 1) \ni \|\widehat{\delta}_{\breve{\gamma}} - \delta\| < \|X - \delta\| \}, \quad (9)$$

  where $\breve{\gamma} \equiv \breve{\gamma}(X - \delta_0, X - \delta)$ may depend on δ.
- This set is exactly H^c.

- Then, since $\Pr[\,\delta_0 \in H^c | X\,] = \frac{1}{2}$ by directional symmetry, this shows that shrinkage toward a randomly chosen δ_0 would have at most a 50-50 chance of moving X closer to δ even when the shrinkage factor is chosen optimally for δ.

The Reverse Stein Effect: Quo Vadis

- For a representative values of δ, the set of all δ_0 such that the James-Stein shrinkage estimator $\widehat{\delta}^{+}_{JS}(X;\delta_0)$ lies closer to δ than the observation X.



K
$\frac{\sigma^2(p-2)}{2\|X-\delta\|}$
$\Pr[\delta_0 \in H \mid X] = \frac{1}{2}$
H
B
δ
X
$\sqrt{p-2}$
K
$\frac{\sigma^2(p-2)}{2\|X-\delta\|}$
$\Pr[\delta_0 \in H \mid X] = \frac{1}{2}$
H
B
δ
X
$\sqrt{p-2}$

(a) $\|X-\delta\| < \frac{\sigma}{2}\sqrt{p-2}$

(b) $\|X-\delta\| > \frac{\sigma}{2}\sqrt{p-2}$

- We consider the set which is complement to the cross-hatched region in the above Figure.

The Reverse Stein Effect (contd.)

- By the directional symmetry of δ_0 about X, it follows from the Figure that

$$\text{Pr}[\,\|\widehat{\delta}_{JS}^{+}(X;\delta_0)-\delta\| > \|X-\delta\| \;\big|\; X\,] \;>\; \text{Pr}[\,\delta_0 \in H \mid X\,] = \frac{1}{2}. \tag{10}$$

- If δ_0 is actually symmetrically distributed about X then it is easy to see that

$$\text{E}[\,\widehat{\delta}_{JS}^{+}(X;\delta_0) \mid X] = X, \tag{11}$$

so by Jensen's inequality,

$$\text{E}[\,\|\widehat{\delta}_{JS}^{+}(X;\delta_0)-\delta\|^2 \;\big|\; X] \;>\; \text{E}[\,\|X-\delta\|^2 \;\big|\; X] \equiv p\sigma^2 \quad \forall\delta \in \mathbb{R}^p. \tag{12}$$

- Furthermore, under additional but still general assumptions,

$$\lim_{p\to\infty} \text{Pr}_\delta[\,\|\widehat{\delta}_{JS}^{+}(X;\delta_0)-\delta\| > \|X-\delta\|\,] = 1. \tag{13}$$

The Reverse Stein Effect (Contd.)

- Don't (10) and (12) contradict (8) and (7)? For example, under *any* probability distribution for δ_0, (7) yields

$$\mathsf{E}_\delta[\, \|\widehat{\delta}^{+}_{JS}(X;\delta_0) - \delta\|^2] \; < \; \mathsf{E}_\delta[\, \|X - \delta\|^2] \equiv p\sigma^2 \quad \forall \delta \in \mathbb{R}^p, \tag{14}$$

  while (12) yields

$$\mathsf{E}_\delta[\, \|\widehat{\delta}^{+}_{JS}(X;\delta_0) - \delta\|^2] \; > \; \mathsf{E}_\delta[\, \|X - \delta\|^2] \equiv p\sigma^2 \quad \forall \delta \in \mathbb{R}^p. \tag{15}$$

- So do we shrink or not?

To Shrink or Not to Shrink – That is the Question

- Actually, no formal contradiction has occurred.

- The probabilities and expectations appearing in (7), (8), (10), and (12) are conditional probabilities and conditional expectations with different conditioning variables.

- Furthermore, the joint distributions of (X, δ_0) in (14) and (15) are different, having joint pdfs of the forms $f_\delta(X)f(\delta_0)$ and $f_\delta(X)f(\delta_0|X)$, respectively.

- In the former, X and δ_0 are independent whereas in the latter, δ_0 is dependent.

The Reverse Stein Effect is real

- The Reverse Stein Effect is just as real as the original Stein Effect itself – both are simply manifestations of the strong curvature of spheres in multi-dimensional Euclidean space.

- The figures and the results (10), (12), and (13) show that without some prior knowledge of the location δ, one should not shrink X.

- If the shrinkage target δ_0 is chosen without reliable prior information but instead is based upon the data X, the minimax/Bayesian robustness property (B) of the JS estimator is lost and no longer guarantees that shrinking is not harmful on average.

- A shrinkage estimator is only as good as, *but no better than*, the prior information upon which it is based. Without reliable *prior*, as opposed to *posterior* (as represented, for example, by "data-dependent" priors), information, *shrinkage is likely to decrease the accuracy of estimation.*

- if the statistical estimation problem is truly invariant under translation then the best invariant estimator should be used, namely X itself.