ColaForm Workshop
Paris 2017

⋆

# Stein's paradox and group rationality

⋆

Jan-Willem Romeijn
Faculty of Philosophy
University of Groningen

## Stein's paradox

Say we estimate a set of means. We can improve the predictive performance of our estimations by nudging them towards the overall mean (Vassend et al, manuscript).

- Separate experts $i$ observe values $X_{ij}$, , with $i = 1, 2, \ldots k$ and $k > 2$, and compute the averages $X_i = 1/N_i \sum_j X_{ij}$.
- They may estimate the means $\theta_i$ of the distributions that generate the observations by the maximum likelihood estimator, $\hat{\theta}_i = X_i$.
- However, the experts can improve the expected accuracy of these estimates by nudging them towards the grand mean $\bar{X} = 1/k \sum_i X_i$. The estimator

  $$\hat{\theta}_i^{\star} = \bar{X} + c(X_i - \bar{X}) = cX_i + (1-c)\bar{X},$$

  with the shrinkage factor $c = 1 - (k-2)\sigma^2 / \sum_i (X_i - \bar{x})^2$, has better overall predictive accuracy.

## What's so weird?

The proof of James and Stein (1957) is entirely formal. So the improvements in predictive performance obtain *independently of the interpretation of the estimates*.

If the $X_i$ are incidence rates of a disease in hospitals $i$ dotted around the country, the nudge towards the grand mean makes sense. But if the estimates are a completely arbitrary collection, the result of Stein seems positively weird.

## Group rationality

In what follows I will explain Stein's result and then apply the insights to another context: deliberating experts.

- By nudging towards the grand mean, the experts are effectively learning from each other, i.e., they put trust in each other's judgments.
- The size of the move towards the opinion of others is determined by considerations of predictive performance. In this sense Stein proposes an independent way of determining mutual trust.
- In Stein's paradox there is no role for a decision maker, someone who collates the opinions of all the experts. But the insights from Stein may help such decision makers as well.

# Contents

**1 Explaining Stein** **6**

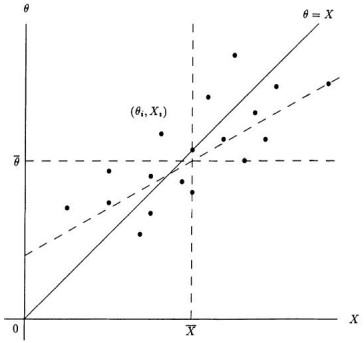
**2 An empirical Bayesian model** **10**

**3 Connections to opinion pooling** **14**

**4 Conclusion** **17**

# 1 Explaining Stein

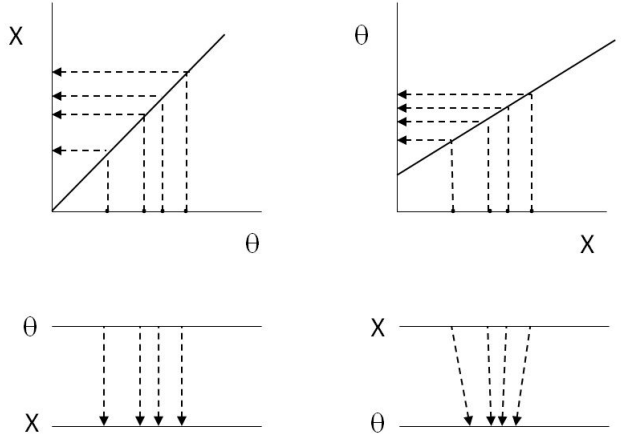
In this exposition I follow Stigler (1990) who offers a geometric explanation of Stein's result. The general idea relates to so-called regression to the mean.




We imagine that a scatter plot of $X$ and $\theta$ is given. Then we try to find the linear relation that minimizes expected error.

**Explaining Stein**

For $k > 2$, regressing $X$ on $\theta$ gives another result than the opposite regression. This roughly explains that the estimators must be nudged together.



**Explaining Stein**

In formulas: given a scatter plot of points $\langle X_i, \theta_i \rangle$, the regression line that minimizes loss in terms of the $X_i$ given the $\theta_i$ is

$$X_i = \theta_i .$$

But to minimize loss for the $\theta_i$, conditional on the $X_i$, we must choose the relation

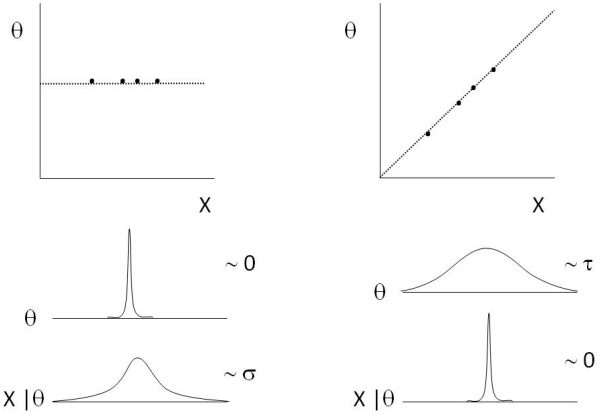
$$\theta_i = \bar{X} + c(X_i - \bar{X}),$$

where the factor $c$ is the shrinkage factor of Stein,

$$c = 1 - \frac{(k-2)\sigma^2}{\sum_i (X_i - \bar{X})^2}.$$

**Explaining Stein**

That the inverse regression line is flattened, can be seen from two extreme cases on how the scatter plot may be generated: no variance in $\theta$, and no variance in $X$ given $\theta$. The inverse regression is a mix of both.



## 2 An empirical Bayesian model

Minimizing the errors when estimating the $\theta_i$ involves an inversion in the roles of $X$ and $\theta$. This suggests that a Bayesian model can provide insights into Stein's results.

- We want to infer the values of the $\theta_i$ that minimize the expected loss, on the basis of the $X_i$.
- Ideally, we derive this expected loss from a posterior over $\theta$. If we had a prior density $P(\theta)$, this would be a simple calculation.
- The estimators of Stein can be understood as the after-the-fact reconstruction of a reasonable prior, which is then used to derive a Bayesian estimator.

This arguably dissolves the paradoxical nature of Stein's estimator: the means $\theta_i$ are implicitly assumed to have a common source, whose statistical characteristics can be reconstructed.

**An empirical Bayesian model**

Following Efron and Morris (1977) we can trace Stein's shrinkage back to a reverse engineered prior over $\theta$. The model is that the means $\theta_i$ are drawn at random from a normal, and that the data $X_i$ are then drawn from normals around those means,

$$P(\theta) \sim \text{Normal}(\bar{\theta}, \tau) \qquad \text{and} \qquad P(X_i|\theta_i) \sim \text{Normal}(\theta_i, \sigma).$$

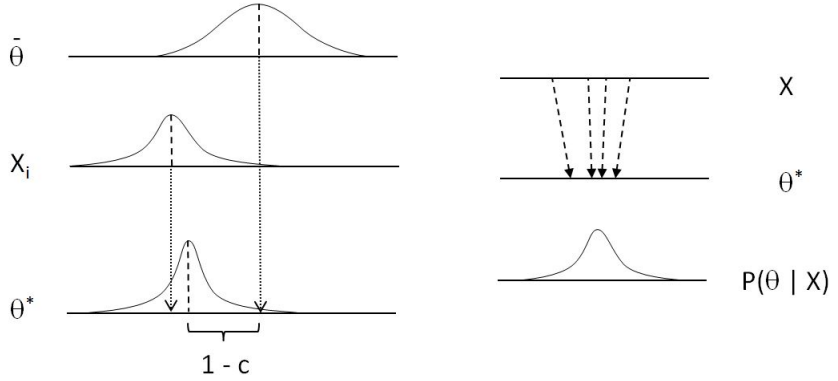
The expressions $\bar{X}$ and $\sum_i (X_i - \bar{X})^2 / k-1$ are sufficient statistics for $\bar{\theta}$ and $\sigma^2 + \tau^2$ respectively. Therefore

$$\hat{\theta_i}^{\star} = \bar{X} + \left(1 - \frac{(k-2)\sigma^2}{\sum_i (X_i - \bar{X})^2}\right)(X_i - \bar{X}) \approx \frac{\tau^2}{\sigma^2 + \tau^2} X_i + \frac{\sigma^2}{\sigma^2 + \tau^2} \bar{X}.$$

This shows that Stein's estimator coincides with the Bayesian estimator using a particular prior for $\theta$.

**An empirical Bayesian model**

Framed as a Bayesian method, Stein's shrinkage factor approximates the Kalman filter. The nudge towards the grand mean is the result of the specific prior that we chose for $\theta$.



**An empirical Bayesian model**

Stein's estimator is best understood as an *empirical* Bayesian method: the prior for $\theta$ is chosen on the basis of the data $X_i$.

- The crucial modeling assumption is that the distribution over $\theta$ has a finite second moment. The squared error loss corresponds with normally distributed $\theta$ but other distributions are possible.
- No assumption is made on the relative sizes of $\sigma$ and $\tau$ as sources of diversity among the estimates. This proportion is derived from the data $X_{ij}$.
- For small $k$ the James-Stein estimator relies a little more on the individual estimation $X_i$ owing to the factor $k-2/k-1$.
- We may take the other estimations as determining the prior over $\theta$, or alternatively as providing further data that impacts the posterior, with an improper prior at the outset.

# 3 Connections to opinion pooling

The foregoing shows that with minor adjustments, the Stein estimators are mixtures of the maximum likelihood estimations by the experts $\hat{\theta}_i = X_i$ and the collated estimations of the other group members. We have

$$\hat{\theta}_i^{\star} = w\hat{\theta}_i + (1-w)\bar{\theta},$$

with $\theta_i$ as chances and $X_i$ as opinions. A story similar to the above can be provided for Beta distributions. Weights for Normals and Beta's are

$$w_{\text{Normal}} = \frac{\tau^2}{\sigma^2 + \tau^2}, \qquad w_{\text{Beta}} = \frac{n_i}{n_i + n},$$

where $n_i$ and $n$, like $\sigma^2$ and $\tau^2$, reflect the relative sizes of uncertainty in the estimations of $\theta_i$ and $\bar{\theta}$.

**Connections to opinion pooling**

Stein's estimator can therefore be taken as a prescription for pooling opinions. Viewing pooling along these lines offers some important lessons.

- The introduction of a latent variable $\theta$, next to the manifest opinions $X_i$, allows for a richer model of social deliberation.
- The revealed opinions of the experts are only an indication of the estimates that they want to get at.
- In the richer model, the diversity of opinions has two sources: the error in the $X_i$ given $\theta_i$, and the spread in the $\theta_i$ themselves.
- The latter source of uncertainty must be kept in place by the group. It expresses the *ambiguity* in the estimation problem.

**Connections to opinion pooling**

Further lessons concern the rationale of pooling and potential iterations of it.

- Experts must pool because information on the prior is contained in the opinions of others. But they must resist full deference because their own information is most salient for their conception of the problem.
- The weight that the experts give to each other is determined by the relative sizes of two uncertainties: ambiguity and error. This offers a new interpretation of the pooling weights.
- The remaining diversity among experts is informative for the decision maker: she must factor in how ambiguous a problem is.

This adds an extra layer to the model of social deliberation. The target of the decision maker is a distribution over $\theta$ that reflects the expert opinions.

# 4 Conclusion

To summarize, I have argued for the following.

- Stein's paradox can be illuminated by focusing on the inverse inference problem involved in the estimation.
- This explanation of the paradox is relevant to rational opinion formation in a group of experts, adding a notion of latent opinion to the model of social deliberation.
- In the Bayesian representation, the shrinkage factor can be related to a pooling weight with a natural interpretation.
- It offers a new motivation for pooling opinions, presents yet another interpretation of weights, and clarifies why experts should treasure their diversity.

## Thank you

The slides for this talk will be available at http://www.philos.rug.nl/ romeyn. For comments and questions, email j.w.romeijn@rug.nl.