Philosophy of Science Seminar
Utrecht 2020

⋆

Shrinking and extremizing

Two studies on Social Deliberation

⋆

Jan-Willem Romeijn

Faculty of Philosophy
University of Groningen

Consulting experts

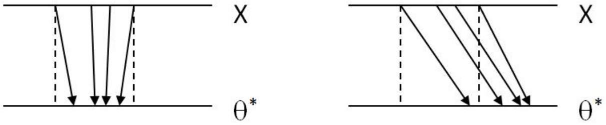
Consider the situation of a decision maker who is consulting a group of experts or research teams on the probability of some proposition.

- Every expert submits a probability value, or a distribution over them.
- All experts record the opinions of their peers and adapt their initial opinion.
- The decision maker aggregates these opinions into a single value or a distribution for further use in a decision.

This situation is well-known from statistics, for example meta-analysis. With the advance of data science, a better grip on meta-analysis is urgently needed.

Two studies on social deliberation

The first part of the talk concerns Stein’s paradox. So-called shrinkage estimators move the experts closer into the convex hull of their opinions.



The second part of the talk is about moving outside of that convex hull. Both moves can be rationalized by appealing to specific statistical modeling assumptions.

1 Shrinking

Say we estimate a set of means. We can improve the predictive performance of our estimations by nudging them towards the overall mean.

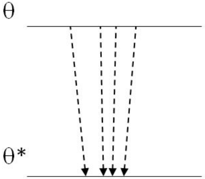
- Separate experts i observe values X_{ij}, , with $i = 1, 2, \ldots k$ and $k > 2$, and compute the averages $X_i = {}^1\!/_{N_i} \sum_j X_{ij}$.
- They may estimate the means θ_i of the distributions that generate the observations by the maximum likelihood estimator, $\hat{\theta}_i = X_i$.
- However, 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.

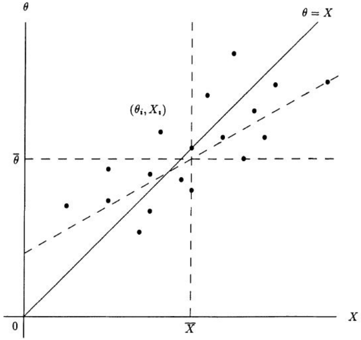
Group rationality

In this first part I will explain Stein's result and then apply the insights to the context of 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 the discussion on Stein little is said about the decision maker, someone who collates the opinions of all the experts. But the insights from Stein may help such decision makers as well.

2 Explaining Stein

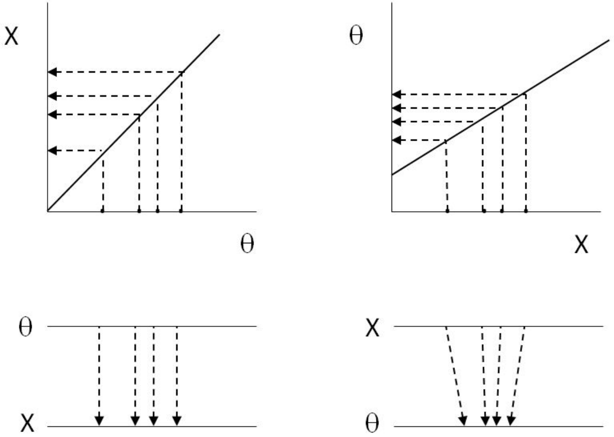
In this exposition I follow Stigler (1990) who offers a geometric explanation of Stein's result. The general idea relates to so-called regression to the mean.



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

Explaining Stein

For $k > 2$, regressing X on θ 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 distance between the line and the X_i, given the θ_i, is

$$X_i = \theta_i \, .$$

But to minimize loss for the θ_i, given the X_i, we must choose the relation

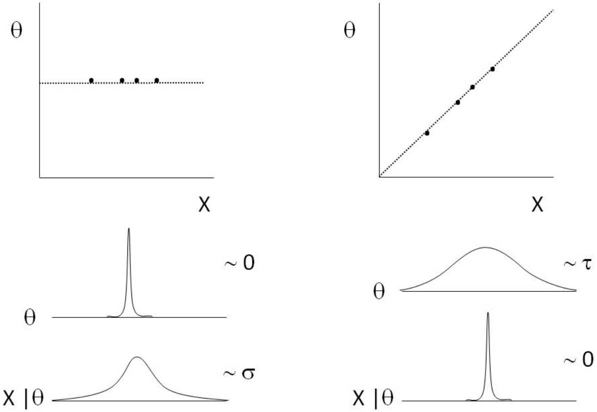
$$\theta_i = \bar{X} + c(X_i - \bar{X}) \, ,$$

where the factor c is precisely 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 θ, and no variance in X given θ. The inverse regression is a mix of both.



3 An empirical Bayesian model

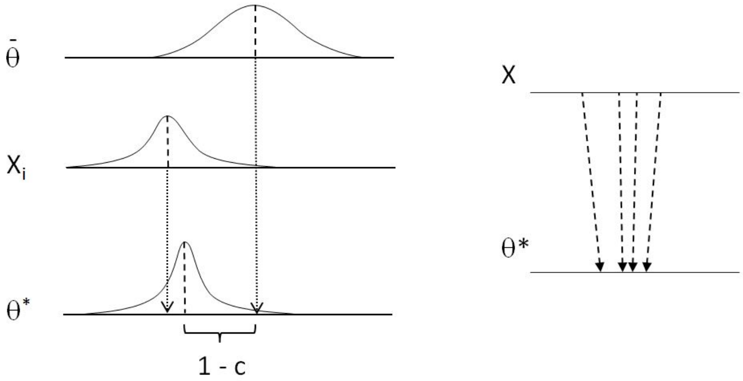
Minimizing the errors when estimating the θ_i involves an inversion in the roles of X and θ. This suggests that a Bayesian model can provide insights into Stein's results.

- We want to infer the values of the θ_i that minimize the expected loss, on the basis of the X_i.
- Ideally, we derive this expected loss from a posterior over θ. 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 θ_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 θ. The shrinkage factor approximates the Kalman filter.



That is, the nudge towards the grand mean is the result of the specific prior that we chose for θ.

An empirical Bayesian model

The Bayesian model is that the means θ_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 θ.

An empirical Bayesian model

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

- The crucial modeling assumption is that the distribution over θ has a finite second moment. The squared error loss corresponds with normally distributed θ but other distributions are possible.
- No assumption is made on the relative sizes of σ and τ 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 θ, or alternatively as providing further data that impacts the posterior, with an improper prior at the outset.

4 Connections to opinion pooling

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

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

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 σ^2 and τ^2, reflect the relative sizes of uncertainty in the estimations of θ_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 θ, 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 θ_i, and the spread in the θ_i themselves.

- The latter source of uncertainty must be kept in place by the group. It expresses a *diversity in the estimation problem*. Clinical variation among hospitals is a case in point.

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 own conception of the problem.
- The weight that the experts give to each other is determined by the relative sizes of two uncertainties: problem diversity 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 θ that reflects the expert opinions.

5 Extremizing

Recall the situation of a decision maker consulting a group of K experts or research teams on the probability of a proposition A.

- Every expert $i > 0$ submits a probability value, θ_i, or a distribution over them, $P_i(\theta)$.
- The decision maker has to collate or aggregate these opinions into a single value, $\theta_0^\star$, or a distribution, $P_0^\star(\theta)$, for further use.
- Experts themselves record the opinions of their peers and adapt their initial opinion to $\theta_i^\star$ or $P_i^\star(\theta)$.

The focus in this part lies on the decision maker's opinions.

Extremizing (cont'd)

In particular we are concerned with cases in which the aggregated opinion lies — in some sense of the word — outside the convex hull of expert opinions.

- For a binary decision target, the aggregate may be more extreme than any of the expert opinions, $\theta_0^\star > \max_{i \leq 0}(\theta_i)$.
- For a distribution $P_0^\star(\theta)$, the aggregate may also be "extremized", e.g., when the mode of $P_0^\star$ is larger than any of the modes of the P_i.

Two-and-a-half models for extremizing will be presented. They provide insight into the conditions under which such audacious forecasting is rational.

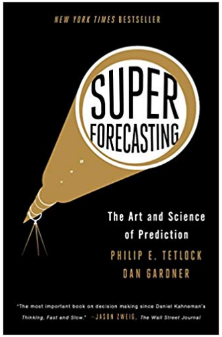
6 Empirical findings

Studies in social psychology reveal that experts sometimes amplify their opinions through social deliberation.

This phenomenon is known as “risky shift”. It is mostly deemed an irrational mechanism but it seems intuitive that the coherence of among experts is mutually confirmatory.

Empirical findings (cont’d)

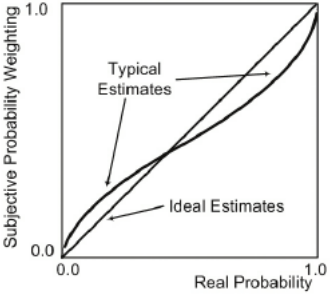
In his book on forecasting, Tetlock describes the benefits of “extremizing” in the context of a forecasting contest for military intelligence.


NEW YORK TIMES BESTSELLER
SUPER FORECASTING
The Art and Science of Prediction
PHILIP E. TETLOCK
DAN GARDNER
"The most important book on decision making since Daniel Kahneman's Thinking, Fast and Slow." —JASON ZWEIG, The Wall Street Journal

Unfortunately this is only informal and anecdotal evidence. But it suggests that extremizing can be rational.

Empirical findings (cont’d)

One potential explanation of the rationality of extremizing may be found in a “horror extremis”, as described in Kahneman’s prospect theory.



Depending on the circumstances, we overestimate low chances and underestimate high ones. Extremizing corrects this bias.

7 Extreme synergy

Assume some partition $\{A_1, ..., A_M\}$, and let $P_i(A_j) = \theta_{ij}$. Multiplicative pooling determines an aggregated probability for all A_j:

$$\theta^{\star}_{0j} \;=\; \mathrm{GP}(\theta_j) \;=\; \frac{\prod_{i\geq 0} \theta_{ij}^{w_i}}{\sum_j \prod_{i\geq 0} \theta_{ij}^{w_i}}.$$

Velasco et al. (2016) show that multiplicative pooling with $w_i = 1$ exhibits “synergy”: for a partition $\{A, \neg A\}$ and experts with $P_i(A) = \theta_i > 1/2$ we have

$$\mathrm{GP}(\theta) \;>\; \max_{i\geq 0} \theta_i.$$

Extreme synergy (cont'd)

There is a Bayesian model that explains and justifies this form of extremizing. Recall the basic format for conditioning one's opinion on the opinion of an expert i:

$$P_0^{\star}(A) = P_0(A|\ulcorner\theta_i\urcorner) = P(A) \times \frac{P_0(\ulcorner\theta_i\urcorner|A)}{P_0(\ulcorner\theta_i\urcorner)}.$$

We can readily see that the multiplicative pool fits this format once we choose $P_0(A) = \theta_0$ and, for all experts i independently,

$$P_0(\ulcorner\theta_i\urcorner|A) \sim \theta_i,$$

$$P_0(\ulcorner\theta_i\urcorner|\neg A) \sim 1-\theta_i.$$

Generalizing to weighted pooling is not straightforward because of the priors but this can be repaired.

Extreme synergy (cont'd)

Notice that the Bayesian representation of multiplicative pooling is somewhat similar to the one of linear pooling. For the latter the likelihoods are

$$P_0(\ulcorner \theta_i \urcorner | A) \sim (1 - w_i) + \frac{w_i}{\theta_0}\theta_i.$$

$$P_0(\ulcorner \theta_i \urcorner | A) \sim \left(1 + \frac{\theta_0}{1-\theta_0} w_i\right) - \frac{w_i}{1-\theta_0}\theta_i.$$

However, the two pooling methods nowhere coincide. And linear pooling does not allow for any form of extremizing: the likelihoods will ensure that the posteriors are included in the convex hull.

Extreme synergy (cont'd)

Multiplicative pooling relates naturally to votes in a Condorcet setting.

- For equal weight voting, an expert who reveals their opinion, $\theta_i > 1/2$, is like a voter who votes for A_1 with competence θ_i:

$$P_0(\ulcorner \theta_i \urcorner | A) \sim \theta_i = P_0(\,\mathrm{Vote}_i(A)\,|\,A\,),$$

- As in the Condorcet Jury Theorem, updating on ever more expert opinions will therefore make the resulting probability $\theta_0^\star$ approach 1.

But is this approach to certainty with an increasing number of experts always adequate?

8 Refined extremizing

The approach to certainty seems right if the goal of the decision maker is to converge on the truth or falsity of the proposition A, and if every new expert brings new information.

What if the goal is to converge on a distribution $P_0^\star(\theta)$? And what if the experts recycle a limited amount of information about A?

Refined extremizing (cont'd)

We assume the experts report a distibution over values of $\theta.$ For convenience we choose Beta-distributions:

$$\text{Beta}(\theta_i, N_i) \sim \theta^{\theta_i N_i}(1-\theta)^{(1-\theta_i)N_i}.$$

We can express the cautiousness or "horror extremis" of experts in the parameters of this Beta-distribution, and modulate it to express the weight of the expert:

$$P_0(\ulcorner \text{Beta}(\theta_i, N_i)\urcorner | \theta) \sim \text{Beta}\left(\theta_i + \alpha_i, w_i N_i\right),$$

where $0 < \alpha_i < 1 - \theta_i$. Larger α_i extremize the impact of the expert more.

Refined extremizing (cont'd)

Owing to the properties of Beta-distributions, we can write down analytic expressions for the resulting aggregated distributions. For weights w_i of the experts,

$$P_0^{\star}(\theta) \;=\; \text{Beta}\left(\frac{1}{K}\sum_i w_i N_i(\theta_i + \alpha_i),\, \sum_i w_i N_i\right).$$

Notice the connection of this aggregation model to Polya's urn models, Carnap's inductive logic, and their statistical foundations. Experts are all offering predictions based on their separate data sets:

$$\text{Beta}(\theta_i, N_i) \;\sim\; \text{Beta}(1/2, N_{0i}) \times \text{Beta}(\theta_i + \alpha_i, N_{1i}),$$

with $N_{0i} + N_{1i} = N_i$. The cautiousness α_i in the revealed opinions arguably derives from a middling prior that experts start with.

Refined extremizing (cont'd)

Several remarks about this model for extremizing are in order.

- Extremizing results from compensating for the cautiousness, and hence the middling priors, of the experts. But it is not clear what motivates the cautious opinions, and hence what the proportion of N_{0i} and N_{1i} is.
- We can express the uncertainty over the causes of cautious expert opinion in the likelihoods as well, in what may be called second-order likelihoods. This leads to more complicated but analytic expressions.
- For incorporating the opinions of the experts, it is also important to gauge to what extent their data sets overlap. This introduces yet another layer of uncertainty.

Refined extremizing (cont'd)

The link to Carnapian inductive logic suggests another model, in which experts report θ_i and the decision maker constructs $P_0^{\star}(\theta)$.

- Carnap's predictions of categorical properties can be generalized to predictions over a continuous attibute space.
- The underlying statistics is the much wider context of Blackwell-McQueen processes and Ferguson distributions.
- This allows us to model absolutely any aggregated distribution on the basis of the revealed opinions θ_i, and offers full control over extremizing.
- The relevant context is that of analogical predictions but this is work in progress.

9 Conclusions

We looked at two different formats for aggregating statistical results. In the part on Stein's paradox, I have argued for the following.

- Shrinkage estimators can be illuminated by focusing on the inverse inference problem involved in the estimation.
- This is relevant to rational opinion formation in a group of experts, adding a notion of problem diversity to the model of social deliberation.
- It offers a new motivation for pooling opinions, presents yet another interpretation of weights, and clarifies why experts should treasure their difference of opinion.

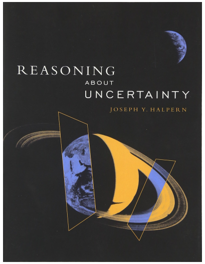
Conclusions (cont'd)

In the part on extremizing we kept the focus on the decision maker, and I argued the following.

- The "extreme synergy" of multiplicative pooling is seen to derive from a particular conception of the revealed expert opinions.
- We can justify extremizing, and control it better, when we aggregate whole distributions in a Bayesian manner. The opinions of the experts should be taken as pieces of information on which the decision maker updates.
- In eliciting advice and setting up the aggregation, we have to take into account whether we have a chance or a binary prediction as target.

Conclusions (cont'd)

Ultimately we can check the adequacy conditions for shrinking or extremizing by expressing the deliberation among the experts in a Bayesian model.



The adequate application of shrinking or extremizing will improve predictive systems and statistical decision making. Meta-analysis may be a case in point.

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.