Formal Social Epistemology Workshop
Lund 2018

⋆

Retaining diversity and seeking the extreme

⋆

Jan-Willem Romeijn
Faculty of Philosophy
University of Groningen

Meta-analysis

Consider the situation of a decision maker who is consulting a group of experts, a collection of research teams, or the like. In the current exposition, the variable of interest is a chance parameter θ.

- Every expert i submits a probability distribution over the chance variable, $P_i(\theta)$.
- The decision maker has to collate or integrate these opinions into a single $P_0(\theta)$ for further use in her practice.

This situation is well-known in statistics, namely as meta-analysis. With the advance of data science, meta-analysis is becoming more and more important.

This talk

The way in which the opinions of the experts are aggregated depends on the decision maker's goals, her interpretation of the target, and her further context.

- The decision maker has to find a midpoint between respecting the variation among experts, and reducing it down.
- If the decision target is binary, the aggregate opinion may be more extreme than any of the expert opinions is.

Both points follow naturally from an analysis of Stein's paradox, when viewed from the perspective of the decision maker, and after deliberation among experts has happened.

Contents

1 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.

- 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.

Why is this a paradox?

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, for instance, 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.

2 An empirical Bayesian model

Following Efron and Morris (1977) we can trace Stein's shrinkage back to a reverse engineered prior over θ. The 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,

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

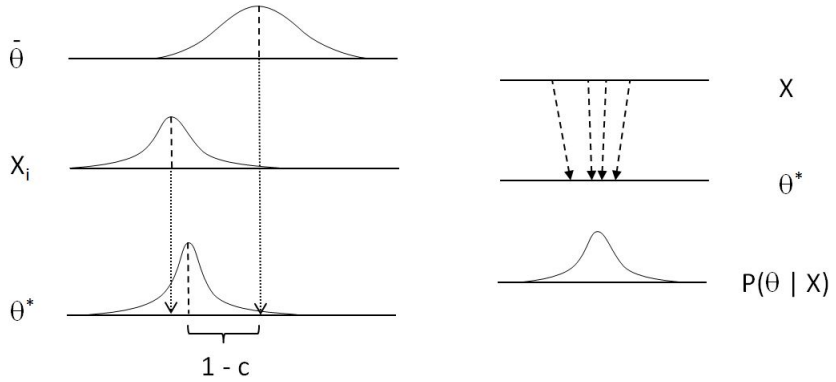
The expressions $\bar{X}$ and $\sum_i (X_i - \bar{X})^2/k-1$ are sufficient statistics for $\bar{\theta}$ and $\sigma^2 + \tau^2$ respectively. Therefore

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

This shows that Stein's estimator approximates the Bayesian estimator using a particular prior for θ.

Kalman filter

Framed as a Bayesian method, Stein's shrinkage factor approximates the Kalman filter. The nudge towards the grand mean is the result of a specific prior for θ that we can reverse-engineer on the basis of the expert opinions.



Crucial assumptions

Stein's estimator is thus 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.
- In a classical treatment we adopt a squared error loss. This 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.

3 Connections to social deliberation

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 θ_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 σ^2 and τ^2, reflect the relative sizes of uncertainty in the estimations of θ_i and $\bar{\theta}$.

Opinion pooling

Stein's estimator can therefore be taken as a prescription for pooling opinions. Viewing pooling along these lines offers some important general 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 the *disparity* in the estimation problem.

Against iterated 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 θ that reflects both uncertainty types.

4 Retaining diversity

The experts do not fully defer to the grand mean because, arguably, they work on different estimation problems, e.g., incidence rates of a disease in different parts of the country.

How much of this diversity should be retained in the opinion of the decision maker? In our example, the variation among the incidence rates of the disease may be very salient.

Pooling methods

There are well-known pooling methods for aggregating probability distributions.

- We can use point-by-point *linear* pooling of the distributions. This comes down to a weighted addition of the distributions of the experts, potentially leading to a very bumpy aggregated distribution.
- Alternatively we may use *geometric* pooling. This comes down to a weighted multiplication of the input distributions, and hence to a smooth aggregated distribution for most kinds of expert distributions.

The methods differ in the degree to which the diversity of the experts comes through in the aggregation result: the linear pool retains diversity and the geometric does not.

Bayesian representations

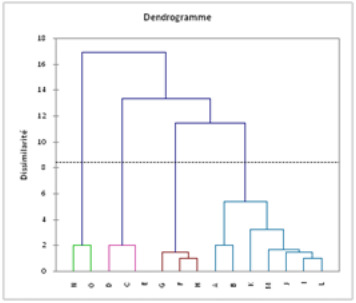
Every such pooling method can be translated into a Bayesian model, in which the probability assignments of the experts are taken as information by which the decision maker updates her opinion.

- Linear pooling corresponds to a Bayesian update for all propositions $P(\theta < t)$ according to likelihoods determined by the relative weights of the experts.
- Geometric pooling corresponds to a Bayesian update in which the distributions of the various experts serve as likelihood functions for the statistical hypotheses θ directly.

The choice to retain a certain level of diversity in the decision maker's distribution can be traced back to the Bayesian modeling assumptions.

Clustering

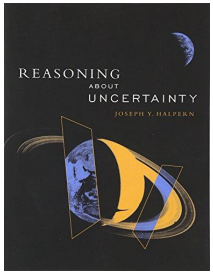
A more systematic approach to determining the level of aggregation is by clustering the opinions of the experts, and choosing the granularity in a data-driven way. This approach can be equated with a model selection approach, e.g., by Bayesian model evaluation.



We could use this clustering stage to determine how much diversity needs to be represented, apply geometric pooling within the clusters, and linear pooling across them.

5 Seeking the extreme

To achieve full generality, the opinions of the experts should be taken as pieces of information on which the decision maker updates. This gives us complete freedom in the aggregation problem.



The pooling and clustering methods are mere “epistemic shortcuts” that correspond to strong constraints on the model used by the decision maker.

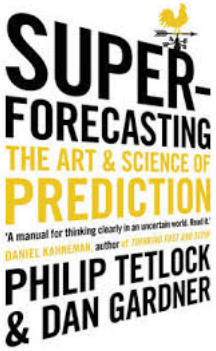
Inspiration from Condorcet voting

Parallel to the Condorcet jury theorem, we might take the opinions of the experts as indications of the truth or falsity of the proposition over which they express their opinion.

This is sometimes called “extremizing”: through Bayesian updates the opinion of the decision maker lands outside the convex hull of the expert opinions.

Interpreting the target parameter

In eliciting expert advice and setting up the aggregation method, we have to acknowledge the distinction between taking a chance and taking a binary prediction as the target.



The empirical results on extremizing indicates that sensitivity to this distinction could improve the performance of prediction and decision based on expert advice.

6 Conclusion

To summarize, I have argued for the following.

- The solution to Stein's paradox invites us to rethink the aggregation of expert judgments.
- Next to diversity in opinions there may be a diversity in problem conceptions.
- The decision maker needs to factor this latter diversity in when coming to her aggregate opinion.
- How this is done best depends on target and context.
- Presuppositions about that can be expressed in the Bayesian modeling assumptions.

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.