The Wisdom of Many in Few: *Finding Individuals Who Are as Wise as the Crowd*

Mark Himmelstein, David V.Budescu & Emily H. Ho

Depart. of Psychology, Fordham University
Depart. of Medical Social Sciences, Nortwestern UNiversity

Slides by José Luis Martínez

Outline

1. Motivation and Intended Approach
2. Skeleton of the paper and Main Items
3. Simulation
4. Forecasting Experiment
5. Meta-Predictions Experiment
6. General Discussion
7. Appendix

Motivation

Ground-truth based forecasting skills studies has a clear disadvantage: Events needs to be results in order to evaluate the forecasters performance.

This may take away from the analysis really distant resolutions events and make impossible to evaluate forecasting skills in real time.

Another complain is that previous studies use data from **forecasting tournaments** and **professional forecaster**; incurring in an obvious non-random assignment and discrepancies about time proximity advantages between forecasters which **interfere** directly **with the inferences procedure**.

Intended Approach

What they present in this paper is a tool based on *proper intersubjective score rules**; that will give us both: (1) an alternative for ground-truth based analysis (no need to wait for resolution) and (2) a real time adjustable measure for individuals' forecasting latent talent.

And also a proper experiment free of previous mentioned caveats.

Structure of the study

The paper is going to be presented as follows:

1. Simulation:
   - To show the potential of this intersubjective score rules
2. Forecasting Experiment:
   - Validity
   - Reliability
3. Meta-Predictions *(continuation)* Experiment:
   - Validity
   - Reliability

Underlying Methodology

Just to give an overview of the framework they will be using during the whole analysis.

Basically we are going to have:

- $N \equiv$ Number of forecasters
- $K \equiv$ Number of events
- $C \equiv$ Number of possible outcomes of each event.

So for a sample of N = 2, K = 2 and C = 3, a possible observation could be:

$$n_1 : [0.6, 0.2, 0.2]_{k=1}; [0.1, 0.7, 0.2]_{k=2}$$

$$n_2 : [0.5, 0.2, 0.3]_{k=1}; [0.2, 0.5, 0.3]_{k=2}$$

where the numbers inside the brackets are the probabilities assigned to each possible outcome c.

Estimators for Latent Forecasting Talent

In order to approximate/model the unobservable forecasting talent they use the following metrics of predictions accuracy:

1. **Ground-truth Scoring Rules**

   - **Brier score (BS)**:

$$\mathrm{BS}_{n,k} = \sum_{c=1}^{C_k} (f_{c,n,k} - r_{c,k})^2$$

where:

$f_{c,n,k} \equiv$ probability assigned by *forecaster n to category c of event k.*
$r_{c,k} =$ *1 if (c of event k contains the ground truth) and 0 o/w.*

$(\text{e.g}: \ r_{k=1} = [0,0,1], \ \textit{if } c = 3 \textit{ is the event's resolution})$

Estimators for Latent Forecasting Talent

2. **Intersubjective Scoring Rules**

   - **Distance Score (DS)**:

$$\mathrm{DS}_{n,k} = \sum_{c=1}^{C_k} \left(f_{c,n,k} - a_{c,k}\right)^2$$

where:

$f_{c,n,k} \equiv$ probability assigned by *forecaster n to category c of event k.*
$a_{c,k} \equiv$ *aggregate crowd forecast of probability distribution for c in event k.*

Estimators for Latent Forecasting Talent

2. **Intersubjective Scoring Rules**

   - **Expected Brier Score (EBS)**:

$$\mathrm{EBS}_{n,k} = \sum_{c=1}^{C} a_{c,k} \mathrm{BS}_{c,n,k}$$

where:

$a_{c,k} \equiv$ *aggregate crowd forecast of probability distribution for c in event k.*
$BS_{c,n,k} \equiv$ *BS that would be obtained for forecaster n if event k resolved such that category c contained the ground truth.*

Estimators for Latent Forecasting Talent

2. **Intersubjective Scoring Rules**

There could be many possibilities for defining the aggregate crowd forecast, but will be using the classic mean:

$$a_k = \frac{1}{N} \sum_{n=1}^{N} f_{n,k}$$

So for our previous example of observation we will have:

$$a_1 = [\frac{0.6 + 0.5}{2}, \frac{0.2 + 0.2}{2}, \frac{0.2 + 0.3}{2}] = [0.55, 0.2, 0.25]$$

$$a_2 = [\frac{0.1 + 0.2}{2}, \frac{0.7 + 0.5}{2}, \frac{0.2 + 0.3}{2}] = [0.15, 0.6, 0.25]$$

Simulation

For the simulation part we will be using:

- $N = 1000$
- $K = 1000$
- $C = 4$
- **Assume** forecasters (and so the crowd average) are **unbiased**. (strong assumption)
  - Based on the **theoretical proof** that intersubjective scoring rules are **proper proxy scores** when the **crowd is unbiased** *(Witkowski et al., 2017)*.

Simulation

- Now, given that, we are going to construct 3 matrices:

1. Latent Probability Distribution Matrix **(P)**:
    - $P = K \times C = 1000 \times 4$
    - where each row $p_k \sim SymmetricDirichelet(0.25)$

      $\alpha\,(= 0.25)$ is the concentration parameter. The bigger, the more uniform (or the higher entropy) shows the probability distribution.

2. Resolution Matrix **(R)**
    - $P = K \times C = 1000 \times 4$
    - where each row $p_k \sim SymmetricCategorical(p_k)$

      this means each row p_k is taking as the probability to pass to the function. Also they say cathegorical but at the end it is just a binomial function, where one item is 1 (representing the true outcome) and the others are 0's.

Simulation

- A reduce example (4×4) of this:

$$P = \begin{pmatrix} 0.0082 & 0.9679 & 0.0213 & 0.0024 \\ 0.1229 & 0.6626 & 0.1990 & 0.0155 \\ 0.6842 & 0.0432 & 0.2723 & 0.0004 \\ 0.2832 & 0.0016 & 0.0655 & 0.6497 \end{pmatrix}$$

$$R = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

The red is just to point out that it could be the case that the actual resolution does not coincide perfectly with the most representative case in the latent probability distribution.
Although it will asymptotically converge to it on average.

Simulation

3. Forecasts Matrix **(F)**:
    - $F = (NK) \times C = 1000000 \times 4$
    - where each row $f_{n,k} \sim Dirichelet(e^{\theta_n}(p_k))$

      General Dirichelet has a vector of $\alpha = [\alpha_1, \alpha_2, ..., \alpha_k]$
      The higher the α_i the bigger probability will be related to that item.

      $\theta_n \equiv$ Forecaster's latent forecasting skills.
      $\theta_n \sim N(0, 1)$

What we achieve with this is that better forecasters have more accurate predictions, or their predictions are closer to those of the latent probability distribution.

Simulation

- Continuing with the reduce example $((4 \cdot 4 =)16 \times 4)$ a possible matrix F is:

$$F = \begin{pmatrix}
0.0000 & \mathbf{1.0000} & 0.0000 & 0.0000 \\
0.0000 & \mathbf{0.9423} & 0.0577 & 0.0000 \\
0.0000 & 0.5639 & 0.4357 & 0.0004 \\
0.0000 & 1.0000 & 0.0000 & 0.0000 \\
0.0001 & \mathbf{0.9667} & 0.0331 & 0.0000 \\
0.0022 & \mathbf{0.9264} & 0.0713 & 0.0000 \\
0.0039 & 0.9470 & 0.0491 & 0.0000 \\
0.0007 & 0.9283 & 0.0710 & 0.0000 \\
0.9805 & 0.0000 & 0.0195 & 0.0000 \\
0.9246 & 0.0453 & 0.0301 & 0.0000 \\
0.2994 & 0.0000 & \mathbf{0.7006} & 0.0000 \\
0.7737 & 0.2262 & 0.0000 & 0.0000 \\
0.0007 & 0.0000 & 0.0000 & 0.9993 \\
0.0000 & 0.0000 & 0.1661 & 0.8339 \\
0.0043 & 0.0000 & 0.0000 & 0.9957 \\
0.9477 & 0.0000 & 0.0000 & 0.0523
\end{pmatrix}$$

$\theta = 1$; $\theta = 0.5$; $\theta = -0.5$; $\theta = -1$.
Bold means they put higher probability to the actual outcome (referred to matrix R)

Simulation

In order to verify that the aggregate crowd forecast effectively measures the probability distributions used in the Latent Probability Distribution matrix (P), they compute the following error:

$$\mathrm{DS}_{a,k} = \sum_{c=1}^{C_k} (a_{c,k} - p_{c,k})^2$$

The mean $DS_{a,k}$ across the 1,000 events was < 0.001.
Proving that the aggregate is a fair enough approximation for this latent probability distribution.

Simulation: Results

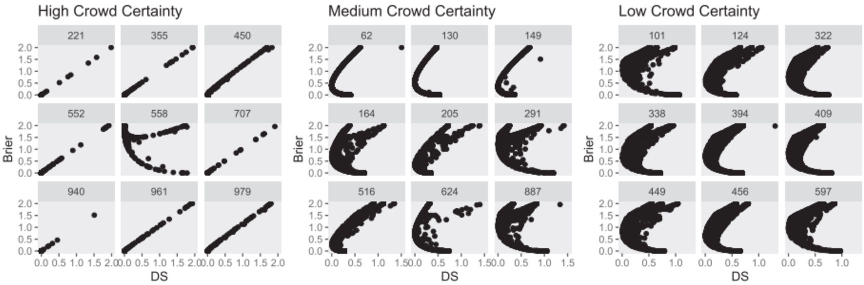


Figure: Event Level Joint Distributions of BS and DS

Levels of Certainty represents how uniform the crowd forecast (a_k) distribution was.

Numbers above graphs just represents the k (among 1000 totals) event selected.

Simulation: Results

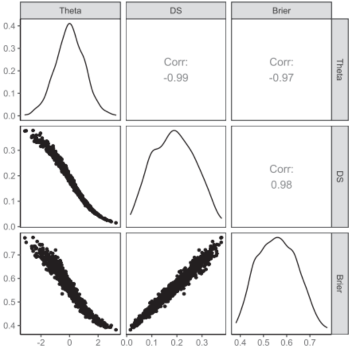


Figure: Distributions of θ_n, and Average BS and DS at the Forecaster Level

Note: *Scatterplots below the diagonal,Pearson correlation above the diagonal, and univariate distributions on the diagonal. BS = Brier score; DS = distance score.*

Simulation: Results

- We see: All the metrics are **extremely highly correlated**.

- This suggests that:

  1. If we have a sufficiently large samples and,

  2. assuming aggregate crowd forecasts are unbiased approximations of the events' true probability distributions.

  - (Under this conditions) **Intersubjective** and **Ground-truth** based scoring rules provide **valid estimates** of the latent forecasting skill.

Forecasting Experiment: Design

"We recruited a group of forecasters to make **repeated predictions** about the **same set of problems** at the **same times**, at regular intervals, allowing for direct comparison of their performance without having to adjust for confounds related to **question-specific variability** or **forecast timing**."

How they actually did that?

Forecasting Experiment: Design

- The experiment is designed as follows:

$$\begin{tikzcd}
W_1 \arrow[r, leftrightarrow, "2 \leftrightarrow 3"] & W_2 \arrow[r, leftrightarrow, "2 \leftrightarrow 3"] & W_3 \arrow[r, leftrightarrow, "2 \leftrightarrow 3"] & W_4 \arrow[r, leftrightarrow, "2 \leftrightarrow 3"] & W_5 \\
\boxed{N_1} \arrow[r, phantom, "\approx"] & \boxed{N_2} \arrow[r, phantom, "\approx"] & \boxed{N_3} \arrow[r, phantom, "\approx"] & \boxed{N_4} \arrow[r, phantom, "\approx"] & \boxed{N_5} \\
K & K & K & K & K
\end{tikzcd}$$

We have several waves (2-3 weeks distanced) of same 12 forecasting events. *(***Disclaimer:** *at beginning were 12 but do to issues they ended up in 11 events)*

People in principle are called back, but possible to recruit new participants in order to reach a minimum of 300 answers per wave.

Forecasting Experiment: Design

They ended up with:

- Full sample of N = 406 (participated at least one wave)
- **Core sample** of f **N = 175** (participated in all waves)

Full sample is going to be take into account for computing the aggregate crowd forecasts.
But the metrics and the actually analysis will be tacked only for the Core sample.

Forecasting Experiment: Results

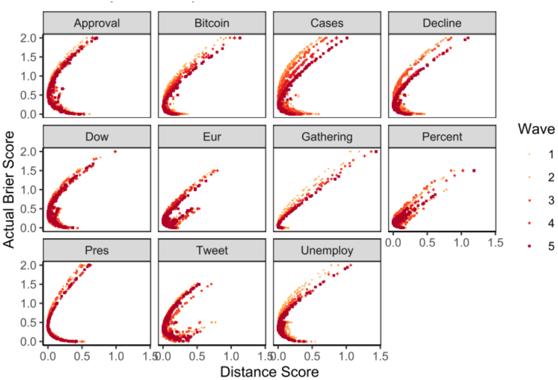


Figure: Joint Distribution of DS and BS by Item and Wave. Overall correlation $r_{DS/BS} = 0.54$

Forecasting Experiment: Validity and Reliability

Now they will be focusing on given evidence about their intersubjective based estimator truly measuring latent forecasting skills **(validity)** and also that the measure is precise **(reliability)**.

Forecasting Experiment: Validity and Reliability

Okey, just to make this clear:

What we are doing with **validity** is testing if this measure that we are using really reflects forecasting talent, so in this sense we test if it is **valid** for our purpose, but is not a test of the estimator per se.

On the other hand **reliability** is about testing if the estimator is accurate so if it is a coherence metric with a proper criteria. In this sense we **rely** on it as an estimator.

Forecasting Experiment: Validity

- **In-Sample Validity:**

  Now this "latent forecasting skills" (θ_n) is not observable anymore (this is real world not a simulation), so we can not use it to prove internal validity. Instead we will be **comparing it to ground truth** accuracy.

  In practice this means compering mean $DS_{n,w}$ with mean $BS_{n,w}$

Forecasting Experiment: Validity

- **Out-Sample Validity:** For this they use a Bootstrapped Cross-Validation.

  (Which i will try to explain with a reduce form example)

Forecasting Experiment: Validity

- Reduce form example (2 waves and 7 events):

$$\underbrace{\begin{array}{cc} \multicolumn{2}{c}{W_1:} \\ \text{training:} & \text{test:} \\ \{1,2,3\}, & \{4,5,6,7\} \\ \{2,3,4\}, & \{5,6,7,1\} \\ \multicolumn{2}{c}{\vdots} \\ \{7,6,5\}, & \{4,3,2,1\} \end{array}}_{35} \underbrace{\begin{array}{cc} \multicolumn{2}{c}{W_2:} \\ \text{training:} & \text{test:} \\ \{1,2,3\}, & \{4,5,6,7\} \\ \{2,3,4\}, & \{5,6,7,1\} \\ \multicolumn{2}{c}{\vdots} \\ \{7,6,5\}, & \{4,3,2,1\} \end{array}}_{35}, = \left[\begin{array}{c|c} \text{training} & \text{test} \\ DS_1 & BS_1 \\ DS_2 & BS_2 \\ \vdots & \vdots \\ DS_{35} & BS_{35} \end{array}\right] = [A \quad B]$$

1. Correlation of DS_i across all 35 partitions (DS intra-reliability)
2. Correlation of BS_i across all 35 partitions (BS intra-reliability)
3. **Cross-correlation of DS_i and BS_i in between partitions.**
   That is cross-correlation of A and B (Actual out-sample validity of DS)

Forecasting Experiment: Reliability

- **Split-form Cross-Validation:**
  Is this measure of consistency across samples
  (the 1 and 2 of previous slide)
- **Internal Consistency:**
  They use this Cronbach's $\alpha \in [0, 1]$ thing. It basically reflects whether this measures of errors (DS, BS) has internal coherence, which means if they are consistence across the 11 events or they outcomes are not related at all.
- **Temporal Consistency:**
  We use the intertemporal correlation of the metric across waves (time)

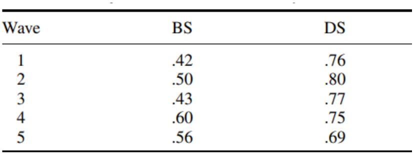
Forecasting Experiment: Reliability Results

- **Split-form Cross-Validation:**

  Across all 462 partitions:

  - Correlation between BS_n was r = 0.49
  - Correlation between DS_n was r = 0.77

- **Internal Consistency:**

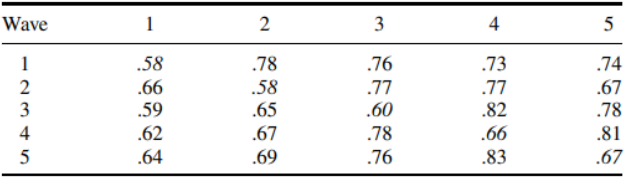


Wave	BS	DS
1	.42	.76
2	.50	.80
3	.43	.77
4	.60	.75
5	.56	.69

Figure: Cronbach's α for BS and DS by Wave

Forecasting Experiment: Reliability Results

- **Temporal Consistency:**



Wave	1	2	3	4	5
1	*.58*	.78	.76	.73	.74
2	.66	*.58*	.77	.77	.67
3	.59	.65	*.60*	.82	.78
4	.62	.67	.78	*.66*	.81
5	.64	.69	.76	.83	*.67*

Figure: Test–Retest Reliability for **BS (Below Diagonal)** and **DS (Above Diagonal)** for Each Wave With **Cross-Correlations** for Each Wave **Along the Diagonal**

Meta-Predictions Experiment

- They introduce a new scoring rule, meta-prediction distance score (MDS):

$$\mathrm{MDS}_{n,k} = \sum_{c=1}^{C_k} (m_{c,n,k} - a_{c,k})^2$$

Meta-Predictions Experiment: Validity Results

- **In-Sample Validity:**
  Correlation between mean MDS_n and mean BS_n was r $= 0.59$

  Note: *given that they are based on different prediction tasks, rather than simply different scoring criteria, the relationship between aggregate MDS and BS is* **remarkably strong***.*

Meta-Predictions Experiment: Validity Results

- **Out-Sample Validity:** The mean cross-correlation between mean MDS_n and mean BS_n was r $= 0.42$

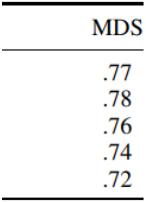
Meta-Predictions Experiment: Reliability Results

- **Split-form Cross-Validation:**

  Across all 462 partitions:

  - Correlation between MDS_n was r = 0.78

- **Internal Consistency:**

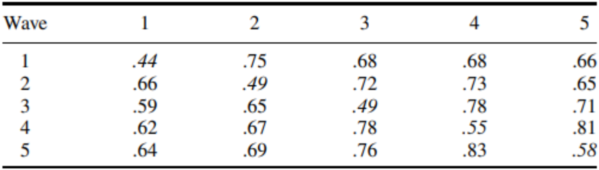


MDS
.77
.78
.76
.74
.72

Figure: Cronbach's α for MDS by Wave

Forecasting Experiment: Reliability Results

- **Temporal Consistency:**



Wave	1	2	3	4	5
1	*.44*	.75	.68	.68	.66
2	.66	*.49*	.72	.73	.65
3	.59	.65	*.49*	.78	.71
4	.62	.67	.78	*.55*	.81
5	.64	.69	.76	.83	*.58*

Figure: Test–Retest Reliability for **BS (Below Diagonal)** and **MDS (Above Diagonal)** for Each Wave With **Cross-Correlations** for Each Wave **Along the Diagonal**

General Discussion

- "Using the aggregate wisdom of the crowd to produce **intersubjective measures** of forecasting skill proved to be a **valid method of assessing forecasting skill** based on both in-sample as well as out-of-sample results"

- **Bias crowd Problem:**

  It could always be the case that forecasters are biased and so their measure is not valid any more.

  - **Potential Solution:**

    MDS, which (unlike previous) does not required unbiased for its accuracy.
    But we need to be aware that even though they both measure some how latent forecasting skills, they are not purely interchangeable.

Fin

Back-up slights

- **Proper Intersubjective score rules:**

  1. **Proper:** incentivize judges to report their true beliefs.
  2. **Intersubjective score rule:** approaches that use the beliefs of others as a proxy for the ground truth.

Go back