Judging forecasting accuracy: How human intuitions can help improving formal models

Katya Tentori

CIMeC, University of Trento

In collaboration with **Crupi**, **V.** & **Passerini**, **A.**

Ancona

Sept 13, 2018

Forecasting is everywhere…

SWIM AT
YOUR OWN RISK

True, your online profile said you were affectionate, loyal and loved to play ball... but somehow I pictured you differently.

Forecasting is not always easy…

June 23, 2016

November 8, 2016

Forecasting can shape the future itself…

INDEPENDENT

News InFact Election 2017 Voices Culture Business Indy/Life Tech

News › UK › UK Politics

Brexit research suggests 1.2 million Leave voters regret their choice in reversal that could change result

The research suggests that if a second referendum were held, the vote would be much closer

"I'm shocked that we voted for Leave, I didn't think that was going to happen," he said. "I didn't think my vote was going to matter too much because I thought we were just going to remain."

More than 4 million people have signed a petition calling for a second EU referendum…

Accurate forecasts are extremely valuable

How should the accuracy of forecasts be quantified and promoted?

Scoring rules

- assume that **forecasts** can be expressed by ***distributions of probabilities*** over future events
- measure the accuracy of forecasts **on the basis** of what **event actually materializes**

There is a lively debate on which *strictly proper* scoring rule should be preferred, and currently none of them is broadly recognized as the "best method" to evaluate forecasting accuracy

The **most popular models** are the following:

$$S_o^Q(x) = 2x_o - \sum_{i=1}^{m} x_i^2 \qquad [-1, 1] \qquad (\textit{Neutrality})$$

$$S_o^L(x) = \log x_o \qquad [-\infty, 0] \qquad (\textit{Locality})$$

$$S_o^S(x) = \frac{x_o}{\sqrt{\sum_{i=1}^{m} x_i^2}} \qquad [0, 1] \qquad (\textit{Proportionality})$$

Note: each prediction (*x*) is modelled as a probability distribution over *m* mutually exclusive and exhaustive hypotheses, the hypothesis which actually materializes is indicated with "o"

Scoring rules are commonly used for **eliciting** subjective probabilities as well as for **assessing** and **rewarding** laypeople and experts for their forecasts in a variety of areas (e.g., strategic games, operations research, …)

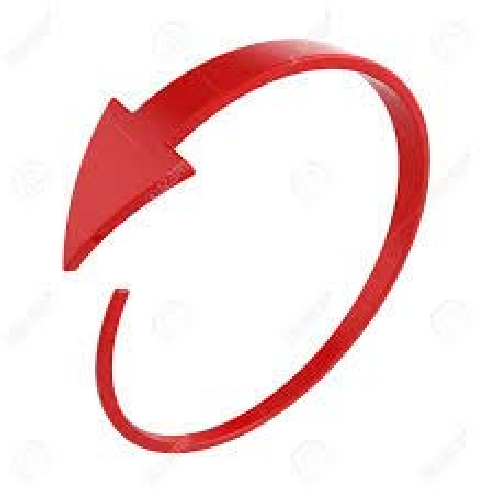
Forecasting problems → **Payoff functions** provided by some scoring rule ("**truth serums**") → Subjective Probabilities

Forecasting problems → Subjective Probabilities → **Evaluation of predictive abilities** according to some scoring rule

Scoring rules are also employed as **learning devices** for professional forecasters (e.g., meteorologists)

But …

- different scoring rules induce significantly different distribution of forecasts (Palfrey & Wang, 2009)
- evaluations based on different scoring rules can be in contradiction with each other (Bickel, 2007, and Merkle & Steyvers, 2013)



Which scoring rule best captures intuitive assessments of forecasting accuracy?

We developed a **new experimental paradigm** for eliciting ordinal judgments (ex-post evaluations) of accuracy concerning pairs of forecasts for which various combinations of associations /dissociations between *Q*, *L*, and *S* are obtained

This allowed us:

- to map the overlap between these models
- to identify which of them is descriptively most accurate
- to find possible situations in which none of them matches people's intuitive assessments of forecasting accuracy

Stimuli (general idea)

Forecasting scenarios consisting of pairs of predictions, *x* and *y*, concerning five mutually exclusive and exhaustive hypotheses, h_1, …, h_5 (N_h= 5), and an observed outcome h_o, that specified which of the five hypotheses at issue came true

More specifically, **each hypothesis h_i** was introduced to participants as referring to the **victory of team *i* in a hypothetical tournament to be played among five teams**, while the outcome indicated what team in the end won the tournament

Example of scenario

	x	Outcome	y	
h_1	20	1	10	h_1
h_2	0	0	40	h_2
h_3	80	0	0	h_3
h_4	0	0	50	h_4
h_5	0	0	0	h_5

prediction *x* proved to be more accurate than prediction *y*	prediction *x* and *y* proved to be equally accurate	prediction *y* proved to be more accurate than prediction *x*

Classification of the scenarios

Dominance: scenarios in which *Q*, *L*, and *S* all agree in evaluating one prediction as better than the other (we will denote this with $x >_{LSQ} [<_{LSQ}] y$)

Indifference: scenarios in which *Q*, *L*, and *S* all agree in evaluating the two predictions as equally good (i.e., $x =_{LSQ} y$)

Dissociation: scenarios in which *Q*, *L*, and *S* do not all agree in evaluating which of the two predictions is better (e.g., $x >_{LS} y$ and $x <_{Q} y$)

DOMINANCE

	Normative										Non-Normative			
	Transparent				Permuted						Contingent (on *Q*, *L*, and *S*)			
	x	Outcome		*y*	*x*		Outcome		*y*		*x*	Outome		*y*
h_1	40	1	>	30	40		1	>	30		40	1	>	30
h_2	30	0		40	30		0		0		30	0	≤	60
h_3	0	0	≤	0	0	≤	0		40		0	0		0
h_4	30	0		30	30		0		0		30	0	>	10
h_5	0	0		0	0		0		30		0	0		0

There is a ***transparent*** ***dominance*** of x over y *iff* $pr_x(h_o) > pr_y(h_o)$ and $pr_x(h_i) \leq pr_y(h_i)$ for all $i \neq o$

There is a ***permuted*** ***dominance*** of x over y *iff* $pr_x(h_o) > pr_y(h_o)$ and there exists a permutation π of the set of indices $i \neq o$ such that $pr_x(h_i) \leq pr_y(h_{\pi i})$ for all $i \neq o$

There is a ***contingent*** ***dominance*** of x over y *iff* $x >_{QLS} y$ but, in principle, there could exist a proper scoring rule M for which the opposite holds (i.e, $x <_M y$)

There is a ***transparent*** ***indifference*** between x and y *iff* $\mathrm{pr}_x(h_i) = \mathrm{pr}_y(h_i)$ for all i

There is a ***permuted*** ***indifference*** between x and y *iff* $\mathrm{pr}_x(h_o) = \mathrm{pr}_y(h_o)$ and there exists a permutation π of the set of indices $i \neq o$ such that $\mathrm{pr}_x(h_i) = \mathrm{pr}_y(h_{\pi i})$ for all $i \neq o$

There is a ***contingent*** ***indifference*** between x and y *iff* $x =_{QLS} y$ but, in principle, there could exist a proper scoring rule M for which $x \neq_M y$

INDIFFERENCE

	Normative						Non-Normative		
	Transparent			**Permuted**			**Contingent** (on Q, L, and S)		
	x	Outcome	y	x	Outcome	y	x	Outcome	y
h_1	40	1	40	40	1	40	40	1	40
h_2	30	0	30	30	0	0	30	0	40
h_3	0	0	0	0	0	30	0	0	10
h_4	30	0	30	30	0	0	30	0	10
h_5	0	0	0	0	0	30	0	0	0

DOUBLE DISSOCIATION

	Q vs. LS			L vs. QS			S vs. QL		
	x	Outcome	y	x	Outcome	y	x	Outcome	y
h_1	20	1	30	50	1	40	50	1	60
h_2	40	0	70	50	0	20	20	0	40
h_3	30	0	0	0	0	20	20	0	0
h_4	10	0	0	0	0	10	10	0	0
h_5	0	0	0	0	0	10	0	0	0

We considered only these three subclasses of dissociation (among the twelve that are theoretically possible) because:

a) we did not want the task to be too long and, since these subclasses involve a **rank reversal**, they appear to be particularly relevant

b) with five hypotheses and probabilities that are multiples of 10%, some subclasses of dissociation are empty

Filtering of redundant scenarios

A completely random sampling from the various subclasses would have easily ended up in many **redundant scenarios**, i.e., scenarios that are *de facto* equivalent and can be obtained from each other by means of one or a combination of the following operations:

- a swap between columns x and y (case *a* below)
- a swap of the true hypothesis row, $pr_x(h_o)$, 1, $pr_y(h_o)$, with any other row (case *b*)
- a swap of two or more values $pr(h_i)$ with $i \neq o$ within column x or y (case *c*)

Target

	x	Outcome	y
h_1	20	1	10
h_2	0	0	40
h_3	80	0	0
h_4	0	0	50
h_5	0	0	0

	a			b			c		
	x	Outcome	y	x	Outcome	y	x	Outcome	y
h_1	10	1	20	0	0	50	20	1	10
h_2	40	0	0	0	0	40	0	0	40
h_3	0	0	80	80	0	0	0	0	0
h_4	50	0	0	20	1	10	80	0	50
h_5	0	0	0	0	0	0	0	0	0

Number of scenarios in each subclass of stimuli that are obtained with our experimental paradigm, before (N) and after (N_f) the filtering procedure, respectively

					N	N_f	%
Dominance	**Transparent**	$x >_{Q,L,S} [<_{Q,L,S}]\ y$			427,570	1728 (Transparent + Permuted)	.3870 (Transparent + Permuted)
	Permuted				1,549,380		
	Contingent				2,028,400	1904	.4264
Indifference	**Transparent**	$x =_{Q,L,S} y$			5,005	94 (Transparent + Permuted)	.0211 (Transparent + Permuted)
	Permuted				66,870		
	Contingent				15,440	14	.0031
Double Dissociation	***Q* vs. *LS***	$x <_Q [>_Q]\ y$	$x >_{L,S} [<_{L,S}]\ y$		63,040	73	.0163
		$x =_Q y$	$x >_{L,S} [<_{L,S}]\ y$		15,680	14	.0031
		$x >_Q [<_Q]\ y$	$x =_{L,S} y$		377,960	246	.0551
	L* vs. *QS	$x <_L [>_L]\ y$	$x >_{Q,S} [<_{Q,S}]\ y$		3,200	12	.0027
		$x =_L y$	$x >_{Q,S} [<_{Q,S}]\ y$		453,500	371	.0831
		$x >_L [<_L]\ y$	$x =_{Q,S} y$		0	0	
	S* vs. *QL	$x <_S [>_S]\ y$	$x >_{Q,L} [<_{Q,L}]\ y$		2,360	6	.0013
		$x =_S y$	$x >_{Q,L} [<_{Q,L}]\ y$		0	0	
		$x >_S [<_S]\ y$	$x =_{Q,L} y$		0	0	
Triple Dissociation	***Q* vs. *L* vs. *S***	$x =_Q y$	$x >_L [<_L]\ y$	$x <_S [>_S]\ y$	1,600	3	.0007
		$x >_Q [<_Q]\ y$	$x =_L y$	$x <_S [>_S]\ y$	0	0	
		$x >_Q [<_Q]\ y$	$x <_L [>_L]\ y$	$x =_S y$	0	0	
					5,010,005	**4465**	**1**

EXPERIMENT 1

Participants

30 students from University of Trento (40% females; M_{age}= 24 years)

None of them had ever heard about scoring rules

They received a carbonium pen drive (€10 in value) for their participation

Procedure and **Stimuli**

For each participant, we randomly drew (without replacement) **30 scenarios**:

- 6 (2 transparent, 2 permuted, and 2 contingent) **dominance scenarios**:

$x >_{Q,L,S} [<_{Q,L,S}]\ y$.

- 6 (2 transparent, 2 permuted, and 2 contingent) **indifference scenarios**:

$x =_{Q,L,S} y$.

- 6 scenarios for each of the following **double dissociations**:

$x >_{Q} [<_{Q}]\ y$ and $x <_{L,S} [>_{L,S}]\ y$. (Q vs. LS)

$x >_{L} [<_{L}]\ y$ and $x <_{Q,S} [>_{Q,S}]\ y$; (L vs. QS)

$x >_{S} [<_{S}]\ y$ and $x <_{Q,L} [>_{Q,L}]\ y$; (S vs. QL)

EXPERIMENT 2

Participants

30 new students from University of Trento (43% females; M_{age}= 25 years)

None of them had ever heard about scoring rules

They received a carbonium pen drive (€10 in value) for their participation

Stimuli

- 3 (1 transparent, 1 permuted, and 1 contingent) dominance scenarios:

  $x >_{Q,L,S} [<_{Q,L,S}] y$

- 6 scenarios for the following double dissociation:

  $x >_{L,S} [<_{L,S}] y$ and $x =_Q y$

- 9 scenarios for each of the following double dissociations:

  $x >_Q [<_Q] y$ and $x =_{L,S} y$

  $x >_{Q,S} [<_{Q,S}] y$ and $x =_L y$

- 3 scenarios (i.e., all) for the (only possible) triple dissociation:

  $x =_Q y$; $x >_L [<_L] y$ and $x <_S [>_S] y$

Number of scenarios in each subclass of stimuli that are obtained with our experimental paradigm, before (N) and after (N_f) the filtering procedure, respectively

					N	N_f	%
Dominance	✓ **Transparent** ✓		$x >_{Q,L,S} [<_{Q,L,S}]\ y$		427,570	1728	.3870
	✓ **Permuted** ✓				1,549,380		
	✓ **Contingent** ✓				2,028,400	1904	.4264
Indifference	✓ **Transparent**		$x =_{Q,L,S}\ y$		5,005	94	.0211
	✓ **Permuted**				66,870		
	✓ **Contingent**				15,440	14	.0031
Double Dissociation	✓ ***Q* vs. *LS***	$x <_Q [>_Q]\ y$		$x >_{L,S} [<_{L,S}]\ y$	63,040	73	.0163
		✓ $x =_Q\ y$		$x >_{L,S} [<_{L,S}]\ y$	15,680	14	.0031
		✓ $x >_Q [<_Q]\ y$		$x =_{L,S}\ y$	377,960	246	.0551
	✓ ***L* vs. *QS***	$x <_L [>_L]\ y$		$x >_{Q,S} [<_{Q,S}]\ y$	3,200	12	.0027
		✓ $x =_L\ y$		$x >_{Q,S} [<_{Q,S}]\ y$	453,500	371	.0831
		$x >_L [<_L]\ y$		$x =_{Q,S}\ y$	0	0	
	✓ ***S* vs. *QL***	✓ $x <_S [>_S]\ y$		$x >_{Q,L} [<_{Q,L}]\ y$	2,360	6	.0013
		$x =_S\ y$		$x >_{Q,L} [<_{Q,L}]\ y$	0	0	
		$x >_S [<_S]\ y$		$x =_{Q,L}\ y$	0	0	
Triple Dissociation	***Q* vs. *L* vs. *S***	✓ $x =_Q\ y$	$x >_L [<_L]\ y$	$x <_S [>_S]\ y$	1,600	3	.0007
		$x >_Q [<_Q]\ y$	$x =_L\ y$	$x <_S [>_S]\ y$	0	0	
		$x >_Q [<_Q]\ y$	$x <_L [>_L]\ y$	$x =_S\ y$	0	0	
					5,010,005	**4465**	**1**

To have a measure of the reliability of participants' judgments and reduce the impact of possible random answers, **we presented each scenario twice** (counterbalancing the left/right position of the two predictions)

Therefore, each participant was presented with two blocks of **30 scenarios** that were identical except for the reversed left/right position of the two predictions in the corresponding scenarios and the order of scenarios (which was randomized)

Results…

EXP 1

Average response times for consistent and inconsistent judgments, and percentages of inconsistent judgments for each class of scenarios

		Consistent judgments	**Inconsistent** judgments	
		RT (sec)	RT (sec)	%
Dominances $x >_{Q,L,S} y$	Transparent	5.34	4.44	3
	Permuted	7.33	-	0
	Contingent	13.15	31.56	5
Indifferences $x =_{Q,L,S} y$	Transparent	3.56	-	0
	Permuted	7.77	29.13	2
	Contingent	20.54	23.25	33
Double Dissociations	$x >_{Q} y$ and $x <_{L,S} y$	11.56	25.16	16
	$x >_{L} y$ and $x <_{Q,S} y$	9.90	16.97	13
	$x >_{S} y$ and $x <_{Q,L} y$	8.14	18.85	10
Overall		**9.70**	**21.34**	**9**

EXP 2

Average response times for consistent and inconsistent judgments, and percentages of inconsistent judgments for each class of scenarios

		Consistent judgments	Inconsistent judgments	
		RT (sec)	RT (sec)	%
Dominances $x >_{Q,L,S} y$	Transparent	4.63	-	0
	Permuted	4.66	3.80	7
	Contingent	4.83	15.47	7
Double Dissociations	$x >_{L,S} y$ and $x =_Q y$	9.51	18.12	18
	$x >_Q y$ and $x =_{L,S} y$	8.81	14.13	21
	$x >_{Q,S} y$ and $x =_L y$	11.03	13.27	30
Triple Dissociation	$x >_L y$ and $x <_S y$ and $x =_Q y$	7.75	17.55	21
	Overall	**7.32**	**13.72**	**15**

EXP 1

Average agreement (in %) between (consistent) judgments and *Q*, *L*, and *S* for each class of scenarios

		Q	L	S	none
Dominances $x >_{Q,L,S} y$	Transparent	100	100	100	0
	Permuted	100	100	100	0
	Contingent	100	100	100	0
Indifferences $x =_{Q,L,S} y$	Transparent	100	100	100	0
	Permuted	98	98	98	2
	Contingent	25	25	25	75
Double Dissociations	$x >_Q y$ and $x <_{L,S} y$	8	86	86	6
	$x >_L y$ and $x <_{Q,S} y$	16	84	16	0
	$x >_S y$ and $x <_{Q,L} y$	94	94	6	0

EXP 2

Average agreement (in %) between (consistent) judgments and Q, L, and S for each class of scenarios

		Q	L	S	none
Dominances $x >_{Q,L,S} y$	Transparent	100	100	100	0
	Permuted	100	100	100	0
	Contingent	96	96	96	4
Double Dissociations	$x >_{L,S} y$ and $x =_Q y$	7	91	91	2
	$x >_Q y$ and $x =_{L,S} y$	22	72	72	6
	$x >_{Q,S} y$ and $x =_L y$	37	25	37	38
Triple Dissociation	$x >_L y$ and $x <_S y$ and $x =_Q y$	0	68	32	0

CONCLUSION

Overall, *L* is the model that best captures intuitive assessments of forecasting accuracy
However, *L* is not perfect and its descriptive limitations/shortcomings are systematic

These results of these experiments might have

interesting implications for

the **development** of new **/**
the **refinement** of the existing
formal models

the development of
"**tailored scoring rules**"
that are effective in improving
forecasting accuracy in various contexts
and for different experts

Suggestions for future research

To generalize our experimental procedure to include **more complex forecasting scenarios** in which:

- **multiple** forecasts have to be evaluated together
- **under-and over-prediction** errors are not equally bad
- the **rank order** of the forecasts matters

To employ **different participants** (e.g., experts or even "superforecasters" (provided they exist :-)

Thanks for your attention! ☺