Evaluating probabilistic COVID19 forecasts under partial missingness: A pairwise comparison approach

Johannes Bracher

Karlsruhe Institute of Technology Heidelberg Institute for Theoretical Studies

Thanks to J. Bien, E. Cramer, T. Gneiting, E. Ray, N. Reich, R. Tibshirani for helpful discussions

October 27, 2020

▲□▶ ▲□▶ ▲□▶ ▲□▶ ■ ●の00

Why take into account uncertainty?

Forecast quality cannot be fully described considering only the central tendency:



- Good forecasts "maximize sharpness subject to calibration"
- Proper scoring rules (Gneiting and Raftery 2007) allow us to compare probabilistic forecasts

Why take into account uncertainty?

Forecast quality cannot be fully described considering only the central tendency:



- Good forecasts "maximize sharpness subject to calibration"
- Proper scoring rules (Gneiting and Raftery 2007) allow us to compare probabilistic forecasts

The interval score

Consider a central $(1 - \alpha) \times 100\%$ prediction interval [I, u] and observation y. The **interval score** is given by



The weighted interval score

Bracher, Ray, Gneiting, Reich (2020), https://arxiv.org/abs/2005.12881 To assess prediction intervals at levels $(1 - \alpha_0, \dots, 1 - \alpha_K)$

simultaneously we can use the weighted interval score:

$$\mathsf{WIS}_{lpha_{0:K}}(F,y) = rac{1}{K+1} imes \sum_{k=0}^{K} rac{lpha}{2} imes \mathsf{IS}_{lpha_k}(F,y).$$

which approximates the CRPS and generalizes the AE.



y

The weighted interval score

Bracher, Ray, Gneiting, Reich (2020), https://arxiv.org/abs/2005.12881 To assess prediction intervals at levels $(1 - \alpha_0, \dots, 1 - \alpha_K)$ simultaneously we can use the weighted interval score:

$$\mathsf{WIS}_{lpha_{0:K}}(F,y) = rac{1}{K+1} imes \sum_{k=0}^{K} rac{lpha}{2} imes \mathsf{IS}_{lpha_k}(F,y).$$

which approximates the CRPS and generalizes the AE.



у

< ∃⇒

Systematic comparison of different forecasters

- Mean weighted interval scores for aggregation across space, time, target
 - preserve propriety (i.e. don't create an incentive to cheat)
 - scores from larger states and weeks with high incidence are influential

Restriction to models fulfilling (E. Cramer / N. Reich):

- covering at least half of all states
- covering at least 12 out of 15 wks between 2020-05-18 and 2020-08-24
- forecast targets with major revisions of truth data removed

Difficulty: Different forecasters cover different subsets of weeks and locations.

- could be dealt with mixed effects regression, but challenging
- my suggestion: use pairwise comparisons
- main assumption: "non-informative missingness"

Availability of forecasts from different models

After removal of weeks/targets with major revisions of truth data

-05-18	-05-25	-06-01	-06-08	-06-15	-06-22	-06-29	-07-06	-07-13	-07-20	-07-27	-08-03	-08-10	-08-17	-08-24		
2020	2020	2020	2020	2020	2020	2020	2020	2020	2020	2020	2020	2020	2020	2020	•	
49.00	49.00	49.00	49.00	50.00	51.00	54.00	54.00		54.00	54.00		55.00			YYG-ParamSearch	
45.00	45.00	45.00	45.00	46.00	47.00	50.00	50.00	50.00	50.00	50.00		0.00	51.00		UT-Mobility	
0.00	0.00	45.00	45.00	46.00	47.00	50.00	50.00	50.00	50.00	50.00					USACE-ERDC_SEIR	
45.00	45.00	45.00	45.00	46.00	47.00	50.00					56.00		56.00		UMass-MechBayes	
0.00	0.00	0.00	0.00	46.00	47.00	54.00	54.00				54.00		54.00	54.00	UCLA-SuEIR	
44.00	44.00	44.00	44.00	45.00	46.00	49.00	49.00	49.00	49.00	49.00	50.00	50.00	50.00	50.00	MOBS-GLEAM_COVID	
0.00	0.00	0.00	0.00	48.00	49.00										LANL-GrowthRate	10
49.00	49.00	49.00	49.00	50.00		54.00	54.00		54.00	54.00					JHU_IDD-CovidSP	20
45.00	0.00	45.00	45.00	46.00	47.00	50.00	50.00	0.00	50.00	50.00			0.00		IHME-CurveFit	30
0.00	25.00	31.00	29.00	32.00	38.00	39.00	40.00	45.00	45.00	44.00	43.00	44.00	46.00	48.00	GT-DeepCOVID	-
0.00	0.00	0.00	49.00	50.00					54.00	54.00					COVIDhub-ensemble	40
50.00	50.00	50.00	50.00	51.00											COVIDhub-baseline	50

Number of covered locations by date and model

Ratios of mean WIS

If all forecasts were available from all forecasters the following would work:

$$\theta_{ij} = \frac{\text{mean WIS model } i}{\text{mean WIS model } j} = \begin{cases} < 1 \text{ if } i \text{ better than } j \\ > 1 \text{ if } i \text{ worse than } j \end{cases}$$
$$\theta_i = \frac{\text{mean WIS model } i}{\left(\prod_{m=1}^{M} \text{mean WIS model } m\right)^{1/M}} = \begin{cases} < 1 \text{ if } i \text{ better than avg} \\ > 1 \text{ if } i \text{ worse than avg} \end{cases}$$
$$\Rightarrow \theta_{ij} = \frac{\theta_i}{\theta_j}$$

 θ_i is a scale-free measure of relative performance (**relative WIS** skill wrt models $1, \ldots, M$). Ratios of average scores are easier to interpret that differences.

Problem: The θ_i can only be evaluated if all forecasters cover all prediction tasks.

Comparisons under partial missingness

If all forecasts were available from all forecasters then

$$\theta_i = \left(\prod_{m=1}^M \theta_{im}\right)^{1/M}$$

If some forecasts are missing we can still compute

$$ilde{ heta}_{ij} = rac{\mathsf{mean WIS model } i \; \mathsf{on} \; \mathcal{A}_{ij}}{\mathsf{mean WIS model } j \; \mathsf{on} \; \mathcal{A}_{ij}}$$

with A_{ij} as the overlap of available forecasts by i and j and

$$ilde{ heta}_i = \left(\prod_{m=1}^M ilde{ heta}_{im}\right)^{1/M}$$

Now generally $\tilde{\theta}_{ij} \neq \tilde{\theta}_i / \tilde{\theta}_j$, but we hope that $\tilde{\theta}_{ij} \approx \tilde{\theta}_i / \tilde{\theta}_j$

Observed WIS ratios $\tilde{\theta}_{ij}$

Direct pairwise comparison: WIS ratio

blue: row model better; red: column model better Pairwise comparisons are *almost* transitive!

Relative WIS skills $\tilde{\theta}_i$

model	$ ilde{ heta}_i$
COVIDhub-ensemble	0.56
YYG-ParamSearch	0.63
UMass-MechBayes	0.72
LANL-GrowthRate	0.89
IHME-CurveFit	0.92
UCLA-SuEIR	0.95
GT-DeepCOVID	0.96
COVIDhub-baseline	1.09
MOBS-GLEAM_COVID	1.18
JHU_IDD-CovidSP	1.46
UT-Mobility	1.56
USACE-ERDC_SEIR	1.78

Checking agreement between $\tilde{ heta}_{ij}$ and $\tilde{ heta}_i/\tilde{ heta}_j$

1.00	0.89	0.79	0.63	0.59	0.59	0.61	0.52	0.48	0.36	0.39	0.32	COVIDhub-ensemble	4
1.13	1.00	0.88	0.71	0.66	0.66	0.69	0.58	0.54	0.41			YYG-ParamSearch	_
1.27	1.13	1.00	0.80	0.75	0.75	0.78	0.66	0.61	0.46			UMass-MechBayes	3
1.59	1.41	1.25	1.00	0.94	0.93	0.97	0.82	0.76	0.57	0.61	0.50	LANL-GrowthRate	_
1.69	1.50	1.33	1.07	1.00	0.99	1.04	0.87	0.81	0.61	0.65	0.54	UCLA-SuEIR	2
1.70	1.51	1.34	1.07	1.01	1.00	1.04	0.88	0.81	0.61	0.66	0.54	GT-DeepCOVID	
1.63	1.45	1.28	1.03	0.96	0.96	1.00	0.84	0.78	0.59	0.63	0.52	IHME-CurveFit	1
1.94	1.72	1.52	1.22	1.14	1.14	1.19	1.00	0.92	0.70	0.75	0.61	COVIDhub-baseline	
2.10	1.87	1.65	1.32	1.24	1.23	1.29	1.08	1.00	0.76	0.81	0.67	MOBS-GLEAM_COVID	- 0
2.77	2.46	2.18	1.74	1.64	1.63	1.70	1.43	1.32	1.00	1.07	0.88	UT-Mobility	
2.59	2.30	2.04	1.63	1.53	1.52	1.59	1.34	1.23	0.94	1.00	0.82	JHU_IDD-CovidSP	
3.15	2.80	2.48	1.98	1.86	1.85	1.93	1.63	1.50	1.14	1.22	1.00	USACE-ERDC_SEIR	
COVIDhub-ensemble	YYG-ParamSearch	UMass-MechBayes	LANL-GrowthRate	UCLA-SuEIR	GT-DeepCOVID	IHME-CurveFit	COVIDhub-baseline	MOBS-GLEAM_COVID	UT-Mobility	JHU_IDD-CovidSP	USACE-ERDC_SEIR		

Pairwise comparisons implied by relative WIS skills

▲□▶▲圖▶▲≣▶▲≣▶ = ● のへで

Checking agreement between $\tilde{\theta}_{ij}$ and $\tilde{\theta}_i/\tilde{\theta}_j$

1.00	0.91	0.78	0.62	0.58	0.60	0.60	0.51	0.47	0.34	0.43	0.32	COVIDhub-ensemble	4
1.10	1.00	0.91	0.69	0.64	0.66	0.72	0.58	0.55	0.43			YYG-ParamSearch	_
1.29	1.10	1.00	0.82	0.77	0.72	0.82	0.64	0.61	0.47			UMass-MechBayes	3
1.62	1.46	1.23	1.00	0.94	0.96	0.92	0.81	0.73	0.52	0.73		LANL-GrowthRate	_
1.73	1.55	1.31	1.06	1.00	1.03	0.98	0.86	0.77	0.55	0.77	0.52	UCLA-SuEIR	2
1.66	1.51	1.38	1.04	0.97	1.00	0.93	1.01	0.86	0.62	0.67	0.51	GT-DeepCOVID	
1.68	1.39	1.22	1.09	1.02	1.07	1.00	0.81	0.74	0.59	0.57	0.52	IHME-CurveFit	1
1.96	1.72	1.56	1.23	1.16	0.99	1.23	1.00	0.94	0.74	0.70	0.63	COVIDhub-baseline	
2.15	1.82	1.65	1.38	1.29	1.17	1.34	1.06	1.00	0.78	0.74	0.68	MOBS-GLEAM_COVID	- 0
2.92	2.34		1.92	1.80	1.60	1.69	1.34	1.29	1.00	0.95	0.93	UT-Mobility	
2.31	2.45	2.23	1.38	1.30	1.50	1.76	1.43	1.35	1.05	1.00	0.78	JHU_IDD-CovidSP	
3.15	2.81	2.35	2.03	1.91	1.96	1.93	1.58	1.47	1.08	1.28	1.00	USACE-ERDC_SEIR	
COVIDhub-ensemble	YYG-ParamSearch	UMass-MechBayes	LANL-GrowthRate	UCLA-SuEIR	GT-DeepCOVID	IHME-CurveFit	COVIDhub-baseline	MOBS-GLEAM_COVID	UT-Mobility	JHU_IDD-CovidSP	USACE-ERDC_SEIR	-	

◆□▶ ◆□▶ ◆三▶ ◆三▶ 三三 のへぐ

Direct pairwise comparison: WIS ratios

Stratified relative WIS skill



Relative WIS skill by forecast date

forecast date

◆□▶ ◆□▶ ◆三▶ ◆三▶ 三三 のへで

Stratified relative WIS skill



Relative WIS skill by target end date

target end date

◆□▶ ◆□▶ ◆ 臣▶ ◆ 臣▶ 三臣 - ∽ � � �

Stratified relative WIS skill





target

イロト 不得 トイヨト イヨト

= √Q ()

Testing for difference in forecast performance

To test for a difference in mean WIS between models i and j:

• test statistic: $\tilde{\theta}_{ij}$ (= ratio of mean WIS of *i* and *j* for A_{ij})

generation of reference distribution:

blockwise permutation of pairs of scores between i and j

to account for dependence between locations and horizons: all forecasts made at one forecast date treated as one block

p-value is given by the proportion of sampled WIS ratios exceeding the observed ratio

Results of performance tests

1.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	COVIDhub-ensemble
0.01	1.00	0.27				0.01	0.00					YYG-ParamSearch
0.00	0.27	1.00	0.05	0.01		0.13	0.00			0.00		UMass-MechBayes
0.00		0.05	1.00	0.25	0.43	0.47	0.07					model1
0.00			0.25	1.00	0.64	0.91	0.03		0.07	0.01		model2
0.00			0.43	0.64	1.00	0.80	0.53	0.14	0.10	0.00	0.02	model3
0.00	0.01	0.13	0.47	0.91	0.80	1.00	0.10	0.04	0.09	0.01		model4
0.00			0.07	0.03	0.53	0.10	1.00	0.84	0.37	0.02	0.04	COVIDhub-baseline
0.00					0.14	0.04	0.84	1.00	0.17	0.08	0.03	model5
0.00				0.07	0.10	0.09	0.37	0.17	1.00	0.75	0.62	model6
0.00							0.02	0.08	0.75	1.00	0.26	model7
0.00					0.02		0.04	0.03	0.62	0.26	1.00	model8
COVIDhub-ensemble	YYG-ParamSearch	UMass-MechBayes	model1	model2	model3	model4	COVIDhub-baseline	model5	model6	model7	model8	

Permutation tests (applied to mean WIS per forecast date)

References

- Johannes Bracher, Evan L. Ray, Tilmann Gneiting, Nicholas G. Reich (2020): Evaluating epidemic forecasts in an interval format. https://arxiv.org/abs/2005.12881
- Tilmann Gneiting and Adrian Raftery (2007): Strictly Proper Scoring Rules, Prediction, and Estimation. JASA, DOI 10.1198/016214506000001437