Evaluating Probabilistic Forecasts for Influenza Using the Probability Integral Transform (PIT)

Rebecca Silva (1,2), Casey Gibson(1), Nicholas Reich(1)

(1) UMASS-Amherst Department of Biostatistics and Epidemiology, Amherst, MA (2) Amherst College, Amherst, MA

Abstract

The Probability Integral Transform (PIT) is a useful metric to assess the behavior and validity of a predicted distribution. This summer the PIT method was used to evaluate five ensemble forecast models for seasonal influenza in the United States. Although none of the models could be validated by PIT, the visual assessment of each model, as a whole and separated by specific targets, give insight into where observations tend to fall within their respective probabilistic distributions. A common trend throughout the histograms of a right skew indicates that fewer observations fall in the right tail of their respective distributions than expected. From a public health standpoint, it could be beneficial for these forecast models to hedge against large outbreaks by placing more weight on observed values that fall above the median of each distribution.

Introduction

Since the 2013-2014 influenza season, the CDC has run a "Forecast the Influenza Season Collaborative Challeng"(FluSight competition) where teams interested in forecasting seasonal influenza in the United States submit real-time models. These models predict weighted influenza-like-illness (wILI) targets and are evaluated based on a log score metric at the end of a season. In 2017, a team of influenza forecasters formed the FluSight Network, a collaborative multi-institutional and multi-disciplinary group whose goal was to build a real-time ensemble model using a diverse array of 21 models. Of the five ensemble models, the four that used past performance to determine the weights of the component models performed better than any other forecast model by itself. The FluSight Network used a log score metric to select the best performing ensemble model, which they then submitted to the CDC for the 2017/2018 season [1]. This project examines another commonly known metric that can be used to evaluate ensemble models: Probability Integral Transform method.

PIT Method

The PIT says that data from any continuous distribution can be transformed into having a standard uniform distribution using its cumulative distribution function (CDF). We use the method as follows:

- Plug data into its own CDF, to obtain an output, known as the PIT value
- Create a histogram plot of the PIT values
- Visually assess the shape of the histogram

Theoretically, the shape of the histogram should be that of a standard uniform distribution if the predictive distribution matches its true, unobservable distribution.

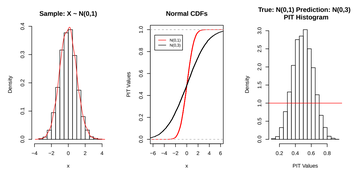


Figure 1: Normal Example

Ensemble Models

Model	No. of weights	description
Equal weights (EW)	1	Every model gets same weight.
Constant weights (CW)	21	Every model gets a single weight, not necessarily the same.
Target-type-based weights (TTW)	42	Two sets of weights, one for seasonal targets and one for weekly wILI targets.
Target-based weights (TW)	147	Seven sets of weights, one for each target separately.
Target-and-region-based weights (TRW)	1,617	Target-based weights estimated separately for each region.

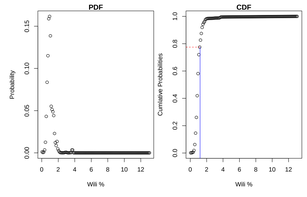


Figure 2: Constant Weight, EW01-2011, 1 wk ahead, Region 1.

In the ensemble models, the predictive distribution is different based on the season, evaluation week, region, and target; there is one observed value for each distribution and therefore one PIT value per distribution. The five targets we studied were the one, two, three, and four weeks ahead wILI, and the peak wILI during the given season.

Results

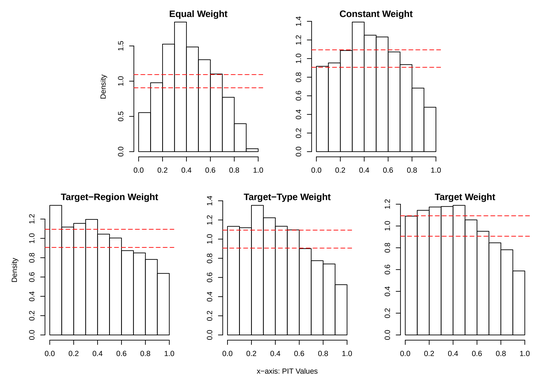


Figure 3: Ensemble PIT Histograms

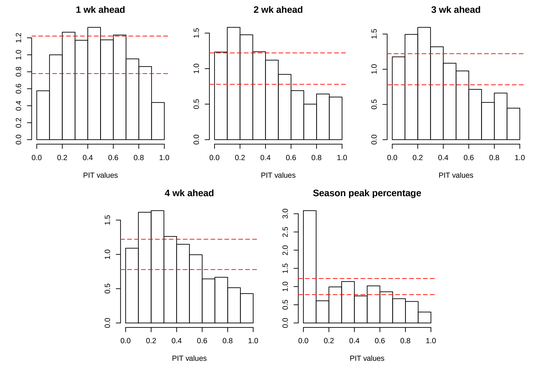
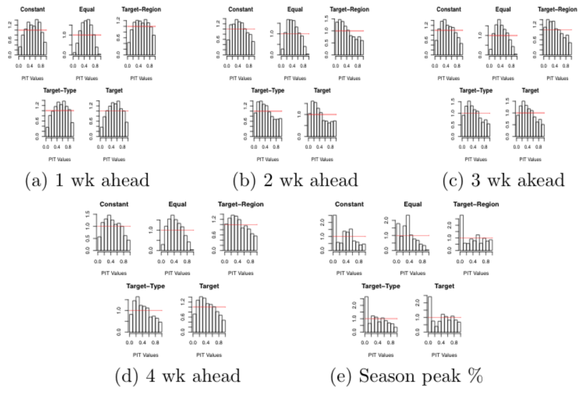


Figure 4: TTW Model by Target



(a) 1 wk ahead (b) 2 wk ahead (c) 3 wk akead

(d) 4 wk ahead (e) Season peak %

Figure 5: Each Target by Ensemble Models

Conclusions

- Predictive distributions generally suggest a greater likelihood of high wILI values than what is observed. Fewer values in the tail could suggest a misspecification of bias in the ensemble models where the models tend to overestimate the target values.
- The new shapes seen in the 1-week-ahead and Season peak percentage histograms suggest that the ensemble models' performance differ between targets.
- Shapes vary more between targets than between the ensemble models themselves. It may be possible to improve models by investigating predictors that factor into forecasts differently depending on the target.
- Hypotheses for why we are not seeing a uniform-shape histogram and why this right skew may not indicate our models are invalid.
  - large number of PIT values for each model (10,500)
  - from a public health standpoint, overestimation of wILI values may be beneficial for hedging against large outbreaks

Additional Work

Study the 2017/2018 season where the ensemble models did not perform as well. Histograms show a different pattern; instead of a low density in high PIT values, there are higher density in the high PIT values.

References

[1] Nicholas G Reich, Logan Brooks, and Sasikaran Kandula.
Forecasting influenza in the u.s. with a collaborative ensemble from the flusight network, Nov 2017.

[2] Tilmann Gneiting, Fadoua Balabdaoui, and Adrian E. Raftery.
Probabilistic forecasts, calibration and sharpness.
Journal of the Royal Statistical Society, Jan 2005.

[3] Francis Diebold, Todd Gunther, and Anthony Tay.
Evaluating density forecasts with applications to financial risk management.
International Economic Review, 1997.

[4] Thomas M. Hamill.
Interpretation of rank histograms for verifying ensemble forecasts.
Monthly Weather Review, 129(3):550–560, 2001.

Acknowledgements

Thank you to Nicholas Reich, Casey Gibson and the Department of Biostatistics and Epidemiology (for all the support) and also to Amherst College for funding my work.