Delphi-Stat Forecasting Methodology

Logan C. Brooks[1] David C. Farrow[1] Shannon Gallagher[2] Sangwon Hyun[2]
Ryan J. Tibshirani[2] Roni Rosenfeld[1]

[1] School of Computer Science [2] Department of Statistics
Carnegie Mellon University,
Pittsburgh, Pennsylvania,
United States of America

Stat Methodology

The Delphi-Stat system is an ensemble of several baselines and statistical forecasting methods. Its forecasts are a linear combination of the forecasts of these individual systems, with a separate set of coefficients determined for each epi week, geographical area (nation + 10 HHS regions), metric (MAE or log score), and target. The methods are outlined below. Note that the term "past epiweeks" refers to epi weeks 21 up to the last epi week in the 2015–2016 season for which there is published weighted ILI data, in any season (not just 2015–2016); "future epiweeks" is used in a similar fashion.

(Nonnegative coefficients summing to 1 are calculated for point predictions using constrained LAD regression (implemented using the linear programming package `lpSolve` [1]), and for distributional predictions with the degenerate EM algorithm [4].)

Empirical prior: ignores all data from the current season, and considers each training season — 2003–2004 to 2014–2015, excluding the pandemic — as equally likely to reoccur.

Pinned baseline: uses the available observations for the current season for previous epi-weeks; for future epi weeks, each training curve is considered equally likely to reoccur.

Basis regression:

1. Aligns training curves with the current season by shifting in time and scaling weighted ILI values until the maximum of each training curve in past epiweeks is the same as that of the current season. (Scaling is performed only above the CDC baseline; if a curve is entirely below the CDC baseline, it is not scaled at all.)
2. Fits a smooth curve to the observed data in past epiweeks and the mean of the aligned training curves in future epiweeks. (The smooth curve is a spline: specifically, a linear combination of B-splines selected with elastic net using the `glmnet` package [3], with a trade-off penalty between the importance of matching past and future epiweeks.)
3. Uses observations from the current season in past epiweeks; considers this single curve as the only possibility for future weeks.

Basis regression with noise:

1. Generates the spline curve above.
2. Considers the spline as estimating the change in weighted ILI from one week to the next; for each epi week, estimates the distribution of errors at that epi week using the training curves. (Distributions are estimated using weighted kernel density estimation: when adding noise to a simulated 2015–2016 curve at some future epiweek, training curves that more closely resemble the simulated curve in previous epiweeks contribute more to the result.)
3. Generates many simulated 2015–2016 curves by taking the observations from the current season so far, and at each week, adding the estimated change from the spline curve, then drawing a value from the estimated error distribution.

Time-parameterized weighted kernel density estimation:

1. Follows the same process as the basis regression with noise; however, it directly estimates the distribution of changes in weighted ILI values, rather than the corresponding distribution of errors in the spline estimate.

Empirical Bayes: We use the procedure described in this document [2], with a few modifications: a smoothed (trend-filtered [5]) curve is never paired with a noise estimate from another smoothed curve, scaling and shifting is performed only in small amounts resulting in "local" transformations, an additional component is added to the likelihood to encourage reasonable predictions at future weeks (by penalizing simulated curves if they deviate too much from all of the training curves), and incorporating a random inflation in the noise parameter to prevent forecast "overconfidence".

Uniform prior: Considers each cell in the spreadsheet to be equally likely. (This component only produces distributional forecasts.) Additional weight is added to this component after the coefficients for each method are determined via CV to prevent any 0 or near-0 probability forecasts.

References

[1] Michel Berkelaar and others. *lpSolve: Interface to 'Lp_solve' v. 5.5 to Solve Linear/Integer Programs*, 2015. R package version 5.6.11.

[2] Logan C. Brooks, David C. Farrow, Sangwon Hyun, Ryan J. Tibshirani, and Roni Rosenfeld. Flexible modeling of epidemics with an empirical bayes framework. *PLoS Comput Biol*, 11(8):e1004382, 08 2015.

[3] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2010.

[4] Roni Rosenfeld. The "degenerate EM" algorithm for finding optimal linear interpolation coefficients λ_i. `http://www.cs.cmu.edu/~roni/11761/Presentations/degenerateEM.pdf`. Accessed: 2015-07-20.

[5] Ryan J Tibshirani et al. Adaptive piecewise polynomial estimation via trend filtering. *The Annals of Statistics*, 42(1):285–323, 2014.