Bayesian Cubic B-Splines to Forecast the 2017-2018 Influenza season

Joseph L. Servadio

Among the years of observation in the defined HHS regions, weekly percentages of Influenza-like illness (ILI) cases show similar longitudinal trends across regions and years. In the majority of influenza seasons among the regions, ILI begins at a low level, then reaches an elevated level for the 5th through 25th weeks of observation. During this time, ILI percentage increases to a peak level and then decreases. Following this period of elevated ILI percentage, ILI percentage typically remains at a low level for the remainder of the season.

There exists opportunity to take advantage of these similarities by using the trends of previous years to inform the trends of current years throughout all regions. Bayesian Hierarchical Models are designed to use data from multiple groups to allow trends among different groups, though distinct, to inform each other. This method can be used to allow all region-years to be considered a distinct, but related.

This model for forecasting the 2017-2018 influenza season, titled "UMNSpl," uses a Bayesian Hierarchical model with cubic B-Splines to use previous seasons' information to inform the current season. B-splines [1] offer flexibility in modeling, removing strong assumptions regarding the shape of the longitudinal trend of ILI percent. The model parameters are believed to belong to a common distribution among all regions and seasons.

The B-spline basis was chosen to be cubic following common convention and allowing two inflection points between knots. Knots were selected a priori to be placed every seven weeks between weeks 7 and 42. These knot points were selected to assure that there were adequate data between knots while allowing adequate flexibility in trends. This will allow late-season increases to occur, which has been seen in previous seasons. The spline bases for time were the only predictors in the model, but other parameters were included.

Based on outbreak thresholds, an indicator variable was created to identify outbreak weeks. Outbreak weeks are defined as week where ILI percentage is above the threshold specified by the CDC and where the ILI percentage of the previous two weeks are both above the threshold. In weeks designated as outbreak weeks, the estimate for expected percentage of ILI is multiplied by 1.2. This value was found by testing values between 1 and 2 by forecasting removed data from the 2016-2017 influenza season. The value of 1.2 was found to lead to the most accurate forecasts of the 2016-2017 season across all regions. The multiplier was included into the model because the model predicts ILI percentages that are lower than observed during the part of the season with increased observed ILI percentage.

The model is specified by:

$$\begin{aligned}
I.\quad & Y_{h,y,w} \sim N(\mu_{h,y,w}, \sigma^2) \\
II.\quad & \mu_{h,y,w} = [\alpha_{h,y} + \sum_{i=1}^{9} \beta_{i,h,y} X_{i,w}] \cdot [1 + 0.2\mathbb{1}(out_{h,y,w} = 1)] \\
III.\quad & out_{h,y,1} = 0 \ \forall h, y, \ out_{h,y,2} = 0 \ \forall h, y, \\
& out_{h,y,w} = \mathbb{1}(Y_{h,y,w} > cut_{h,y}, \ Y_{h,y,w-1} > cut_{h,y}, \ Y_{h,y,w-2} > cut_{h,y}), \ 3 \leq w \leq 52 \\
& \alpha_{h,y} \sim N(\mu_\alpha, \sigma_\alpha^2), \ \beta_{i,h,y} \sim N(\mu_{\beta_i}, \sigma_{\beta_i}^2) \\
IV.\quad & \mu_\alpha \sim N(0, 100^2), \ \mu_{\beta_i} \sim N(0, 100^2), \\
& \sigma_\alpha \sim U(0, 1000), \ \sigma_{\beta_i} \ U(0, 1000) \ \forall i, \ \sigma \sim U(0, 1000)
\end{aligned}$$

where h denotes the HHS region (or nation), y denotes the year, w denotes the week number, and i denotes the spline bases. The function $\mathbb{1}(\cdot)$ is the indicator function, returning a value of 1 if the statement is true and 0 if the statement is false. The variable Y represents ILI percentage, X represents the B-spline bases, *out* is an indicator variable for presence of an outbreak, and *cut* is the matrix of cutoffs for an outbreak.

The model is computed with the 'rjags' [2] package in R version 3.4.2 [3]. The package connects with JAGS (Just Another Gibbs Sampler) version 4.3.0 [4] to run Markov Chain Monte Carlo (MCMC) simulations. After compiling the model, the model runs 10,000 burn-in simulations and then collects 1,000 samples to calculate estimated forecast targets and probabilities. Samples are generated for all unknown values, including missing ILI data and model parameters.

This model assumes that all region-years are independent of each other, and that their model parameters are independent realizations from a common distribution. Also assumed are 52-week years. Years with 53 weeks were truncated to 52 weeks to prevent years with 52 weeks having a 53rd week imputed, which would introduce more extraneous unknown values into the model. An additional assumption is that the national ILI values can be treated as another independent group, essentially creating 11 regions.

A conceptual strength of this model is the use of information in other regions, both in the current year and in previous years, to inform forecasts. Observed data provide information regarding the trends that are likely to be observed during the current influenza season. While this is a conceptual strength, it also provides a weakness. If a region during the current year does not follow trends observed in other regions and years, these deviations are less likely to be captured in the model. The combination of the different regions and years. The model will likely perform well for regions that have an influenza season that closely follows trends observed in previous years, but will not perform well for regions that have notably different trends.

Creation of this model is motivated by interest in the ability of previous trends alone to predict the trends of an ongoing influenza season. The results from this model will inform how well forecasts can be generated using no other sources of data. Other factors can be hypothesized to influence influenza seasons such as climate variables or human behaviors; the results of a model that does not incorporate any of these factors can provide insight into the potential usefulness of other factors.

References

[1] Patrikalakis N, Maekawa T, Cho W. B-spline curves and surfaces. MIT Hyperbook. Dec 2009. Available at web.mit.edu/hyperbook/Patrikalakis-Maekawa-Cho/node-15.html. Last Accessed 2 Nov 2017.

[2] Plummer M. rjags: Bayesian Graphical Models using MCMC. 2016. Version 4-6. https://CRAN.R-project.org/package=rjags

[3] R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

[4] Plummer M. JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. 2003.