Dean Gladish and Nick Reich
Final Report for the Reich Lab
July 20, 2018

The Kolmogorov-Smirnov Test – A First Pass

Quantifying Differences Between Models used in the Ensemble.

Previously at the Reich Lab, I had looked at a variety of topics including the way in which the weighting was done (the EM algorithm, which tends to find local maximums in order to determine the optimal weights), the way in which missing data was handled (by substituting data points with similar characteristics in some cases), and how models like SARIMA look at moving averages in order to approximate what is going to happen in the future. In this report, I would like to introduce a cursory look at the Kolmogorov-Smirnov test – this is a test that has been frequently used to compare models in part because of its strength over the χ^2 goodness-of-fit test. I hope that this document provides a framework for finding a single metric for quantifying model similarities. Although, as Steve said, there is no single metric for comparing models (we could use the four moments for instance), I would like to present some graphs that I have made and conclusions that can be made about them based on how the models work.

In short, this is how the test works conventionally: we are given an independent and identically distributed sample of data from an unknown distribution $\mathbf{P}$ and a proposed distribution $\mathbf{P}_0$.

What the Kolmogorov-Smirnov test does is tell us which of the following hypotheses,

$$H_0 : \mathbf{P} = \mathbf{P}_0,\ H_a : \mathbf{P} \neq \mathbf{P}_0$$

are true. This test is stronger precisely because it does not create discrete groups based off the sample. Rather, it considers each data point on a continuous distribution[1]. The comparisons are done entirely using the Kolmogorov-Smirnov test statistic. What is the K-S test statistic?

$$D_x = sup_n[\,|F_n(x)\ -\ F_0(x)|\,]$$

The test statistic is the supremum, over the domain, of the vertical difference between a data point and the distribution that we are proposing. This difference is calculated between the empirical CDF (empirical because it comes from the data) and the theoretical CDF (say, the CDF of the normal distribution)[2]. Essentially, what we want to do in this document is outline the way in which I implemented the Kolmogorov-Smirnov test and describe ways in which it could be improved, possibly by myself or future interns as we delve into this peculiar realm of statistics.

The Kolmogorov-Smirnov test statistic is then compared to a table of threshold values. Below the threshold, the distributions are considered to be the same with X amount of certainty (where X is the threshold percentile). Above the threshold, the distributions can be considered to be different. This is the way in which the test works. These thresholds are determined entirely by the size of the sample from which our data comes.

In this case, my project was originally about quantifying similarities between the predictive models. My focus was on the models implemented and developed by the Reich Lab such as KDE, KCDE, and SARIMA. Therefore, I looked at the data from week 43 of 2017 to week 18 of 2018 and calculated Kolmogorov-Smirnov test statistics. In the early stages of my project there was a crucial mistake that had to be corrected; I did not understand that the

[1] https://ocw.mit.edu/courses/mathematics/18-443-statistics-for-applications-fall-2006/lecture-notes/lecture14.pdf

[2] https://newonlinecourses.science.psu.edu/stat414/node/322/

Kolmogorov-Smirnov test statistic was supposed to be calculated between the cumulative distribution functions. Believing that it was the maximum difference between the probability mass and probability density functions, I implemented it as the maximum absolute magnitude of the difference between the PMFs rather than the CDFs of the two models respectively.

The way in which I implemented this test was quite simple – within the .csv files of the real-time-component-models folder available on GitHub, there are point predictions (also included in my analysis) for the actual %ILI (each year, the CDC collects data on reports of influenza-like illness in different US regions). What I looked at was exclusively the region labeled US National (I only looked at the total %ILI in the entire United States). Furthermore my PMFs (and from there, eCDFs) were determined by the probabilities assigned to each bin; the models (KDE, for example) were faced with four series of bins (one week, two weeks, three weeks, and four weeks ahead). These bins are associated with different outcomes and represented our predictive distributions – they capture the %ILI rounded to a single decimal place. This means that for one week ahead (or four weeks ahead, etc.), the bins – [0, 0.05), [0.05, 0.15), …, [12.95, 13.05), [13.05, ∞) are assigned probabilities which sum to 1[3].

I essentially calculated the eCDFs by summing all the probabilities for each bin up to and including the current bin. In this way I was able to generate a vector with 131 values, each value of which corresponded to the y-value of the empirical CDF for 131 x-values defined by the first bin ([0, 0.05)), the second bin ([0.05, 0.15)), and so on. From there, I was able to use the abs and max functions in R to essentially derive what are Kolmogorov-Smirnov test statistics.

[3] https://arxiv.org/pdf/1703.10936.pdf

Let's say I was looking at one week ahead. Since there are 28 weeks (28 .csv files) between week 43 of 2017 and week 18 of 2018, I am able to derive 28 Kolmogorov-Smirnov test statistics derived from the eCDFs (which are derived from the PMFs of the predictive distributions' probabilities assigned to all bins). Then, I can do the same for two weeks ahead, three weeks ahead, and four weeks ahead. In this way, I am able to create four graphs of Kolmogorov-Smirnov test statistics between two models of my choice (in this case, I mostly focused on comparing Reich Lab's models). The results are as follows:

RESULTS:

As Steve said, the KDE model uses Kernel Density Estimation to find a distribution for each of the target number of weeks ahead. 2017 being the first year for the ensemble model's prediction-making (it was trained on prior years), it was quite difficult to locate data for previous years so my analysis focuses entirely on a 28-week window. Right now, I'm going to focus on the results that I have between KDE and KCDE.

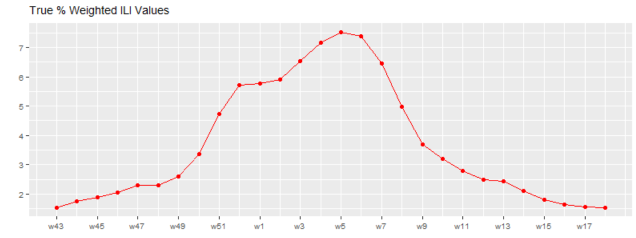


Figure 1. The first graph represents the actual percent ILI recorded by the CDC[4]. The 43rd week of 2017 would be October 23 to 29. The 18th week of 2018 would be April 30 to May 6.

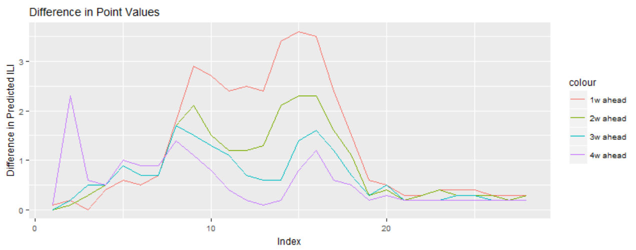


Figure 2. Each model makes a point-value prediction for %ILI for X weeks ahead. Index 1 is at week 43 of 2017. Index 28 is at week 18 of 2018.

[4] https://gis.cdc.gov/grasp/fluview/fluportaldashboard.html

Steve also explained that the KCDE model uses weighted Gaussian kernels (previous %ILI values) to determine the predicted %ILI for week X. Kernel conditional density estimation was used to obtain separate predictive distributions for flu incidence in each week of the flu season. The KCDE model bases its predictions on recent observations of incidence and the current week of the season. There are many ways to interpret the second graph. One thing to note is that as both models extrapolate further into the future (one week ahead to four weeks ahead), their point-value predictions become more similar. Another thing to note is that nearing the end of 2017, there is a sudden rise indicating that the KCDE model makes a very high prediction for %ILI at 4 weeks ahead. More broadly, the differences decrease substantially right as 2017 ends, which means that the KCDE model behaves more similarly to the KDE model at this time of the year. Another interesting thing to note is that right as the next year begins, their differences increase substantially for all weeks ahead. This indicates that the start of the next year could play a greater role in the KCDE model's weighting system for predicting the actual %ILI. Furthermore, larger values of actual %ILI seem to be correlated with larger differences in general.

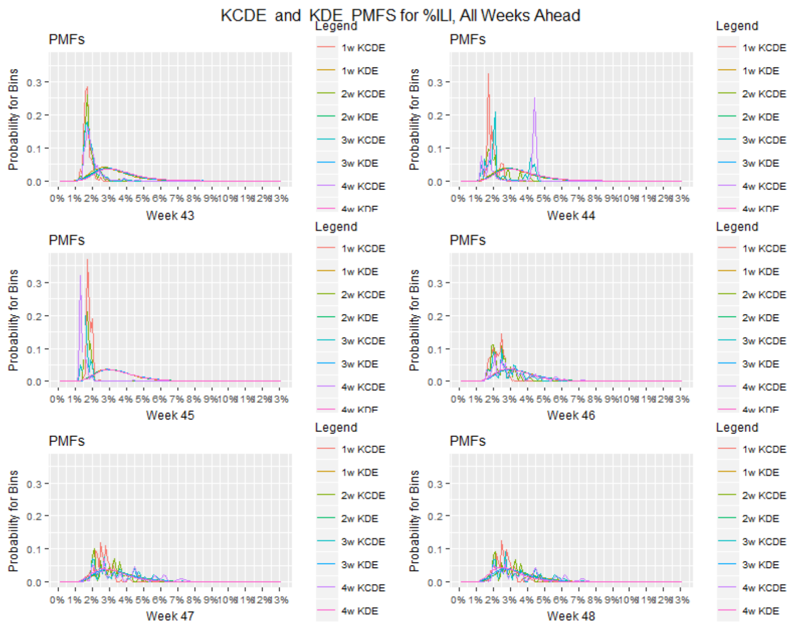


Figure 3. These are the probability mass functions for each model respectively. They are given for weeks 43 through 48, 2017. 1w KCDE, for example, describes the PMF for the KCDE model, one week ahead.

As we can see, the KCDE model has a substantial spike for the end of 2017. As shown by the previous graph of point-value predictions, this graph basically shows that as the beginning of the next year approaches, the KCDE model behaves more similarly to the KDE. Why is this the case? As I said in my conversation with Steve earlier, the KCDE model will put more weight on the seasonal average when it encounters less certainty. Since the seasonal average is covered by the KDE model, it makes sense that the models would have more similar behavior when the KCDE model perceives less certainty.

As we know, KDE makes different predictions for say, week 50 when the time-zero is week 43 or 44, etc. But these predictions are very close.

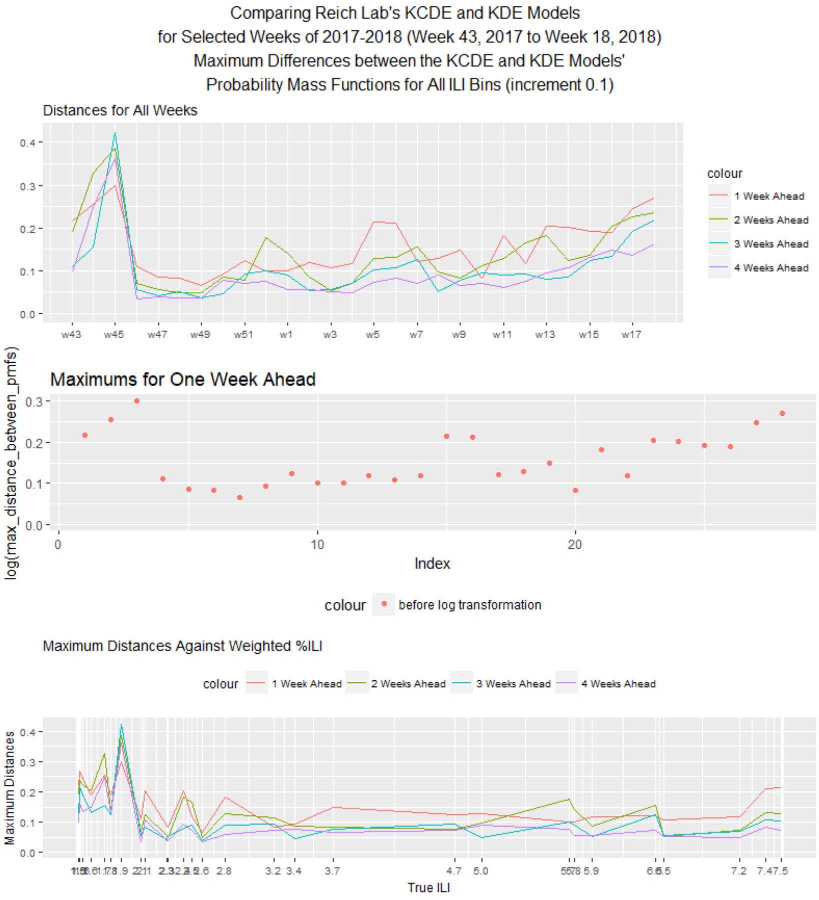


Figure 4. For these plots, I looked at each individual bin and found the magnitude of the difference between the probabilities that each model (KCDE and KDE) assigned to it. For the I asked the question, over all bins what is the maximum difference of this type? The second plot portrays the isolated maximum difference for one week ahead. Because it was not useful for analysis purposes, the window excludes those points that are log-transformed (the log-transformed y-variable is likely to be negative and is not displayed in the plot although it is included in the actual graph). The third plot describes these maximum differences not against the time of year (week) but against the CDC's weighted %ILI.

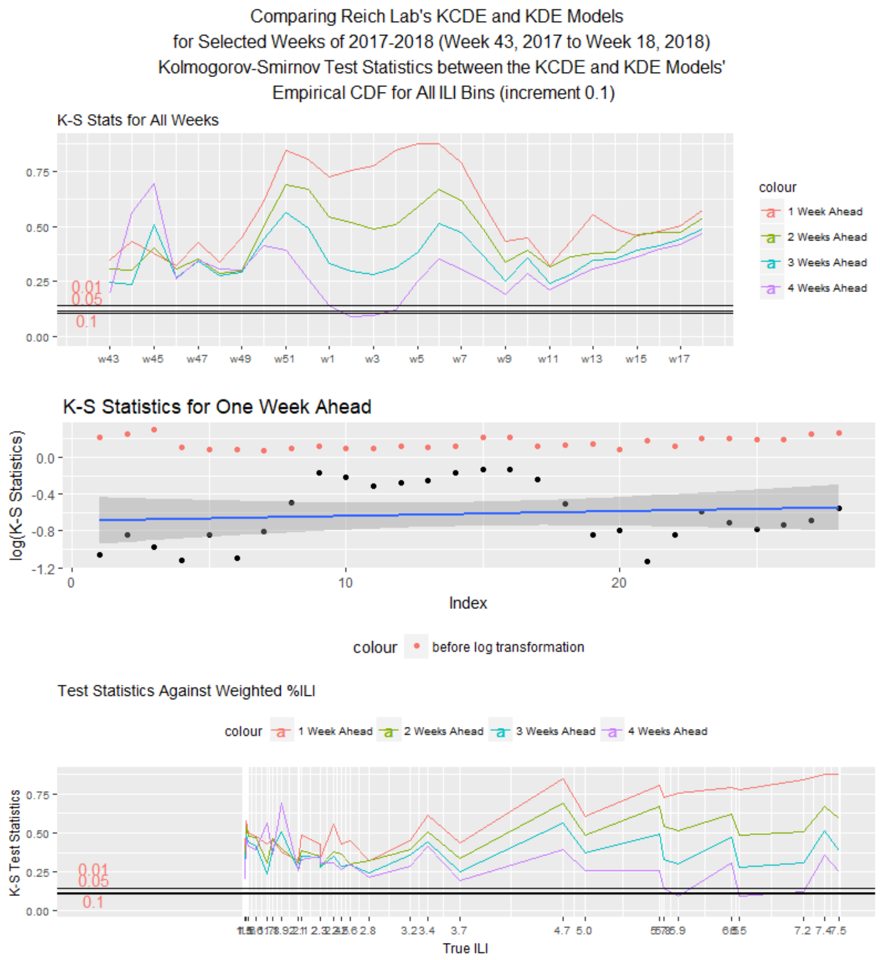
As with the actual Kolmogorov-Smirnov test statistics, looking at the maximum difference across the probabilities assigned to each bin yields some insightful information. As before, the models KDE and KCDE become much more similar as they go from predicting 1 week ahead to 4 weeks ahead. The KCDE and KDE models are constructed on the same basis of kernel density estimation. This means that, as Steve and I have concluded, Gaussian distributions surrounding different kernels from past years' information about percent %ILI at each week of the year are used in order to estimate the density of the occurrences of percent %ILIs at particular times of the year. This density is continuous and thus can be used to estimate the probabilities to assign to %ILI for the upcoming flu season. Of course, there is one crucial difference that is described by the "conditional" aspect of the KCDE model. This conditionality means that the predictive distributions are not only affected by the density of %ILIs as determined by Gaussian distributions surrounding kernels. They are conditional on recent observations of flu incidence and the current week of the season[5]. Now, what these graphs show (and the following graphs also confirm) is that the KCDE model recognizes that it can rely less on recent observations of flu incidence when it is making predictions further into the future from the time-zero. The current week of the season, of course, does not change so we can make this inference.

Figure 5. The first plot displayed is truly essential as it describes the actual Kolmogorov-Smirnov test statistics. The thresholds for significance are displayed as well[6]. Except for the period between week 1 and week 3, it is evident that only 0.01 or 1% of the time (given some data) will the KCDE and KDE models reach such high levels of difference if they were in fact the same model. Essentially, if a Kolmogorov-Smirnov test statistic is above the threshold then the proportion of times that the K-S test statistic would be at such a value assuming that the models are the same is the threshold value.

The rest of the plots are quite similar. The second plot displays the K-S statistics for one week ahead and provides a log transformation on the y-axis in order to determine whether the K-S test statistics happen to follow an exponential distribution (if they did then the log-transformed points would be linear). The third plot displays the test statistics against the weighted %ILI in

[5] https://arxiv.org/pdf/1703.10936.pdf, 5.

[6] http://www.real-statistics.com/statistics-tables/kolmogorov-smirnov-table/



increasing order. As we can see, the KCDE and KDE models become quite similar when they are predicting for four weeks ahead.

Discussion:

All plots were made using ggplot2 and gridExtra. All of the source code is available in the file that I have included. I chose the Kolmogorov-Smirnov test due to its simplicity and ease of implementation. There are many directions one could go from here. You could use Kuiper's test to find which models are the best fit for the data. You could also compare mean, variance, skewness, and kurtosis between the models[7]. I would look at how the models vary from each other from bin to bin rather than at once. Other things to look at would be how much models tend to increase or decrease with each other. The Kruskal-Wallis One-Way Analysis of Variance might also be useful[8].

[7] Stephen Lauer

[8] https://en.wikipedia.org/wiki/Kruskal%E2%80%93Wallis_one-way_analysis_of_variance

Finally, for cases such as KCDE versus SARIMA with seasonal differencing we might make different conclusions. In that case, the differences between models seem to increase as we extrapolate further into the future. If we're basing our forecasts on the differences between the predictions made in the past between two points 52 weeks apart, then we could possibly conclude that the SARIMA model regresses less to the seasonal average than the KCDE model does. While the K-S statistics increase as the SARIMA and KCDE models predict further into the future, other measures such as the maximum difference between the probability mass functions decrease as they predict further into the future.

I consider the K-S test statistics to be a more reliable measure of models' differences. For example, a specific model, which is quite similar to another model otherwise, might make an incredibly large prediction for a specific bin and thus skew the maximum difference value. On the other hand, the K-S test statistic captures the extent of the deviations in probabilities assigned between models that cumulatively develop over time. The Kolmogorov-Smirnov test is less affected by temporary jumps in the probabilities assigned to certain bins as was displayed by the KCDE model in the first three weeks (43, 44, 45) of 2017 in my PMF graphs. Although the Kolmogorov-Smirnov test statistic is only a single metric of model similarities and does not address some of the concerns regarding misspecification raised by Noceti et al.[9], I believe that this project should serve as a good starting point for comparing models and inferring their characteristics.

[9] https://vdocuments.mx/an-evaluation-of-tests-of-distributional-forecasts.html