Local Regression Algorithm for Non-Constant Reporting Rates in Epidemic Dynamics

Kaitlyn Stocker

5/11/2017

Introduction:

In my previous work on susceptible reconstruction, I made a crucial assumption about the rate at which infected individuals are reported: I assumed that the reporting rate was constant across time. Under the assumption of constant reporting rate, reconstructing the susceptible dynamics can be accomplished via global linear regression. This assumption simplified the process of reconstructing susceptible and infected dynamics, but it rarely holds true in the world of real data.

As an example, let's look at measles data from pre-vaccination era New York City, from 1920-1940 (Willem G. van Panhuis 2013). In figure 1A is a graph of the reconstructed susceptible dynamics, obtained as previously described using global linear regression. It is evident from figure 1A that the Z dynamics suffer from local shifts away from the mean of zero. This indicates that the previously held assumption of constant reporting rate, ρ, has been violated.

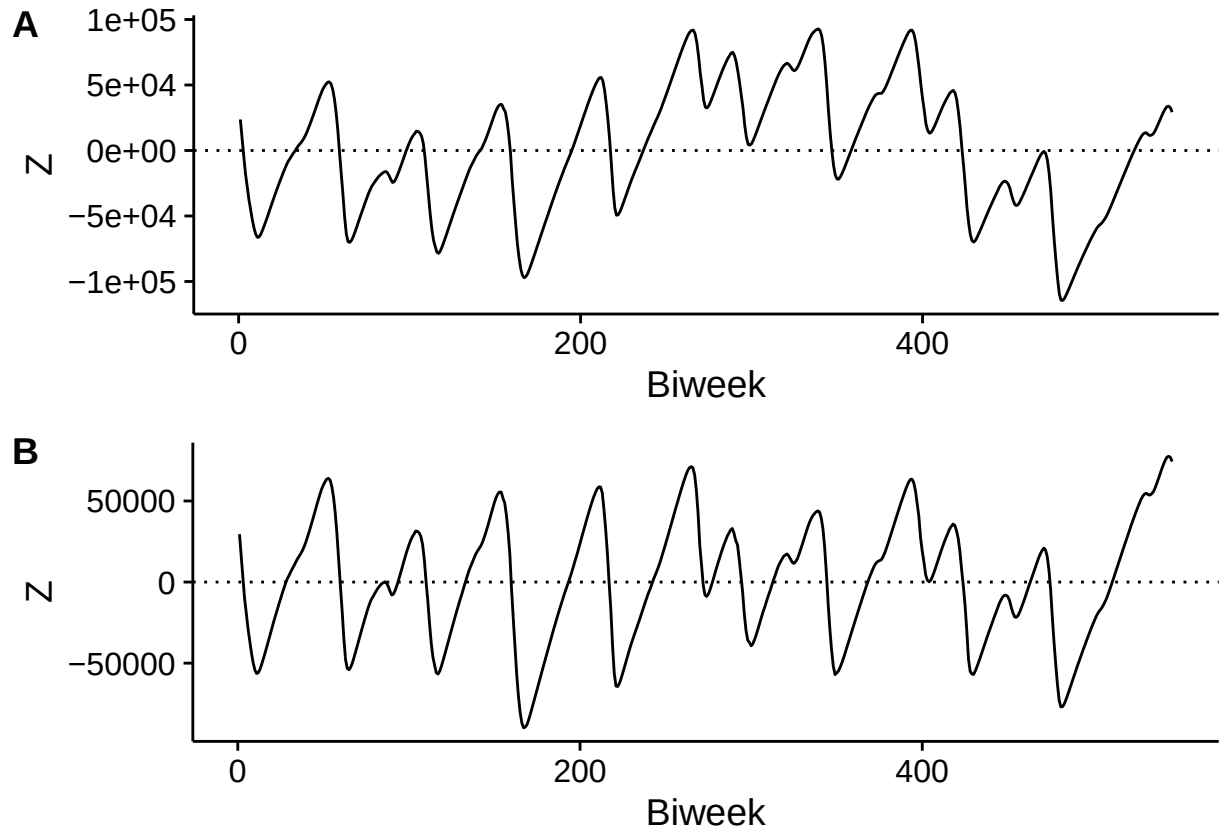

Figure 1: Comparison of reconstructed Z dynamics from the New York data. Plot (A) shows Z dynamics obtained via global linear regression. Plot (B) shows Z dynamics obtained via local linear regression.

When local shifts in the mean of the Z dynamics are observed, such as those in Figure 1A, it indicates

a nonconstant reporting rate. Following the work of Finkenstadt and Grenfell (FinkenstaÈdt and Grenfell 2000), I addressed the issue of time-varying reporting rate by performing local linear regression with gaussian smoothers. To give an overview before I break down the process in detail, I split the data into overlapping chunks (or neighborhoods) centered around each $\{X_t, ..., X_T\}$, then assigned a gaussian weight to each observation in each neighborhood. Then I performed T weighted linear regressions using the previously defined gaussian weights, and derived the value of ρ for each time point by pulling the slope from each of these weighted linear regressions.

For the sake of simplicity as I continue, I will use the following notation: -$\{C_t, ..., C_T\}$ for the reported cases at each biweekly timepoint, t. -$\{B_t, ..., B_T\}$ for the number of births at each timepoint, t.

Note that in this example, I will be performing a local regression in which the cumulative cases is the predictor and the cumulative births is the response. It is possible to perform the regression with cumulative births as the predictor and cumulative cases as the response, as in the work of Bjornstad et. al (2016).

Algorithm

Step 1: Define the Local Regression Function

Begin by defining a function with one parameter, *scale*, which should be a value between 0 and 1. You can also add a *data* parameter if you want to create a general case for use with different sets of epidemic data.

Compute $\sum_{t=1}^{T} C_t$ and $\sum_{t=1}^{T} B_t$, which I will call c.cases and c.births respectively. In this example, c.cases will be the predictor (X) and c.births will be the response (Y).

Step 2: Define the Neighborhoods

Define a variable m to be $T * scale$. Make sure that m is an integer value, as it is the varibale that defines the number of observations in each neighborhood.

Next, we want to select the m closest observations to each X_t. The goal here is to create a matrix, $x.n$, with T rows and m columns, for which each row is a neighborhood of size m. The focal X for the t^{th} row of $x.n$ will be $c.cases_t$. We will also construct a corresponding $y.n$ matrix, with the same dimensions as $x.n$, that will contain corresponding values from $c.births$.

To do this, begin by intializing the $x.n$ and $y.n$ matrices. Additionally, initialize the variables $x.focal$ and $which.diff$ to 0, and create an empty matrix called $diffs$ with T columns and T rows.

Now we can construct a loop that will iterate from 1 to T. Within the loop, begin by setting $x.focal_t$ equal to $c.cases_t$. Now that we have identified the focal x for neighborhood t, define the t^{th} row of the $diffs$ matrix to be the absolute value of the difference between each value in $c.cases$ and $x.focal_t$. Each row of $diffs$ now contains the difference between the focal X_t and each value of $\{X_t, ..., X_T\}$. We want to keep the m closest obvservations to each of our focal X_t's, so we will set $which.diff_t$ to be the m^{th} smallest value of the t^{th} row of $diffs$.

Now we can fill in our $x.n$ by defining the t^{th} row to be every value of $c.cases$ for which the corresponding value of the t^{th} row of $diffs$ is less than or equal to $which.diff_t$. To construct the $y.n$ matrix, do the same for values of $c.births$. Note than since m was rounded to be an integer, you may be one observation short or one observation long in some neighborhoods. I dealt with this by taking the first m values of rows of $x.n$ and $y.n$ that were one observation long, and by adding a zero to the end of rows that were one observation short.

The final product is two matricies, $x.n$ and $y.n$ for which each row, t, corresponds to the neighborhood centered around the t^th value of $c.cases$.

Step 3: Apply Gaussian Weights

The next step in perfroming local linear regression is to apply gaussian weights to each observation for each of your T neighborhoods. The gaussian weight function for the i^{th} observation in the t^{th} neighborhood is as follows:

$$w_i(x.focal_t) = \frac{K(\frac{x.focal_t - x.n_{t,i}}{which.diff_t})}{\sum_{i=1}^{m} K(\frac{x.focal_t - x.n_{t,i}}{which.diff_t})} \tag{1}$$

I applied this weight function to each observation i in each neighborhood t. In other words, I applied the weight function to each of the m columns in each of the T rows of the $x.n$ matrix, and stored the results in a matrix with the same dimensions as $x.n$ called $x.weights$.

Step 4: Complete the Regression

The final step is to perform T weighted linear regressions. To do this, I regressed each row of $y.n_{t,m}$ onto each row of $x.n_{t,m}$, with weights pulled from the corresponding row of $x.weights_{t,m}$.

To get a vector of $\{\rho_t ... \rho_T\}$, I pulled the slopes from each of the T regressions.

To obtain fitted values, I took the fitted value from each regression for the focal X_t of that regression.

Step 5: Complete the Reconstruction

The final step is to reconstruct the infected and susceptible dynamics.

I obtained the reconstructed infected dynamics, $\{I_t ... I_T\}$ by multiplying C_T by ρ_T.

I obtained the reconstructed susceptible dynamcis, $\{Z_t ... Z_T\}$ by finding the residuals of the regression.

At this point, the reconstruction is complete and you can move on to inference.

References

FinkenstaÈdt, BaÈrbel F., and Bryan T. Grenfell. 2000. "Time Series Modelling of Childhood Diseases: A Dynamical Systems Approach." *Applied Statistics* 49 (2): 187–205.

Willem G. van Panhuis, Su Yon Jung, John Grefenstette. 2013. "Contagious Diseases in the United States from 1888 to the Present." *NEJM* 369 (22): 2152–8.