Journal Title
XX(X):1–21
©The Author(s) 0000
Reprints and permission:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/ToBeAssigned
www.sagepub.com/

Infectious disease prediction with kernel conditional density estimation

Evan L. Ray[1], Krzysztof Sakrejda[1], Stephen A. Lauer[1], Michael Johansen[2] and Nicholas G. Reich[1]

Abstract
Abstract

Keywords
copula, infectious disease, kernel conditional density estimation, prediction

Introduction

Accurate prediction of infectious disease incidence is important for public health officials planning disease prevention and control measures such as vector control and increased use of personal protective equipment by medical personnel during periods of high disease incidence (cite ***). Several quantities have emerged as being of particular utility in making these planning decisions (cite ***); in this article we focus on measures of weekly incidence, the timing of the season peak, and incidence in the peak week. Predictive distributions for these quantities are preferred to point predictions because they communicate uncertainty in the predictions and give decision makers more information in cases where the predictive distribution is skewed or has multiple modes. In this work, we employ a non-parametric approach referred to as kernel conditional density estimation (KCDE) to obtain separate predictive distributions for disease incidence in each week of the season, and then combine those marginal distributions using copulas to obtain joint predictive distributions for the trajectory of incidence over the course of multiple weeks. Predictive distributions relating to the timing of and incidence at the peak week can be obtained from this joint predictive distribution for the trajectory of disease incidence. In addition to the novel application of these methods to predicting disease incidence, our contributions include the

[1]Department of Biostatistics and Epidemiology, University of Massachusetts, Amherst
[2]CDC, Puerto Rico

Corresponding author:
Evan Ray, UMass Address Here
Email: elray@umass.edu

use of a periodic kernel specification to capture seasonality in disease incidence and a method for obtaining multivariate kernel functions that handle discrete data while allowing for a fully parameterized bandwidth matrix.

KCDE is a method for estimating the conditional distribution of a random vector $\mathbf{Y}$ given observations of another vector $\mathbf{X}$. In our work, $\mathbf{Y}$ is a measure of disease incidence at some future date (the prediction target) and $\mathbf{X}$ is a vector of predictive variables that we condition on in order to make our prediction. In our example applications, $\mathbf{X}$ includes observations of incidence over several recent time points and variables indicating the time of year at which we are making a prediction; in general, it would be possible to include other variables such as weather covariates.

KCDE has not previously been applied to obtain predictive distributions in the context of infectious disease, but it has been successfully used for prediction in other settings such as survival time of lung cancer patients[?] , female labor force participation[?] , bond yields and value at risk in financial markets[?] , and wind power[?] among others. Although KCDE has not previously been applied to predicting infectious disease, closely related methods for obtaining point predictions have been employed for diseases such as measles[?] and influenza[?] . In the infectious disease literature these methods have been referred to as state space reconstruction and the method of analogues, but they amount to an application of nearest neighbors regression methods. The point prediction obtained from nearest neighbors regression is equal to the expected value of the predictive distribution obtained from KCDE if a particular kernel function is used in the formulation of KCDE[?] . However, KCDE offers the advantage of providing a complete predictive distribution rather than only a point prediction. Methods similar to those we explore in this article can also be formulated in the Bayesian framework. One example along these lines is Zhou et al.[?] , who model the time to arrival of a disease in amphibian populations using Dirichlet processes and copulas.

There is also a long history of using other modeling approaches such as compartmental models for infectious disease prediction. A full discussion of those methods is beyond the scope of this article; see *** for a recent review. KCDE is distinguished from these alternative approaches in that it makes minimal assumptions about the data generating process. This can be either an advantage and a disadvantage of KCDE. In general, flexible non-parametric methods such as KCDE exhibit low bias but high variance. If they are correctly specified, models with more structure may achieve reduced variance without introducing bias. On the other hand, because non-parametric approaches such as KCDE make fewer assumptions they may outperform incorrectly specified parametric models. An evaluation of the benefits of an approach such as KCDE is therefore dependent on the particular characteristics of the system being modeled, the data that are available, and the quality of the more structured parametric models that are considered as alternatives. We will return to this point in our conclusions.

Need to find a review of prediction methods for infectious disease.

To our knowledge, all previous authors using kernel methods to estimate multivariate densities involving discrete variables have employed a kernel function that is a product of univariate kernel functions[?, ?, ?, ?] . A variety of functional forms have been proposed for this purpose, including geometric, triangular and Poisson among others[?, ?, ?, ?, ?] .

Using a product kernel simplifies the mathemetical formulation of the kernel function when discrete variables are present, but has the effect of forcing the kernel function to be orientied in line with the coordinate axes. In settings with only continuous variables, asymptotic analysis and experience with applications have shown that using a multivariate kernel function with a bandwidth parameterization that allows for other orientations can result in improved density estimates in many cases (cite ***). We introduce an approach to allowing for discrete kernels with orientation by discretizing an underlying continuous kernel function.

A limitation of kernel-based density estimation methods is that they may not scale well with the dimension of the vector whose distribution is being estimated. This is particularly relevant in our application, where it is desired to obtain joint predictive distributions for disease incidence over the course of many weeks. Copulas present one strategy for estimating the joint distribution of moderate to high dimensional random vectors, and work by specifying a relatively simple parametric model for the dependence relations among those variables. Specifically, we model the joint distribution of $Y_1, \ldots, Y_D$ by $F_{Y_1,\ldots,Y_D}(y_1, \ldots, y_D) = C(F_{Y_1}(y_1), \ldots, F_{Y_D}(y_D); \boldsymbol{\xi})$. Here $C : [0, 1]^D \rightarrow [0, 1]$ is the copula function depending on parameters $\boldsymbol{\xi}$ and mapping the vector of marginal c.d.f. values to the joint c.d.f. value.

It would be possible to handle this task using just the formulation of KCDE we discussed above, but a direct application of this approach has some limitations. First, the performance of kernel-based density estimation methods scales poorly with the dimension of the random vector whose density is being estimated (cite ***). Second, we have found that different information is available in the data at different prediction horizons. For example, we will demonstrate in our applications below that recently observed incidence is important for making short-term predictions, but terms capturing seasonality are more important for making long-term predictions.

The remainder of this article is organized as follows. We: - describe how kernel density estimation with a non-diagonal bandwidth can be achieved using a partially discretized multivariate normal distribution for the kernel functions.

- simulation study comparing product and non-product formulations for marginal and conditional density estimation

- applications

Method Description

In this Section, we give a detailed discussion of our methods. Suppose we observe a measure z_t of disease incidence at each point in time $t = 1, \ldots, T$. At time t^*, our model gives a predictive distribution $f(z_{t^*+1}, \ldots, z_{t^*+H}|t^*, z_{t^*-l_1}, \ldots, z_{t^*-l_M})$ for the trajectory of disease incidence over a range of prediction horizons from 1 to H weeks in the future. The time t^* will typically be equal to T when we are applying the method to perform prediction, but takes other values in the estimation procedure we describe below. The predictive distribution is conditional on the time at which we are making the predictions and observed incidence at a few recent time points with lags given by the non-negative integers $l_1, \ldots, l_M$. It would also be possible to condition on other covariates such as weather, but we have not pursued that line in this work.

Our model represents this density as follows:

$$f(z_{t^*+1}, \ldots, z_{t^*+H}|t^*, z_{t^*-l_1}, \ldots, z_{t^*-l_M}) = \\ c^H\{f^1(z_{t^*+1}|t^*, z_{t^*-l_1}, \ldots, z_{t^*-l_M}; \boldsymbol{\theta}^1), \ldots, f^H(z_{t^*+H}|t^*, z_{t^*-l_1}, \ldots, z_{t^*-l_M}; \boldsymbol{\theta}^H); \boldsymbol{\xi}^H\}.$$

Here, each $f^h(z_{t^*+h}|t^*, z_{t^*-l_1}, \ldots, z_{t^*-l_M}; \boldsymbol{\theta}^h)$ is a predictive density for one prediction horizon obtained through KCDE. The distribution for each prediction horizon depends on a separate parameter vector $\boldsymbol{\theta}^h$. The function $c^H(\cdot)$ is a copula used to tie these marginal predictive densities together into a joint predictive density, and depends on parameters $\boldsymbol{\xi}^H$. In our applications, we will obtain a separate copula fit for each trajectory length H of interest for the prediction task.

Broadly, estimation for the model parameters proceeds in two stages: first we estimate the parameters for KCDE separately for each prediction horizon h, and second we estimate the copula parameters while holding the KCDE parameters fixed. The efficiency of two-stage estimation procedures for copula models has been studied in the literature both theoretically and through simulation studies. In general the two-stage approach may result in some loss of efficiency relative to one-stage methods, but this efficiency loss is be small for some model specifications[?] . We pursue the two-stage strategy in this work because it results in a large reduction in the computational cost of parameter estimation.

In the following subsections we describe the formulations of KCDE and the copula in more detail and give our estimation strategy for each set of model parameters.

KCDE for Predictive Densities at Individual Prediction Horizons

We now discuss the methods we use to obtain the predictive density $f^h(z_{t^*+h}|t^*, z_{t^*-l_1}, \ldots, z_{t^*-l_M}; \boldsymbol{\theta}^h)$ for disease incidence at a particular horizon h after time t^*. In order to simplify the notation we define two new variables: $Y_t^h = Z_{t+h}$ represents the prediction target relative to time t, and $\mathbf{X}_t = (t, Z_{t-l_1}, \ldots, Z_{t-l_M})$ represents the vector of predictive variables relative to time t. With this notation, the distribution we wish to estimate is $f^h(y_{t^*}^h|\mathbf{x}_{t^*}; \boldsymbol{\theta}^h)$.

In order to estimate this distribution, we use the observed data to form the pairs $(\mathbf{x}_t, y_t^h)$ for all $t = 1 + \max_m l_m, \ldots, T - h$; for smaller values of t there are not enough observations before t to form $\mathbf{x}_t$ and for larger values of t there are not enough observations after t to form y_t^h. We then regard these pairs as a (dependent) sample from the joint distribution of $(\mathbf{X}, Y^h)$ and estimate the conditional distribution of $Y^h|\mathbf{X}$ via KCDE:

$$\widehat{f}^h(y_{t^*}^h|\mathbf{x}_{t^*}) = \frac{\sum_{t \in \boldsymbol{\tau}} K^{\mathbf{X},Y}\left\{(\mathbf{x}_{t^*}, y_{t^*}^h), (\mathbf{x}_t, y_t^h); \boldsymbol{\theta}^h\right\}}{\sum_{t \in \boldsymbol{\tau}} K^{\mathbf{X}}(\mathbf{x}_{t^*}, \mathbf{x}_t; \boldsymbol{\theta}^h)} \tag{1}$$

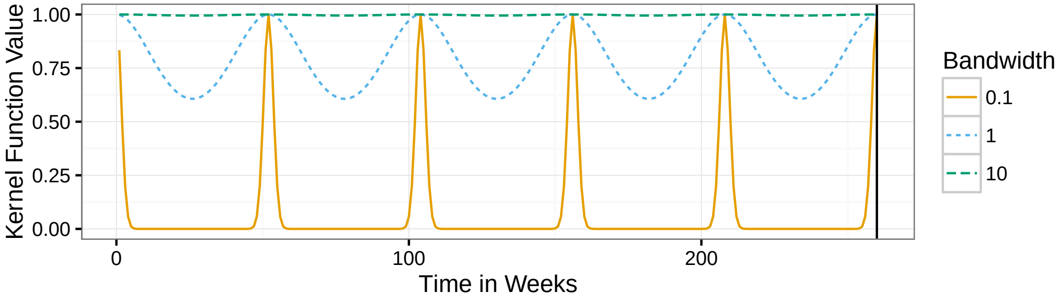
$$= \frac{\sum_{t \in \boldsymbol{\tau}} K^{Y|\mathbf{X}}(y_{t^*}^h, y_t^h|\mathbf{x}_{t^*}, \mathbf{x}_t; \boldsymbol{\theta}^h) K^{\mathbf{X}}(\mathbf{x}_{t^*}, \mathbf{x}_t; \boldsymbol{\theta}^h)}{\sum_{t \in \boldsymbol{\tau}} K^{\mathbf{X}}(\mathbf{x}_{t^*}, \mathbf{x}_t; \boldsymbol{\theta}^h)} \tag{2}$$

$$= \sum_{t \in \boldsymbol{\tau}} w_t^h K^{Y|\mathbf{X}}(y_{t^*}^h, y_t^h|\mathbf{x}_{t^*}, \mathbf{x}_t; \boldsymbol{\theta}^h), \text{ where} \tag{3}$$

$$w_t^h = \frac{K^{\mathbf{X}}(\mathbf{x}_{t^*}, \mathbf{x}_t; \boldsymbol{\theta}^h)}{\sum_{s \in \boldsymbol{\tau}} K^{\mathbf{X}}(\mathbf{x}_{t^*}, \mathbf{x}_s; \boldsymbol{\theta}^h)} \tag{4}$$

Here we are working with a slightly restricted specification in which the kernel function $K^{\mathbf{X},Y}$ can be written as the product of $K^{\mathbf{X}}$ and a "conditional kernel" $K^{\mathbf{Y}|\mathbf{X}}$. With this restriction, we can interpret $K^{\mathbf{X}}$ as a weighting function determining how much each observation $(\mathbf{x}_t, y_t^h)$

Figure 1. The periodic kernel function illustrated as a function of time in weeks with $\rho = \pi/52$ and three possible values for the bandwidth parameter h.



contributes to our final density estimate according to how similar $\mathbf{x}_t$ is to the value $\mathbf{x}_{t^*}$ that we are conditioning on. For each y_t^h, $K^{\mathbf{Y}|\mathbf{X}}$ is a density function that contributes mass to the final density estimate near y_t^h. The parameters $\boldsymbol{\theta}^h$ control the locality and orientation of the weighting function and the contributions to the density estimate from each observation. In Equations (1) through (4), $\boldsymbol{\tau} \subseteq \{1 + \max_m l_m, \ldots, T - h\}$ indexes the subset of observations used in obtaining the conditional density estimate; we return to how this subset of observations is defined in the discussion of estimation below.

We take the kernel function $K^{Y,\mathbf{X}}$ to be a product kernel with one component being a periodic kernel in time and the other component capturing the remaining covariates:

$$\begin{aligned}
&K^{\mathbf{X},Y}\left\{(\mathbf{x}_{t^*}, y_{t^*}^h), (\mathbf{x}_t, y_t^h); \boldsymbol{\theta}^h\right\} \\
&\quad = K^{\mathbf{X},Y}\left\{(t^*, z_{t^*-l_1}, \ldots, z_{t^*-l_M}, z_{t^*+h}), (t, z_{t-l_1}, \ldots, z_{t-l_M}, z_{t+h}); \boldsymbol{\theta}^h\right\} \\
&\quad = K^{Periodic}(t^*, t; \boldsymbol{\theta}^h) K^{Incidence}\{(z_{t^*-l_1}, \ldots, z_{t^*-l_M}, z_{t^*+h}), (z_{t-l_1}, \ldots, z_{t-l_M}, z_{t+h}); \boldsymbol{\theta}^h\}
\end{aligned}$$

The periodic kernel function was originally developed in the literature on Gaussian Processes[?], and is defined by

$$K^{Periodic}(t^*, t; \rho, b) = \exp\left[-\frac{\sin^2\{\rho(t^* - t)\}}{2b^2}\right]. \tag{5}$$

We illustrate this kernel function in Figure 1. The kernel has two parameters: ρ, which determines the length of the periodicity, and b, which determines the strength and locality of of this periodic component in computing the observation weights w_t^h. In our applications, we have fixed $\rho = \pi/52$, so that the kernel has period of length 1 year with weekly data. Using this periodic kernel provides a mechanism to capture seasonality in disease incidence by allowing the observation weights to depend on the similarity of the time of year that an observation was collected and the time of year at which we are making a prediction.

The second component of our kernel is a multivariate kernel incorporating all of the other variables in $\mathbf{x}_t$ and y_t^h. In our applications, these variables are measures of incidence, and are continuous in the application to Influenza and discrete case counts in the application to Dengue fever. In the continuous case, we have used a multivariate log-normal kernel function. This kernel specification automatically handles the restriction that counts are non-negative, and approximately captures the long tail in disease incidence that we will illustrate in the applications Section below.

functional form for multivariate log-normal? discussion of bandwidth matrix – locality, orientation, parameterization in terms of $\boldsymbol{\theta}$?

In the discrete case, we obtain the kernel function by discretizing an underlying continuous kernel function:

$$\begin{aligned} &K^{Incidence}\{(z_{t^*-l_1}, \ldots, z_{t^*-l_M}, z_{t^*+h}), (z_{t-l_1}, \ldots, z_{t-l_M}, z_{t+h}); \boldsymbol{\theta}^h\} \\ &\quad = \int_{a_{z_{t^*-l_1}}}^{b_{z_{t^*-l_1}}} \cdots \int_{a_{z_{t^*+h}}}^{b_{z_{t^*+h}}} L\{(w_1, \ldots, w_{M+1}), (z_{t-l_1}, \ldots, z_{t-l_M}, z_{t+h}); \boldsymbol{\theta}^h\}\, dw_1 \cdots dw_{M+1} \end{aligned}$$

Here, the w_m are dummy integration variables and $L(\cdot)$ is a continuous multivariate kernel function; in our application we have used the multivariate log-normal kernel as above. For each component variable in $(z_{t^*-l_1}, \ldots, z_{t^*-l_M}, z_{t^*+h})$, we associate lower and upper bounds of integration a_{z_j} and b_{z_j} with each value in the domain of that random variable. The value of the kernel function K at $\mathbf{w}$ is obtained by integrating over the hyper-rectangle specified by these bounds. In our application, the possible values of the random variables are integer counts and the corresponding integration bounds are the half-integers.

We use cross-validation to estimate the bandwidth parameters by numerically minimizing a cross-validation measure of the quality of the predictions obtained from the model. Specifically, we use the log-score. We leave out a year of data before and after the time t^*. Hart and Vieu[?] show that when kernel density estimation is used to estimate a marginal density with dependent observations, leaving out multiple time points around the target time point in cross validation can yield small improvements in the ISE under certain assumptions about the form of the dependence.

Talk about proper scoring rules and our particular choice of Q. Criticism of bandwidth estimation by likelihood crossvalidation. Relationship to evaluation criteria.

Combining Marginal Predictive Distributions with Copulas

The approach we take for some of the prediction targets we examine in our applications is to obtain a joint predictive distribution for disease incidence over a sequence of multiple prediction horizons. We do this by using a copula to combine marginal predictive densities for each of those prediction horizons. Specifically, we us the isotropic Gaussian copula implemented in the R[?] package `copula`[?] .

This copula function is given by

$$c(u_1, \ldots, u_J; \boldsymbol{\theta}_c) = \Phi_\Sigma(\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_J)), \tag{6}$$

where Φ^{-1} is the inverse c.d.f. of a standard univariate Gaussian distribution and Φ_Σ is the c.d.f. of a multivariate Gaussian distribution with mean $\underline{0}$ and covariance matrix Σ. We set $\Sigma = [\sigma^2_{i,j}]$, where

$$\sigma^2_{i,j} = \begin{cases} 1 \text{ if } i = j, \\ \rho_d \text{ if } |i - j| = d \end{cases} \tag{7}$$

Intuitively, ρ_d captures the correlation of incidence at future times that are d weeks apart.

Estimation by maximum likelihood.

Simulation Study

In this Section, we conduct two sets of simulation studies designed to answer two separate questions:

1. How much does using a kernel function with a non-diagonal bandwidth matrix contribute to the quality of conditional density estimates relative to density estimates obtained through KCDE using diagonal bandwidth matrices?
2. How does our method perform in the context of seasonal time series data? Specifically, how does the method perform relative to common alternatives, and how much do each of our three contributions (non-diagonal bandwidth matrices for discrete data, using a periodic function of time as predictive variable, and use of low band-pass filtered observatiosn as predictive variables) contribute to predictive performance?

Comparison of KCDE approaches

Our first set of simulation studies is based closely on those conducted in[?]; their examples demonstrate the utility of using a fully parameterized bandwidth matrix in kernel density estimation of continuous distributions. We modify their simulation study to examine the benefits of fully parameterized bandwidth matrices in the context of conditional density estimation with discrete variables.

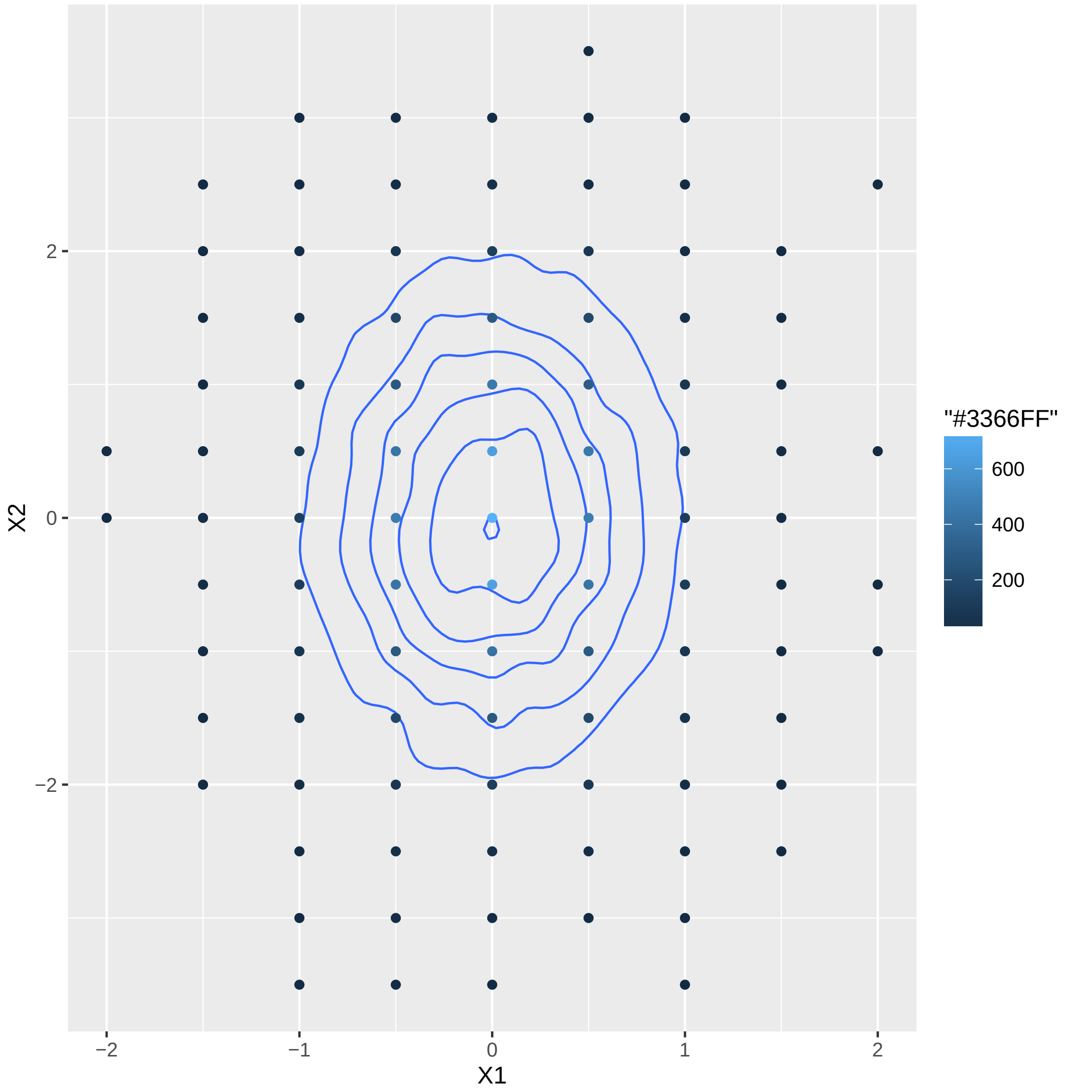
We simulate observations from each of seven distributions. The first five of these are plotted in Figure ***.

```
library(ggplot2)
library(grid)
library(plyr)
library(dplyr)

##
## Attaching package: 'dplyr'
## The following objects are masked from 'package:plyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize
## The following objects are masked from 'package:stats':
##
##     filter, lag
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

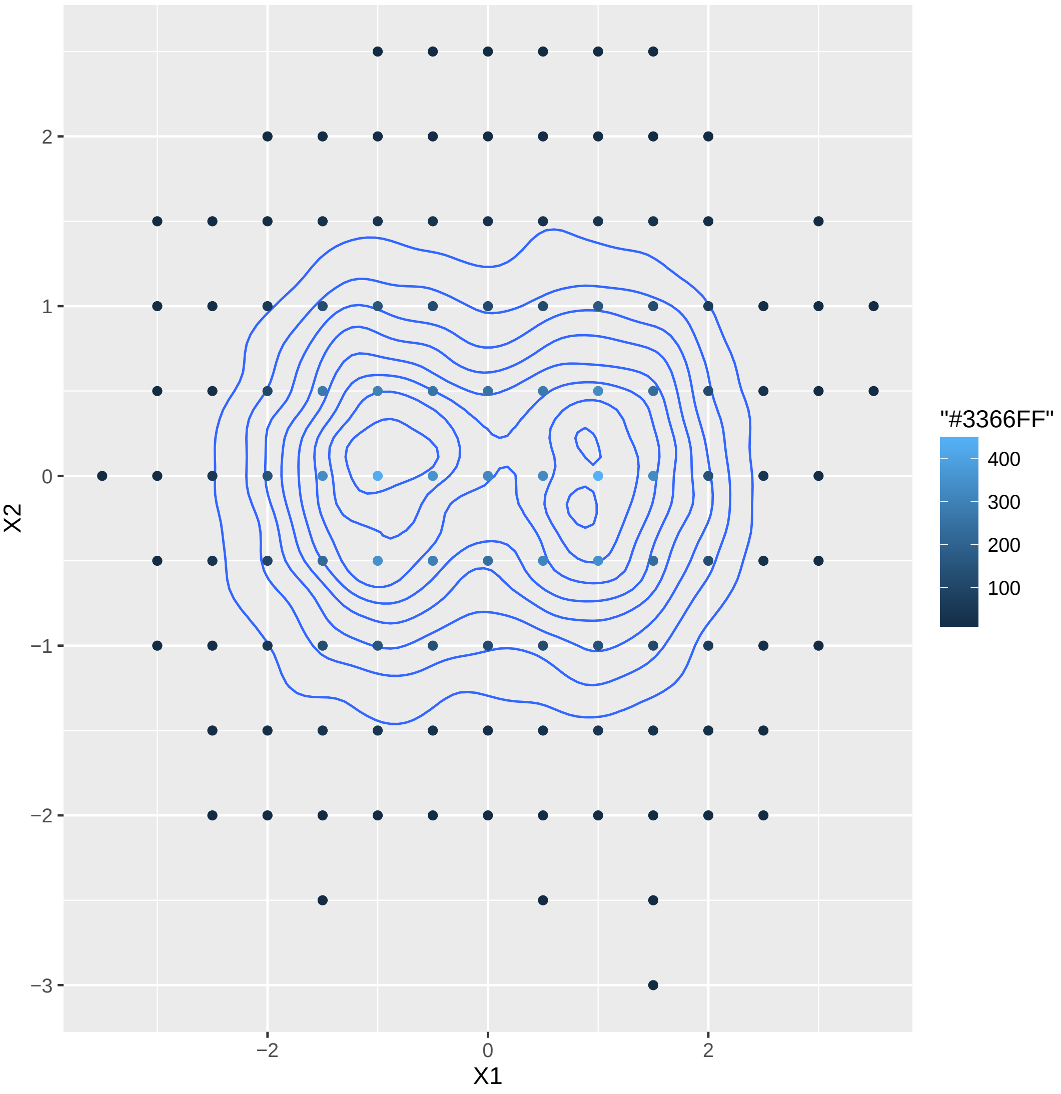
```
library(tidyr)
library(pdtmvn)
library(kcde)
```

```
## Warning: failed to assign RegisteredNativeSymbol for logspace_add_C to
logspace_add_C since logspace_add_C is already defined in the 'kcde' namespace
## Warning: failed to assign RegisteredNativeSymbol for
logspace_sum_matrix_rows_C to logspace_sum_matrix_rows_C since
logspace_sum_matrix_rows_C is already defined in the 'kcde' namespace
## Warning: failed to assign RegisteredNativeSymbol for logspace_sub_C to
logspace_sub_C since logspace_sub_C is already defined in the 'kcde' namespace
## Warning: failed to assign RegisteredNativeSymbol for
logspace_sub_matrix_rows_C to logspace_sub_matrix_rows_C since
logspace_sub_matrix_rows_C is already defined in the 'kcde' namespace
##
## Attaching package: 'kcde'
## The following objects are masked from 'package:pdtmvn':
##
##     logspace_add, logspace_sub, logspace_sub_matrix_rows,
##     logspace_sum_matrix_rows
```

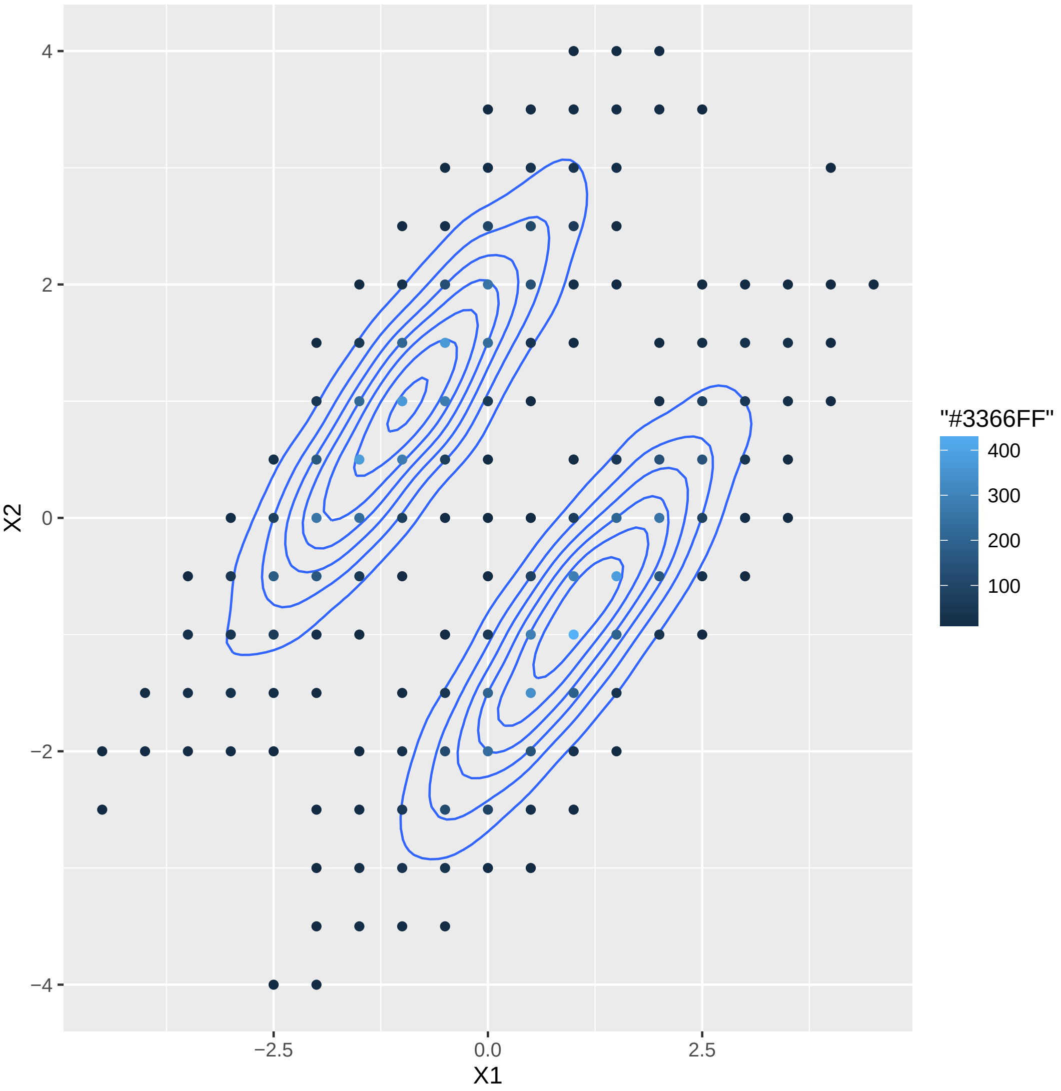
```
source("/media/evan/data/Reich/infectious-disease-prediction-with-kcde/inst/code/sim-densi

## Density family bivariate-A
n_sim <- 10000
discrete_sample <- sim_from_pdtmvn_mixt(n = n_sim, sim_family = "bivariate-A-discretized")
    as.data.frame()
continuous_sample <- sim_from_pdtmvn_mixt(n = n_sim, sim_family = "bivariate-A") %>%
    as.data.frame()
discrete_sample_counts <- discrete_sample %>%
    count(X1, X2)

pa <- ggplot() +
    geom_density_2d(aes(x = X1, y = X2), data = continuous_sample) +
    geom_point(aes(x = X1, y = X2, colour = n), data = discrete_sample_counts)
pa
```



```
## Density family bivariate-B
n_sim <- 10000
discrete_sample <- sim_from_pdtmvn_mixt(n = n_sim, sim_family = "bivariate-B-discretized")
    as.data.frame()
continuous_sample <- sim_from_pdtmvn_mixt(n = n_sim, sim_family = "bivariate-B") %>%
    as.data.frame()
discrete_sample_counts <- discrete_sample %>%
    count(X1, X2)

pb <- ggplot() +
    geom_density_2d(aes(x = X1, y = X2), data = continuous_sample) +
    geom_point(aes(x = X1, y = X2, colour = n), data = discrete_sample_counts)
pb
```



```
## Density family bivariate-C
n_sim <- 10000
discrete_sample <- sim_from_pdtmvn_mixt(n = n_sim, sim_family = "bivariate-C-discretized")
    as.data.frame()
continuous_sample <- sim_from_pdtmvn_mixt(n = n_sim, sim_family = "bivariate-C") %>%
    as.data.frame()
discrete_sample_counts <- discrete_sample %>%
    count(X1, X2)

pc <- ggplot() +
    geom_density_2d(aes(x = X1, y = X2), data = continuous_sample) +
    geom_point(aes(x = X1, y = X2, colour = n), data = discrete_sample_counts)
pc
```

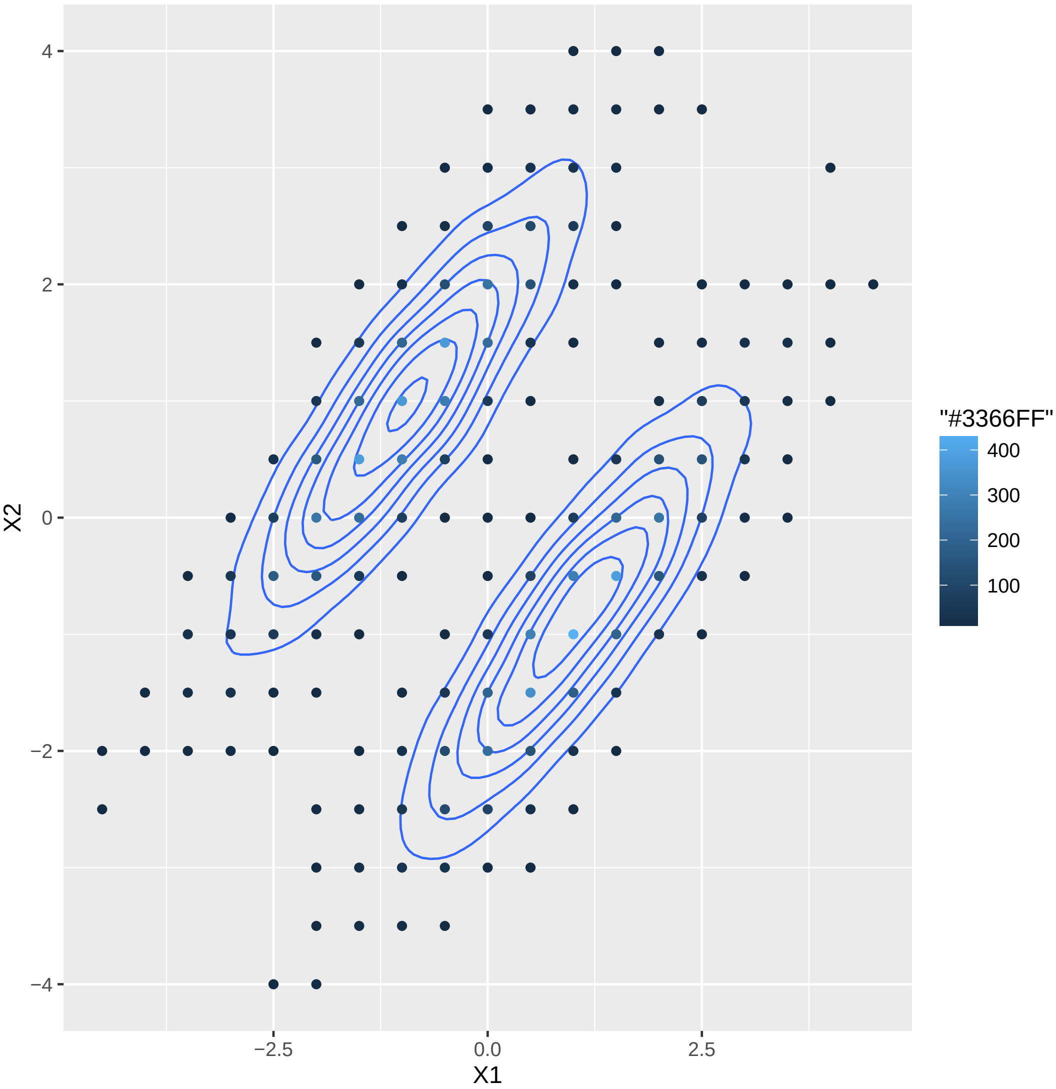


```
## Density family bivariate-D
n_sim <- 10000
discrete_sample <- sim_from_pdtmvn_mixt(n = n_sim, sim_family = "bivariate-D-discretized")
    as.data.frame()

## Error in get_dist_component_params_for_sim_family(sim_family): Invalid
sim_family
```

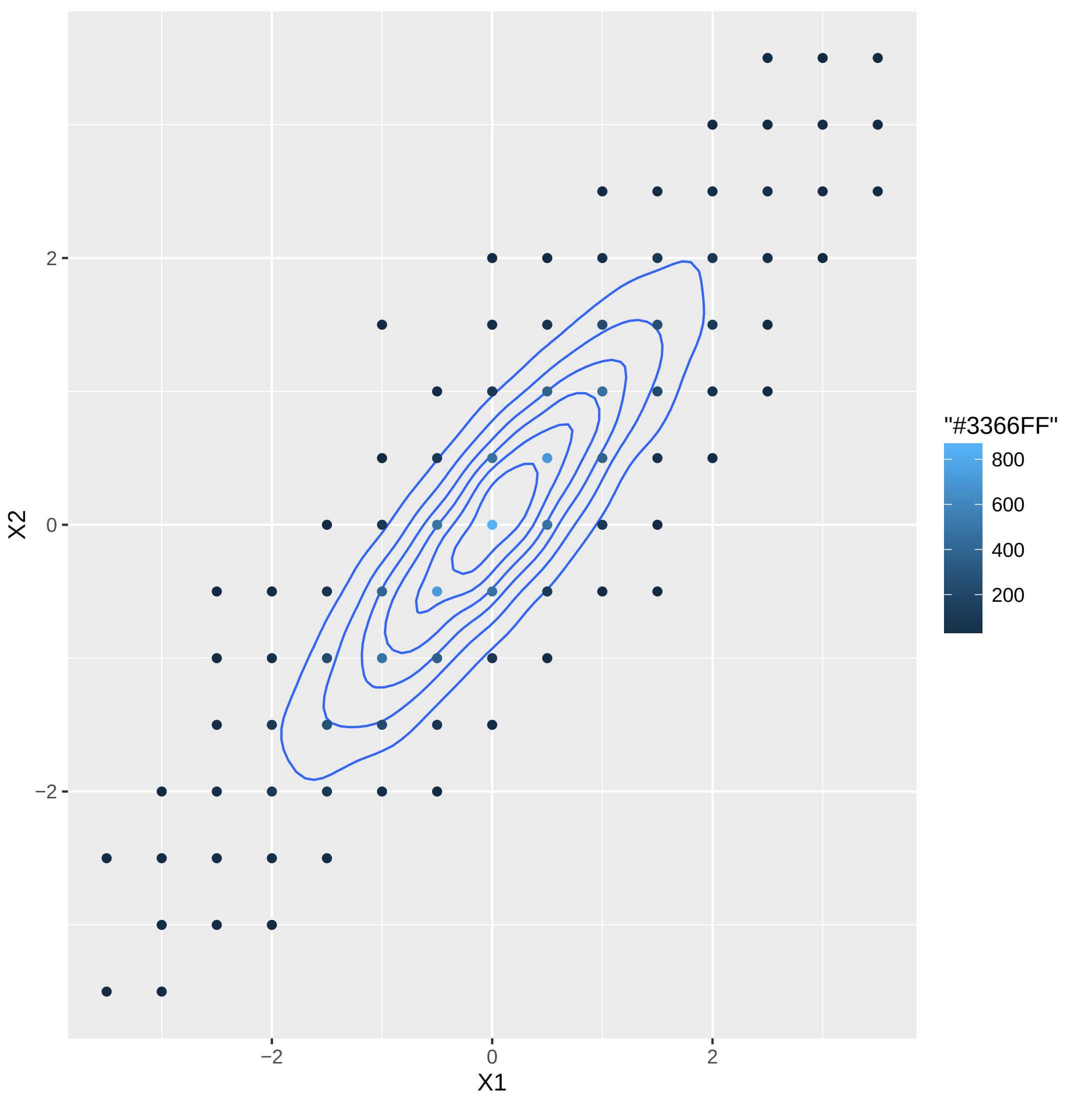
```
continuous_sample <- sim_from_pdtmvn_mixt(n = n_sim, sim_family = "bivariate-D") %>%
    as.data.frame()
```

```
## Error in get_dist_component_params_for_sim_family(sim_family): Invalid
sim_family
```

```
discrete_sample_counts <- discrete_sample %>%
    count(X1, X2)

pd <- ggplot() +
    geom_density_2d(aes(x = X1, y = X2), data = continuous_sample) +
    geom_point(aes(x = X1, y = X2, colour = n), data = discrete_sample_counts)
pd
```



```
## Density family multivariate-2d
n_sim <- 10000
discrete_sample <- sim_from_pdtmvn_mixt(n = n_sim, sim_family = "multivariate-2d-discretiz
    as.data.frame()
continuous_sample <- sim_from_pdtmvn_mixt(n = n_sim, sim_family = "multivariate-2d") %>%
    as.data.frame()
discrete_sample_counts <- discrete_sample %>%
    count(X1, X2)

pd <- ggplot() +
    geom_density_2d(aes(x = X1, y = X2), data = continuous_sample) +
    geom_point(aes(x = X1, y = X2, colour = n), data = discrete_sample_counts)
pd
```



Applications

In this Section, we illustrate our methods through applications to prediction of infectious disease in two examples with real disease incidence data sets: one with a weekly measure of incidence of influenza like illness in the United States, and a second with a weekly measure of incidence of Dengue fever in San Juan, Puerto Rico. These data sets were used in two recent prediction competitions sponsored by the United States federal government. (cite something???)

We plot the data in Figure ***. As indicated in the figure, we have divided each data set into two subsets. The first period is used as a training set in estimating the model parameters. The last four years of each data set are reserved as a test set for evaluating model performance.

There are three prediction targets for each data set, based closely on the prediction targets that were used in those competitions. First, for each week in the test data, we obtain a predictive distribution for the incidence measure in that week at each prediction horizon from 1 to 52 weeks ahead. Second, in each season of the test data set, we make predictions for the timing of the peak week. Third, we predict the incidence measure in the peak week. In all cases, we compare the models using log score.

We use a seasonal ARIMA model as a baseline to compare our approach to. In fitting this model, we first transformed the observed incidence measure to the log scale (after adding 1 in the Dengue data set, which included some observations of 0 cases); this transformation makes the normality assumptions of the ARIMA model more plausible. We then performed first-order seasonal differencing, and obtained the final model fits using the `auto.arima` function in `R`'s `forecast` package? ; this function uses a stepwise procedure to determine the terms to include in the model. This procedure resulted in a SARIMA$(2,0,0)(2,1,0)_{52}$ model for the influenza data and a SARIMA$(3,0,2)(1,1,0)_{52}$ model for the Dengue data. We note that a different SARIMA model was used as a baseline in the Dengue competition, but the SARIMA model we obtained using this procedure performed slightly better on the test set than that previous baseline model.

Discuss variations on KCDE models.

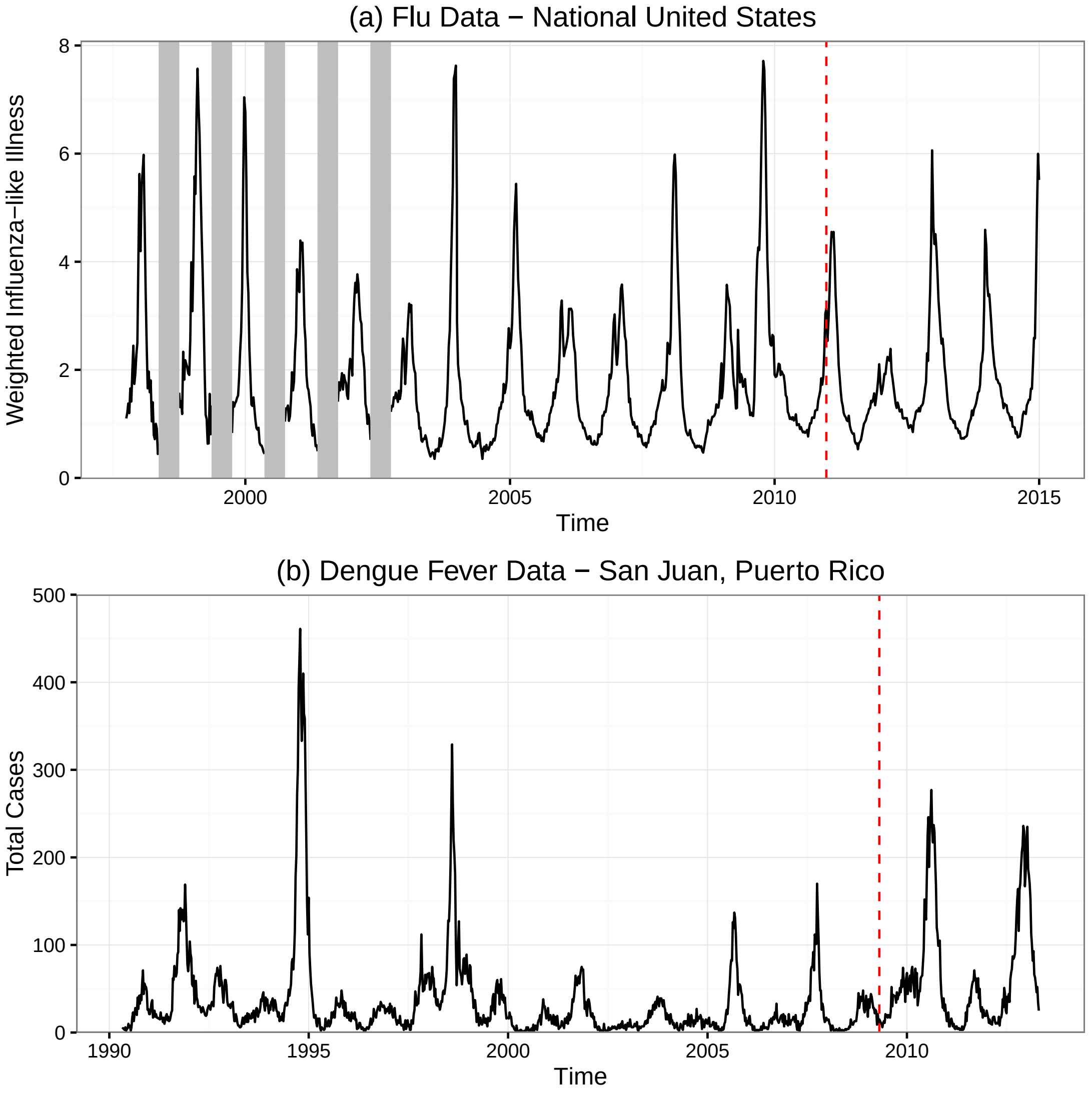
Example 1: Influenza

Future Work

Hall, Racine, and Li? show that when cross-validation is used to select the bandwidth parameters in KCDE using product kernels, the estimated bandwidths corresponding to irrelevant conditioning variables tend to infinity asymptotically as the sample size increases. They discuss the fact that similar results could be obtained for linear combinations of continuous variables if a full bandwidth matrix were used. Our approach for obtaining kernels that can be used with mixed discrete and continuous variables opens up an opportunity to extend this analysis to that case; we have not pursued this mathematical analysis here.

The above results regarding the inclusion of irrelevant conditioning variables hold asymptotically as the sample size increases. However, in practice, data set sizes are often limited. In other modeling settings where some conditioning variables may not be informative, shrinkage methods are often helpful. These methods could be incorporated into a kernel-based approach by imposing a penalty on the elements of the bandwidth matrix; in particular, we suggest that a penalty on the inverse of the bandwidth matrix encouraging it to have small eigenvalues could be

Figure 2. Plots of the data sets we apply our methods to. In each case, the last four years of data are held out as a test data set; this cutoff is indicated with a vertical dashed line. For the flu data set, low-season incidence was not recorded in early years of data collection; these missing data are indicated with vertical grey bars.



helpful. Another alternative would be to pursue the Bayesian framework, using Dirichlet process mixtures with an informative prior on the mixture component covariance matrices.

Figure 3. Differences in log scores for the weekly predictive distributions among pairs of models across all combinations of prediction horizon and prediction time in the test period. In panel (a) positive values indicate cases when KCDE outperformed SARIMA. In panel (b) positive values indicate cases when the specification of KCDE with the periodic kernel outperformed the corresponding specification without the periodic kernel. In panel (c) positive values indicate cases when the specification of KCDE with a fully parameterized bandwidth outperformed the KCDE specification with a diagonal bandwidth matrix.

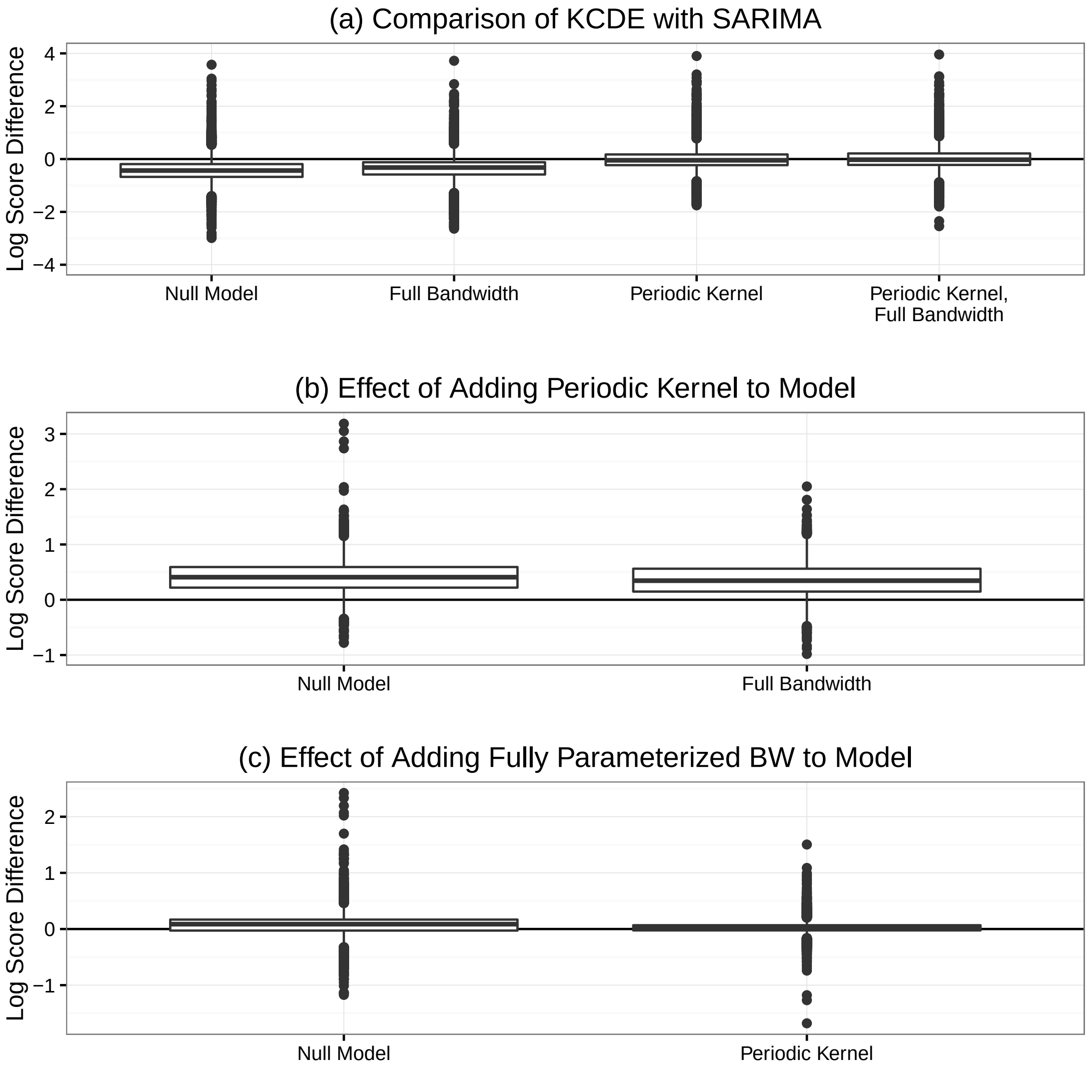
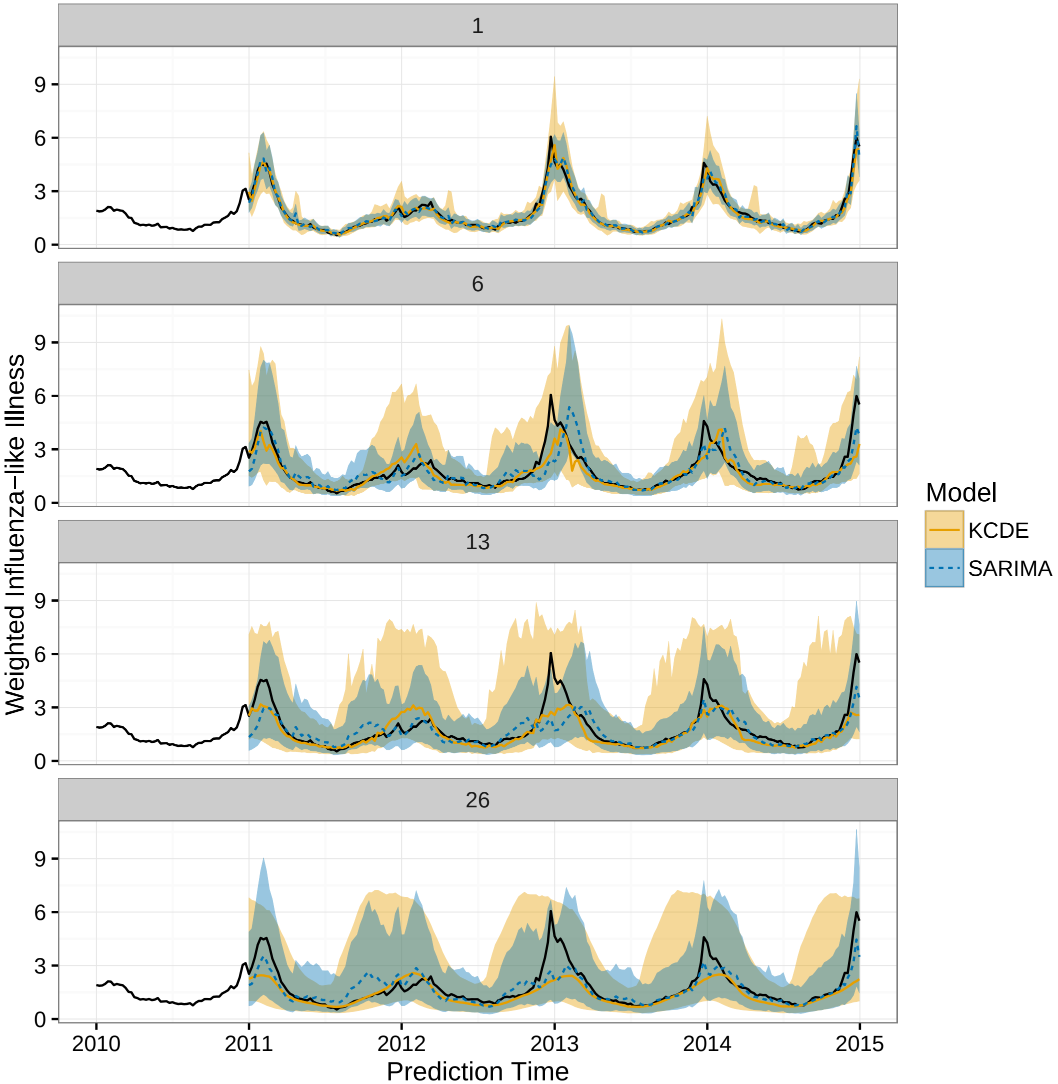
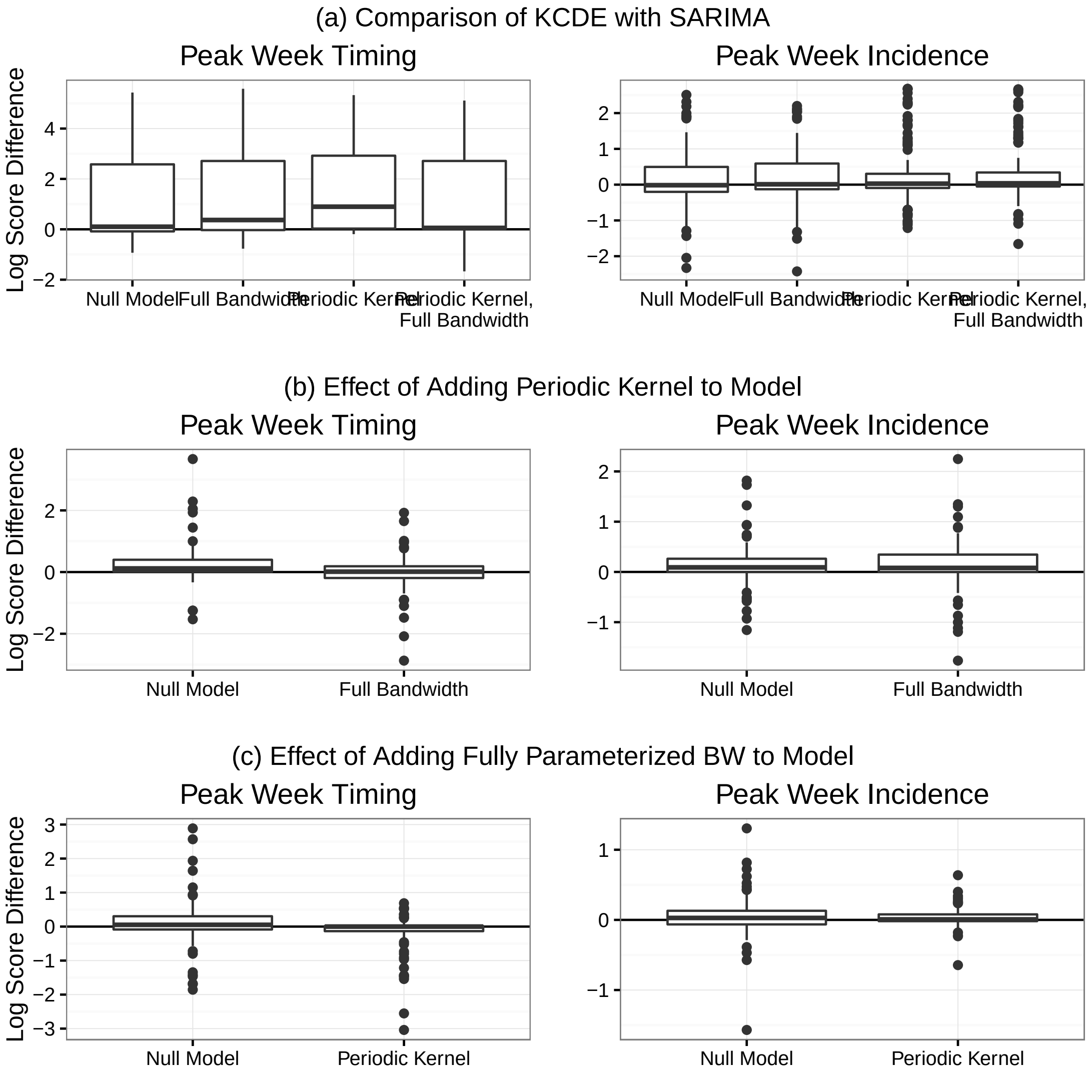


Figure 4. Plots of point and interval predictions from SARIMA and KCDE.



We could also make some tweaks to our implementation of KCDE. Locally linear – help address edge effects. Cite Hyndman, Bashtannyk, Grunwald - "Estimating and Visualizing Conditional Densities", maybe also Fan and Yim - "A crossvaildation method for estimating conditional densities" and Fan et al. 1996 "Estimation of conditional densities and sensitivity measures in nonlinear dynamical systems."

Figure 5. Differences in log scores for the predictive distributions for the peak week and incidence at the peak week among pairs of models across all analysis times in the test period. In panel (a) positive values indicate cases when KCDE outperformed SARIMA. In panel (b) positive values indicate cases when the specification of KCDE with the periodic kernel outperformed the corresponding specification without the periodic kernel. In panel (c) positive values indicate cases when the specification of KCDE with a fully parameterized bandwidth outperformed the KCDE specification with a diagonal bandwidth matrix. In the plot for peak week timing in panel (a), the log score differences are not displayed for one analysis time when none of the simulated trajectories from SARIMA peaked at the true peak week. In that case, our monte carlo estimate of the difference in log scores is infinity.

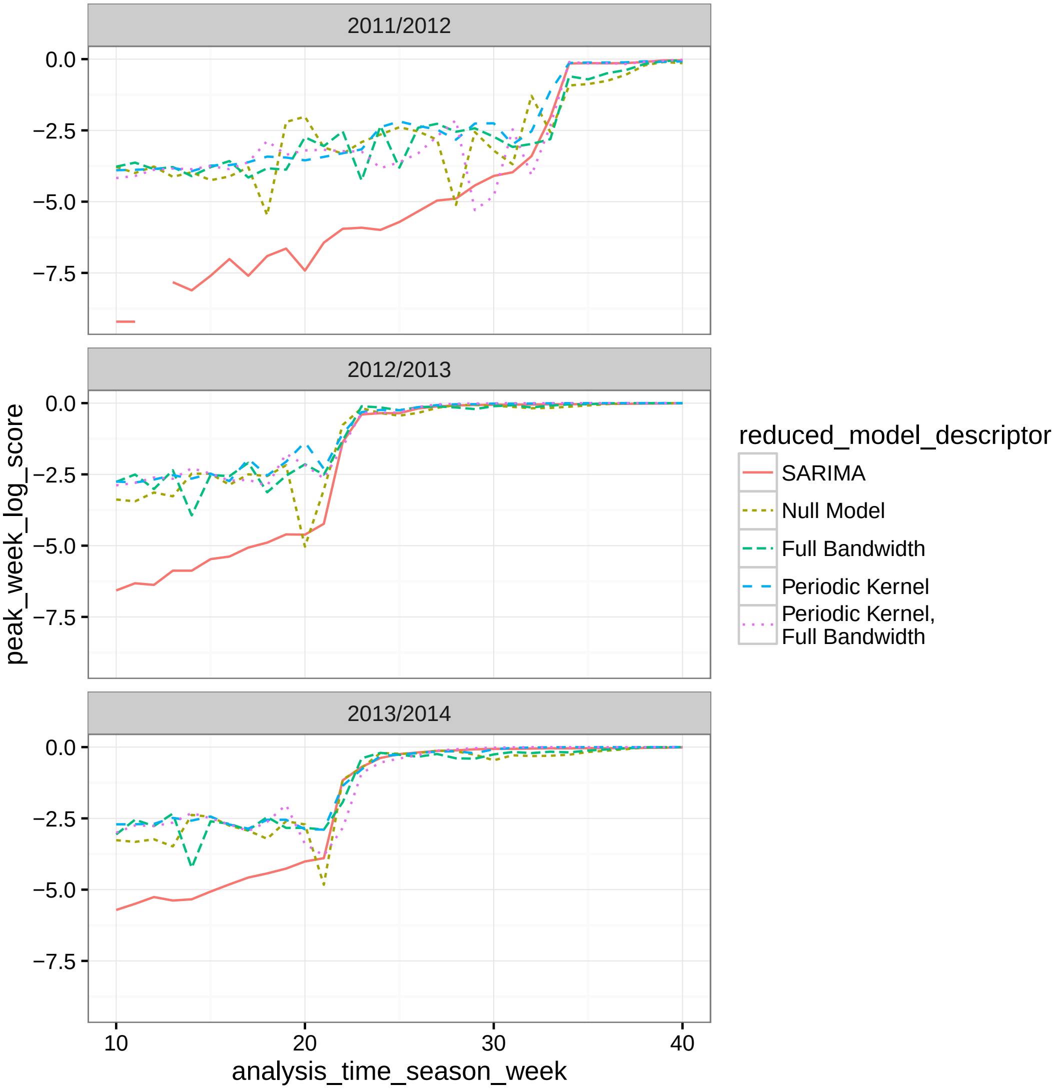


(a) Comparison of KCDE with SARIMA

(b) Effect of Adding Periodic Kernel to Model

(c) Effect of Adding Fully Parameterized BW to Model

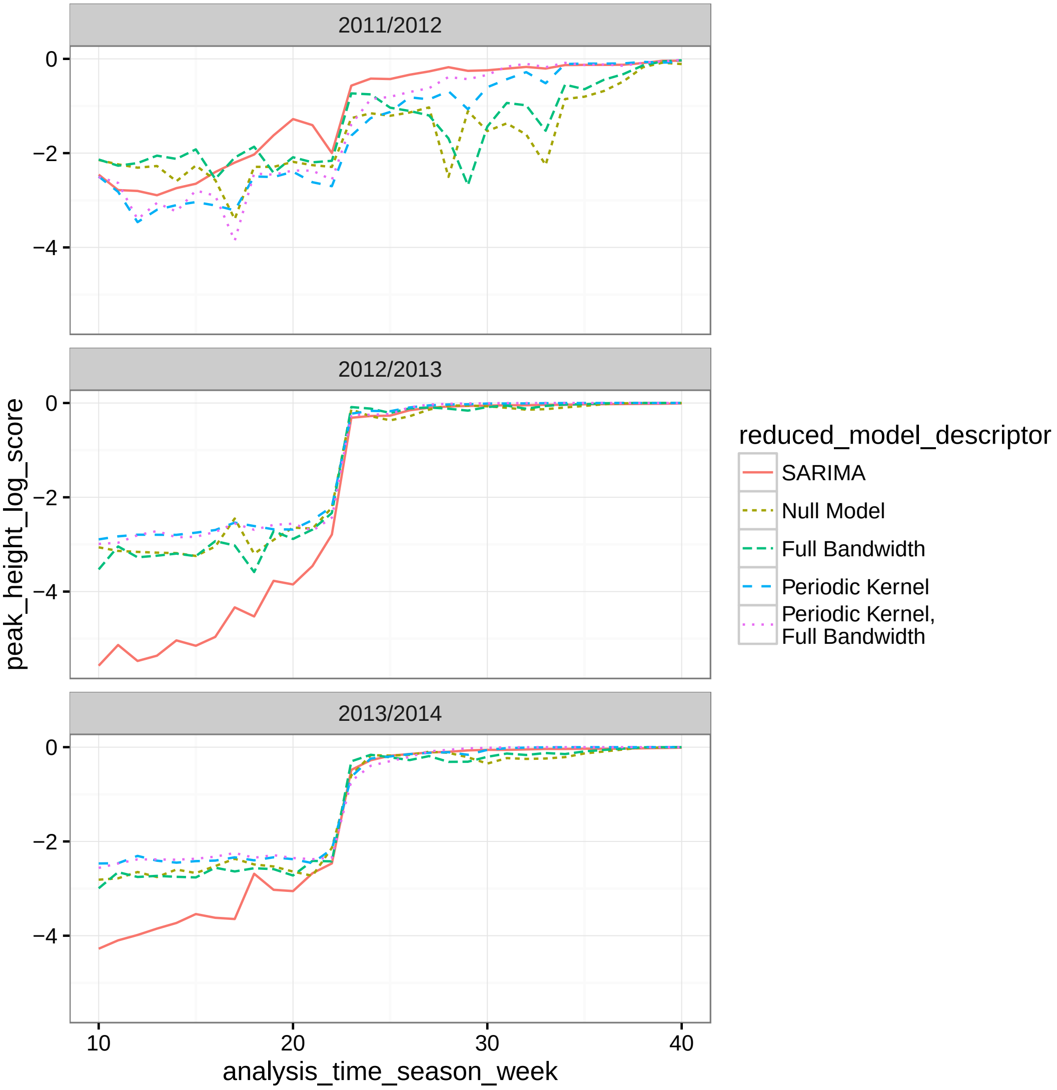
Figure 6. Log scores for predictions of peak week timing by predictive model and analysis time.



Ensembles – either ensembles of KCDE and/or include as a component in an ensemble. Also Bayesian model averaging. Return to discussion of bias/variance trade-off?

Other covariates

Figure 7. Log scores for predictions of incidence in the peak week by predictive model and analysis time.


2011/2012
2012/2013
2013/2014
peak_height_log_score
analysis_time_season_week
reduced_model_descriptor
SARIMA
Null Model
Full Bandwidth
Periodic Kernel
Periodic Kernel, Full Bandwidth