Journal Title
XX(X):1–20

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] and Nicholas G. Reich[1]

Abstract

We develop a novel approach to prediction of infectious disease incidence using Kernel Conditional Density Estimation (KCDE). This method obtains predictive distributions for incidence in individual weeks and uses copulas to tie those distributions together into joint distributions in order to make predictions for the timing of and incidence in the peak week of the season. Our implementation of KCDE incorporates two novel kernel components: a periodic component that captures seasonality in disease incidence, and a component that is appropriate for use with discrete variables but also allows for a full parameterization of the bandwidth matrix. A simulation study demonstrates that allowing for a fully parameterized bandwidth matrix can improve conditional density estimates. In applications to predicting dengue fever and influenza, our method outperforms a baseline seasonal autoregressive integrated moving average (SARIMA) model for predictions of dengue incidence in individual weeks, and is comparable to the SARIMA model on the other prediction targets. The periodic kernel function leads to improved predictions of incidence in both applications. Our approach and extensions of it could yield improved predictions for public health decision makers, particularly in diseases with heterogeneous seasonal dynamics such as dengue fever.

Keywords

copula, dengue fever, infectious disease, influenza, 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

[1]Department of Biostatistics and Epidemiology, University of Massachusetts, Amherst

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

Prepared using sagej.cls *[Version: 2015/06/09 v1.01]*

protective equipment by medical personnel during periods of high disease incidence[8;23]. Several quantities have emerged as being of particular utility in making these planning decisions; in this article we focus on measures of weekly incidence, the timing of the season peak, and incidence in the peak week[3;16]. 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. We 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 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.

At its heart, KCDE is a local method in the sense that the conditional density estimate for future incidence given conditioning variables is a weighted combination of contributions from previous observations of incidence with similar conditioning values. Using such local methods is a natural idea in predicting nonlinear dynamical systems. For example, in the infectious disease literature nearest neighbors regression has been used to make point predictions for incidence of measles[21] and influenza[22]. 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[7]. However, KCDE offers the advantage of providing a complete predictive distribution rather than only a point prediction. KCDE has not previously been applied to obtain predictive distributions for infectious disease incidence, but it has been successfully used for prediction in other settings such as survival time of lung cancer patients[6], female labor force participation[6], bond yields and value at risk in financial markets[4], and wind power[10] among others. 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.[25], who model the time to arrival of a disease in amphibian populations using Dirichlet processes and copulas.

To our knowledge, previous implementations of kernel methods for estimating multivariate densities involving discrete variables have employed a kernel function that is a product of univariate kernel functions[1;13;15;24]. 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.

add remaining citations for first sentence in above paragraph

A limitation of kernel-based density estimation methods is that their performance may not scale well with the dimension of the vector whose distribution is being estimated (cite cite). 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. This simple dependence model ties separate marginal distribution estimates together into a joint distribution. In our case, we obtain those marginal distribution estimates through KCDE.

Add 1 sentence with citations for how copulas have been used in similar settings/ways before.

The remainder of this article is organized as follows. First, we describe our approach to prediction using KCDE and copulas. Next, we present the results of a simulation study comparing the performance of KCDE for estimating discrete distributions using a fully parameterized bandwidth matrix and a diagonal bandwidth matrix. We then illustrate our methods by applying them to predicting disease incidence in two data sets: one with a measure of weekly incidence of influenza in the United States and a second with a measure of weekly incidence of dengue fever in San Juan, Puerto Rico. We conclude with a discussion of these results.

Method Description

Suppose we observe a measure z_t of disease incidence at evenly spaced times indexed by $t = 1, \ldots, T$. Our goal is to obtain predictions relating to incidence after time T. We allow the incidence measure to be either continuous or discrete and use the term density to refer to the Radon-Nikodym derivative of a (conditional) probability measure with respect to an appropriately defined reference measure. We will use a colon notation to specify vectors: for example, $\mathbf{z}_{s:t} = (z_s, \ldots, z_t)$. The variable $t^* \in \{1, \ldots, T\}$ will be used to represent a time at which we desire to form a predictive distribution, using observed data up through t^* to predict incidence after t^*. When we apply the method to perform prediction for incidence after time T, t^* is equal to T; however, t^* takes other values in the estimation procedure we describe below. Let W denote the number of time points in a disease season (typically $W = 52$ if we have weekly data). For each time t^*, let S_{t^*} denote the time index of the last time point in the *previous* season, so that the times in the same season as t^* are indexed by $S_{t^*} + 1, \ldots, S_{t^*} + W$. Finally, let $H_{t^*} = W - (t^* - S_{t^*})$ denote the number of time points after t^* that are in the same season as t^*. H_{t^*} gives the largest prediction horizon for which we need to make a prediction in order to obtain predictions for all remaining time points in the season.

We obtain predictive distributions for each of three prediction targets. We will model the first of these prediction targets directly and frame the second and third as suitable integrals of a predictive distribution $f(\mathbf{z}_{(t^*+1):(t^*+H_{t^*})} | t^*, \mathbf{z}_{1:t^*})$ for the trajectory of incidence over all remaining weeks in the season:

1. Incidence in a single future week with prediction horizon $h \in \{1, ..., W\}$:

$$f(z_{t^*+h}|t^*, \mathbf{z}_{1:t^*})$$

2. Timing of the peak week of the current season, $w^* \in \{1, \ldots, W\}$:

$$\begin{aligned} P(\text{Peak Week} = w^*) &= P(Z_{S_{t^*}+w^*} = \max_w Z_{S_{t^*}+w}|t^*, \mathbf{z}_{1:t^*}) \\ &= \int_{\{\mathbf{z}_{(t^*+1):(t^*+H_{t^*})}: z_{S_{t^*}+w^*} = \max_w z_{S_{t^*}+w}\}} f(\mathbf{z}_{(t^*+1):(t^*+H_{t^*})}|t^*, \mathbf{z}_{1:t^*})\, d\mathbf{z}_{(t^*+1):(t^*+H_{t^*})}. \end{aligned} \tag{1}$$

3. Binned incidence in the peak week of the current season:

$$\begin{aligned} P(\text{Incidence in Peak Week} \in [a, b]) &= P(a \leq \max_w Z_{S_{t^*}+w} \leq b|t^*, \mathbf{z}_{1:t^*}) \\ &= \int_{\{(\mathbf{z}_{(t^*+1):(t^*+H_{t^*})}): a \leq \max_w z_{S_{t^*}+w} \leq b\}} f(\mathbf{z}_{(t^*+1):(t^*+H_{t^*})}|t^*, \mathbf{z}_{1:t^*})\, d\mathbf{z}_{(t^*+1):(t^*+H_{t^*})}. \end{aligned} \tag{2}$$

In practice, we use Monte Carlo integration to evaluate the integrals in Equations (1) and (2) by sampling incidence trajectories from the joint predictive distribution.

At time t^*, our model approximates $f(\mathbf{z}_{(t^*+1):(t^*+H_{t^*})}|t^*, \mathbf{z}_{1:t^*})$ by conditioning only 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 $l_1, \ldots, l_M$: $f(\mathbf{z}_{(t^*+1):(t^*+H_{t^*})}|t^*, z_{t^*-l_1}, \ldots, z_{t^*-l_M})$. For notational simplicity, we take l_M to be the largest of these lags. The model represents this density as follows:

$$\begin{aligned} &f(z_{(t^*+1):(t^*+H_{t^*})}|t^*, z_{t^*-l_1}, \ldots, z_{t^*-l_M}) = \\ &\quad c^{H_{t^*}}\{f^1(z_{t^*+1}|t^*, z_{t^*-l_1}, \ldots, z_{t^*-l_M}; \boldsymbol{\theta}^1), \ldots, f^{H_{t^*}}(z_{t^*+H_{t^*}}|t^*, z_{t^*-l_1}, \ldots, z_{t^*-l_M}; \boldsymbol{\theta}^H); \boldsymbol{\xi}^{H_{t^*}}\}. \end{aligned} \tag{3}$$

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 $c^{H_{t^*}}(\cdot)$ is a copula used to tie these marginal predictive densities together into a joint predictive density, and depends on parameters $\boldsymbol{\xi}^{H_{t^*}}$. In our applications, we will obtain a separate copula fit for each trajectory length H_{t^*} 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 $h = 1, \ldots, H_{t^*}$, and second we estimate the copula parameters while holding the KCDE parameters fixed. In general the two-stage approach may result in some loss of efficiency relative to one-stage methods, but this efficiency loss is small for some model specifications [11]. We pursue the two-stage strategy in this work because it results in a large reduction in the computational cost of parameter estimation.

The estimation methods were implemented in R[17] and C. All source code as well as the data we used in the applications below are available in R packages hosted on GitHub[18].

Permission to put data in our package? Also, current code just uses the flu data package in R to pull in data. Maybe I should get a stable/permanent copy to put in the package?

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 $t = 1 + l_M, \ldots, T - h$ (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:

$$\widehat{f}^h(y_{t^*}^h|\mathbf{x}_{t^*}) = \frac{\sum_{t\in\boldsymbol{\tau}} K^{\mathbf{X},Y}\left\{(\mathbf{x}_{t^*}, y_{t^*}^h), (\mathbf{x}_t, y_t^h); \boldsymbol{\theta}^h\right\}}{\sum_{t\in\boldsymbol{\tau}} K^{\mathbf{X}}(\mathbf{x}_{t^*}, \mathbf{x}_t; \boldsymbol{\theta}^h)} \tag{4}$$

$$= \sum_{t\in\boldsymbol{\tau}} \zeta_{t^*,t}^h K^{Y|\mathbf{X}}(y_{t^*}^h, y_t^h|\mathbf{x}_{t^*}, \mathbf{x}_t; \boldsymbol{\theta}^h), \text{ where} \tag{5}$$

$$\zeta_{t^*,t}^h = \frac{K^{\mathbf{X}}(\mathbf{x}_{t^*}, \mathbf{x}_t; \boldsymbol{\theta}^h)}{\sum_{s\in\boldsymbol{\tau}} K^{\mathbf{X}}(\mathbf{x}_{t^*}, \mathbf{x}_s; \boldsymbol{\theta}^h)}. \tag{6}$$

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 $(\mathbf{x}_t, y_t^h)$ 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 (4) through (6), $\boldsymbol{\tau} \subseteq \{(1 + l_M), \ldots, (T - h)\}$ 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, which are measures

of disease incidence:

$$K^{\mathbf{X},Y}\left\{(\mathbf{x}_{t^*}, y^h_{t^*}), (\mathbf{x}_t, y^h_t); \boldsymbol{\theta}^h\right\} = K^{\text{per}}(t^*, t; \boldsymbol{\theta}^h_{\text{per}}) K^{\text{inc}}\{(z_{t^*-l_1}, \ldots, z_{t^*-l_M}, z_{t^*+h}), (z_{t-l_1}, \ldots, z_{t-l_M}, z_{t+h}); \boldsymbol{\theta}^h_{\text{inc}}\}.$$

Here we have set $\boldsymbol{\theta}^h = (\boldsymbol{\theta}^h_{\text{per}}, \boldsymbol{\theta}^h_{\text{inc}})$.

The periodic kernel function was originally developed in the literature on Gaussian Processes[14], and is defined by

$$K^{\text{per}}(t^*, t; \rho^h, \eta^h) = \exp\left[-\frac{\sin^2\{\rho^h(t^* - t)\}}{2(\eta^h)^2}\right]. \tag{7}$$

We illustrate this kernel function in Figure 1. It has two parameters: $\boldsymbol{\theta}^h_{\text{per}} = (\rho^h, \eta^h)$, where ρ^h determines the length of the periodicity and η^h determines the strength and locality of this periodic component in computing the observation weights $\zeta^h_{t^*,t}$. In our applications, we have fixed $\rho^h = \pi/52$, 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^h_t. In our applications, these variables are measures of incidence; for brevity of notation, we collect them in the column vector $\tilde{\mathbf{z}}_t = (z_{t-l_1}, \ldots, z_{t-l_M}, z_{t+h})'$. These incidence measures 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 (Figure 1). 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. This kernel function has the following functional form:

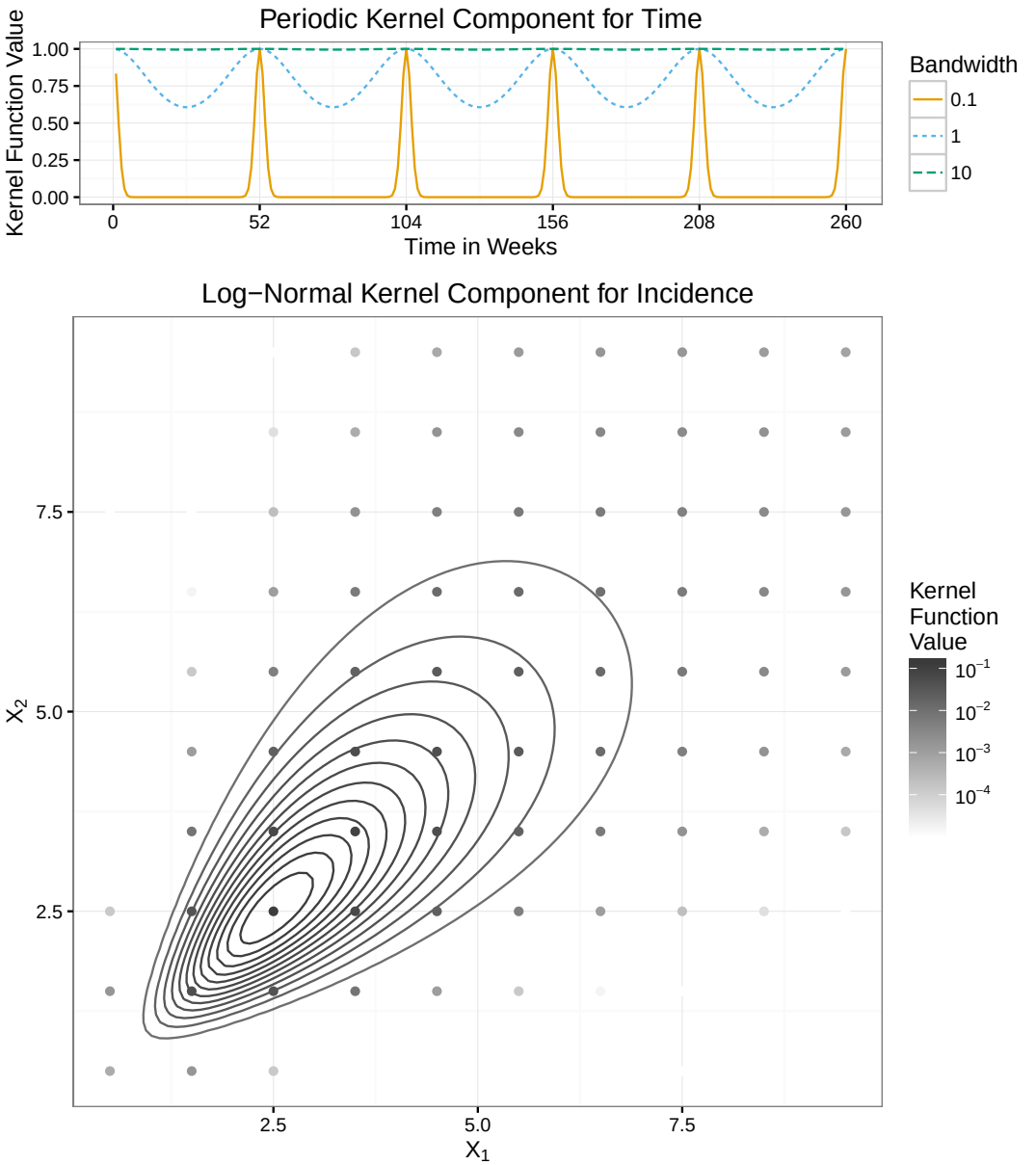
$$K^{\text{inc}}_{\text{cont}}(\tilde{\mathbf{z}}_{t^*}, \tilde{\mathbf{z}}_t; \mathbf{B}^h) = \frac{\exp\left[-\frac{1}{2}\{\log(\tilde{\mathbf{z}}_{t^*}) - \log(\tilde{\mathbf{z}}_t)\}'\mathbf{B}^{-1}\{\log(\tilde{\mathbf{z}}_{t^*}) - \log(\tilde{\mathbf{z}}_t)\}\right]}{(2\pi)^{\frac{M+1}{2}}|\mathbf{B}|^{\frac{1}{2}} z_{t^*+h}\prod_{m=1}^{M} z_{t^*-l_m}} \tag{8}$$

fix functional form for incidence kernel

In this expression, the log operator applied to a vector takes the log of each component of that vector. The matrix $\mathbf{B}$ is a bandwidth matrix that controls the orientation and scale of the kernel function. This bandwidth matrix is parameterized by $\boldsymbol{\theta}^h_{\text{inc}}$. In this work we have considered two parameterizations: a diagonal bandwidth matrix, and a fully parameterized bandwidth based on the Cholesky decomposition. In order to obtain the discrete kernel (Figure 1), we integrate an underlying continuous kernel function over hyper-rectangles containing the points in the range of the discrete random variable (see supplement for details).

We estimate the bandwidth parameters $\boldsymbol{\theta}^h$ by numerically maximizing the cross-validated log score of the predictive distributions for the observations in the training data. For a random variable Y with observed value y the log score of the predictive distribution f_Y is $log\{f_Y(y)\}$. A larger log score indicates better model performance. In obtaining the cross-validated log score for

Figure 1. The components of the kernel function. The top panel shows the periodic kernel function illustrated as a function of time in weeks with $\rho = \pi/52$ and three possible values for the bandwidth parameter η. The lower panel shows the log-normal kernel function in the bivariate case. The curves indicate contours of the continuous kernel function and the points indicate the discrete kernel function, which is obtained by integrating the continuous kernel function. The kernel is centered at $(2.5, 2.5)$ and has bandwidth matrix $\begin{bmatrix} 0.2 & 0.15 \\ 0.15 & 0.2 \end{bmatrix}$.



the predictive distribution at time t^*, we leave the year of training data before and after the time t^* out of the set $\boldsymbol{\tau}$ in Equations (4) through (6). Our primary motivation for using the log score as the optimization target during estimation is that this is the criteria that has been used to evaluate and compare prediction methods in two recent government-sponsored infectious disease prediction contests[3;16]. We will apply our method to the data sets from those competitions in the applications section below, and will report log scores in order to facilitate comparisons with other results from those competitions that may be published in the future. In general, the log score is a strictly proper scoring rule; i.e., its expectation is uniquely maximized by the true predictive distribution[5]. However, its use as an optimization criterion has been criticised for being sensitive to outliers[5]. In the kernel density estimation literature, this approach to estimation is referred to as likelihood cross-validation, and similar criticisms have been made regarding its performance in handling outliers and estimating heavy-tailed distributions[19;20].

Combining Marginal Predictive Distributions with Copulas

We use copulas (cite cite) to tie the marginal predictive distributions for individual prediction horizons obtained from KCDE together into a joint predictive distribution for the trajectory of incidence over multiple time points. In order to describe our methods for both continuous and discrete distributions, it is most convenient to frame the discussion in this section in terms of c.d.f.s instead of density functions. We will use a capital C to denote the copula function for c.d.f.s and a lower case c to denote the copula function for densities. Similarly, the predictive densities $f^h(y^h_{t^*}|\mathbf{x}_{t^*};\boldsymbol{\theta}^h)$ we obtained in the previous section naturally yield corresponding predictive c.d.f.s $F^h(y^h_{t^*}|\mathbf{x}_{t^*};\boldsymbol{\theta}^h)$.

Our model specifies the joint c.d.f. for $(Y^1_{t^*},\ldots,Y^{H_{t^*}}_{t^*})$ as follows:

$$\begin{aligned} &F^{H_{t^*}}(y^1_{t^*},\ldots,y^{H_{t^*}}_{t^*}|\mathbf{x}_{t^*};\boldsymbol{\theta}^1,\ldots,\boldsymbol{\theta}^{H_{t^*}},\boldsymbol{\xi}^{H_{t^*}}) = \\ &\quad C\{F^1(y^1_{t^*}|\mathbf{x}_{t^*};\boldsymbol{\theta}^1),\ldots,F^{H_{t^*}}(y^{H_{t^*}}_{t^*}|\mathbf{x}_{t^*};\boldsymbol{\theta}^{H_{t^*}});\boldsymbol{\xi}^{H_{t^*}}\} \end{aligned} \tag{9}$$

The copula function C maps the marginal c.d.f. values to the joint c.d.f. value. We use the isotropic normal copula implemented in the `R` package `copula`[9]. The copula function is given by

$$C(u_1,\ldots,u_H;\boldsymbol{\xi}^H) = \Phi_{\Sigma^H}(\Phi^{-1}(u_1),\ldots,\Phi^{-1}(u_H)), \tag{10}$$

where Φ^{-1} is the inverse c.d.f. of a univariate normal distribution with mean 0 and variance 1 and Φ_{Σ^H} is the c.d.f. of a multivariate Gaussian distribution with mean $\underline{0}$ and covariance matrix Σ^H. The isotropic specification sets $\Sigma^H = [\sigma^H_{i,j}]$, where

$$\sigma^H_{i,j} = \begin{cases} 1 \text{ if } i = j, \\ \xi^H_d \text{ if } |i-j| = d \end{cases} \tag{11}$$

Intuitively, ξ^H_d captures the amount of dependence between incidence levels at future times that are d weeks apart.

We obtain a separate copula fit for each value of H from 2 to W (note that a copula is not required for "trajectories" of length $H = 1$). In order to do this, we follow a two-stage estimation strategy[11]:

1. Estimate the parameters for marginal predictive distributions using the procedures described in the previous subsection.
2. Estimate the copula parameters, holding the parameters for the marginal predictive distributions fixed:
   (a) Form vectors of "pseudo-observations" by passing observed incidence trajectories from previous seasons through the marginal predictive c.d.f.s obtained in step 1:

$$(u_{k,1}, \ldots, u_{k,H}) = \{F^1(z_{t_k^*+1}|t_k^*, z_{t_k^*-l_1}, \ldots, z_{t_k^*-l_M}; \boldsymbol{\theta}^1), \ldots, F^H(z_{t_k^*+H}|t_k^*, z_{t_k^*-l_1}, \ldots, z_{t_k^*-l_M}; \boldsymbol{\theta}^H)\}$$

   We form one such vector of pseudo-observations for each season in the training data; in the notation here, these seasons are indexed by k. The relevant time points t_k^* are the times in those previous seasons falling H time steps before the end of the season.
   (b) Estimate the copula parameters $\boldsymbol{\xi}^H$ by maximizing the likelihood of the pseudo-observations.

Simulation Study

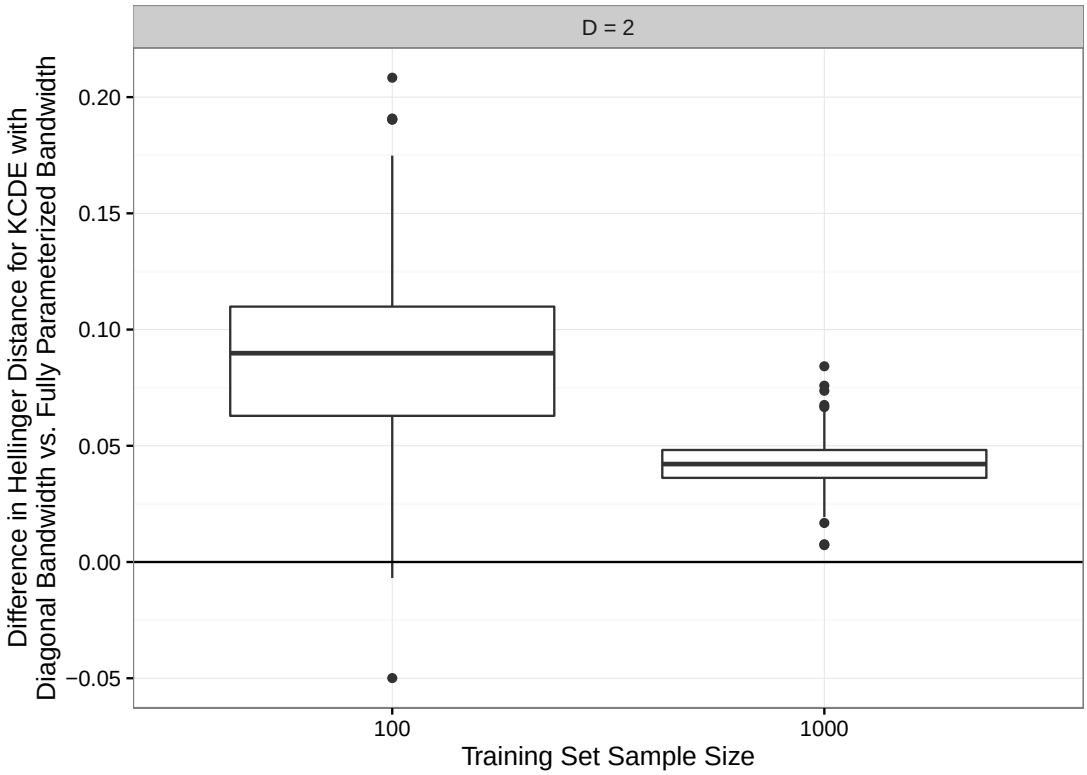
We conducted a simulation study to examine the utility of using a non-diagonal bandwidth matrix specification when estimating conditional distributions with KCDE. There are many factors that determine the relative performance of KCDE estimators with different bandwidth parameterizations. In this simulation study, we vary just two of these factors: the number of conditioning variables (either 1 or 3) and the sample size ($N = 100$ or $N = 1000$). We hold other factors that may be related to the relative performance of different bandwidth specifications fixed.

The distributions that we simulate from are discretized multivariate normal distributions of dimension either $D = 2$ or $D = 4$. To define this distribution, let $\mathbf{U} \sim MVN(\underline{0}, \Sigma)$ where Σ is a $D \times D$ matrix with 1 on the diagonal and 0.9 off of the diagonal. We treat $\mathbf{U}$ as a latent variable and discretize it to obtain the random variable $\mathbf{X}$ using the approach described in the supplement.

We conducted 500 simulation trials for each combination of the sample size N and dimension D. In each trial, we simulated N observations of the discretized multivariate normal random variable $\mathbf{X}$. Using these observations as a training data set, we estimated the bandwidth parameters for two variations on a KCDE model for the conditional distribution of $X_1|X_2, \ldots, X_D$: one with a diagonal bandwidth matrix specification and one with a fully parameterized bandwidth matrix. In this simulation study, the kernel function was obtained by discretizing a multivariate normal kernel function rather than a log-normal kernel function as in our applications below. Otherwise, the method is as described previously.

We evaluated the conditional density estimates by an importance sampling approximation of the Hellinger distance of the conditional density estimate from the true conditional density, integrated over the range of the conditioning variables (see supplement). The Hellinger distance lies between 0 and 1, with smaller values indicating that the density estimate is better. It has been argued that the Hellinger distance is preferred to other measures of the quality of kernel

Figure 2. Results from the simulation study. The top facet of the plot shows results from simulation trials with dimension $D = 2$, and the lower facet has results from dimension $D = 4$. Positive values indicate simulation trials where the full bandwidth specification outperformed the diagonal bandwidth specification with the same training data set, as measured by Hellinger distance from the target conditional density.



density estimates such as integrated squared error[12]. For each combination of the training set sample size, dimension, and simulation trial, we compute the difference between the Hellinger distance from the true conditional distribution achieved with a diagonal bandwidth matrix and with a fully parameterized bandwidth matrix.

The results indicate that using a fully parameterized bandwidth matrix instead of a diagonal bandwidth generally yields improved density estimates as measured by the integrated Hellinger distance (Figure 2). The average improvement from using a fully parameterized bandwidth matrix is larger with a sample size of $N = 100$ instead of $N = 1000$, but there is also more variation in performance with the smaller sample size.

To do: update simulation study results with results from 4 dimensional case. Put in a sentence here about whether or not the simulation study results are the same with $D = 4$.

Applications

In this Section, we illustrate our methods through applications to prediction of infectious disease incidence in two examples with real disease incidence data sets: one with a weekly measure of incidence of dengue fever in San Juan, Puerto Rico, and a second with a weekly measure of incidence of influenza like illness in the United States. These data sets were used in two recent prediction competitions sponsored by the United States federal government [3;16]. In the dengue data set, the incidence measure is an integer number of reported cases in the given week. In the influenza data set the incidence measure is continuous, a weighted proportion of doctor visits with influenza-like illness.

Figure 3 displays each time series. 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. All predictions are made as though in real time assuming that once cases are reported, they are never revised. Specifically, we use only data up through a given week in order to make predictions for incidence after that week.

There are three prediction targets for each data set, based closely on the prediction targets that were used in the original 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 week of the test data set, we make predictions for the timing of the peak week of the corresponding season. Third, in each week of the test data set we predict incidence in the peak week for the corresponding season. Following the precedent set in the competitions, we make predictions for *binned* incidence in the peak week. For the dengue data set, the bins are $[0, 50), [50, 100), \ldots, [500, \infty)$. For the influenza data set, the bins are $[0, 0.5), [0.5, 1), \ldots, [13, \infty)$. Our predictions for incidence in individual weeks are for the raw, unbinned, incidence measure. These prediction targets are illustrated in the supplement.

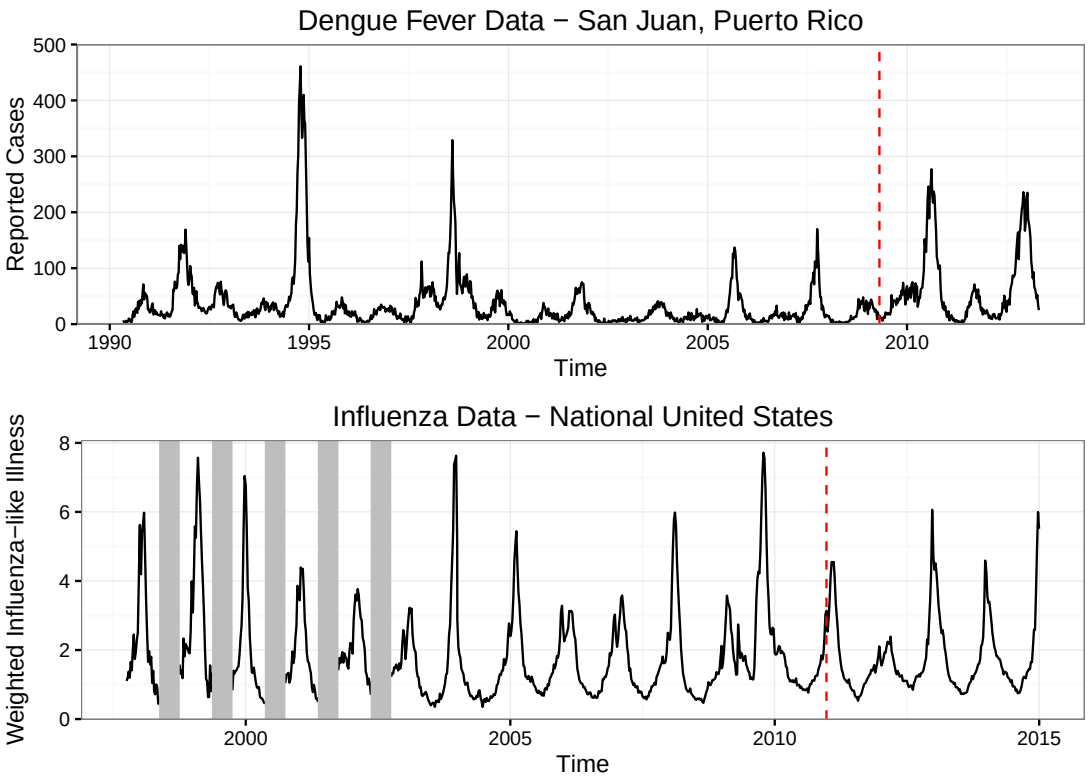
Our applications include four variations on KCDE model specifications:

1. The "Null KCDE Model" omits the periodic component of the kernel function and uses a diagonal bandwidth matrix specification.
2. The "Full Bandwidth KCDE Model" omits the periodic component of the kernel function and uses a fully parameterized bandwidth matrix specification.
3. The "Periodic KCDE Model" includes the periodic component of the kernel function and uses a diagonal bandwidth matrix specifications.
4. The "Periodic, Full Bandwidth KCDE Model" includes the periodic component of the kernel function and uses a fully parameterized bandwidth matrix specification.

We use a seasonal autoregressive integrated moving average (SARIMA) model for the log-transformed incidence measure as a baseline to compare our methods to. We obtained 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. Details of the SARIMA model specification and estimation procedure are in the supplemental materials.

We begin with a discussion of predictive distributions for incidence at individual time points. Figure 4 displays the median and 95% interval limits for the predictive distributions obtained at prediction horizons of 1, 6, or 26 weeks from SARIMA and from the KCDE specification with

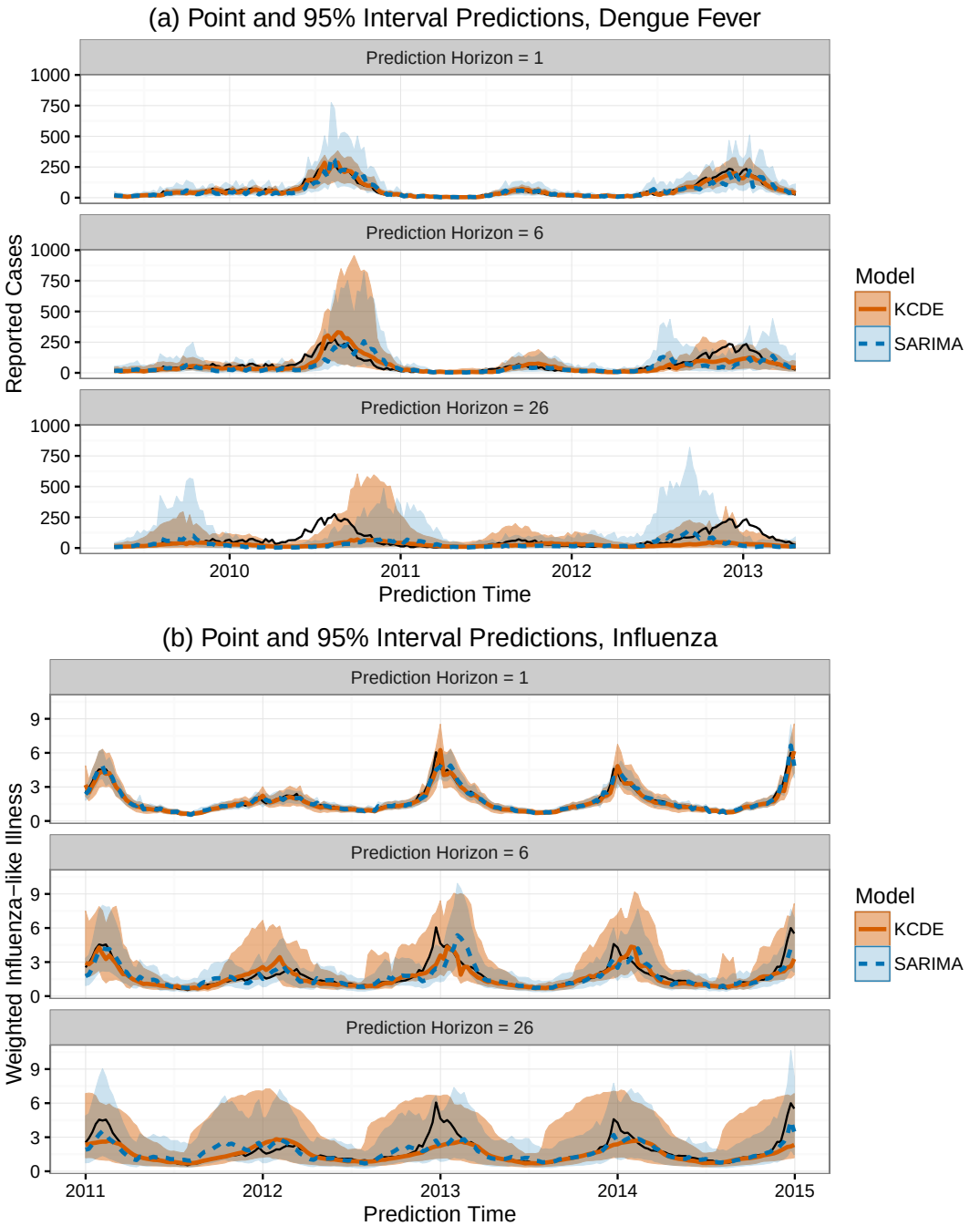
Figure 3. Plots of the data sets we apply our methods to. In each case, the last four years of data are held out as a test data set; this cutoff is indicated with a vertical dashed line. For the flu data set, low-season incidence was not recorded in early years of data collection; these missing data are indicated with vertical grey bars.



a fully parameterized bandwidth matrix and a periodic kernel component. For predictions of dengue fever incidence, at a prediction horizon of one week the point predictions from SARIMA and KCDE are similar, but the distribution from KCDE is much more concentrated around its center. Both methods struggle with larger prediction horizons, but it appears that the SARIMA model has more difficulty with aligning the predictive distribution with the season's peak, particularly in the two seasons with higher incidence. For the predictions of influenza incidence, which exhibits more regular seasonality, there is much less of a noticeable distinction between the predictions given by the two methods. Throughout, the point predictions and intervals are similar.

Figure 5 offers a more quantitative summary of these results in terms of log scores. In the application to predicting dengue fever, the KCDE specifications including a periodic kernel component had slightly larger log scores than the SARIMA model on average, with many outlying

Figure 4. Plots of point and interval predictions from SARIMA and the KCDE specification with a fully parameterized bandwidth and periodic kernel component.



Prepared using sagej.cls

cases where KCDE did much better than SARIMA. For the specifications without the periodic kernel component, the median performance was similar to the SARIMA model, but again there were outlying cases where KCDE did much better than SARIMA. Figure 5 (b) shows that KCDE particularly outperformed SARIMA in its predictions for times of high incidence near the season peaks. In the application to predicting influenza, the KCDE specifications including a periodic kernel did about as well as the SARIMA model, while the median performance of the specifications without a periodic kernel was slightly worse than SARIMA. In both applications, including the periodic kernel component led to improved predictions; this can be seen directly in Figure *** of the supplement. Using the fully parameterized bandwidth matrix generally had little impact on the quality of the predictive distributions as measured by the log score.

Figure 6 displays the log score of the predictive distributions for incidence in the peak week obtained from SARIMA and KCDE models over the course of each season in the test data sets, and Figure *** in the supplement displays log scores for predictions of peak week timing. For both of these prediction targets, there is no consistent pattern of KCDE either outperforming or underperforming relative to SARIMA.

Conclusions

Prediction of infectious disease incidence at horizons of more than a few weeks is a challenging task. We have presented a non-parametric approach to doing this based on KCDE and copulas and found that it is a viable method that can yield improved predictions relative to commonly employed methods in this field. In predicting incidence of dengue fever in individual weeks, we saw that our approach offered consistent and substantial performance gains relative to a SARIMA model. These improvements were particularly concentrated in the times that are of most interest to public health decision makers: periods of high incidence near the season peak. For predicting influenza-like illness, our method did about as well as SARIMA when predicting incidence in individual weeks.

We believe that the difference in relative performance of KCDE and SARIMA for prediction in the dengue and influenza data sets can be explained to a great extent by differences in the underlying disease processes and how they relate to the differing model specifications. The most salient difference between the two time series depicted in Figure 3 is the much greater season-to-season variablility in the dengue data set relative to the influenza data set. For dengue, the peak incidence in the largest season is about 30 times larger than the peak incidence in the smallest season; this ratio is only about 3 for influenza. It may be the case that the restrictive linear structure of the SARIMA model means that it is not able to capture the dynamics of dengue incidence accurately. Relaxing that structure by using a nonparametric approach such as KCDE may yield improved capability to represent the disease dynamics. This is less of an issue in predicting influenza where there is much more consistency across different seasons.

Another more subtle effect is present in the influenza data: there is a consistent short-term peak in influenza incidence on Christmas week. This is visible in Figure 3, and is highlighted in Figure *** in the supplementary materials. This “Christmas effect” sometimes coincides with the season peak, but sometimes occurs before the season peak. We have observed evidence that the seasonal structure of the SARIMA model picks up on this structure, and SARIMA tended

Figure 5. Differences in log scores for the weekly predictive distributions obtained from KCDE specifications and SARIMA. For reference, a log score difference of 2.3 (4.6) indicates that the predictive density from KCDE was about 10 (100) times as large as the predictive density from SARIMA at the realized outcome. In Panel (a) we summarize the results across all combinations of prediction horizon and prediction time in the test period. In Panel (b) we display more detailed results for just the application to predicting dengue fever and the KCDE specification with a periodic kernel component and fully parameterized bandwidth. Each point corresponds to a unique combination of prediction target week and prediction horizon.

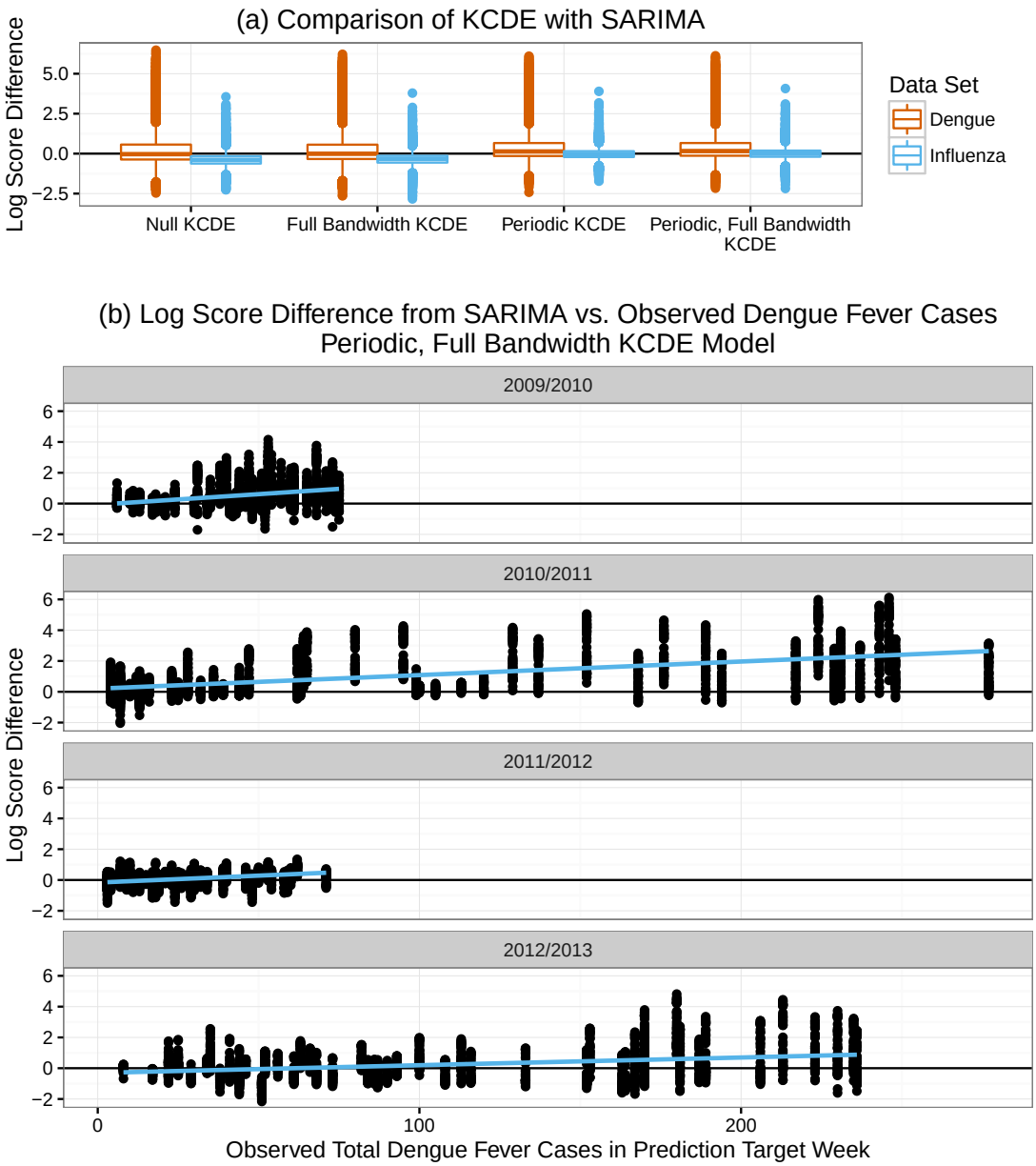
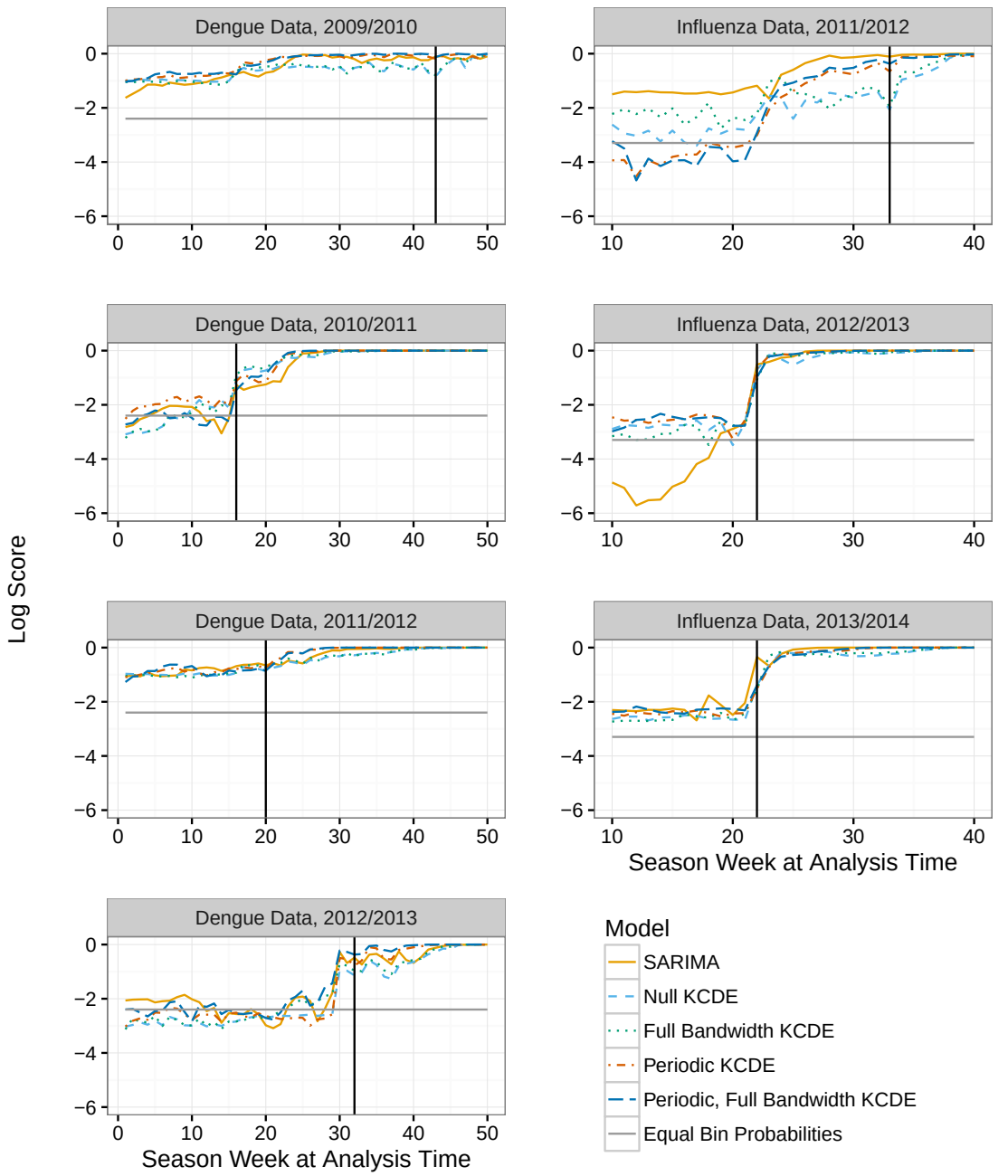


Figure 6. Log scores for predictions of peak week incidence by predictive model and analysis time. The vertical line is placed at the peak week for each season. The log score for "Equal Bin Probabilities" is obtained by assigning equal probability that the peak incidence will be in each of the specified incidence bins. There are 11 incidence bins for dengue and 27 bins for influenza.



to outperform KCDE on Christmas week and the weeks immediately thereafter on the influenza data set. We believe that it would be possible to construct a variation on KCDE that captures this effect, for example by including indicator variables for the weeks around Christmas as conditioning variables. However, we have not explored that avenue in this work.

We have also demonstrated that it is feasible to use KCDE in combination with copulas to obtain predictions for the timing of and incidence in the week of the season with the highest incidence. For those prediction targets, our method was competitive with SARIMA; we did better in some of the test seasons and worse in others, with no clear indication that either model was better than the other.

One explanation for the difference in relative performance of the methods on these different prediction tasks may lie in the connection between the objective function used in parameter estimation and the prediction task. We estimated the bandwidth parameters for KCDE by optimizing the log score of predictive distributions for incidence in individual weeks in the training data set. This is the prediction target where KCDE outperformed SARIMA on the dengue data set. It may be the case that performance on the other two prediction tasks could be improved by implementing a combined one-stage estimation strategy for both the KCDE and copula parameters that optimizes a measure of performance on the specific prediction task at hand.

Our implementation of KCDE offers two main methodological improvements relative to previous implementations. Most importantly in the context of modeling infectious disease, we have introduced the use of a periodic kernel component that captures seasonality. In both of our applications, including this periodic kernel component in the KCDE specification led to substantial improvements in the predictive distributions for incidence in individual weeks. We also introduced a method for obtaining kernel functions that are appropriate for use with discrete data while allowing for a fully parameterized bandwidth matrix. In our applications, using a fully parameterized bandwidth matrix did not lead to consistent improvements in predictions. However, we have demonstrated through a simulation study that the fully parameterized bandwidth can be helpful in some conditional density estimation tasks. This general method for obtaining discrete kernel functions may be beneficial in other applications of KCDE.

There is a great deal of room for extensions and improvements to the methods we have outlined in this article. One major limitation of our work lies in the selection of conditioning variables for the predictive model. We have simply used incidence at the two most recent time points, and possibly the observation time, as conditioning variables. We considered using a stepwise variable selection approach to select the model specification as in (cite ***), but we found this to be too computationally expensive to be practical.

Another possibility for addressing this problem would be to replace variable selection with shrinkage. Hall *et al.*[6] 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. We conjecture that by introducing an appropriate penalty on the elements bandwidth matrix, we could include more (possibly irrelevant) conditioning variables in the model without requiring a dramatically larger sample size. In particular, we hypothesize that a penalty on the inverse of the bandwidth matrix encouraging it to have small eigenvalues could be helpful. If successful,

this would also enable further exploration of using other predictive variables (such as weather) in the model.

Another aspect of our method that should be explored further is the use of log score in estimation. We used log scores in this work in order to match the use of log scores in evaluating and comparing the performance of different models. The log score has the advantage of defining a proper scoring rule, but it has the disadvantage of being sensitive to outlying values. Previous authors have suggested the use of other loss functions in estimation for kernel-based density estimation methods that reduce these effects, such as variations on integrated squared error that were used by **** cite cite cite.

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 Brown et al.[2] for a recent review. KCDE is distinguished from these approaches in that it makes minimal assumptions about the data generating process. This can be either an advantage or a disadvantage of KCDE. In general, we would expect a well-specified parametric model to outperform KCDE. On the other hand, because non-parametric approaches such as KCDE make fewer assumptions about the data generating process, 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 models that are considered as alternatives.

FIX CITATION INFO FOR LEXI'S REVIEW PAPER OR FIND AN ALTERNATIVE

However, rather than selecting one "preferred" modeling framework or model formulation, we believe it may be fruitful to incorporate the models developed in this paper as components of an ensemble method with several different types of models. For example, in our application to influenza, we saw that the SARIMA model captured some features of the data generating process, such as the Christmas-week effect, that KCDE did not capture. On the other hand, the KCDE approach was more flexible and yielded better predictions than SARIMA at other times – most notably, in periods of high incidence in the application to dengue. An appropriately constructed ensemble incorporating predictions from both SARIMA and KCDE as well as other methods such as mechanistic models might perform better than any of these models on its own, and would be a valuable approach for maximizing the utility of these predictions to public health decision makers.

References

1. John Aitchison and Colin GG Aitken. Multivariate binary discrimination by the kernel method. *Biometrika*, 63(3):413–420, 1976.
2. Alexandria Brown, Stephen A. Lauer, Evan L. Ray, Xi Meng, and Nicholas G. Reich. A systematic review of prediction for infectious disease. *Journal*, submitted.
3. Epidemic Prediction Initiative. FluSight: Seasonal Influenza Forecasting, January 2016. URL `http://dengueforecasting.noaa.gov/`.
4. Jianqing Fan and Tsz Ho Yim. A crossvalidation method for estimating conditional densities. *Biometrika*, 91(4):819–834, 2004.

5. Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.
6. Peter Hall, Jeff Racine, and Qi Li. Cross-validation and the estimation of conditional probability densities. *Journal of the American Statistical Association*, 99(468):1015–1026, 2004.
7. Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Science + Business Media, 2 edition, 2009.
8. Richard J Hatchett, Carter E Mecher, and Marc Lipsitch. Public health interventions and epidemic intensity during the 1918 influenza pandemic. *Proceedings of the National Academy of Sciences*, 104 (18):7582–7587, 2007.
9. Marius Hofert, Ivan Kojadinovic, Martin Maechler, and Jun Yan. *copula: Multivariate Dependence with Copulas*, 2015. URL `http://CRAN.R-project.org/package=copula`. R package version 0.999-14.
10. Jooyoung Jeon and James W Taylor. Using conditional kernel density estimation for wind power density forecasting. *Journal of the American Statistical Association*, 107(497):66–79, 2012.
11. Harry Joe. Asymptotic efficiency of the two-stage estimation method for copula-based models. *Journal of Multivariate Analysis*, 94(2):401–419, 2005.
12. Yuichiro Kanazawa. Hellinger distance and kullbackleibler loss for the kernel density estimator. *Statistics & probability letters*, 18(4):315–321, 1993.
13. Qi Li and Jeff Racine. Nonparametric estimation of distributions with categorical and continuous data. *journal of multivariate analysis*, 86(2):266–292, 2003.
14. David JC MacKay. Introduction to gaussian processes. *NATO ASI Series F Computer and Systems Sciences*, 168:133–166, 1998.
15. Desheng Ouyang, Qi Li, and Jeffrey Racine. Cross-validation and the estimation of probability distributions with categorical data. *Journal of Nonparametric Statistics*, 18(1):69–100, 2006.
16. Pandemic Prediction and Forecasting Science and Technology Interagency Working Group. Dengue Forecasting, July 2015. URL `http://dengueforecasting.noaa.gov/`.
17. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2016. URL `https://www.R-project.org/`.
18. Evan Ray, Krzysztof Sakrejda, Stephen A. Lauer, and Nicholas G. Reich. The Reich Lab at UMass-Amherst, May 2016. URL `https://github.com/reichlab/article-disease-pred-with-kcde`.
19. Eugene F Schuster and Gavin G Gregory. On the nonconsistency of maximum likelihood nonparametric density estimators. In *Computer Science and Statistics: Proceedings of the 13th Symposium on the interface*, pages 295–298. Springer, 1981.
20. David W Scott and Lynette E Factor. Monte carlo study of three data-based nonparametric probability density estimators. *Journal of the American Statistical Association*, 76(373):9–15, 1981.
21. George Sugihara and Robert M. May. Nonlinear forecasting as a way of distinguishing chaos from measurement error in time series, April 1990.
22. Cécile Viboud, Pierre-Yves Boëlle, Fabrice Carrat, Alain-Jacques Valleron, and Antoine Flahault. Prediction of the spread of influenza epidemics by the method of analogues. *American Journal of Epidemiology*, 158(10):996–1006, 2003.
23. Jacco Wallinga, Michiel van Boven, and Marc Lipsitch. Optimizing infectious disease interventions during an emerging epidemic. *Proceedings of the National Academy of Sciences*, 107(2):923–928, 2010.

24. Min-Chiang Wang and John van Ryzin. A class of smooth estimators for discrete distributions. *Biometrika*, 68(1):301–309, 1981.
25. Haiming Zhou, Timothy Hanson, and Roland Knapp. Marginal bayesian nonparametric model for time to disease arrival of threatened amphibian populations. *Biometrics*, 71(4):1101–1110, 2015.