XI Congreso Galego de Estatística e Investigación de Operacións
A Coruña, 24–25–26 de outubro de 2013

Dependent Functional Regression Models for detecting influenza epidemics

Manuel Oviedo de la Fuente[1], Manuel Febrero-Bande[1] and M. Pilar Muñoz Gracia[2]

[1]Department of Statistics and Operations Research, Universidade de Santiago de Compostela
[2]Department of Statistics and Operations Research, Universitat Politcnica de Catalunya

ABSTRACT

The objective of this work is to estimate the space-time components of a contagious disease, such as the flu, using functional linear models (FLM). We study the conditions for the fit of the parameters of FLM by analyzing spatial dependence (each curve is indexed in a grid), or temporal dependence (each curve is indexed in time) and evaluate the forecasting performance.

keywords: Linear Models, influenza, count data, basis representation, epidemiological surveillance, least squares, functional dependent data.

1. INTRODUCTION

Surveillance systems must critically have precise indicators that detect in advance various epidemics that may occur. One of the problems that concern epidemiologists the most on a global level is the outbreak of flu, both for sick leave caused by this illness as well as the number of deaths. In addition, this disease is highly contagious. To have accurate estimates of the incidence of flu is vitally important both for public health services and citizens in general. This information should allow patients to be informed of possible contagions in advance while also reducing the spred of possible contagions. Controlling this contagion is very important because influenza causes more morbidity than any other vaccine-preventable illness Monto et al. (2002).

The statistical methodology for forecasting incidences of influenza in particular and contagious diseases in general has changed over time. One of the first works on time series was that proposed by Choi and Tacker (1981), where they fit an ARIMA model for estimating pneumonia and influenza mortality in order to know the number of deaths caused by these diseases. Recently, Dushoff et al. (2006) investigated how cold temperatures contribute to excess seasonal mortality through a regression model and Muñoz et al. (2011) studied the relationships between influenza morbidity and all causes of mortality, taking into account influenza vaccination coverage.

In order to monitor infectious diseases, an alternative to these procedures is that proposed by Hohle and Paul (2008), which consists of applying count data charts to monitor these kinds of time series. An approach that could also be applied to infectious diseases is to take into account the geographical component in statistical models in addition to the temporal evolution, i.e., how far apart or close together the detected cases are. The number of contributions to this methodology is growing steadily through disease mapping. As examples, we can cite the recent papers of Ugarte et al (2010) and Paul and Held (2011). The common denominator to all of them is that they apply different statistical methodologies to multivariate time series (hierarchical Bayesian space-time, mixed models, P–splines, conditional autoregressive (CAR), and a long etcetera) of infectious disease counts, collected in different geographic areas.

Functional data analysis (FDA) is a very active area of research in recent years. As a starting point, we discuss the functional linear models (FLM) treated in the literature by Ramsay and Silverman, (2005). In recent years, several methods have been proposed for analyzing high-dimensional data; some of them fall under the category of functional data analysis, and only a few of them believe this functional data presents either spatial Delicado et al (2010) or temporal Damon and Guillas (2005) dependencies. This work introduces a functional model which includes the two components, spatial and temporal.

The objective of this work is to estimate the space-time components of a contagious disease, such as the flu, using functional regression models and to predict the rate of incidence of this disease out of sample for horizons of one week to a quarter of a year. In particularity, we are inter

2. Methodology: Functional Regression Models

Regression models are those techniques for modeling and analyzing the relationship between a dependent variable and one or more independent variables. When one of the variables have a functional nature, we have functional regression models. This section is devoted to all the functional regression models where the response variable y is scalar and at least, there is one functional covariate $X(t)$.

Let $X(t)$ be a functional r.v. taking values in $\mathcal{H} = \mathcal{L}_2[0, T]$. In the functional linear model (FLM), the scalar response y is modeled as a linear function of covariate $X(t)$ as follows,

$$y = \langle X(t), \beta(t) \rangle + \epsilon \tag{1}$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product on $\mathcal{L}_2$ and ϵ is the error term with covariance matrix Σ. The estimation of the functional parameter β is done by minimizing the RSS:

$$\hat{\beta} = arg\ min_{\beta \in \mathcal{H}} \sum_{i=1}^{n} \left(Y_i - \langle X_i, \beta \rangle\right)^2 .$$

Different methods have been proposed to search for the $\hat{\beta}$ that minimizes the RSS (see Ferraty and Romain (2011)).

The general form of Σ is given by a function of unknown parameter θ, $\Sigma = \sigma^2 \Sigma(\theta)$. If Σ is known, the best linear unbiased estimator (BLUE) of functional parameter $\beta(t)$ is given by a functional version of the generalized least squares estimator (GLSE) of β

$$\hat{\beta} = \left(\tilde{X}^{\top} \Sigma^{-1} \tilde{X}\right)^{-1} \tilde{X}^{\top} \Sigma^{-1} y \tag{2}$$

where $\tilde{\mathbf{X}}$ is the representation of functional data $X(t)$ in a basis expansion. Note that, if $\Sigma = I$, the model includes the classical functional linear regression models (FLM) which entails independent errors structure, i.e., the GLSE in equation 2 reduces to the ordinary least squares estimator (OLSE).

3. Correlation structures for Σ

An extension of the FLM model for dependent data require to study the structure of Σ: analyzing temporal dependence (each curve is indexed in time), spatial dependence (each curve is indexed in a grid) or both.

Serial correlation structures

In temporal dependences typically assumes that errors are observed at integer time points. A common method of modeling serial correlation structures is to estimate an autoregressive moving average model ARMA of the errors.

$$\epsilon_t = \phi \epsilon_{t-1} + \ldots + \phi^q \epsilon_{t-p} + a_s + \theta a_{t-1} + \ldots + \theta^q a_{t-q}$$

where a_t are a independent and centered errors, $E\left[a_t\right] = 0$, $Var\left[a_t\right] = \sigma_a^2$.

In this case, the covariance matrix of the errors $\Sigma = \sigma^2 diag\left(\Sigma(\Phi)\right)$ depends of the $p+q$ parameters $\Phi = \{\phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q\}$. For example, we can model serial correlation using the simplest and useful AR(1) model. Its covariance matrix is $\Sigma = \sigma^2 \Sigma(\phi)$ where $\Sigma_{i,j} = \phi^{|i-j|}$.

Spatial Correlation Structures

This structure appear when the data are measured at spatial location vector, to simplify the data are in a 2-dimensional vector (x, y). There estimation of spatial correlation structures is based on shape of semivariogram models as Exponential, Spherical, Linear, among others, see Cressie, (1993). For example, the exponential semivariogram $\gamma(\cdot)$ without nugget effect is given by $\gamma(\sigma^2,\rho)=\sigma^2\left(1-e^{s/\rho}\right)$ where σ^2 is the sill, ρ is the range (correlation parameter) and s the distance $s_{i,j}=dist\left((x_i,y_i)-(x_j,y_j)\right)$. The semivariogram is related the covariance matrix of the errors as $\gamma\left(\cdot\right)=\sigma^2 I-\Sigma$ so in this case we need know $\left(\sigma^2,\rho\right)$ to compose $\Sigma=\sigma^2 e^{s/\rho}$.

Group covariance structures

The correlation structures are used to model dependence among the within-group errors. Now, observations in different groups are assumed independent, so Σ is a block diagonal matrix of the correlation matrices per group, $\Sigma=\sigma^2 diag\left(\Sigma_1(\Phi),\ldots,\Sigma_k(\Phi)\right)$.

Heterogeneous group covariance structures

This structure gives rise to independent observations, however with different variability according group level since different levels are mutually independent. The heteroscedastic model require know different variances for each level of the grouping variable. The covariance matrix of errors is $\Sigma=diag\left(\sigma_1^2\Sigma_1(\Phi),\ldots,\sigma_k^2\Sigma_k(\Phi)\right)$, where σ_i^2 is the variance of the group level i. We can also model different dependence parameter by group level $\Sigma=diag\left(\sigma_1^2\Sigma_1(\Phi^1),\ldots,\sigma_k^2\Sigma_k(\Phi^k)\right)$.

3. Illustrative example

We analyze the weekly influenza reported at $s=53$ regions during the period $2001-2011$ in Galicia. The data has been obtained from the Galician Influenza Surveillance Program. For each region, temperatures (downloaded from http://www.meteogalicia.es/) and incidence rate of flu of previous days and weeks respectively are used as covariates, see Figure 1. Figure 2 shows the influenza rate for epidemic period and non–epidemic period.

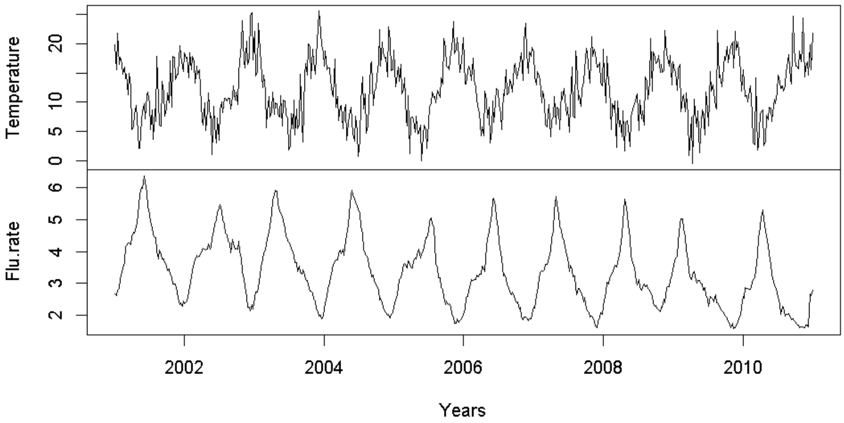

Figure 1: Weekly average temperature (up) and influenza rate (down) on the Galician regions.

We apply the functional dynamic models for detecting the outbreak of influenza, where the functional regression model is required to use influenza circulation data at observed time t in region s. We use the first n times to fit the model. The proposed model estimates the incidence in the region s and time $n+5$, $Rate_{n+5,s}$. The explanatory variables to introduce in the models are: the incidence of previous weeks $Rate_{n,s}(t)=(Rate_{n-13,s},\ldots,Rate_{n,s})$, the temperature and it first derivative of the previous 14 days $Temp(s), Temp.d1(s)$ respectively. Figure 3 shows the estimation of the functional parameter β for each covariate. All functional covariate effects are significant. The incidence rate is increased by reason of the last values of $\hat{\beta}_{rate}$ (approximately the last 3 weeks). When the temperature decreases the incidence rate increases.

The next 52 times has been used to check the predictions for model which assumes independent errors and autoregressive errors AR(1) respectively, see table 1. Akaike information criterion (AIC),

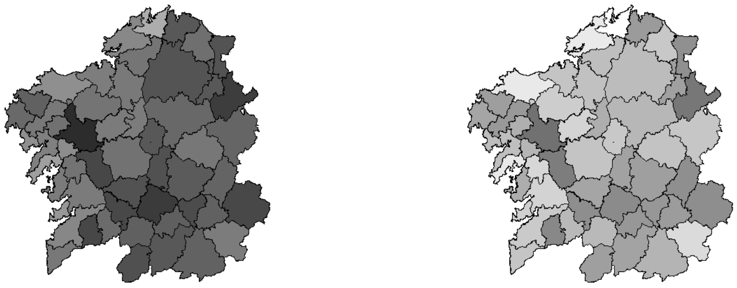

Figure 2: In left, influenza rate for epidemic period (weeks $40-20$) and in right, non–epidemic period (weeks $21-39$).

Model	AR(1)	*df*	AIC	MSE	MRE
$Rate_{t+5,s} \sim Rate_{n,s}(t) + Temp(s) + Temp.d1(s)$	–	16	20248	2.23	0.33
$Rate_{t+5,s} \sim Rate_{n,s}(t) + Temp(s) + Temp.d1(s)$	$\phi = 0.77$	17	16482	1.12	0.21

Table 1: Predictions errors for functional regression models.

mean square error of prediction (MSE) and mean relative error of prediction (MRE) have been computed for each model.

4. Conclusions

We propose a flexible and generic procedure to model influenza using functional models. This work considers the extension of functional linear models with independent errors to dependent errors. We study how to incorporate the temporal and spatial dependence structures into functional model and we evaluate the forecasting performance. Furthermore, the results might be more precise, if we introduce other epidemiological data such as data of influenza virus type, vaccination status, age or gender of infected, for example.

Acknowledgments

This work was supported by grants MTM2008-03010 from the Ministerio de Ciencia e Innovación, 10MDS207015PR from the Xunta de Galicia and GI-1914 MODESTYA-Modelos de optimización, decisión, estadística y aplicaciones.

REFERENCES

Efron, B. and Tibshirani, R. (1994) An introduction to the Bootstrap. Chapman & Hall.

Monto, A, Pichichero, M. Blanckenberg, S. et al. *Zanamivir Prophylaxis: An Effective Strategy for the Prevention of Influenza Types A and B within Households.* J Infect Dis. (2002) 186(11), 1582–1588.

Choi, K. adn Thacker, S. (1981) *An evaluation of influenza mortality surveillance, 1962–1979.* Am J Epidemiol 113, 215–26.

Dushoff, J., Plotkin, J.B., Viboud, C, Earn, D. and Simonsen, L. (2006). *Mortality due to Influenza in the United States, An Annualized Regression Approach Using Multiple Cause Mortality Data.* Am J Epidemiol, 163, 181-187.

Muñoz, M.P., Soldevila, N., Martnez, A., Carmona, G., Batalla, J., Acosta, L. and Domnguez, A. (2011) *Influenza vaccine coverage, influenza-associated morbidity and all-cause mortality in Catalonia (Spain).* Vaccine 29, 5047–5052.

Hohle, M. and Paul, M. (2008) *Count data regression charts for the monitoring of surveillance time series.* Computational Statistics and Data Analysis 52, 4357-68.

Ugarte, M.D., Goicoa, T. and Militino, A.F. (2010) *Spatio-temporal modeling of mortality risks using penalized splines.* Environmetrics, 21, 270–289.

Paul, M. and Held, L. *Predictive assessment of a non-linear random effects model for multivariate time series of infectious disease counts.* Statistics in Medicine (2011), 30, 1118–1136.

Delicado,P. ,Giraldo, R., Comas C. and Mateu J. (2010). *Statistics for spatial functional data: some recent contributions.* Environmetrics 21, 224–239.

Damon, J. and Guillas, S. (2005) *Estimation and Simulation of Autoregressive Hilbertian Processes with Exogenous Variables.* Statistical Inference for Stochastic Processes 8, 185-204.

Febrero Bande, M. and Oviedo de la Fuente, M. (2012). *Statistical Computing in Functional Data Analysis: The R Package fda.usc*, Journal of Statistical Software,51(4), 1–28,
http://www.jstatsoft.org/v51/i04/.