Research Statement

Bree Ettinger

1. Introduction

My research is in the application of multivariate splines to functional linear models. My interests include spatial and autoregressive functional linear models, time series, splines and scattered data fitting. I am particularly interested in environmental applications such as low level ozone prediction and climate models. My expertise is in multivariate splines. For the functional linear regression models, the explanatory variable is a random surface and the response is a real random variable. In these models, we use bivariate splines over triangulations represent the random surfaces. The spline representations are then used to construct a least squares estimator of the regression function or an autoregressive estimator based on a principal component analysis.

2. Current Research

For problems with large data sets, like low level ozone prediction, we can use Functional Data Analysis (FDA) as an approach for prediction. For FDA, collections of consecutive discrete recordings are viewed as sample values of a random curve or surface. A functional linear model is defined as a regression model with a random function as the explanatory variable and a real random variable as the response, (see Ramsay and Silverman, 2005 (19)). In a series of papers [Cardot, Ferraty, and Sarda'99 (2)], [Cardot, Ferraty, and Sarda'03(3)], and [Cardot and Sarda'05 (4)], Cardot and his collaborators study the autoregressive approach for random curves and a functional associated with the random curves by using univariate splines to approximate the empirical estimator for the function associated with the random functional. In particular, they introduce consistent estimates based on functional principal components, and decompositions in univariate splines spaces. Bivariate splines over triangulations can be used to extend Cardot and his collaborators' univariate autoregressive approach to the bivariate setting. Bivariate splines over triangulations are an excellent tool to approximate the random surfaces that are only observed at scattered locations over a domain of irregular shape, (see [Lai'08, (16)]). Hence bivariate splines are used as the explanatory variable in the autoregressive functional linear model. The univariate autoregressive model is extended to the bivariate setting as follows. Let Y be a real valued random variable and take $\mathcal{D}$ be a polygonal domain in $\mathbb{R}^2$. The regression model is:

$$Y = \int_{\mathcal{D}} g(s)X(s)ds + \varepsilon, \tag{1}$$

where X is a random surface over the domain $\mathcal{D}$, $g \in H$ where H is usually $L^2(\mathcal{D})$, and ε is a real random variable that satisfies $E\varepsilon = 0$ and $EX(s)\varepsilon = 0$ for all $s \in \mathcal{D}$. The objective is to approximate the function g defined on the two dimensional spatial domain $\mathcal{D}$ from a given set of design points in $\mathcal{D}$. We rephrase the problem, by looking for the function $g \in H$ that solves the following minimization problem:

$$\min_{\beta \in H} \mathcal{E}\left[(f(X) + \epsilon - \langle \beta, X\rangle)^2\right]. \tag{2}$$

We want to approximate the function, g, defined on the two dimensional spatial domain $\mathcal{D}$ based on the observations on X from a set of design points in $\mathcal{D}$ and the random variable Y. We cannot assume that the random surfaces at various time steps are completely independent; hence we want to write the model in terms of the covariance operator of the H-valued random variable X, $\Gamma := \mathcal{E}(X(s)X(t))$ and the cross covariance of (X, Y), $\Delta := \mathcal{E}(X(s)Y)$. Then solution to the minimization problem (2) is given by $\Gamma g = \Delta$. However, Γ many not be invertible therefore an estimate of g is given by a principle component analysis.

An alternative brute force approach to solving (2) is offered in [Guillas and Lai'08, (14)]. I outline the brute force results here because I implemented their method for the 2D numerical experiments in [Guillas and Lai'08, (14)] and in my dissertation I offer an extension of this method. The brute force approach exploits the optimal approximation property of splines [Lai and Schumaker'98, (17)]. The brute force approach starts by trying to approximate a bounded and continuous functional f. By the Riesz representation theorem, the functional can be written as $f(X) = \langle g, X\rangle$ for some function $g \in H$. However it is impossible to solve the minimization problem (2) because we have an infinite dimensional Hilbert space H. Hence we find an approximation to the solution by choosing a finite dimensional spline space $S_d^r(\triangle)$ of smoothness r and degree $d \geq 3r+2$ which is dense in H as $|\triangle| \to 0$. This reduces the original problem (2) to the spline estimate

$$S_\alpha = \arg \min_{\beta \in S_d^r(\triangle)} \mathcal{E}\left[(f(X) - \epsilon - \langle \beta, X \rangle)^2\right]. \tag{3}$$

From the spline estimate we develop an empirical estimate for a given set of observed surfaces $\{X_1, \ldots, X_n\}$

$$\widehat{S_{\alpha,n}} = \arg \min_{\beta \in S_d^r(\triangle)} \frac{1}{n} \sum_{i=1}^{n} (f(X_i) - \epsilon_i - \langle \beta, X_i \rangle)^2. \tag{4}$$

In practice we are not able to observe the entire random surface. Instead, we observe a random surface X over a set of design points $s_k, k = 1, \ldots, N$ over $\mathcal{D}$. We create a spline S_X by finding the penalized least square fit of X from the given data by using splines. To do this we assume that $s_k, k = 1, \ldots, N$ are evenly distributed over $\triangle$ of $\mathcal{D}$ with respect to $S_d^r(\triangle)$. Then we repeat the above theory using our approximated random surfaces. We start with the original problem (2) using S_X instead of X:

$$\alpha_D = \arg \min_{\beta \in H} \mathcal{E}\left[(f(X) - \epsilon - \langle \beta, S_X \rangle)^2\right]. \tag{5}$$

Again we have the same issue, that the Hilbert space H is infinite and we cannot find the minimum but we can find an approximate solution in the finite spline space $S_d^r(\triangle)$. Hence we obtain a spline estimate based on the approximated random surfaces:

$$S_{\alpha_D} = \arg \min_{\beta \in S_d^r(\triangle)} \mathcal{E}\left[(f(X) - \epsilon - \langle \beta, S_X \rangle)^2\right]. \tag{6}$$

From this spline estimate we finally reach something we can compute from real data, the empirical estimate based the approximated random surfaces:

$$\widetilde{S_{\alpha,n}} = \arg \min_{\beta \in S_d^r(\triangle)} \frac{1}{n} \sum_{i=1}^{n} (f(X_i) - \epsilon_i - \langle \beta, S_{X_i} \rangle)^2. \tag{7}$$

Both the autoregressive and brute force methods do a good job of predicting values for locations within the set of design points. However, we may want to predict values for locations where there are no measurements. To do this I offer an extension of the brute force method where we consider the case when both the explanatory and response variables are random surfaces. We define the model by convolution:

$$Y(s) = G(s,t) * X(t) = \int_{\mathcal{D}} G(s,t) X(t) dt, \tag{8}$$

where $\mathcal{D}$ a polygonal domain in $\mathbb{R}^2$ and $G(s,t)$ is some function in $H \times H$. We usually take H to be $L^2(\mathcal{D})$. For this application, we assume we are given a function F such that $F = G * X$ for some function G. The objective is to recover the function G. In my dissertation I give a tensor product version of optimal approximation property of splines and a similar derivation for the extended model (see [Ettinger' 09, (7)]).

3. Application to Low Level Ozone Forecasting

The autoregressive functional linear model is a popular approach for low level ozone prediction. We use the autoregressive and brute force methods to forecast the low level ozone concentrations at three locations: Atlanta, Boston and Cincinnati. We assume that the level of ozone for a given time of day in one specific city is a linear functional of the previous days' ozone concentrations measured over the geographical region containing the city of interest. For example, we would assume that today's ozone concentration in Atlanta at 9:00 a.m. is a linear functional of all the ozone values in the southeast up to 9:00 a.m. yesterday. Thus we can implement our two methods to make our forecasts. We let $f(X)$ be the hourly ozone concentration at a particular city of interest and X is the ozone concentration distribution over a geographical region containing the city of interest at the same hour of the previous day. Our domain is the continental United States and our design points are the Environment Protection Agency (EPA) stations where the ozone values are collected.

4. Future Research

For future research, we are looking for different applications. One idea is to implement the brute force models to predict the paths of hurricanes. For this application, we want to use a model where both the explanatory and response variables are random surfaces. Using such a model allows us to predict values for locations where there are no measurements. For hurricane tracking, we could use barometric pressure data to create an input surface and then implement the model the brute force model extension to acquire an output surface of barometric pressures. Or we could run the model functional brute for model for several locations and fit a surface through the prediction results. The path of the hurricane is tracked by identifying eye of hurricane, where the barometric pressure is lowest. As new measurements become available, the models are quickly and easily updated to consider the new information. If we can improve the prediction of hurricane movement within fifty miles of its exact position, we could save peoples' lives, towns and money. The improved predictions give people more time to plan efficient evacuations and secure their homes and businesses for the storm surges and flooding that follow hurricane landfall.

References

[1] Awanou, G. and Lai, M. J. and Wenston, P., The multivariate spline method for numerical solution of partial differential equations and scattered data interpolation, in Wavelets and Splines: Athens 2005, G. Chen and M. J. Lai (eds), Nashboro Press, 2006, 24–74.

[2] Cardot, H. and Ferraty, F. and Sarda, P., Functional linear model, Stat. Probab. Lett., (1999) 45, 11–22.

[3] Cardot, H. and Ferraty, F. and Sarda, P., Spline estimators for the functional linear model, Stat. Sin., 2003, 13, 571-591.

[4] Cardot, H. and Sarda, P., Estimation in generalized linear models for functional data via penalized likelihood, J. Multivar. Anal., 92 (2005), 24-41.

[5] Crambes, C., A. Kneip, and P. Sarda, Smoothing splines estimators for functional linear regression, The Annals of Statistics (2009) 37, 3572.

[6] Damon, J. and S. Guillas, The inclusion of exogenous variables in functional autoregressive ozone forecasting, *Environmetrics 13* (2002), 759–774.

[7] Ettinger, B., *Bivariate Splines for Ozone Concentration Predictions.* PhD thesis, Univ. of Georgia, 2009.

[8] Ettinger, B., Guillas, S. and Lai, M. J., *Bivariate Splines for Ozone Concentration Prediction, under preparation, 2007.*

[9] Fasshauer, G., and Schumaker, L. L., *Scattered data fitting on the sphere*, in Mathematical Methods for Curves and Surfaces II, M. Daehlen, T. Lyche, L. Schumaker, Vanderbilt University Press, 1998, pp. 117–166.

[10] Ferraty, F. and Vieu, P., *Nonparametric Functional Data Analsysis: Theory and Practice*, Springer-Verlag, London, 2006

[11] von Golitschek, M., Lai, M. J., Schumaker,L. L., *Error bounds for minimal energy bivariate polynomial splines*, Numer. Math. 93(2002), 315–331.

[12] von Golitschek, M., and Schumaker, L. L., *Bounds on projections onto bivariate polynomial spline spaces with stable local bases*, Const. Approx. 18(2002), 241–254.

[13] von Golitschek, M., and Schumaker, L. L., *Penalized least squares fitting*, Serdica 18 (2002), 1001–1020.

[14] Guillas, S. and M. J. Lai, Bivariate Splines for Spatial Functional Regression Models, Journal of Nonparametric Statistics, 2009.

[15] Golub, G. H., and Van Loan, C.F., *Matrix Computions* , John Hopkins Univeristy Press, Baltimore, MD., 1989

[16] Lai, M.-J., Multivariate splines for data fitting and approximation, in *Approximation Theory XII: San Antonio 2007*, M. Neamtu and L. L. Schumaker (eds.), Nashboro Press (Brentwood), 2007, 210–228.

[17] Lai, M. J. and Schumaker,L. L., Approximation power of bivariate splines, Advances in Comput. Math., 9(1998), pp. 251–279.

[18] Lai, M. J. and Schumaker,L. L., *Spline Functions over Triangulation*, Cambridge University Press, Cambridge, U.K., 2007.

[19] Ramsay, J. and Silverman, B.W., *Functional Data Analysis*, Springer-Verlag, 2005.