Computational Statistics and Data Analysis 81 (2015) 38–51

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Penalized scalar-on-functions regression with interaction term☆

Karen Fuchs [a,b,*], Fabian Scheipl [b], Sonja Greven [b]

[a] Siemens AG, CT RTC SET CPS-DE, Otto-Hahn-Ring 6, D – 81739 Munich, Germany
[b] Department of Statistics, Ludwig–Maximilians-University Munich, Ludwigstr. 33, D – 80539 Munich, Germany

ARTICLE INFO

Article history:
Received 15 November 2013
Received in revised form 1 July 2014
Accepted 2 July 2014
Available online 16 July 2014

Keywords:
Functional data analysis
Covariate interaction
Penalized likelihood
Cell chip data
Signal regression
Generalized functional linear model

ABSTRACT

Generalized models for scalar responses with functional covariates are extended to include linear functional interaction terms. The coefficient functions are estimated using basis expansions and maximization of a log-likelihood, which is penalized to impose smoothness upon the coefficient functions. The respective smoothing parameters for the penalties are estimated from the data, e.g. via generalized cross-validation. Further functional or scalar terms as well as functional interactions of higher order can be added within the same framework. The performance of the introduced approach is tested in simulations. Additionally, it is applied to the two motivating data sets, to spectroscopic data of a fossil fuel and to cell chip sensor data, where three functional signals are measured over time. The main aim is to predict the respective response, namely the heat value of the fossil fuel and the concentration of paracetamol in the cell chip medium.

1. Introduction

Functional data analysis (Ramsay and Silverman, 2005; Ferraty and Vieu, 2006; Horvath and Kokoszca, 2012) is an active field of research. The amount and diversity of functional data are growing due to technical developments in signal recording and inspire researchers in the field of statistics as well as practitioners. Methods range from (semi-) parametric to nonparametric regression to classification approaches (e.g. Ferraty and Vieu, 2003, 2006). Overviews of established and recent methods in functional data analysis can be found in Gonzalez Manteiga and Vieu (2007), Ferraty and Romain (2011) or Bongiorno et al. (2014).

For a scalar response and functional covariates, many regression models include only a single functional covariate, such as the non-parametric functional regression models of Burba et al. (2009); Wang et al. (2012) and Kudraszow and Vieu (2013). The work of Ferraty and Vieu (2009) introduces a non-parametric additive model including two or more functional covariates.

☆ Please see the online Appendix A for supplementary material.

* Corresponding author at: Siemens AG, CT RTC SET CPS-DE, Otto-Hahn-Ring 6, D – 81739 Munich, Germany. Tel.: +49 0 89 636 40709.
E-mail addresses: karenfuchs@gmx.de, karen.fuchs.ext@siemens.com (K. Fuchs), fabian.scheipl@stat.uni-muenchen.de (F. Scheipl), sonja.greven@stat.uni-muenchen.de (S. Greven).
URL: http://www.statistik.lmu.de/institut/ag/fda/index_e.html (S. Greven).

http://dx.doi.org/10.1016/j.csda.2014.07.001
0167-9473/

The most common parametric model is the generalized functional linear model, for which several methods for estimation have been proposed. One strain of research expands both the functional covariate and the coefficient function in a principal component basis (e.g. Müller and Stadtmüller, 2005; Reiss and Ogden, 2007). Other approaches use a spline basis expansion of the coefficient function or the functional covariate and a smoothness penalty approach (e.g. James, 2002; Wood, 2011; Goldsmith et al., 2011).

Although some of the above methods include effects of more than one functional covariate, the estimation of interaction effects between functional covariates does not seem to have received any attention until now. If the assumption of additivity of the effects of multiple functional covariates is questionable, a sensible way to extend the generalized functional linear model is to add covariate interaction effects. This paper introduces a functional interaction term $\iint x_{i1}(s)x_{i2}(t)\beta(s,t)dsdt$ of functional covariates $x_{i1}(s)$ and $x_{i2}(t)$ with bivariate parameter function $\beta(s,t)$, extending the model with only main effects in Wood (2011).

Our bivariate parameter function $\beta(s,t)$ for the interaction term is represented in terms of a tensor product spline basis. A similar representation of a bivariate coefficient function can be found in Marx and Eilers (2005) in the context of scalar-on-image regression. Marx and Eilers (2005) also examine a generalized linear model and use a penalized log-likelihood approach for estimation. The main difference lies in the fact that Marx and Eilers (2005) consider a single image covariate $x_i(s,t)$, while we have two covariates $x_{i1}(s)$ and $x_{i2}(t)$ and consider their main effects as well as their interaction. Our model is also related to Yao and Müller (2010), who consider a pth-order polynomial model, where the scalar mean response depends on two-way up to p-way interaction effects of the centered predictor process with itself. Our approach, on the other hand, allows for interaction effects between different functional covariates. Yao and Müller (2010) expand the functional regression parameters as well as the centered functional covariate in the empirical eigenfunction basis of the functional covariate. By contrast, we do not assume the interaction effect surface to lie in the space spanned by the eigenfunctions of the two covariate processes, but to be smooth, and use penalized splines for estimation. Bivariate parameter functions can also be found for example in Antoch et al. (2010) or Ivanescu et al. (2013) in the context of function-on-function regression.

Our method, although general, is motivated by two data sets. The first data set contains spectra of fossil fuel samples measured at the ultraviolet–visible (UV–VIS) and near infrared (NIR) range. The main goal here is the prediction of the heat value of a sample based on its spectrum. The second data set consists of cell chip data, where three different and concurrently measured sensor signal types reflect the metabolism of a layer of living cells growing on the chip surface. Especially the prediction of the concentration of probably bioactive substances contained in the cell nutrient medium is of interest.

In Section 2, we present our model and the estimation method used. Section 3 presents an extensive simulation study. Our method is applied to the two motivating data sets in Sections 4 and 5. We close with a short discussion and outlook in Section 6.

The two data sets and code fully reproducing our analyses are provided in an online Appendix A.

2. Method

2.1. Scalar-on-function regression with interaction term

We extend the generalized functional linear model to include interactions for functional covariates. We assume the scalar responses y_i, $i = 1, \ldots, n$, to be (conditionally) mutually independent and to follow an exponential family distribution with a known link function $g(\cdot)$ linking the expected value μ_i of y_i to the linear predictor η_i,

$$g(\mu_i) = \eta_i = \beta_0 + \int x_{i1}(s)\xi_1(s)ds + \int x_{i2}(t)\xi_2(t)dt + \iint x_{i1}(s)x_{i2}(t)\beta(s,t)dsdt. \tag{1}$$

Here, β_0 is the intercept term, and $x_{i1}(s)$ and $x_{i2}(t)$ are two functional covariates that are expected to influence y_i. The covariate values $x_{i1}(s)$ are observed without error in the interval $\mathbb{D}$ with discrete observation points $\{s_1, \ldots, s_J\} \subset \mathbb{D}$. Likewise, $x_{i2}(t)$ is observed without error in the interval $\mathbb{E}$ with discrete observation points $\{t_1, \ldots, t_K\} \subset \mathbb{E}$. $\xi_1(s)$, $\xi_2(t)$ and $\beta(s,t)$ are unknown functional coefficients corresponding to the main and interaction terms. In the linear case $y_i = \mu_i + \varepsilon_i$, we assume ε_i to be independent and identically distributed normal errors with zero mean and variance σ^2. Following Wood (2011) in approximating the integrals of Model (1) by quadrature sums, the model can be expressed as

$$g(\mu_i) \approx \beta_0 + h_1 \sum_{j=1}^{J} x_{i1}(s_j)\xi_1(s_j) + h_2 \sum_{k=1}^{K} x_{i2}(t_k)\xi_2(t_k) + h_1h_2 \sum_{j=1}^{J}\sum_{k=1}^{K} x_{i1}(s_j)x_{i2}(t_k)\beta(s_j,t_k),$$

with h_1, h_2 being the lengths of the intervals between two observation points in $\mathbb{D}$ and $\mathbb{E}$, respectively, assuming a regular grid of observations on both intervals. In the case of unequal spacing, the sums could be replaced by appropriate weighted sums from quadrature rules.