Published in final edited form as:

Comput Stat Data Anal. 2007 June 15; 51(10): 4911–4925.

Penalized solutions to functional regression problems

Jaroslav Harezlak^{1,*}, Brent A. Coull¹, Nan M. Laird¹, Shannon R. Magari², and David C. Christiani³

¹ Department of Biostatistics, Harvard School of Public Health, 655 Huntington Ave. Boston, MA 02115 USA

² Colden Corporation 100 North 17th Street, 9th Floor Philadelphia, PA 19103 USA

³ Department of Environmental Health, Harvard School of Public Health, 655 Huntington Ave. Boston, MA 02115 USA

SUMMARY

Recent technological advances in continuous biological monitoring and personal exposure assessment have led to the collection of subject-specific functional data. A primary goal in such studies is to assess the relationship between the functional predictors and the functional responses. The historical functional linear model (HFLM) can be used to model such dependencies of the response on the history of the predictor values. An estimation procedure for the regression coefficients that uses a variety of regularization techniques is proposed. An approximation of the regression surface relating the predictor to the outcome by a finite-dimensional basis expansion is used, followed by penalization of the coefficients of the neighboring basis functions by restricting the size of the coefficient differences to be small. Penalties based on the absolute values of the basis function coefficient differences (corresponding to the LASSO) and the squares of these differences (corresponding to the penalized spline methodology) are studied. The fits are compared using an extension of the Akaike Information Criterion that combines the error variance estimate, degrees of freedom of the fit and the norm of the bases function coefficients. The performance of the proposed methods is evaluated via simulations. The LASSO penalty applied to the linearly transformed coefficients yields sparser representations of the estimated regression surface, while the quadratic penalty provides solutions with the smallest L_2 -norm of the basis functions coefficients. Finally, the new estimation procedure is applied to the analysis of the effects of occupational particulate matter (PM) exposure on the heart rate variability (HRV) in a cohort of boilermaker workers. Results suggest that the strongest association between PM exposure and HRV in these workers occurs as a result of point exposures to the increased levels of particulate matter corresponding to smoking breaks.

Keywords

environmental assessment; functional data; heart rate variability; LASSO; penalized regression splines

1. Introduction

Ramsay and Dalzell (1991) have termed data where either the predictor or the response variables are measured continuously as “functional data”. We consider the case, common in

* email: jharezla@hsph.harvard.edu; phone: 617-432-3172, fax: 617-432-5619.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

many biological fields, where both independent and dependent variables are functions. One example is the study of growth curves where more than one characteristic of growth is observed, e.g. height and lung function. Other examples include continuous monitoring of biological functions where the predictor is an exogenous or an endogenous variable also varying over time.

Our research is motivated by a study of the effects of particulate matter on heart functioning. Particulate matter has been shown to be associated with mortality and morbidity (Dockery et al., 1993; Pope et al., 2004). Recent attention has focused on effects of pollution exposure on cardiac function, with current research suggesting that there are two components of action of the particulate matter on cardiac function: a long-lasting component through inflammatory cytokines and a short-lasting component via control of the autonomic nervous system.

Here we consider the data collected on an occupational cohort of boilermaker construction workers who were exposed to residual oil fly ash (ROFA) and metal fumes (Magari et al., 2001). Continuous, personal $PM_{2.5}$ measurements were taken during workdays and on most non-work days. Measurements were taken every 10 seconds and five-minute averages were reported. In addition, continuous heart rate (HR) monitoring was performed using personal monitors. Five-minute averages of heart rate (HR) and SDNN (standard deviation of the normal-to-normal intervals) were obtained. The SDNN outcome was used as the main heart rate variability (HRV) indicator which measures the cardiac autonomic tone. Early analysis of these data (Magari et al., 2001) used models that assume constant effects of exposure on HRV and hence did not take into account possibly different cumulative and lag effects of particulate exposure on HRV. We relax those assumption in our modeling strategy in order to study complex patterns of association between the environmental exposure and HRV by utilizing a flexible functional data approach.

Interest in modeling functional data has grown in the past decade. Ramsay and Silverman (1997) provide an excellent summary of functional linear models. Ramsay and Silverman (1997, Chapter 11) consider the case where both the predictor $x(t)$ and the response $y(t)$ are observed over the same time period $[0, T]$, and both functions are periodic.

A key feature of functional regression models is that the response values at time t , $y(t)$, may be influenced by the predictors measured at times different from t . Ramsay and Silverman (1997) allowed $x(\cdot)$ observed at any time point $s \in [0, T]$ to influence the response at any time $t \in [0, T]$. Therefore their model was defined by

$$Y_i(t) = \alpha(t) + \int_0^T X_i(s)\beta(s, t)ds + \varepsilon_i(t), \quad (1)$$

where $Y_i(t)$ and $X_i(s)$ were, respectively, response and predictor functions for subject $i = 1, \dots, m$, $\varepsilon_i(t)$ was a residual function with $E[\varepsilon_i(t)] = 0$ and $\text{Cov}(\varepsilon_i(t), \varepsilon_i(t')) = \sigma^2$ if and only if $i = i'$ and $t = t'$ and zero otherwise, $\alpha(t)$ was a time-varying intercept function and $\beta(s, t)$ was an unknown regression surface describing how the predictor values $X_i(s)$ influenced the response at time t , $Y_i(t)$.

Because of its complexity and difficult interpretation, the regression theory for functional data has not been as well-developed as corresponding theoretical results for classical regression with continuous or categorical predictors. Cuevas et al. (2002) and He et al. (2003) studied the theoretical properties of model (1). They concentrated on the asymptotic properties of estimators of $\beta(s, t)$ in functional spaces. Ramsay and Silverman (1997) and Ramsay and Silverman (2002) studied approaches oriented towards practical examples and computational issues related to model (1). Malfait and Ramsay (2003) tailored model (1) to situations where the influence of the predictor $x(\cdot)$ on the response function $y(\cdot)$ could occur only in a feed-

forward matter, so that the outcome $y(t)$ at time t depended only on the predictors $x(s)$ at times $s \leq t$. Therefore model (1) was replaced by model (2) by restricting the integration region

$$Y_i(t) = \alpha(t) + \int_{s_0(t)}^t X_i(s)\beta(s, t)ds + \varepsilon_i(t), \quad (2)$$

where $s_0(t) = \max(0, t - \delta)$ (where δ was the maximum considered lag).

Ramsay and Silverman (1997) discussed identifiability and estimability of regression surface $\beta(s, t)$ in model (1). They pointed out that even though $\beta(s, t)$ was identifiable, the estimation involved a model with infinite number of parameters. Our primary goal in the estimation of the regression surface $\beta(\cdot, \cdot)$ is the stability of solutions when there are small perturbations in the data. A general approach of dealing with such ill-posed problems is a method of regularization (Tikhonov and Arsenin, 1977), which provides meaningful and stable solutions. Malfait and Ramsay (2003) regularized the fit by expansion into a small number of known basis functions. Their idea is equivalent to the regression spline methods in scatterplot smoothing. The approach of Malfait and Ramsay (2003) was satisfactory when the smoothness of the regression surface was uniform over the region of integration and the basis functions used approximated the regression surface $\beta(\cdot, \cdot)$ well. However, the fit could be poor when either one of the aforementioned conditions did not hold. In order to recover possible fine features of the regression surface, one would need to use a large number of basis functions, leading to an unstable estimation procedure. On the other hand, a small number of basis functions suitable for a stable estimation will decrease the goodness of fit.

We use the historical functional linear model (HFLM) considered in Malfait and Ramsay (2003) and propose an estimation procedure for the model parameters that uses regularization techniques based on penalty functions. We consider variations of the L_1 - and L_2 -norm penalties that incorporate the spatial relationship between the coefficients of the estimated regression surface $\beta(s, t)$. Our approach forces the coefficients of the neighboring basis functions to be similar through a discrete penalty on their differences.

The paper is organized as follows. In Section 2, we review the procedures used for fitting the historical functional linear model in Malfait and Ramsay (2003). In Section 3, we describe our penalized fitting procedures and in Section 4, we present simulation studies and their results. In Section 5, we analyze the data from the boilermaker study and we present conclusions and discuss possible future research directions in Section 6.

2. Historical functional linear model

This section follows closely Malfait and Ramsay (2003). We assume without loss of generality that both responses and predictors are pointwise centered. The regression surface $\beta(s, t)$ is approximated by an expansion $\widehat{\beta}(s, t)$ defined as

$$\widehat{\beta}(s, t) = \sum_{k=1}^K b_k \phi_k(s, t),$$

where $\phi_k(s, t)$, $k = 1, \dots, K$ are known basis functions. Let

$$\psi_{ik}(t) = \int_{s_0(t)}^t x_i(s) \phi_k(s, t) ds,$$

and consider the alternative formulation of model (2)

$$\begin{aligned}
 y_i(t) &= \sum_{k=1}^K b_k \int_{s_0(t)}^t x_i(s) \phi_k(s, t) ds + \int_{s_0(t)}^t x_i(s) \varepsilon_a(s, t) ds + \varepsilon_i(t) \\
 &= \sum_{k=1}^K b_k \psi_{ik}(t) + \varepsilon'_i(t),
 \end{aligned} \tag{3}$$

where $\varepsilon_a(s, t) = \beta(s, t) - \beta(s, t)$, and $\varepsilon'_i(t)$ combines the random error $\varepsilon_i(t)$ and the approximation error.

In order to estimate the vector of coefficients $\mathbf{b} = [b_1, \dots, b_K]^\top$, we minimize the squared error loss criterion, defined as

$$\text{SSE} = \int_0^T \sum_{i=1}^n \{\varepsilon'_i(t)\}^2 dt,$$

which is equivalent to solving an integral normal equation

$$\left\{ \int_0^T \mathbf{\Psi}^\top(t) \mathbf{\Psi}(t) dt \right\} \mathbf{b} = \int_0^T \mathbf{\Psi}^\top(t) \mathbf{y}(t) dt, \tag{4}$$

where $\mathbf{y}(t) = [y_1(t), \dots, y_m(t)]^\top$ and $\mathbf{\Psi}(t)$ is an $m \times K$ matrix with $\psi_{ik}(t)$ entries.

The estimation procedure still involves approximations of integrals. Simplifications proposed include representing the prediction process $x(\cdot)$ in terms of basis functions, and approximation of the model (3) by a multivariate linear model based on evaluation of $y_i(t)$ at finite number of time points t_q , $q = 0, \dots, Q$.

Malfait and Ramsay (2003) used a computational technique called the finite element method (FEM) to estimate the vector of coefficients $\mathbf{b}$. In general, FEM divides the problem domain into small regions (“finite elements”), and then constructs separate low-order polynomials on the finite elements in such a way that the union of these pieces closely approximates the solution to the problem. Technical details on FEMs are provided in the books by Brenner and Scott (2002) and Braess (2001).

In our application, the region is a triangle (when $\delta = T$), or a trapezoid (when $0 < \delta < T$). For such regions a natural “finite element” is a triangle, and the shape of a basis function is tent-like spanning 6 triangles adjacent to a node of interest (see Figure 1).

3.1 Penalties

We construct the horizontal penalty matrix $D_{(1)}^H$ by stacking the difference matrices for every time point where the node of the basis function $\phi_k(s, t)$ is defined in the block-diagonal structure. Figure 2 presents a simple example of the horizontal direction of penalization with $M = 3$ subintervals of the time interval $[0, 1]$ and $B = 2$ lag periods. Let $\mathbf{b} = [b_1, \dots, b_9]$ denote the basis function expansion of the regression surface $\beta(s, t)$. We can write the penalty term $P(\mathbf{b})$ in the following form

$$\begin{aligned}
 P(\mathbf{b}) &= [b_3 - b_2]^2 \\
 &+ [b_5 - b_4]^2 + [b_6 - b_5]^2 \\
 &+ [b_8 - b_7]^2 + [b_9 - b_8]^2
 \end{aligned}$$

Basis functions defined in such a way are an extension of linear B-spline basis functions defined on a line. Each basis function $\phi_k(s, t)$ is piecewise linear and continuous, with $\max \phi_k(s, t) = 1$ at the node v that defines it and $\phi_k(s, t) = 0$ at the nodes of the triangles adjacent to v . The estimation of $\beta(s, t)$ involves approximation by $\beta(s, t)$ linearly on each triangle defining the

region of integration. Each approximant can be written locally as a linear combination of 3 basis functions defined on each of the triangle vertices using the system of area coordinates. The global interpolant is found by fitting together the local approximations.

3. Penalized functional linear model

Our proposal provides a more flexible approach to estimation of the regression surface $\beta(s, t)$ in model (2) than that of least squares. We work with the representation (3), but impose the penalties on the size of the coefficients $\mathbf{b}$. We extend the idea of discrete roughness penalties on the coefficients of a basis function expansion (Eilers and Marx, 1996) to the historical functional linear model setting by penalizing the interpretable directions of the regression surface. The penalty matrix for the coefficients $\mathbf{b}$ contains the information about neighboring finite elements in a two-dimensional space. In our prediction problem, there are three important directions in which we wish to control the changes in the coefficients $\mathbf{b}$. The first direction is along the s axis, which represents the cumulative effect of exposure from time $s_0(t)$ to t ; the second direction is along the t axis, which reflects the point exposure effect at time s on the future times $t > s$, and the third direction is parallel to the diagonal $s = t$, which corresponds to the lagged effect of exposure at the constant time difference $(t - s)$. Following the graphical representation of the regression surface $\beta(s, t)$, we call the penalty applied along the s axis a “horizontal” penalty (see Figure 2), the penalty applied along the t axis a “vertical” penalty and the penalty applied in the direction parallel to the diagonal a “parallel” penalty.

An alternative expression for $P(\mathbf{b})$ is $P(\mathbf{b}) = \mathbf{b}^\top (D_{(1)}^H)^\top D_{(1)}^H \mathbf{b}$ where $D_{(1)}^H$ is presented in Table 1. In general, for a fixed time t , we penalize the differences of the coefficients corresponding to the neighboring basis functions.

The vertical difference matrix ($D_{(1)}^V$) penalizes the differences of the coefficients corresponding to the neighboring basis functions for a fixed s , whereas the parallel difference matrix ($D_{(1)}^P$) penalizes the differences of the coefficients corresponding to the neighboring basis functions for a fixed difference $(t - s)$. We drop the subscript (1) of the difference matrix $D_{(1)}$ in further discussion, since we work only with the first order difference matrices in our research. It is worth noticing that the null space or unpenalized part of the horizontal penalty D^H consists of the constant functions for a given t , i.e. $\beta^H(s, t) = \beta(t)$. Similarly, in the case of vertical penalty $\beta^V(s, t) = \beta(s)$, and in the case of parallel penalty $\beta^P(s, t) = \beta(t - s)$.

In the L_2 -norm approach, we write the minimization criterion for estimation as

$$\min_{\mathbf{b}=(b_1, \dots, b_K)} \sum_{i=1}^m \sum_{j=1}^J (y_i(t_j) - \sum_{k=1}^K \Psi_{ik}(t_j) b_k)^2 + \lambda \mathbf{b}^\top \mathbf{K} \mathbf{b}, \quad (5)$$

where m is the number of independent units, J is the number of observations per unit, $t_j, j = 1, \dots, J$ are the observation times and $\Psi_{ik}(t_j) = \int_{s_0(t_j)}^{t_j} \phi_k(s, t_j) x_i(s) ds$. We create the penalty matrix $\mathbf{K}$ as a weighted sum of the penalty matrices with difference operators along the three directions,

$$\mathbf{K} = w_H \mathbf{K}^H + w_V \mathbf{K}^V + w_P \mathbf{K}^P, \quad (6)$$

where $\mathbf{K}^H = (D^H)^\top D^H$, $\mathbf{K}^V = (D^V)^\top D^V$, $\mathbf{K}^P = (D^P)^\top D^P$, and w_H, w_V, w_P are the adjustments to smoothing parameter λ in the horizontal, vertical, and parallel directions respectively.

Analogously, in the L_1 -norm approach, we work with the first-order differences on the coefficients $\mathbf{b}$. We can write the minimization criterion as

$$\min_{\mathbf{b}=(b_1, \dots, b_K)} \sum_{i=1}^m \sum_{j=1}^J [y_i(t_j) - \sum_{k=1}^K \Psi_k(t_j) b_k]^2 + \lambda \|\mathbf{P}(\mathbf{b})\|_1, \quad (7)$$

where $\mathbf{P}(\mathbf{b}) = w_H D^H(\mathbf{b}) + w_V D^V(\mathbf{b}) + w_P D^P(\mathbf{b})$ is a function of absolute values of differences of neighboring coefficients in the respective directions: horizontal, vertical and parallel.

3.2 Computational issues

In the regression splines approach, the choice of the number and the position of the knots plays an important role in the fitting procedure. Similarly the approach of Malfait and Ramsay (2003) depends in a crucial way on the choice of the parameters M and B controlling the resolution and the maximum lag respectively. The total number of basis functions for the trapezoidal region is $K = (B+1)(M+1 - B/2)$ and for a fixed ratio B/M it grows quadratically in M . In our fitting approach, we have the ability to estimate the approximate number of required basis functions via the notion of equivalent degrees of freedom of the fit. We start with a fixed M and increase B from 2 to about $M/2$. We repeat the fitting procedure with an increased M and select the model maximizing one of the criteria defined in Section 3.3.

Our proposed estimation procedure for the L_2 -norm approach is based on the equivalence of the penalized regression spline methods and linear mixed models (Brumback et al., 1999; Ruppert et al., 2003). In particular, parameter estimates found using the minimization criterion (5) are equivalent to those from a linear mixed model: $\mathbf{y}_i = \mathbf{X}_i \boldsymbol{\beta} + \mathbf{Z}_i \mathbf{u}_i + \boldsymbol{\varepsilon}$, where $\mathbf{y}_i$ is the vector of responses for subject i , $\mathbf{X}_i$ contains covariates corresponding to unpenalized coefficients, and $\mathbf{Z}_i$ contains covariates corresponding to penalized coefficients. Placing distributional assumptions on $\mathbf{u}_i$ allows us to introduce random effects covariance structure into the model (5). In our approach, the matrix $\mathbf{X}_i$ is either empty or it consists of the vector of ones, and $\mathbf{Z}_i = \Psi_i[(D^A)^T D^A]^{-1} (D^A)^T$, where D^A is one of the penalty matrices D^H, D^V, D^P or their weighted combination as defined in Section 3. The equivalence of the minimization criterion (5) and maximization of the mixed model log-likelihood enables us to use well-known mixed model algorithms and theoretical results in our estimation approach. Also, the choice of the smoothing parameter λ and the weights w_H, w_V and w_P is equivalent to the estimation of the variance components associated with the random effects $\mathbf{u}_i$ and the error terms, which we estimate using a restricted maximum likelihood (REML) approach.

The L_1 -norm approach does not have a simple analytical solution. In our computations, we utilize the result of Grandvalet (1998), who proved that a modified LASSO scheme is equivalent to a weighted ridge regression. We choose the smoothing parameter λ based on the Akaike Information Criterion (AIC), and the weights w_H, w_V and w_P using a grid search.

3.3 Model assessment

Due to no generally agreed upon definition of R^2 in the functional regression setting, we considered two proposals that are natural extensions of the definition of R^2 in the linear regression setting:

$$R_{int}^2 = (SSY - SSE)/SSY,$$

$$\text{and } R_{ave}^2 = \frac{\int_0^T \sum_{i=1}^n R_i^2(t) dt}{nT},$$

$$\text{where } SSY = \int_0^T \sum_{i=1}^n \{Y_i(t)\}^2 dt \text{ and } R_i^2(t) = 1 - \frac{\sum_i (y_i(t) - \bar{y}_i(t))^2}{\sum_i (y_i(t))^2}.$$

The norm of the estimated coefficients $\|\mathbf{b}\|$ gave a good summary of “wiggleness” of surface $\beta(s, t)$ when the values of R^2 are very close to each other for all penalization techniques considered. In general, smaller values of $\|\mathbf{b}\|$ gave rise to the estimates of the regression surface with less fluctuation.

Equivalent degrees of freedom (EDF) is defined as a function of a smoothing parameter λ . It is a generalization of the number of parameters in parametric fitting. In the L_2 -norm penalization framework, EDF is defined as the trace of the smoother matrix S_λ . In the L_1 -norm, EDF is defined as the number of non-zero parameters (Zou et al., 2005).

A proposal by Demidenko (2004) of hAIC (healthy Akaike Information Criterion), a modification of AIC, combines the previously discussed measures of the fit quality. Here, we give a formula for L_2 - and L_1 -norm hAIC which equals

$$\text{hAIC} = H_q - 2l_{\max} + 2p, \quad (8)$$

where $l_{\max}$ is the maximum value of log-likelihood, p stands for the equivalent degrees of

freedom and $H_1 = 2p \ln(\|\mathbf{b}\|_1/p)$, $H_2 = p(\ln \|\mathbf{b}\|_2^2/p) - 1$. We present the derivations of hAIC for both cases in the Appendix.

4. Simulation study

In order to evaluate our penalized fitting procedures, we conducted a Monte Carlo simulation study. We examined the performance of the various penalization techniques under several different scenarios of the data generating mechanism. We varied the combinations of the parameters defining the regression surface $\beta(s, t)$ and lag δ . For each data generating scheme, we simulated 100 data sets and examined a number of possible estimation procedures. We varied the number of basis functions, applied penalties based on L_1 -norm and L_2 -norm and specified the direction of penalization. We also tested our fitting procedures for sensitivity to misspecification of the lag δ and direction of penalization.

4.1 Data generation

In our simulation study, we defined the functions $X_i(t)$ and $Y_i(t)$ on the closed interval $t \in [0, 1]$ and $s \in [\max(0, t - \delta), t]$. We chose our predictor and response to resemble the data from the motivating boilermaker study. We defined subject-specific exposures curves

$$X_i(t) = \gamma_0^i + \gamma_1^i t + \gamma_2^i \sin(\gamma_3^i t), \quad (9)$$

where $\gamma_0^i \sim N(5, 1.5^2)$, $\gamma_1^i \sim N(6, 2^2)$, $\gamma_2^i \sim N(6, 2^2)$, and $\gamma_3^i \sim N(18, 0.8^2)$.

As depicted in Figure 3, we specified the regression surface as

$$\beta(s, t) = a_1 \sin(a_2 t - a_3 s) + a_4, \quad (10)$$

and the intercept function $\alpha(t) = 1.5 * (4 - 7t + 6.5t^2)$.

Specification of values for the coefficients a_2 and a_3 of the regression surface $\beta(s, t)$ enabled us to keep the regression surface varying slowly in the primary directions of interest: horizontal ($|a_3| \ll |a_2|$), vertical ($|a_2| \ll |a_3|$) and parallel ($a_2 \approx a_3$). We selected the $a_2 = 20$, $a_3 = 4$ (horizontal), $a_2 = 4$, $a_3 = 20$ (vertical), and $a_2 = 30$, $a_3 = 28$ (parallel) for the maximum lag $\delta = 0.2$, while for the lag $\delta = 0.4$ we used the same coefficients for the horizontal and vertical directions and specified $a_2 = 60$, $a_3 = 56$ for the parallel direction.

In our Monte Carlo simulations, we specified δ and calculated the noisy version of $y_i(t)$, $i = 1, \dots, m$, by applying the integral transform (2) with $\varepsilon_i(t) \sim N(0, 0.2^2)$. We show a sample of generated curves for one simulation setting with the lag $\delta = 0.2$ in Figure 4. It is worth noticing that the responses depend on the predictors for the maximum lag of $\delta = 0.2$. However, it is not immediately obvious from the plot in Figure 4 that the relationship does not extend for times $s < t - 0.2$.

In order to compare the estimation methods, we discretized the problem. We sampled the curves $y_i(\cdot)$ and $x_i(\cdot)$ at 101 equally-spaced points on the interval $[0, 1]$. As described in Section 2, we approximated $\beta(\cdot, \cdot)$ by a finite dimensional expansion in terms of known basis functions $\phi_k(\cdot, \cdot)$, $k = 1, \dots, K$.

4.2 Model fitting

We divided the interval $[0, 1]$ into M subintervals. In the set of simulations with known and fixed value of δ , we specified the number of the lag subintervals B to match the lag δ in the data generating step. The total number of basis functions $\phi_k(\cdot, \cdot)$ is $K = (B + 1)(M + 1 - B/2)$. Thus, for instance, when $\delta = 0.2$ and $M = 20$, we specified $B \equiv M\delta = 4$ and $K = 95$.

In addition, in order to test the sensitivity of our fitting procedure to misspecification of the lag B we conducted a smaller simulation study with $B/M \neq \delta$. We chose $M = 20$ and $B = 2, 3, 5$ and 6 corresponding to underestimating the lag δ for the $B = 2$ and $B = 3$, and overestimating δ for $B = 5$ and $B = 6$.

We compared the fits under the considered penalization methods to the least squares approach of Malfait and Ramsay (2003) using the criteria described in Section 3.3. These criteria included the explained variation (R^2), L_q norms of the estimated coefficients $\|\mathbf{b}\|$, equivalent number of degrees of freedom of the estimated surface and the healthy Akaike information criterion (hAIC).

4.3 Simulation results

For each regression surface $\beta(s, t)$, we found the estimates of the coefficients $\mathbf{b}$ of basis functions ϕ using least squares method (LS) and both L_1 -norm and L_2 -norm regularization techniques. For the L_q norms, we compared simplified penalization where either one or two directions of variation are penalized. For each simulation setup, one of the penalization directions corresponded to the primary mode of constancy of the regression surface. In total, we fit 36 models with varying number of basis functions and different penalization techniques to each generated regression surface for a lag $\delta = 0.2$ and 27 models for a lag $\delta = 0.4$.

We present only selected results of simulations for the considered setups with a lag $\delta = 0.2$. In Table 2, we summarize the hAIC with L_1 -norm for all the setups. In general, the best fits were obtained with the correct penalization of the primary direction of constancy in regression surface $\beta(s, t)$. L_1 penalization techniques produced estimates with the smaller number of equivalent degrees of freedom and smaller norm $\|\mathbf{b}\|_1$, while L_2 penalization techniques were more robust to misspecification of the direction of penalization, and outperformed L_1 penalties especially in the case of parallel direction of changes in the regression surface $\beta(s, t)$. Also, in most simulation scenarios, the quality of the fit could not be evaluated using the R^2 criterion, since most estimates were approximately the same. For example, for all penalized techniques and the number of subdivisions $M = 15$ or greater, all estimates of R^2 were within 1% of each other.

The equivalent number of degrees of freedom indicated that for a fixed M and B , L_1 penalization required fewer basis functions than the corresponding L_2 penalization to provide a good fit. When EDF was compared to the total number of basis functions used in the least squares fitting,

the proportion of degrees of freedom, defined as EDF/K , decreased from 87.6% (for $M = 10$, $B = 2$, $K = 30$) to 32.7% (for $M = 25$, $B = 5$, $K = 141$) for the correct penalty with the L_1 -norm penalty (see Table 3). Simulation results for the lag $\delta = 0.4$ were very similar to the results summarized above for the $\delta = 0.2$ lag. The least squares fit was comparable to the penalized fits only for the smallest number of basis functions with $M = 10$ and $B = 4$. However, according to the hAIC, penalized fits with correct direction of penalization and greater number of basis functions outperformed the least squares fit.

Figures 5 and 6 show the plot of the estimated coefficients $\mathbf{b}$ using least squares and L_1 penalization, respectively. We could assess visually that the least squares fit is very wiggly with the values of the estimated coefficients ranging from -300 to 300 , while the neighboring coefficients obtained via L_1 penalization were close to each other.

We also tested the sensitivity of the fitting methods to misspecification of the lag δ . We present the summaries of the hAIC with L_1 -norm for misspecified lags in Table 2. In all cases, the hAIC criterion was the smallest for the correct lag used for fitting. When the empirical lag used in fitting (B/M) was smaller than the lag δ used for data generation, the quality of the fit was poor. Overspecification of the lag introduced more nonzero parameters than should be used in the fit. However, the values of the extraneous parameters were very close to zero, which resulted in the fit being only marginally worse than that obtained using the correct lag.

5. Analysis of the boilermaker data set

Here, we analyze a subset of the boilermaker data set consisting of observations during the workday starting at 8:30am and ending at 5pm for 14 subjects. For each boilermaker, it was the first workday they were participating in the study. The data set consisted of HRV and $PM_{2.5}$ measurements obtained every 5 minutes for a total of 103 observations per subject. It is worth noticing that the composition and the toxicity of the $PM_{2.5}$ particles changed during the workday. For instance, all the boilermakers were required to congregate in closed quarters during the work breaks where the main source of the particulate pollution was the cigarette smoke. We present the plot of the raw data in Figure 7 with the break times indicated by vertical lines.

We analyze the data using the methods described in Section 4. The total observation period was 510 minutes. In order to explore possible long- and short-term effects of $PM_{2.5}$ on HRV, we specified three values for the minimum considered resolution: 30 minutes, 15 minutes and 10 minutes corresponding to the discretization of observation period into $M = 17$, 34, 51 subintervals respectively. Of course, we do not know the value of lag δ in a given application. We perform the analysis using the values of $B = 3, 5, 7, 9$ for the number of subintervals $M = 17$. For the other values of $M = 34$ and $M = 51$, we doubled and tripled the number of subintervals B respectively. Note that the longest lags considered (e.g. $M = 17$, $B = 9$) corresponded to the maximum lag effect of 4.5 hours, which enabled us to explore long-term effects of $PM_{2.5}$ exposure on HRV. The summary of the results is presented in Table 4.

When the analysis was performed on the coarse grid of 30 minutes ($M = 17$), the least squares method outperformed the penalization techniques. However, the hAIC value using L_1 penalization was close to that obtained using least squares. When the resolution was increased to 15 minutes ($M = 34$), the L_1 penalty with the vertical direction of penalization yielded the best fit. Increasing the resolution further to 10 minutes ($M = 51$) indicated that the vertical direction of penalization corresponded to the primary mode of constancy in the relationship between the exposure to $PM_{2.5}$ and HRV in the occupational exposure of boilermakers. The minimum value of hAIC was achieved for the lag of 260 minutes ($B = 26$) when the 10-minute minimum resolution ($M = 51$) was used in the analysis. The plot of the estimated regression

surface $\beta(s, t)$ (Figure 8) indicates that the environmental tobacco exposure during the breaks had the strongest association with the variation in the HRV responses.

Magari et al. (2001) analyzed a larger sample of 33 boilermakers over a time span extending to the nonworking hours. They assumed constant effects of the 4-hour moving average exposure to $PM_{2.5}$ on HRV and found a significant decrease in HRV for every $1\mu g/m^3$ increase in $PM_{2.5}$ levels. We applied the mixed models approach taken in Magari et al. (2001) to our sample of 14 boilermakers. We compared the results of this approach to our method via the explained variation measure R_{int}^2 which was estimated to be $\widehat{R}_{int}^2 = 0.61$ in the mixed model approach and $\widehat{R}_{int}^2 = 0.76$ in our model having the minimum hAIC. This big improvement can be attributed to the relaxation of the assumption of constant effects of the pollution on the HRV during the study period. Our approach allowed us to explore various modes of association between the exposure and the outcome.

6. Conclusions and discussion

We have developed a procedure for fitting penalized functional linear models under various penalties, taking into account spatial relationships between the coefficients of the basis function representation of the regression surface. We have compared the performance of our method to the regression spline approach proposed by Malfait and Ramsay (2003).

Expression of an L_2 -norm penalization in a mixed model formulation has enabled us to obtain smoothing parameter estimates using the restricted maximum likelihood. Also, equivalence of L_1 -norm penalties to weighted ridge regression penalization simplified the computations in the case of modified LASSO procedures. The methods were implemented in MATLAB, and sample programs are available from the first author.

Our simulation study indicated that among the penalization techniques considered, L_1 penalization provided fits with smaller number of degrees of freedom than L_2 fits, while the variation explained was the same up to 3 or 4 significant digits. Least squares fitting was comparable with the penalization techniques only when the number of basis functions was the smallest considered ($M = 10$). However, hAIC values obtained from the models using a larger number of basis functions were much smaller than those from models with few basis functions, indicating the need for finer discretization of the region of integration for the functions considered.

Finally, we applied our methods to assess the relationship between the exposure to pollutants and heart rate variability in the occupational exposure setting of boiler-makers. The analysis indicated that primarily the point exposure to $PM_{2.5}$ during the breaks showed the strongest association with the changes in the HRV in this sample of boilermakers. As in the simulation studies, the penalized techniques outperformed the least squares fitting method in most cases. In addition, the comparison with the mixed model methodology used in the Magari et al. (2001) favored our approach according to the explained variation criterion.

It is worth noticing that model (2) is a generalization of the varying-coefficient model (Hastie and Tibshirani, 1993) where the response $Y_i(t)$ depends on the predictor $X_i(t)$ only concurrently at time t :

$$Y_i(t) = \alpha(t) + \beta(t)X_i(t) + \varepsilon_i(t)$$

Another special case of the model (2) is a distributed lag model (Almon, 1965; Zanobetti et al., 2000) defined as

$$Y_i(t) = \alpha + \sum_{d=0}^D \beta_d X_i(t-d) + \varepsilon_i(t),$$

where the intercept, α , is constant over time t , and the coefficients β_d , $0 \leq d \leq D$, depend only the lag time $d = t - s$ between the observed $Y_i(t)$ and $X_i(s)$ values. Application of the Zanobetti et al. (2000) approach to the boilermaker data set provided a minimal improvement over the mixed model approach of Magari et al. (2001), but still yielded a poorer fit than that provided by our proposed historical model.

Possible extensions of our methods include incorporation of dependence of response variables on additional covariates either scalar or time-varying. Also, an extension including higher-order finite elements and penalization of additional neighboring coefficients might provide a better fit to smooth surfaces than the piecewise linear basis functions.

Acknowledgements

This work was supported in part by NIH grants ES07142, GM29745, ES012044, and ES00002. The authors thank the editor and the two reviewers for comments and suggestions for improvement.

References

- Almon S. The distributed lag between capital appropriations and expenditures. *Econometrica* 1965;33:178–196.
- Braess, D. Finite elements: Theory, fast solvers, and applications in solid mechanics. Cambridge University Press; 2001.
- Brenner, SC.; Scott, RL. The mathematical theory of finite element methods. Springer Verlag; 2002.
- Brumback BA, Ruppert D, Wand MP. Comment on “Variable selection and function estimation in additive nonparametric regression using a data-based prior”. *Journal of the American Statistical Association* 1999;94:794–797.
- Cuevas A, Febrero M, Fraiman R. Linear functional regression: The case of fixed design and functional response. *The Canadian Journal of Statistics* 2002;30:285–300.
- Demidenko, E. Mixed Models. Wiley Interscience; 2004.
- Dockery DW, Pope CA, Xu X, Spengler JD, Ware JH, Fay ME, Ferris BGJ, Speizer FE. An association between air pollution and mortality in six U.S. cities. *New England Journal of Medicine* 1993;329:1753–9. [PubMed: 8179653]
- Eilers PHC, Marx BD. Flexible smoothing with *B*-splines and penalties (Disc: P102–121). *Statistical Science* 1996;11:89–102.
- Grandvalet, Y. Perspectives in Neural Computing. Springer; 1998. Least absolute shrinkage is equivalent to quadratic penalization; p. 201–206.
- Hastie T, Tibshirani R. Varying-coefficient models. *Journal of the Royal Statistical Society, Series B, Methodological* 1993;55:757–796.
- He G, Müller HG, Wang J. Functional canonical analysis for square integrable stochastic processes. *Journal of Multivariate Analysis* 2003;85:54–77.
- Magari SR, Hauser R, Schwartz J, Williams PL, Smith TJ, Christiani DC. Association of heart rate variability with occupational and environmental exposure to particulate air pollution. *Circulation* 2001;104:986–991. [PubMed: 11524390]
- Malfait N, Ramsay JO. The historical functional linear model. *The Canadian Journal of Statistics* 2003;31:185–201.
- Pope CA, Hansen ML, Long RW, Nielsen KR, Eatough NL, Wilson WE, Eatough DJ. Ambient particulate air pollution, heart rate variability, and blood markers of inflammation in a panel of elderly subjects. *Environmental Health Perspectives* 2004;112:339–345. [PubMed: 14998750]
- Ramsay JO, Dalzell CJ. Some tools for functional data analysis (Disc: p561–572). *Journal of the Royal Statistical Society, Series B, Methodological* 1991;53:539–561.
- Ramsay, JO.; Silverman, BW. Functional Data Analysis. Springer-Verlag Inc; 1997.

- Ramsay, JO.; Silverman, BW. Applied Functional Data Analysis: Methods and Case Studies. Springer-Verlag Inc; 2002.
- Ruppert, D.; Wand, MP.; Carroll, R. Semiparametric regression. Cambridge University Press; 2003.
- Tikhonov, A.; Arsenin, V. Solutions of Ill-Posed Problems. Wiley; New York: 1977.
- Zanobetti A, Wand MP, Schwartz J, Ryan L. Generalized additive distributed lag models: quantifying mortality displacement. *Biostatistics* 2000;1:279–292. [PubMed: 12933509]
- Zou, H.; Hastie, T.; Tibshirani, R. On the “degrees of freedom” of the lasso Technical report. Stanford University; 2005.

Appendix

We present here the justification beyond the healthy Akaike Information Criterion (hAIC) of Demidenko (2004), which is based on AIC - a popular method for model selection, especially for nonnested models. It is a well-known fact that in the linear regression models $AIC = -2l_{max} + 2p$, where l_{max} is the value of the maximized log-likelihood and p is the number of parameters. The best model according to AIC is the model with a minimum AIC. There are however weaknesses of AIC that has been pointed out by Demidenko (2004). Major weakness of AIC relevant to our research is a poor discrimination of models with multicollinear predictors. Let $L(\mathbf{b}; \mathbf{y})$ be conditional likelihood of data $\mathbf{y}$ on a p -dimensional random coefficients $\mathbf{b}$. If we assume $b_i \sim G$ for $i = 1, \dots, p$ than the density of $\mathbf{b}$ is $\tau^{-p}G(\tau^{-1}\mathbf{b})$, where τ is a scale parameter. Let $g = \ln G$, then the marginal likelihood of τ takes the form

$$l(\tau) = -p \ln \tau + \ln \int_{\mathbb{R}^p} e^{l(\mathbf{b}; \mathbf{y}) + g(\tau^{-1}\mathbf{b})} d\mathbf{b}, \quad (11)$$

and MLE $\hat{\tau}$ maximizes the Equation (11). Using Laplace approximation for the integral in the Equation (11), we can obtain the following approximate expression for $l(\tau)$:

$$l(\tau) \approx -p \ln \tau + l(\mathbf{b}; \mathbf{y}) + g(\tau^{-1}\mathbf{b}). \quad (12)$$

Assuming that the marginal distribution of $\mathbf{b}$ is Gaussian, we obtain a simplified expression for the likelihood of τ

$$l(\sigma^2) \approx -\frac{p}{2} \ln \sigma^2 + l(\mathbf{b}; \mathbf{y}) - \frac{1}{2\sigma^2} \|\mathbf{b}\|^2. \quad (13)$$

Equation (13) is maximized over the variance at $\hat{\sigma}^2 = \|\mathbf{b}\|^2/p$ giving

$$l_{\max} = -\frac{p}{2} \ln(\|\mathbf{b}_{ML}\|^2/p) + l(\mathbf{b}_{ML}; \mathbf{y}) - \frac{p}{2}. \quad (14)$$

Now, the healthy AIC takes the form

$$\begin{aligned} hAIC &= H - 2l_{\max} + 2p \\ &= H + AIC, \end{aligned} \quad (15)$$

where

$$H \equiv H_2 = p(\ln(\|\mathbf{b}_{ML}\|_2^2/p) - 1).$$

Another popular choice for the marginal distribution of $\mathbf{b}$ is double-exponential with a scale parameter ω . In this case, the H in the healthy AIC takes the form

$$H \equiv H_1 = 2p \ln(\|\mathbf{b}_{ML}\|_1/p).$$

Figure 1.
Plot of one finite element in the integration region

Figure 2. Horizontal penalization. Arrows indicate the direction of penalization of the basis function coefficients b_k .

Figure 3.
Function $\beta(s, t)$ used for a simulation study with $a_2 = 20$, $a_3 = 4$.

Figure 4.
Plot of a 50 simulated $X_i(t)$ (top panel) and $Y_i(t)$ (bottom panel) curves.

Figure 5.
Coefficients $\mathbf{b}$ estimated using least squares method

Figure 6.
Coefficients $\mathbf{b}$ estimated using L_1 penalization technique with horizontal penalty

Figure 7. Plot of HRV and $PM_{2.5}$ values for 14 boilermakers. Vertical lines indicate the compulsory breaks.

Figure 8.
 β estimated using L_1 penalization technique relating exposure to $PM_{2.5}$ and HRV for the boilmaker data set with vertical penalty - D^V .

Table 1

Horizontal difference operator - D^H .

s	0	1	0	1	2	1	2	1	2	3
t	1	1	2	2	2	2	2	3	3	3
D^H	-1	1	0	0	0	0	0	0	0	0
	0	0	-1	1	0	0	0	0	0	0
	0	0	0	-1	1	0	0	0	0	0
	0	0	0	0	0	-1	1	-1	1	0
	0	0	0	0	0	0	0	0	-1	1

Table 2

Simulation results for all directions of constancy with lags $\delta = 0.2$, $\delta = 0.4$ and misspecified discretized lags for $\delta = 0.2$. Summary of the hAIC with the L_1 -norm penalty on the basis functions coefficients. **Bold** numbers indicate the smallest values of hAIC in each row and ***Bold Italic*** the second smallest (for fixed M and B). NA denotes the entries with computationally non-estimable coefficients and * denotes the entries based on a limited number of the convergent runs.

M(B)	D	LS	L_2 -norm			L_1 -norm		
			D^H	D^V	D^{HV}	D^H	D^V	D^{HV}
10 (2)	H	-8798	-8846	-8827	-8814	-8843	-8815	-8749
15 (3)		-8943	-9201	-9154	-9129	-9196	-9175	-8931
20 (4)		-8523	-9277	-9201	-9180	-9286	-9217	-9131
25 (5)	V	-6690	-9266	-9165	-9134	-9280	-9196	-9157
10 (2)		-9189	-9214	-9249	-9219	-9207	-9247	-9037
15 (3)		-8985	-9211	-9278	-9208	-9209	-9278	-8809
20 (4)	P	-8536	-9220	-9307	-9205	-9222	-9315	-8594
25 (5)		-6701	-9177	-9287	-9152	-9185	-9304	-8334
10 (2)	H	-8902	-9064	-9063	-9067	-9058	-9053	-8873
15 (3)		-8572	-9195	-9195	-9166	-9056	-9048	-9044
20 (4)		-7792	-9230	-9226	-9172	-8864	-8839	-8928
25 (5)	V	-6588	-9231	-9220	-9146	-8835	-8885	-8703
20 (2)		-8124	-8100	-8103	NA	-8093	-8096	-8230
20 (3)		-8665	-8909	-8915	-8830*	-8917	-8900	-8829
20 (5)	P	-8055	-9185	-9147	-9079	-9200	-9168	-8966
20 (6)		-7399	-9175	-9133	-9110	-9199	-9162	-8972
20 (2)		-8282	-8269	-8263	NA	-8257	-8262	-8372
20 (3)	V	-8672	-8929	-8934	-8834*	-8946	-8953	-8841
20 (5)		-8071	-9167	-9202	-9101	-9179	-9223	-8750
20 (6)		-7421	-9164	-9194	-9117	-9185	-9218	-8744
20 (2)	P	-8553	-8525	-8528	NA	-8529	-8531	-8657
20 (3)		-8387	-8795	-8795	-8701*	-8613	-8609	-8587
20 (5)		-7278	-9060	-9063	-8977	-8205	-8240	-8057
20 (6)		-5470	-9052	-9053	-8988	-7972	-7827	-7853

Table 3

Simulation results for the horizontal direction of constancy with lag $\delta = 0.2$ based on 100 simulated data sets. Top 4 rows give the summary of the equivalent degrees of freedom (EDF) and the bottom 4 rows give the summary of the hAIC with L_2 norm. Directions of penalization: D^H - horizontal, D^V - vertical, D^{HV} - horizontal and vertical, D^P - parallel.

LS		L_2 -norm				L_1 -norm			
M(B)		D^H	D^V	D^{HV}	D^P	D^H	D^V	D^{HV}	D^P
10(2)	30	29.01	29.54	28.27	29.59	26.28	28.61	28.49	28.50
15(3)	58	44.80	50.35	41.11	52.74	36.02	37.53	46.82	43.98
20(4)	95	54.92	66.29	49.96	68.72	40.86	54.10	52.15	52.99
25(5)	141	63.25	78.59	58.08	80.89	46.17	57.16	60.51	59.32
10(2)	-8818	-8861	-8850	-8837	-8846	-8857	-8836	-8760	-8835
15(3)	-8977	-9238	-9197	-9179	-9175	-9215	-9201	-8934	-9161
20(4)	-8576	-9331	-9268	-9249	-9242	-9305	-9251	-9228	-9218
25(5)	-6754	-9341	-9259	-9227	-9231	-9299	-9228	-9171	-9194

Table 4

The hAIC values from the analysis of the boilermaker data set. **Bold** numbers indicate the smallest values of hAIC in each row and **Bold Italic** the second smallest (for fixed M and B). Directions of penalization: D^H - horizontal, D^V - vertical, D^{HV} - horizontal and vertical, D^P - parallel. NA denotes the entries with computationally non-estimable coefficients.

M	B	LS		L_2 -norm			L_1 -norm			D^P
				D^H	D^V	D^{HV}	D^H	D^V	D^{HV}	
17	3	-2185		-1624	-1624	NA	-1624	-2041	-1965	-1832
	5	-2406		-1624	-1624	NA	-1624	-2324	-2146	-1408
	7	-2611		-1624	-1962	NA	-1624	-2507	-2347	-1307
	9	-2773		-1624	-2269	NA	-1624	-2615	-2391	-779
34	6	-3090		-1624	-2140	NA	-1624	-2465	-2787	-2062
	10	-3776		-1624	-2908	-2470	-1624	-3565	-3544	907
	14	-3869		-1624	-3603	-2911	-1624	-3450	-4640	1677
	18	-4668		-1624	-4020	-3206	-1624	-4546	-4693	2119
51	9	-4006		-1940	-2859	-2338	-1624	-3733	-3393	-734
	12	-4282		-1624	-3530	-2707	-1624	-3800	-5046	2355
	15	-4580		-1624	-4406	-3156	-1624	-4014	-5785	-233
	18	-5715		-1624	-5213	-3622	-1624	-5269	-6458	3010
	21	-5484		-1624	-5257	-3654	-1624	-5702	-6680	4189
	24	5379		-1624	-5225	-3628	-1624	-6006	-6873	3397
	25	12010		-1624	-5240	-3638	-1624	-6110	-6930	3503
	26	69371		-1624	-5271	NA	-1624	-6151	-7013	NA
	27	56509		-1624	-5214	NA	-1624	-6236	-6648	NA