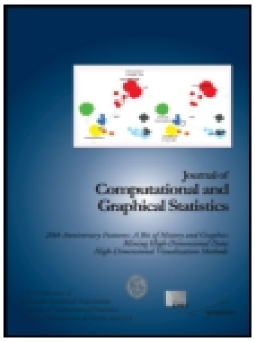

Journal of Computational and Graphical Statistics

ISSN: 1061-8600 (Print) 1537-2715 (Online) Journal homepage: http://www.tandfonline.com/loi/ucgs20

Wavelet-Based LASSO in Functional Linear Regression

Yihong Zhao , R. Todd Ogden & Philip T. Reiss

To cite this article: Yihong Zhao , R. Todd Ogden & Philip T. Reiss (2012) Wavelet-Based LASSO in Functional Linear Regression, Journal of Computational and Graphical Statistics, 21:3, 600-617, DOI: 10.1080/10618600.2012.679241

To link to this article: https://doi.org/10.1080/10618600.2012.679241

Supplementary materials for this article are available online.
Please go to www.tandfonline.com/r/JCGS

Wavelet-Based LASSO in Functional Linear Regression

Yihong ZHAO, R. Todd OGDEN, and Philip T. REISS

In linear regression with functional predictors and scalar responses, it may be advantageous, particularly if the function is thought to contain features at many scales, to restrict the coefficient function to the span of a wavelet basis, thereby converting the problem into one of variable selection. If the coefficient function is sparsely represented in the wavelet domain, we may employ the well-known LASSO to select a relatively small number of nonzero wavelet coefficients. This is a natural approach to take but to date, the properties of such an estimator have not been studied. In this article we describe the wavelet-based LASSO approach to regressing scalars on functions and investigate both its asymptotic convergence and its finite-sample performance through both simulation and real-data application. We compare the performance of this approach with existing methods and find that the wavelet-based LASSO performs relatively well, particularly when the true coefficient function is spiky. Source code to implement the method and datasets used in the study are provided as supplementary materials available online.

Key Words: Functional data analysis; Penalized linear regression; Variable selection; Wavelet regression.

1. INTRODUCTION

With advances in technology, it is increasingly common to encounter data that are functional or curves in nature. In the recent literature, extensive research has focused on problems with regression of a scalar response on a functional predictor. We will consider the functional linear regression model

$$y_i = \alpha + \int x_i(t)\omega(t)\,dt + \varepsilon_i, \quad i = 1, \ldots, n, \tag{1}$$

where the intercept parameter α is scalar, the error terms are iid $N(0, \sigma^2)$ random variables, and the coefficient ω is a square-integrable function on a compact interval $I \subset R$. As with

Yihong Zhao is Assistant Professor, Department of Psychiatry, Columbia University, New York, NY (E-mail: *yz2135@columbia.edu*). R. Todd Ogden is Professor, Department of Biostatistics, Columbia University, New York, NY (E-mail: *to166@columbia.edu*). Philip T. Reiss is Assistant Professor, Department of Child and Adolescent Psychiatry, New York University School of Medicine, New York, NY and Nathan S. Kline is Research Scientist, Institute for Psychiatric Research, Orangeburg, NY (E-mail: *phil.reiss@nyumc.org*).

Journal of Computational and Graphical Statistics, Volume 21, Number 3, Pages 600–617
DOI: 10.1080/10618600.2012.679241

the standard linear regression model, ω describes the relationship between X and Y. The regions where $|\omega(t)|$ is large indicate that changes in $X(t)$ have greater predictive power on Y.

This type of model has applications in many fields including chemometrics, where direct measurement of some chemical contents often requires expensive and time-consuming analysis in a laboratory. However, near-infrared (NIR) spectra of the samples are easy and inexpensive to obtain, and so it is of interest to use NIR spectra to predict chemical contents in a sample. In spectroscopic applications, X represents the spectrum (typically NIR) of a sample and Y is some chemical compound. For instance, Kalivas (1997) used the NIR spectrum of a sample (X) to determine moisture and protein contents (Y) in wheat. Recently, Saeys, Ketelaere, and Darius (2008) discussed the potential use of functional linear regression with a scalar response in the field of chemometrics. Interested readers are referred to Ramsay and Silverman (2002) for an overview about other applications in functional linear regression.

In this article, our primary interest is in estimating the regression coefficient function ω, as the intercept α can be easily calculated via $\hat{\alpha} = \bar{y} - \int \bar{x}(t)\omega(t)dt$, where $\bar{y}$ and $\bar{x}(t)$ are the sample means of y_is and $x_i(t)$s, respectively (Cai and Hall 2006). Thus, we will drop the term α from Equation (1) for ease of presentation. In practice, the functional predictor x_i's are observed at N discrete points, and quite often N is much larger than n. A classical multiple regression approach to this problem fails to provide a meaningful or consistent estimator of the coefficient function in the model (1). Therefore, regularization or dimension reduction is necessary. Classical multivariate methods include Principal Component Regression (PCR; Massy 1965) and Partial Least Squares (PLS; Wold 1975; de Jong 1993; Goutis and Fearn 1996). Alternatively, dimension reduction can be achieved through sufficient dimension reduction (SDR) approaches in which the goal is to retain all the relevant information in X that can be used to predict Y. However, if any of these methods are applied directly to the discretized data, then they are effectively ignoring the functional nature of the data.

Many procedures have been proposed to estimate the coefficient function ω. For instance, a functional principal component regression (FPCR) based approach has been taken by Cardot, Ferraty, and Sarda (1999); James, Hastie, and Sugar (2000); Müller and Stadtmüller (2005); Müller and Yao (2008); and Delaigle, Hall, and Apanasovich (2009). Cai and Hall (2006) and Hall and Horowitz (2007) discussed the convergence rates of the FPCR-based estimator. Marx and Eilers (1999) and Cardot, Ferraty, and Sarda (2003) derived estimators of ω based on B-splines applying different penalties to control the roughness of the estimators, while Reiss and Ogden (2007) proposed to use a combination of FPCR and penalization to estimate ω. Ferre and Yao (2003) introduced a functional version of sliced inverse regression. James, Wang, and Zhu (2009) assumed sparsity in the derivatives of ω and used variable selection ideas to produce interpretable and accurate estimators.

For a number of reasons, we prefer to use wavelets as a tool for dimension reduction. First, the wavelet transform is computationally efficient. Second, the wavelet transform allows us to estimate functions with features at a variety of different scales. Third, wavelets are especially good at handling sharp, highly localized features. Fourth and most important, the wavelet transform provides a sparse representation of a variety of functional forms that occurred in real-life applications. That is, for a large variety of functions, the wavelet decomposition allows good representation of the function using only a relatively small number of wavelet coefficients. A comprehensive survey of wavelet applications in statistics can be found in literature by Ogden (1997) or Vidakovic (1999).

A wavelet-based approach to functional linear regression has received some attention in recent literature. For example, Brown, Fearn, and Vannucci (2001) used a Bayesian variable selection approach to select a subset of important wavelet coefficients that can explain the response well. Amato, Antoniadis, and De Feis (2006) extended the multivariate dimension reduction method, minimum average variance estimation (MAVE), to the functional setting through the use of wavelets. The key idea of their method relied on replacing each functional predictor by SDR in the wavelet domain. More recently, Malloy et al. (2010) developed wavelet-based linear mixed distributed lag models incorporating functional data as covariates into a linear mixed model.

In this article, we apply the least absolute shrinkage and selection operator (LASSO) developed by Tibshirani (1996) in the wavelet domain to the functional regression situation. The central idea is to transform each functional predictor into a set of wavelet coefficients and then to select good predictors of the response variable from among these. This converts the functional regression problem into a high-dimensional variable selection problem. We assume that only a relatively small number of wavelet coefficients are useful for predicting the response, as the wavelet transform provides a sparse representation for many functions encountered in real-life application. A natural way to handle this situation is through the penalized least squares approach with an L_1-type penalty, since this penalty enforces sparse solutions. Tibshirani et al. (2005) suggested that such an approach might be useful in some situations, but briefly mentioned that the result of such an approach in their proteomic study was disappointing. However, their analysis was based on the Haar wavelet transform of proteomic spectra, which is not likely to express spiky data well, suggesting that this approach may merit further study. Here, our interest lies in predictors resulting from sampling a continuous curve at equally spaced points, such as those of NIR spectra, and investigating whether wavelet-based LASSO is viable for regression of a scalar on functions. The main contributions of our work are as follows. First, we describe the wavelet-based LASSO approach to regressing scalar responses to functional predictors. Second, we study the asymptotic convergence rate of the wavelet-based LASSO estimator. Finally, we investigate its finite-sample performance through both simulation and real-data application.

The rest of the article is organized as follows. In Section 2, we provide some necessary background on wavelets and review the LASSO method. We also describe the implementation of the wavelet-based LASSO method and discuss various methods for determining the tuning parameter. Section 3 gives asymptotic properties of the estimated regression function. We illustrate the wavelet-based LASSO on both simulated and real data in Section 4. Finally, Section 5 provides some discussion and directions for future research.

2. WAVELET-BASED LASSO

2.1 SOME BACKGROUND ON WAVELETS

Wavelets are basis functions that can be used to efficiently approximate other functions with relatively few nonzero wavelet coefficients. The construction of a wavelet family starts with two related and suitably chosen orthonormal basic functions: the scaling function ϕ and the mother wavelet ψ. A wavelet system is generated by dilation and translation of ϕ and ψ through

$$\phi_{j,k}(t) = 2^{j/2}\phi(2^j t - k), \quad \psi_{j,k}(t) = 2^{j/2}\psi(2^j t - k),$$

where the dilation index j and translation index k are integers. A wavelet has p vanishing moments iff its scaling function ϕ can generate polynomials of degree smaller than or equal to p. The above definitions give wavelet functions defined on all of $\mathbb{R}$, but in our application we are interested only in a finite interval of $\mathbb{R}$, taken without loss of generality to be [0, 1]. We require an orthonormal wavelet basis on [0, 1] such as that resulting from applying the so-called periodic boundary handling method.

Given a primary decomposition level j_0, the wavelet decomposition of a function ω in $L^2[0, 1]$ can be represented as

$$\omega(t) = \sum_{k=0}^{2^{j_0}-1} \beta'_{j_0,k}\phi_{j_0,k}(t) + \sum_{j=j_0}^{\infty}\sum_{k=0}^{2^j-1} \beta_{j,k}\psi_{j,k}(t),$$

where the wavelet coefficients are defined by

$$\beta'_{j_0,k} = \int \omega(t)\phi_{j_0,k}(t)dt, \quad \beta_{j,k} = \int \omega(t)\psi_{j,k}(t)dt.$$

Coefficients at coarser levels capture global features of the data while those at finer levels describe local characteristics.

2.2 Functional Linear Regression in the Wavelet Domain

We begin by expressing both the $\omega(t)$ function and the functional predictor in model (1) in terms of their wavelet components, where $\omega(t)$ is as in Section 2.1 and $x_i(t)$ is expressed as

$$x_i(t) = \sum_{k=0}^{2^{j_0}-1} z'_{i,j_0,k}\phi_{j_0,k}(t) + \sum_{j=j_0}^{\infty}\sum_{k=0}^{2^j-1} z_{i,j,k}\psi_{j,k}(t),$$

where

$$z'_{i,j_0,k} = \int x_i(t)\phi_{j_0,k}(t)dt, \quad z_{i,j,k} = \int x_i(t)\psi_{j,k}(t)dt.$$

In practice, the functional predictors $x_i(t)$ are discretely sampled at N equally spaced points, thus, a maximum of $J = \log_2(N) - 1$ levels of decomposition can be performed via discrete wavelet transform (DWT), resulting in an $N \times 1$ vector of wavelet coefficients z_i. When N is a power of 2, the DWT can be performed in only $O(N)$ operations through a fast pyramid algorithm developed by Mallat (1989). The reconstruction of the function can be obtained through a fast inverse DWT.

Due to the orthonormality of the wavelet basis, model (1) becomes simply

$$y_i = \sum_{k=0}^{2^{j_0}-1} z'_{i,j_0,k}\beta'_{j_0,k} + \sum_{j=j_0}^{J}\sum_{k=0}^{2^j-1} z_{i,j,k}\beta_{j,k} + \varepsilon_i, \quad i = 1, \ldots, n.$$

In matrix form we have

$$Y = Z\beta + \varepsilon, \tag{2}$$

where $Y = (y_1, y_2, \ldots, y_n)^T$, $\varepsilon = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)^T$ with ε_i defined in (1), β is an $N \times 1$ vector of coefficients, and Z is an $n \times N$ design matrix whose ith row contains the $z_{i,j,k}$ values in the same order as β. We note that the model (2) depends on the choice of the primary decomposition level j_0, but we suppress this dependency for ease in notation. The derived variables, the wavelet coefficients, $z_{i,j,k}$s become the predictors in the transformed space. We now have N potential predictors and among those only a relatively small subset of these predictors is important for the prediction of the response Y.

The functional regression problem may be considered as a variable selection problem in that we must identify the few important wavelet coefficients in (2), which will determine the fitted $\hat{\omega}$ function. To find a subset of the wavelet coefficients with good predictivity, a number of approaches have been proposed in the literature. Brown, Fearn, and Vannucci (2001) investigated a wavelet-based Bayesian variable selection method. Specifically, with the use of mixture priors, a single subset of wavelet coefficients that best predicts the response was chosen through a Metropolis search. This method is computationally intensive due to the use of a large-scale Markov chain Monte Carlo (MCMC) algorithm. Inspired by a similar problem, Amato, Antoniadis, and De Feis (2006) proposed a method based on the SDR technique MAVE (Xia et al. 2002).

2.3 VARIABLE SELECTION BY LASSO

We take advantage of sparse representation of the functions in the wavelet domain and tackle this problem by the so-called LASSO method, that is, applying an L_1 penalty to the coefficient vector β when fitting the regression model (2). The LASSO, introduced by Tibshirani (1996), is a regularization technique that performs model estimation and variable selection simultaneously and which has become popular for several reasons. First, LASSO results in sparse solutions because of the nature of the L_1 penalty on coefficient vector and hence gives more interpretable fitted models. In addition, the LASSO is computationally feasible even for high-dimensional data. Using the homotopy method (Osborne, Presnell, and Turlach 2000) or the LARS algorithm (Efron et al. 2004), the entire regularization path of LASSO can be computed with the same computational effort that is required to perform one single ordinary least squares fit. Furthermore, LASSO can be applied to the situation where there are more predictor variables than samples.

The LASSO estimate for the coefficients of the regression model (2) is obtained by

$$\hat{\beta}_\lambda = \text{argmin}_\beta \frac{1}{2}(Y - Z\beta)^T(Y - Z\beta) + \lambda \sum_{j=1}^{N} |\beta_j|, \tag{3}$$

where λ is a nonnegative tuning parameter. The fitted coefficient function $\hat{\omega}$ can be easily obtained by mapping $\hat{\beta}$ back to the original domain using the inverse DWT.

2.4 CHOICE OF THE TUNING PARAMETERS

The tuning parameter λ in Equation (3) controls the complexity of the fitted models. When $\lambda = 0$, the LASSO solution becomes an ordinary least square estimate and will typically be badly overparameterized. Conversely, a very large value of λ will shrink almost all β_j estimates to zero. Selecting a model with either too few or too many predictors can degrade the efficiency of the resulting estimator and yield less accurate predictions. Therefore, choosing the appropriate amount of regularization in the LASSO estimates is of critical importance.

Another tuning parameter that must be considered is the lowest level of decomposition j_0. The total number of coefficients is determined by N, but among those, j_0 controls how many are "averaging" coefficients and how many are "differencing" coefficients, corresponding to scaling function(ϕ) and wavelet function(ψ), respectively. The lowest level of decomposition j_0 ranges from 0 to $\log_2(N) - 1$. For a given j_0, the number of scaling and wavelet coefficients is 2^{j_0} and $N - 2^{j_0}$, respectively. Based on our experience we have found that this choice does impact the resulting estimate and should be carefully considered. In this study, the tuning parameters λ and j_0 are chosen by K-fold cross-validation (CV).

3. CONSISTENCY OF THE WAVELET-BASED LASSO ESTIMATOR

In this section, we investigate the behavior of our wavelet-based LASSO estimator when both $n \to \infty$ and $N \to \infty$, meaning thereby that as the sample size n increases, the curves are also more densely observed. Let N_n be the number of discrete points at which the functional predictors are observed when the sample size is n. Let z_i be the $N_n \times 1$ vector of the observed wavelet coefficients for ith sample, $i = 1, 2, \dots, n$. To simplify the treatment, we reindex the wavelet functions with a single index, starting with the lowest (coarsest) level. Here, we assume that the functional predictors are fixed, thus, z_i's are also fixed. However, we note that for random functional predictors, the result holds conditionally on the predictors. Let $Z_n = (z_1, z_2, \dots, z_n)^T$. Let $\Sigma_n = \frac{1}{n} Z_n^T Z_n$, and ρ_n be the smallest eigenvalue of Σ_n. Let λ_n denote the penalty parameter λ when the sample size is n. To derive the convergence rate of $\hat{\omega}$ to ω, we also need to make the following assumptions.

A1. $\varepsilon_1, \dots, \varepsilon_n$ are independent and identically distributed variables with zero mean and finite variance σ^2.

A2. There exists M such that $\|X_i\| < M$ for all i.

A3. ω is a q times differentiable function in the Sobolev sense (i.e., $\omega \in W^q[0, 1]$), and the wavelet basis has p vanishing moments, where $p > q$.

A4. $(\sum_{j=0}^{\infty} |\langle \omega, \psi_j \rangle|^r)^{1/r} < \infty$ for some $r < 2$.

A5. $\lambda_n = o(n)$ and $n/\lambda_n = O(N_n^{2/r})$.

A6. $\frac{N_n}{n\rho_n} \to 0$, $\frac{(\lambda_n/n)^{1-r/2}}{\rho_n} \to 0$, and $\frac{1}{N_n^{2q}\rho_n} \to 0$.

Assumption (A6), which ensures that the order of the L^2 error as given by Theorem 1 tends to zero, requires that the smallest eigenvalue of Σ_n not go to zero too quickly. This may be interpreted as requiring that as both n and N_n grow large (i.e., as both the sample size and the resolution at which the predictors are sampled increase), the improvement of resolution brings with it sufficient "new information" to allow for estimation of ω. We note that A6 implies n grows faster than N.

Theorem 1. Let $\hat{\omega}_n$ be the estimator resulting from (3). If conditions (A1)–(A6) hold, then $||\hat{\omega}_n - \omega||^2 = O_p(\frac{(\lambda_n/n)^{1-r/2}}{\rho_n}) + O_p(\frac{N_n}{n\rho_n}) + o_p(\frac{1}{N_n^{2q}\rho_n})$.

A detailed proof of the theorem is provided in the supplementary material (available online).

4. NUMERICAL STUDIES

The performance of our wavelet-based LASSO in finite sample cases is studied through simulations and a real data example. We compare the results with the penalized B-splines expansion approach of Cardot, Ferraty, and Sarda (2003). We use Daubechies' least asymmetric wavelet with eight vanishing moments for both the simulation study and the real data analysis.

4.1 SIMULATION STUDY

In the simulation, each functional predictor $x_i(t)$ is a Brownian bridge stochastic process for $t \in (0, 1)$. That is, $X(t)$ is chosen to be a zero-mean Gaussian process with continuous sample functions and $\mathrm{cov}(X(t), X(s)) = s(1-t)$ for $s < t$, with $X(0) = X(1) = 0$. We apply DWT with periodic boundary correction on each of the predictors. We investigate both smooth and bumpy coefficient functions. The smooth function was defined by $\omega(t) = 0.03f(t, 20, 60) - 0.05f(t, 50, 20)$, where $f(t, \alpha, \beta) = (\Gamma(\alpha+\beta))/(\Gamma(\alpha)\Gamma(\beta))t^{\alpha-1}(1-t)^{\beta-1}$. The bumpy function is modified from the "bumps" test function of Donoho and Johnstone (1994). To demonstrate the performance of the proposed method for different noise levels, we set the variance of the error term σ^2 so that the signal-to-noise ratio (SNR), measured by the squared multiple correlation coefficient of the true model $(R^2) = 0.5, 0.7$, and 0.9. Each curve is calculated at $N = 128$ equally spaced time points. The sample sizes n are 100 and 200. We carried out 200 simulations for each setting of the parameters.

The tuning parameters λ and j_0 are selected by fivefold CV. For the penalized B-splines method, the degrees of the spline functions and the number of derivatives controlling for the smoothness of the estimated function are set to four and two, respectively. As suggested by Cardot, Ferraty, and Sarda (2003), the number of knots of the spline functions and the regularization parameter λ for the model are selected by generalized cross-validation (GCV).

The performance of the methods is compared according to two criteria: the mean squared error (MSE) of prediction defined by $\frac{1}{n}(\hat{\omega}-\omega)^T X^T X(\hat{\omega}-\omega)$ and the mean of the integrated squared error (ISE) $\frac{1}{N}(\hat{\omega}-\omega)^T(\hat{\omega}-\omega)$. The means and boxplots of

Table 1. Mean of mean squared prediction errors for simulation

Method	n	$R^2 = 0.9$	$R^2 = 0.7$	$R^2 = 0.5$
Bumpy function				
LASSO	200	0.0046	0.0125	0.0246
B-splines	200	0.0082	0.0192	0.0377
LASSO	100	0.0074	0.0204	0.0390
B-splines	100	0.0120	0.0372	0.0456
Smooth function				
LASSO	200	0.0197	0.0640	0.1306
B-splines	200	0.0206	0.0679	0.1447
LASSO	100	0.0336	0.1114	0.2315
B-splines	100	0.0376	0.1283	0.3431

mean squared prediction errors of each method under different settings are presented in Table 1 and Figure 1. The wavelet-based LASSO approach tends to predict better than the penalized spline estimator under the given settings. We note that the wavelet-based LASSO obtained a lower MSE. However, the B-splines method obtained a lower ISE for the smooth function. This could be due to CV sometimes selecting a too-large value for j_0. Because wavelet coefficients of the regression function are constrained to be sparse, this essentially eliminates large-scale features and relies instead on local smaller scale features, and this is evidently more detrimental to estimation (ISE) than to prediction. In separate simulations, we have confirmed that restricting the choice of j_0 in fact circumvents this problem, resulting in an estimation method that outperforms the B-splines method in both the smooth case and the bumpy case in both prediction and estimation.

The means and boxplots of the ISE for each method are summarized in Table 2 and Figure 2. We observe similar patterns as for the MSEs for the prediction. The mean estimated functions at $R^2 = 0.9$ from 200 simulated datasets are plotted in Figure 3. The penalized B-spline approach outperforms the wavelet-based approaches for the smooth function, while wavelet-based approaches show a clear advantage over the penalized B-spline method for the bumpy function. It should be noted that there is a high variability in the estimates of ISE for both methods. For the B-splines method, this might be due to an inclination of GCV to undersmooth (Cardot, Ferraty, and Sarda 2003). We acknowledge that the restricted maximum likelihood-based smoothness selection might be less susceptible to overfitting than GCV (Reiss and Ogden 2009).

The boxplots in Figure 4 indicate that the selected decomposition level depends on the smoothness of the underlying function, demonstrating that the lowest level of decomposition j_0 is another key factor in fitting these models.

Figure 5 displays the distributions of nonzero coefficients in the final models from 200 simulated data for both the bumpy and the smooth functions at $R^2 = 0.9$ and $R^2 = 0.5$. Clearly, the number of nonzero coefficients depends on the smoothness of the underlying function. As would be expected, there are more nonzero coefficients in the bumpy function setting than in the smooth function setting, and the number of nonzero coefficients increases as the sample size increases. It should also be noted that the model tends to pick up more nonzero coefficients (possibly noise coefficients) as the SNR decreases.

Table 2. Mean of integrated squared errors for simulation

Method	n	$R^2 = 0.9$	$R^2 = 0.7$	$R^2 = 0.5$
Bumpy function				
LASSO	200	0.0050	0.0065	0.0076
B-splines	200	0.0067	0.0076	0.0077
LASSO	100	0.0060	0.0075	0.0077
B-splines	100	0.0074	0.0076	0.0076
Smooth function				
LASSO	200	0.0019	0.0050	0.0083
B-splines	200	0.0012	0.0029	0.0054
LASSO	100	0.0030	0.0074	0.0136
B-splines	100	0.0024	0.0065	0.0082

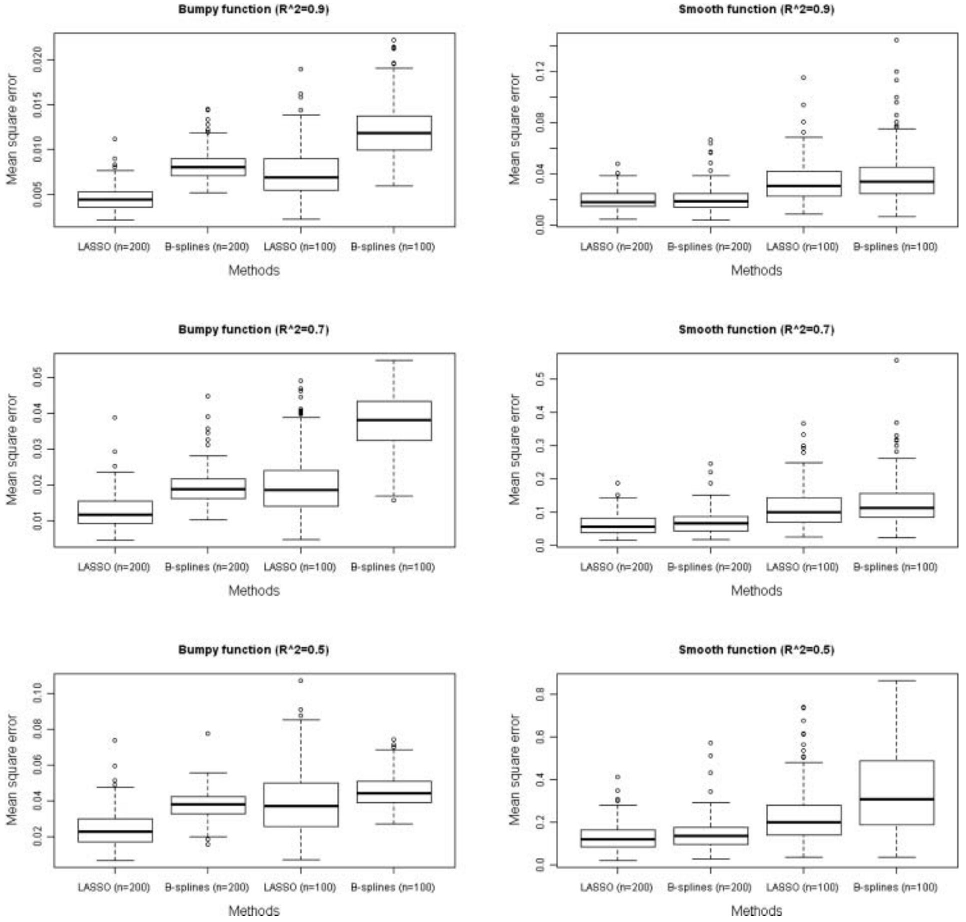

Figure 1. Boxplots of MSE for bumpy (left column) and smooth (right column) coefficient functions based on 200 simulated datasets (top row: $R^2 = 0.9$, middle row: $R^2 = 0.7$, bottom row: $R^2 = 0.5$).

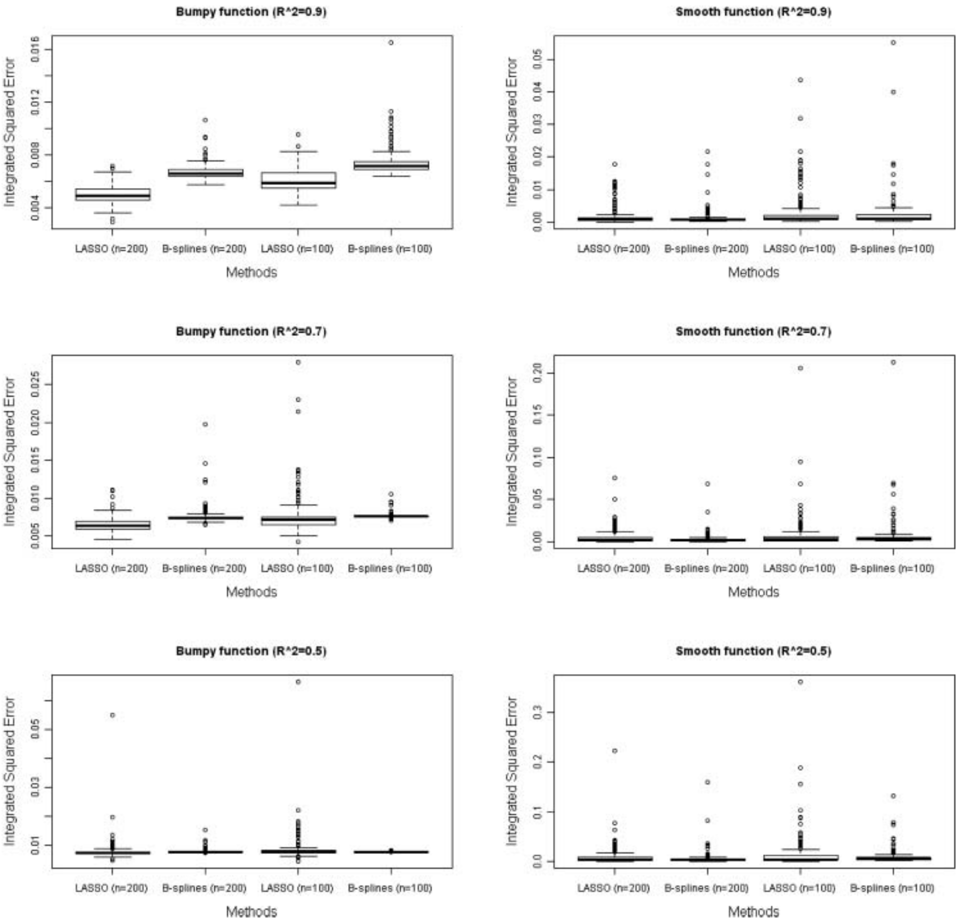

Figure 2. Boxplots of integrated squared error for bumpy (left column) and smooth (right column) coefficient functions based on 200 simulated datasets (top row: $R^2 = 0.9$, middle row: $R^2 = 0.7$, bottom row: $R^2 = 0.5$).

4.2 APPLICATION TO REAL DATA

In this section, we apply the wavelet-based LASSO approach to real datasets representing two situations: $n < N$ (the "wheat data") and $n > N$ (the "Tecator data"). The wheat data consist of NIR spectra of 100 wheat samples (Figure 6(a)) and two associated response variables: the moisture and the protein contents (Kalivas 1997). NIR spectra are measured using diffuse reflectance as $\log(1/R)$ at wavelength from 1100 to 2500 nm in 2 nm intervals. In our analysis, we use the once-differenced spectra (Figure 6(b)) to correct for a baseline shift. We apply linear interpolation to approximate the original curves at 512 equally spaced points (Figure 6(c)). Then, we apply the DWT to each interpolated curve and the sample variance of each wavelet coefficient is plotted in Figure 6(d), where the resolution levels, ordered from finest to coarsest, are separated by vertical lines.

The Tecator data consist of NIR spectra of 240 samples of ground pork (Thodberg 1996). It is of interest to predict the wet-chemistry measurements, using the corresponding NIR spectra, on the fat, water, and protein contents, respectively. The NIR spectra

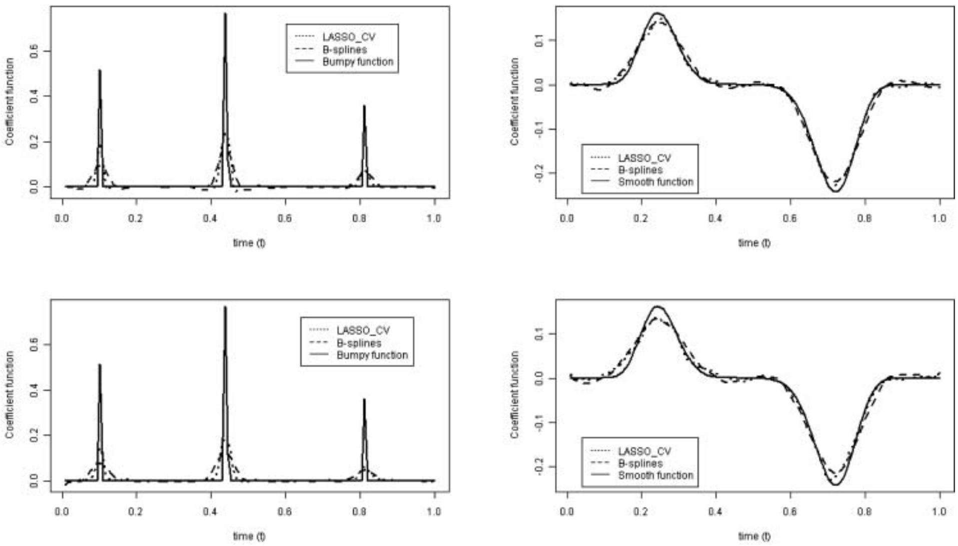

Figure 3. Mean estimated functions for bumpy (left column) and smooth coefficient functions (right column) at $R^2 = 0.9$ based on 200 simulated datasets (top row: $n = 200$, bottom row: $n = 100$).

are recorded on a Tecator Infrared spectrometer that measures the absorbance at 100 wavelengths in the region 850–1050 nm. In this study, we focus on the absorbance at 64 wavelengths in the region of 902–1028 nm, because this region is considered to be the most informative part of the spectrum for the three contents (Amato, Antoniadis, and

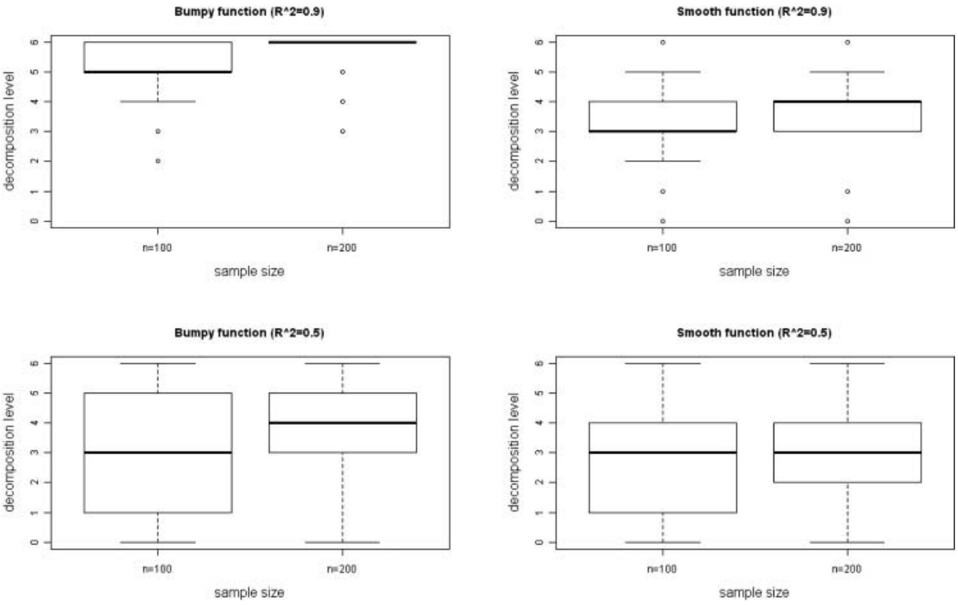

Figure 4. Selected level of decomposition for bumpy (left column) and smooth (right column) coefficient functions based on 200 simulated datasets (top row: $R^2 = 0.9$, bottom row: $R^2 = 0.5$).

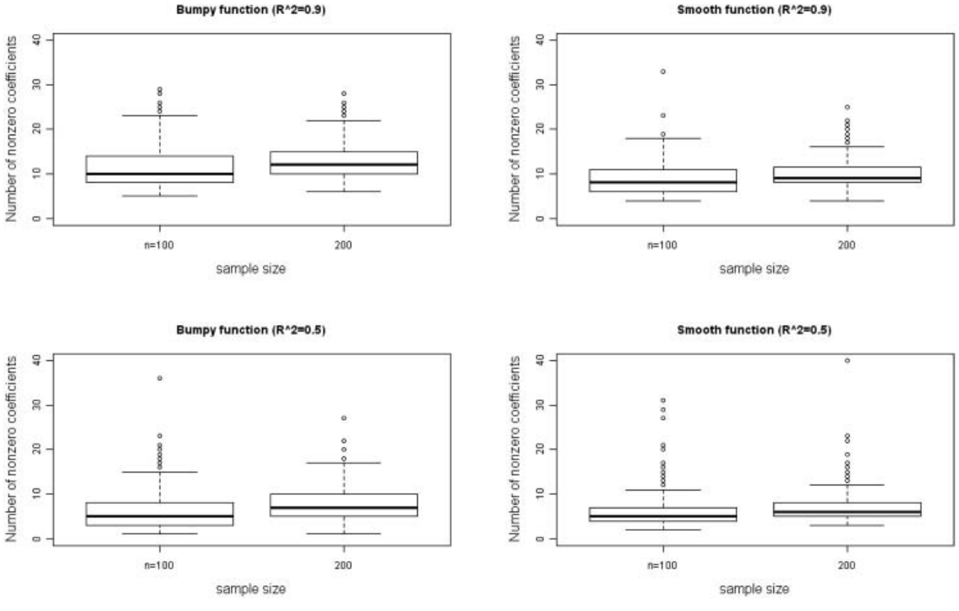

Figure 5. Number of nonzero coefficients for bumpy (left column) and smooth (right column) coefficient functions in the final model based on 200 simulated datasets (top row: $R^2 = 0.9$, bottom row: $R^2 = 0.5$).

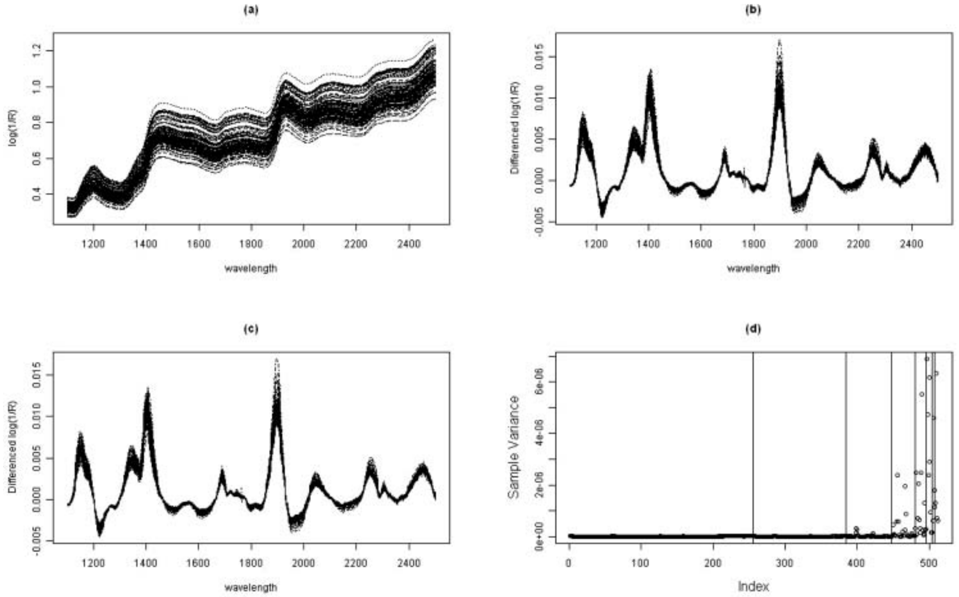

Figure 6. NIR spectra from the wheat data and the sample variance in wavelet domain. (a): original spectra; (b): once-differenced spectra; (c): linearly interpolated once-differentiated spectra; (d): sample variance of each wavelet coefficient with $j_0 = 2$. The wavelet coefficients at various resolution levels, ordered from finest to coarsest, are separated by vertical lines.

Table 3. Mean sum of squared errors of prediction

Method	Moisture	Protein	
	(a) The wheat data		
LASSO	0.2882	0.8871	
B-splines	0.3346	1.9804	
Method	Water	Fat	Protein
	(b) The Tecator data		
LASSO	0.0035	0.0050	0.0008
B-splines	0.0051	0.0059	0.0008

De Feis 2006). We use the first 215 samples in our study, as suggested by the weblink (*http://lib.stat.cmu.edu/datasets/tecator*).

To assess the predictive ability of each method, we randomly split the samples into five subsets. We fit the model to four of these subsets (the training sample) and predicted the outcomes of the other subset (the validation sample), repeating this process until each subset has been the validation sample once. The mean sum of squared errors of prediction over five validation sets is calculated as $\sum_{i=1}^{5}\sum_{j=1}^{n_i}(y_j - \hat{y}_j)^2$, where n_i is the number of samples in ith validation set. The results are summarized in Table 3. The wavelet-based approach outperforms the penalized B-spline approach in both studies.

Figure 7 shows the estimated coefficient functions from the five fitted models for each method with the moisture as response variable in the wheat data and the moisture, fat, and protein contents in the Tecator data. The gray curve in each plot is the mean estimated curve from five training models. For the wheat data, the wavelet-based LASSO approach tends to produce very sparse models. There are only seven or eight nonzero wavelet coefficients at low resolution levels in the final models. The estimated coefficient functions for the model with protein values as the response are less consistent and interpretable, regardless of the methods used. This may be partly because the protein values are known to have poor precision (Centner et al. 2000). For the Tecator data, the estimated coefficient functions tend to be rather spiky for all three response variables. In general, the wavelet-based approach produces very stable coefficient function estimates across the training sets.

5. DISCUSSION

In this article, we consider a wavelet-based LASSO approach for estimating the coefficient function in functional linear regression and derive its asymptotic properties. Compared to the B-splines method, our results suggest that the wavelet-based LASSO approach performs better, particularly when the coefficient function is characterized by local features (e.g., peaks). The new approach also shows desirable properties in modeling the smooth function under certain settings. This method is implemented as "wnet" function in the publicly available R package "refund."

We also compared our approach to the wavelet-based MAVE method (Amato, Antoniadis, and De Feis 2006). The wavelet LASSO tends to do considerably better in terms of estimation accuracy and prediction ability in both a simulation study and a real-data analysis (data not shown). It should be noted that $\hat{\omega}$ for the wavelet-based MAVE method in our study is obtained by QR decomposition, as recommended by Amato, Antoniadis,

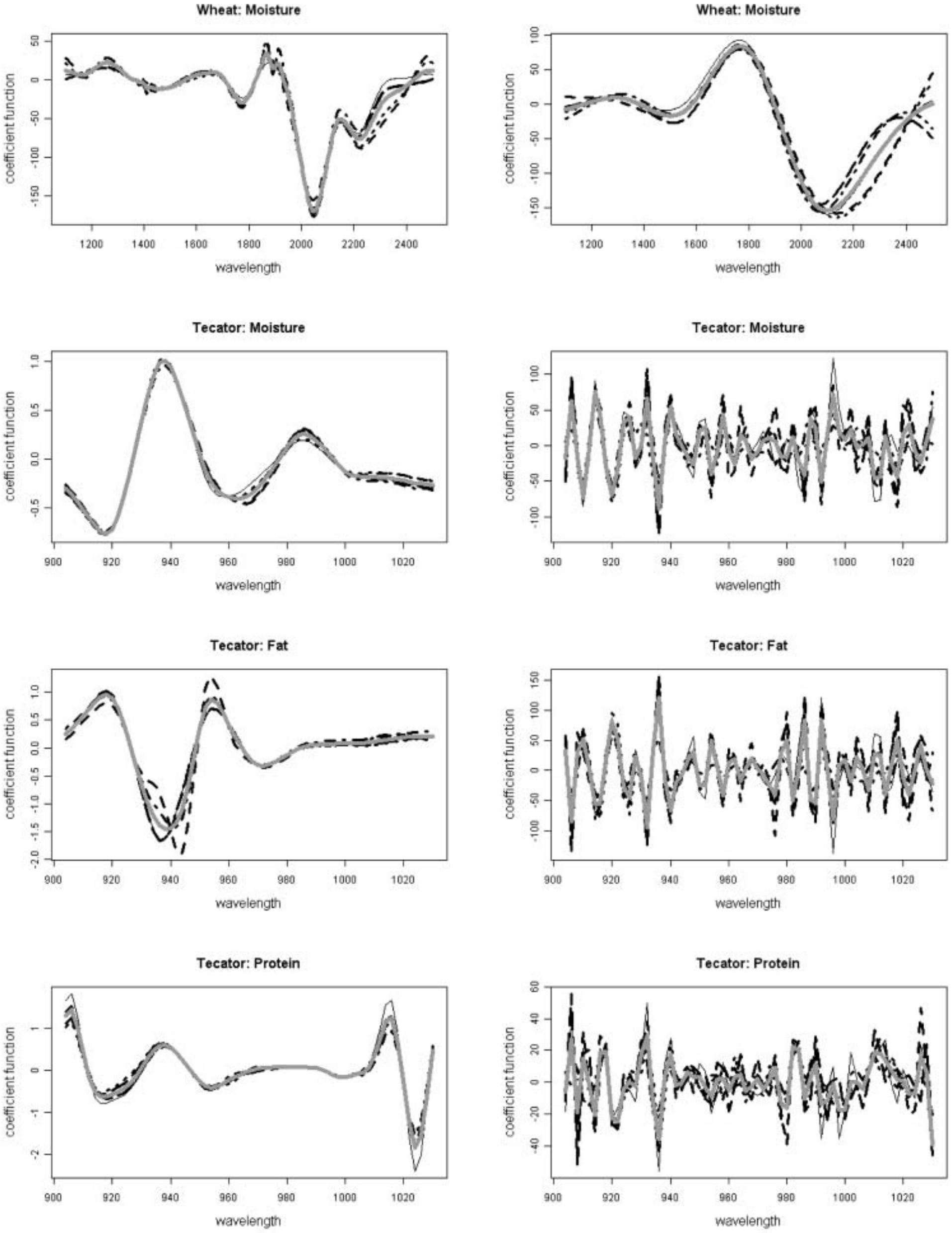

Figure 7. Estimated coefficient functions from five training models. The first row is for the wheat data and the other rows for the Tecator data (left column: wavelet-based LASSO; right column: B-splines).

and De Feis (2006). We acknowledge that the purpose of MAVE was really determining the effective direction reduction space (EDR space). The wavelet-based MAVE may be improved by using other estimates such as kernel estimates given by Xia et al. (2002). On the other hand, the subspace spanned by the variables selected by LASSO has been suggested to be close to an optimal subspace in recent work (Zhang and Huang 2008; Meinshausen and Yu 2009). In addition, the wavelet-based MAVE does not provide sparse estimation.

Each SDR predictor is a linear combination of all the original variables, making interpretation somewhat difficult. These might partly explain why the wavelet-based MAVE is less efficient than wavelet-based LASSO in terms of both prediction and estimation.

Several factors may be important to the performance of wavelet-based LASSO. First, arbitrary choice of a wavelet basis function for DWT might not be desirable for the efficient use of our proposed method. Ideally, we would like to obtain a sparse representation of the functional predictor so that all relevant information is concentrated in relatively few wavelet coefficients. In our study, we choose wavelet basis from the Daubechies family as those are considered to have good localizing properties both in temporal and frequency domains. Second, it is important to appropriately handle the boundaries in wavelet decomposition when dealing with a signal defined over an interval. In practice, we can impose extra constraint on the underlying coefficient function by assuming the function to be either periodic or symmetric reflection around the boundaries. Artifacts might be created at boundaries when the extra constraints of the coefficient function are not satisfied, resulting in bias in the model estimates around the edges. Other methods proposed to correct the boundary problems in wavelet regression (see Oh and Lee 2005 and references therein) might be extended to wavelet-based functional linear regression model setting. Third, like all other regularization methods, performance of the wavelet-based LASSO depends heavily on an appropriate choice of the regularization parameter (see Figure 8 for an example). The tuning parameters in this study are selected by a fivefold

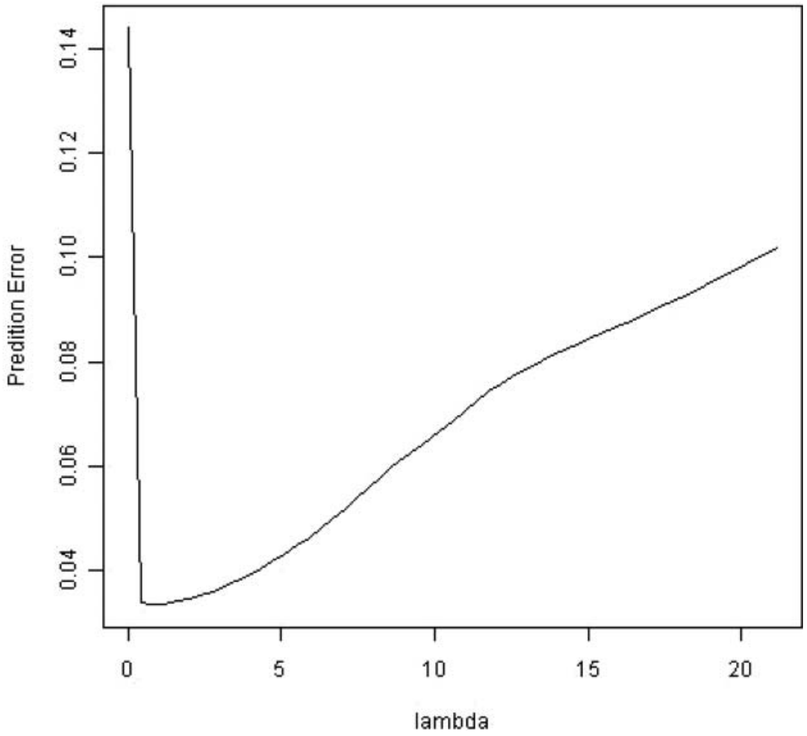

Figure 8. Prediction errors as a function of λ on the performance of the wavelet-based LASSO from single simulation.

CV criterion. This criterion is effective in choosing tuning parameters that minimize mean squared prediction errors, but performs less well in terms of minimizing ISE. We tried the Bayes' Information Criterion (BIC) type criterion in our study. BIC score is defined as $\mathrm{BIC}_\lambda = \log(\mathrm{RSS}(\lambda)/n) + (\log(n) df(\lambda))/n$, where $\mathrm{RSS}(\lambda) = (Y - \hat{Y}_\lambda)^T (Y - \hat{Y}_\lambda)$ and $df(\lambda)$ is the number of nonzero coefficients in the fitted model (Zou, Hastie, and Tibshirani 2007). However, BIC does not generally perform as well as a fivefold CV in our study.

SUPPLEMENTARY MATERIALS

Appendix: The derivation of Theorem 1 can be found in the appendix. (Appendix.pdf)

supp.zip: The zip file includes datasets used in the study and R code to perform the methods proposed in the article. (supp.zip, zip archive)

ACKNOWLEDGMENTS

The authors thank the editor, the associate editor, and the reviewers for their helpful suggestions and comments that greatly improve the manuscript. They also thank Prof. Hervé Cardot for kindly providing his computer program and Prof. Richard P. Sloan for his generous support through this project. Yihong Zhao was partly supported by the National Cancer Institute grant 5T32CA09529, R. Todd Ogden's research was supported in part by the NIH grant 5R01EB009744, and Philip T. Reiss was supported in part by the National Science Foundation grant DMS-0907017 and the National Institutes of Health grant 5R01EB009744.

[Received October 2010. Revised September 2011.]

REFERENCES

Amato, U., Antoniadis, A., and De Feis, I. (2006), "Dimension Reduction in Functional Regression With Applications," *Computational Statistics and Data Analysis*, 50, 2422–2446. [602,604,612,612]

Brown, P. J., Fearn, T., and Vannucci, M. (2001), "Bayesian Wavelet Regression on Curves With Application to a Spectroscopic Calibration Problem," *Journal of the American Statistical Association*, 96, 398–408. [602,604]

Cai, T., and Hall, P. (2006), "Prediction in Functional Linear Regression," *The Annals of Statistics*, 34, 2159–2179. [601]

Cardot, H., Ferraty, F., and Sarda, P. (1999), "Functional Linear Model," *Statistics & Probability Letters*, 45, 11–22. [601]

——— (2003), "Spline Estimators for the Functional Linear Model," *Statistica Sinica*, 13, 571–591. [601,606,607]

Centner, V., Verdú-Andrés, J., Walczak, B., Jouan-Rimbaud, D., Despagne, F., Pasti, L., Poppi, R., Massart, D. L., and de Noord, O. E. (2000), "Comparison of Multivariate Calibration Techniques Applied to Experimental NIR Data Sets," *Applied Spectroscopy*, 54, 608–623. [612]

de Jong, S. (1993), "SIMPLS: An Alternative Approach to Partial Least Squares Regression," *Chemometrics and Intelligent Laboratory Systems*, 18, 251–263. [601]

Delaigle, A., Hall, P., and Apanasovich, T. V. (2009), "Weighted Least Squares Methods for Prediction in the Functional Linear Model," *Electronic Journal of Statistics*, 3, 865–885. [601]

Donoho, D., and Johnstone, I. (1994), "Ideal Spatial Adaptation via Wavelet Shrinkage," *Biometrika*, 81, 425–455. [606]

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004), "Least Angle Regression," *The Annals of Statistics*, 32, 407–499. [604]

Ferre, L., and Yao, A. F. (2003), "Functional Sliced Inverse Regression Analysis," *Statistics*, 37, 475–488. [601]

Goutis, C., and Fearn, T. (1996), "Partial Least Squares Regression on Smooth Factors," *Journal of the American Statistical Association*, 91, 627–632. [601]

Hall, P., and Horowitz, J. L. (2007), "Methodology and Convergence Rates for Functional Linear Regression," *The Annals of Statistics*, 35, 70–91. [601]

James, G., Hastie, T., and Sugar, C. (2000), "Principal Component Models for Sparse Functional Data," *Biometrika*, 87, 587–602. [601]

James, G., Wang, J., and Zhu, J. (2009), "Functional Linear Regression That's Interpretable," *The Annals of Statistics*, 37, 2083–2108. [601]

Kalivas, J. H. (1997), "Two Data Sets of Near-Infrared Spectra," *Chemometrics and Intelligent Laboratory Systems*, 37, 255–259. [601,609]

Mallat, S. G. (1989), "A Theory for Multiresolution Signal Decomposition: The Wavelet Representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11, 674–693. [603]

Malloy, E., Morris, J., Adar, S., Suh, H., Gold, D., and Coull, B. (2010), "Wavelet-Based Functional Linear Mixed Models: An Application to Measurement Error-Corrected Distributed Lag Models," *Biostatistics*, 11, 432–452. [602]

Massy, W. F. (1965), "Principal Components Regression in Exploratory Statistical Research," *Journal of the American Statistical Association*, 60, 234–256. [601]

Marx, B. D., and Eilers, P. H. C. (1999), "Generalized Linear Regression for Sampled Signals or Curves: A P-Spline Approach," *Technometrics*, 41, 1–13. [601]

Meinshausen, N., and Yu, B. (2009), "LASSO-Type Recovery of Sparse Representations for High-dimensional Data," *The Annals of Statistics*, 37, 246–270. [613]

Müller, H., and Stadtmüller, U. (2005), "Generalized Functional Linear Models," *The Annals of Statistics*, 33, 774–805. [601]

Müller, H., and Yao, F. (2008), "Functional Additive Models," *Journal of the American Statistical Association*, 103, 1534–1544. [601]

Ogden, R. T. (1997), *Essential Wavelets for Statistical Applications and Data Analysis*, Boston, MA: Birkhäuser. [602]

Oh, H., and Lee, T. (2005), "Hybrid Local Polynomial Wavelet Shrinkage: Wavelet Regression With Automatic Boundary Adjustment," *Computational Statistics & Data Analysis*, 48, 809–819. [614]

Osborne, M. R., Presnell, B., and Turlach, B. A. (2000), "A New Approach to Variable Selection in Least Squares Problems," *IMA Journal of Numerical Analysis*, 20, 389–404. [604]

Ramsay, J. O., and Silverman, B. W. (2002), *Applied Functional Data Analysis*, New York: Springer. [601]

Reiss, P. T., and Ogden, R. T. (2007), "Functional Principal Component Regression and Functional Partial Least Squares," *Journal of the American Statistical Association*, 102, 984–996. [601]

——— (2009), "Smoothing Parameter Selection for a Class of Semiparametric Linear Models," *Journal of the Royal Statistical Society,* Series B, 71, 505–523. [607]

Saeys, W., Ketelaere, B., and Darius, P. (2008), "Potential Applications of Functional Data Analysis in Chemometrics," *Journal of Chemometrics*, 22, 335–344. [601]

Thodberg, H. H. (1996), "A Review of Bayesian Neural Networks With an Application to Near Infrared Spectroscopy," *IEEE Transactions on Neural Networks*, 7, 56–72. [609]

Tibshirani, R. (1996), "Regression Shrinkage and Selection via the LASSO," *Journal of the Royal Statistical Society,* Series B, 58, 267–288. [602,604]

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005), "Sparsity and Smoothness via the Fused LASSO," *Journal of the Royal Statistical Society,* Series B, 67, 91–108. [602]

Vidakovic, B. (1999), *Statistical Modeling by Wavelets*, New York, NY: Wiley. [602]

Wold, H. (1975), "Soft Modeling by Latent Variables: The Nonlinear Iterative Partial Least Squares Approach," in *Perspectives in Probability and Statistics, Papers in Honour of M.S. Bartlett*, ed. J. Gani, London: Academic Press, pp. 117–142. [601]

Xia, Y., Tong, H., Li, W., and Zhu, L. (2002), “An Adaptive Estimation of Dimension Reduction Space,” *Journal of the Royal Statistical Society,* Series B, 64, 363–410. [604,613]

Zhang, C., and Huang, J. (2008), “The Sparsity and Bias of the LASSO Selection in High Dimensional Linear Regression,” *The Annals of Statistics*, 36, 1567–1594. [613]

Zou, H., Hastie, T., and Tibshirani, R. (2007), “On the “Degrees of Freedom” of the LASSO,” *The Annals of Statistics*, 35, 2173–2192. [615]