

Functional Additive Regression

YINGYING FAN AND GARETH M. JAMES *

April 8, 2011

Abstract

We suggest a new method, called “Functional Additive Regression”, or FAR, for efficiently performing high dimensional functional regression. FAR extends the usual linear regression model involving a functional predictor, $X(t)$, and a scalar response, Y , in two key respects. First, FAR uses a penalized least squares optimization approach to efficiently deal with high dimensional problems involving a large number of different functional predictors. Second, FAR extends beyond the standard linear regression setting to fit general non-linear additive models. We demonstrate that FAR can be implemented with a wide range of penalty functions using a highly efficient coordinate descent algorithm. Theoretical results are developed which provide motivation for the FAR optimization criterion. Finally, we show through simulations and two real data sets that FAR can significantly outperform competing methods.

Some key words: Shrinkage; Variable Selection; Single Index Model

*Marshall School of Business, University of Southern California. This work was partially supported by NSF Grant DMS-0906784. Fan’s research was also partially supported by 2010 Zumberge Individual Award from USC’s James H. Zumberge Faculty Research and Innovation Fund. We would like to thank the Center for Clinical Neurosciences, University of Texas Health Science Center at Houston for the use of their MEG data.

1 Introduction

The univariate functional regression situation, where one models the relationship between a scalar response, Y , and a functional predictor, $X(t)$, has recently received a great deal of attention. A few examples include Hastie and Mallows (1993); Hall *et al.* (2000); Alter *et al.* (2000); Hall *et al.* (2001); James (2002); Cardot *et al.* (2003); Ferraty and Vieu (2003); James and Silverman (2005), and Muller and Stadtmuller (2005). See Chapter 15 of Ramsay and Silverman (2005) for a thorough discussion of the issues involved with fitting such data.

Most work in this area involves different approaches for fitting the functional linear regression model,

$$Y_i = \beta_0 + \int \beta(t)X_i(t)dt + \varepsilon_i, \quad i = 1, \dots, n. \quad (1)$$

Model (1) provides a natural extension of linear regression to the functional domain. However, it has two significant limitations. First, it assumes a single functional predictor so can not be applied in a situation involving p different predictors, $X_1(t), X_2(t), \dots, X_p(t)$. Second, the model is relatively inflexible because it assumes a linear relationship between the predictor and response.

The first of these limitations can, in principle, be easily solved using the multiple functional linear regression model,

$$Y_i = \beta_0 + \sum_{j=1}^p \int \beta_j(t)X_{ij}(t)dt + \varepsilon_i \quad i = 1, \dots, n. \quad (2)$$

Fitting Model (2) poses few additional complications over Model (1) provided p is relatively small. However, functional situations where p is very large are becoming increasingly common. For example, Storey *et al.* (2005) analyzes two gene expression data sets measured over time. These data sets only involve a very small number of patients, but tens of thousands of functional predictors. In settings like this some form of functional variable selection is required and there has been little research on this problem.

The second limitation, that of linearity, can be addressed using a functional addi-

tive regression framework of the form,

$$Y_i = \beta_0 + \sum_{j=1}^p f_j(X_{ij}) + \varepsilon_i, \quad i = 1, \dots, n, \quad (3)$$

where the f_j 's are general non-linear functions of $X_{ij}(t)$. Fitting Model (3), in the high dimensional, i.e. large p , setting poses a couple of significant complications. First, in order to make the problem feasible, we must assume sparsity in the predictor space, i.e. that most of the predictors are unrelated to the response. Thus we need an approach that can automatically perform high dimensional variable selection on non-linear functions. Second, Model (3) involves estimating functions, f_j , of functional predictors, $X_j(t)$. Even in the bivariate situation, involving only two predictors, there has been little research on this problem and the best approach is unclear. Most methods involve using the first few functional principal component scores of $X(t)$ as a finite dimensional predictor space (Muller and Yao, 2008). However, the principal component scores are computed independently from the response so there is no a priori reason to believe that these scores will correspond to the best dimensions for the regression problem.

In this paper we suggest a new penalized least squares method called ‘‘Functional Additive Regression’’, or FAR, for fitting (3). FAR makes three important contributions. First, it efficiently fits high dimensional functional models while simultaneously performing variable selection to identify the relevant predictors; an area that has received very little attention in the functional domain. Second, FAR extends beyond the standard linear regression setting to fit general non-linear additive models such as (3). FAR automatically computes the optimal space to project the functional predictors into using a supervised single index model approach while simultaneously estimating the f_j 's. We believe this is an important distinction because projecting into the unsupervised PCA space is currently the dominant approach in functional regressions, even though it is well known that this space need not be optimal for predicting the response. Hence, FAR provides a superior supervised method for estimating a finite-dimensional representation of the predictors. Third, FAR can be implemented using a wide range of penalty functions and a highly efficient coordinate descent algorithm. We present theoretical results which show that, under suitable conditions and for an appropriately chosen penalty function, FAR is guaranteed to asymptotically choose

the correct model as n and p go to infinity.

Our paper is set out as follows. In Section 2 we develop FAR in the high dimensional functional linear regression setting. An optimization criterion is presented which can be solved using an efficient coordinate descent algorithm. Our theoretical results are provided in Section 3. Here we show that, under appropriate conditions, FAR will asymptotically include all the true signal variables and remove all the noise predictors from the model. We also provide an asymptotic bound on the L_2 error in the estimate of the signal functions, f_j , and show that the FAR estimator enjoys the asymptotic normality. Section 4 extends FAR to the general non-linear additive model framework using a supervised approach for projecting the predictors into a finite dimensional space. Extensive simulation results, both for the linear and non-linear versions of FAR, are presented in Section 5. We compare FAR to other functional regression methods and demonstrate its superior performance in many settings. Finally, we apply FAR to both low and high dimensional real data sets in Section 6, and end with a discussion in Section 7.

2 Linear Functional Additive Regression

We assume, without loss of generality, that each predictor is observed over the range $0 \leq t \leq 1$. Hence, in the linear setting,

$$f_j(X_j) = \int_0^1 \beta_j(t)X_j(t)dt, \quad (4)$$

where $\beta_j(t)$ is an unknown smooth coefficient function. Our goal is to fit the additive linear model (3) where p is large but the true model is sparse in the sense that for most j , $\|f_j\|_{L_2(F_j)} \equiv \{E f_j^2(X_j)\}^{1/2} = 0$, where the expectation is taken with respect to the distribution F_j of $X_j(t)$. Since in practice F_j is unknown, we can approximate $\|f_j\|_{L_2(F)}$ by replacing F_j with its empirical distribution \hat{F}_j estimated from the training data, i.e.,

$$\|f_j\|_{L_2(F_j)} \approx \|f_j\|_{L_2(\hat{F}_j)} = \left\{ \frac{1}{n} \sum_{i=1}^n f_j^2(X_{ij}) \right\}^{1/2} = \frac{1}{\sqrt{n}} \|\mathbf{f}_j\|_2,$$

where $\mathbf{f}_j = (f_j(X_{1j}), \dots, f_j(X_{nj}))^T \in R^n$ and $\|\mathbf{f}_j\|_2$ represents the 2-norm of the vector \mathbf{f}_j . To reduce notation we drop the subscript and use $\|\cdot\|$ to represent the 2-norm in

the remainder of the paper. We also center both the response, Y_i , and the f_j 's. Using the centered formulation, (3) can be expressed as,

$$\mathbf{Y} = \sum_{j=1}^p \mathbf{f}_j + \boldsymbol{\epsilon}, \quad (5)$$

where \mathbf{Y} and $\boldsymbol{\epsilon}$ are n -dimensional vectors respectively corresponding to the response and error terms. The estimate for the intercept in the uncentered model can be easily computed from the fitted values of (5).

Our general approach for fitting (3) is to minimize the following penalized regression criterion over the \mathbf{f}_j 's,

$$\frac{1}{2n} \left\| \mathbf{Y} - \sum_{j=1}^p \mathbf{f}_j \right\|^2 + \sum_{j=1}^p \rho_{\lambda_n} \left(\frac{1}{\sqrt{n}} \|\mathbf{f}_j\| \right), \quad (6)$$

where $\rho_{\lambda_n}(t)$ is a penalty function and λ_n is the regularization parameter. In this article we explore general concave penalty functions, with the L_1 penalty, $\rho_{\lambda}(t) = \lambda t$, considered as a special case. It has been justified both theoretically and empirically by many researchers that concave penalty functions have advantages in model selection and are preferred in high dimensional problems. See, for example, Fan and Li (2001), Lv and Fan (2009), and Fan and Lv (2011).

2.1 Linear FAR Criterion

Combining equations (4) and (6) gives the following linear FAR optimization criterion,

$$\frac{1}{2n} \left\| \mathbf{Y} - \sum_{j=1}^p \int_0^1 \beta_j(t) \mathbf{X}_j(t) dt \right\|^2 + \sum_{j=1}^p \rho_{\lambda_n} \left(\frac{1}{\sqrt{n}} \|\mathbf{f}_j\| \right), \quad (7)$$

where $\mathbf{X}_j(t) = (X_{1j}(t), \dots, X_{nj}(t))^T$. In order for the function optimizing (7) to have a non-trivial solution some form of smoothness constraint must be imposed on the $\beta_j(t)$'s. Two standard approaches are to include a smoothness penalty in the optimization criterion or alternatively to restrict the functions to some finite-dimensional class. In this setting either approach could be adopted but we use the latter method. Specifically we select a q -dimensional orthonormal basis function on the unit interval, $\mathbf{b}(t)$, satisfying $\int_0^1 \mathbf{b}(t) \mathbf{b}^T(t) dt = I_q$ with I_q the $q \times q$ identity matrix. Hence the

coefficient functions and predictors can be expressed as $\beta_j(t) = \mathbf{b}(t)^T \boldsymbol{\eta}_j + \tilde{r}_j(t)$ and $X_{ij}(t) = \mathbf{b}(t)^T \boldsymbol{\theta}_{ij} + r_{ij}(t)$, where $\tilde{r}_j(t)$ and $r_{ij}(t)$ are the remainders, respectively. Here the $\boldsymbol{\eta}_j$'s must be estimated but the $\boldsymbol{\theta}_{ij}$'s are assumed known because the $X_{ij}(t)$'s are observed. The assumption we are making is that $\beta_j(t)$ and $X_{ij}(t)$ are well approximated by this basis. The exact details of this assumption are given in the theory section. The assumption that $\beta(t)$ and $X(t)$ can be expressed using the same basis is made for simplicity of exposition. All the FAR calculations can be extended to the situation where the bases differ at the cost of some additional notation.

Using the orthonormal basis function $\mathbf{b}(t)$, $f_j(X_{ij})$ can be approximated as

$$f_j(X_{ij}) = \boldsymbol{\theta}_{ij}^T \left(\int_0^1 \mathbf{b}(t)^T \mathbf{b}(t) dt \right) \boldsymbol{\eta}_j + e_{ij} = \boldsymbol{\theta}_{ij}^T \boldsymbol{\eta}_j + e_{ij}, \quad (8)$$

where $e_{ij} = \int_0^1 r_{ij}(t) \beta_j(t) dt + \int_0^1 \tilde{r}_j(t) X_{ij}(t) dt - \int_0^1 r_{ij}(t) \tilde{r}_j(t) dt$. For $j = 1, \dots, p$, let Θ_j be an $n \times q$ matrix whose rows are formed by $\{\boldsymbol{\theta}_{ij}, i = 1, \dots, n\}$. Then the model can be written as

$$Y_i = \sum_{j=1}^p \Theta_j \boldsymbol{\eta}_j + \varepsilon_i^*, \quad (9)$$

where $\varepsilon_i^* = \varepsilon_i + \sum_{j=1}^p e_{ij}$. Thus, the quadratic loss function in (7) can be approximated by $\|\mathbf{Y} - \sum_{j=1}^p \Theta_j \boldsymbol{\eta}_j\|^2$. So instead of minimizing (7), we estimate the $\boldsymbol{\eta}_j$'s by minimizing,

$$Q(\boldsymbol{\eta}) = \frac{1}{2n} \|\mathbf{Y} - \sum_{j=1}^p \Theta_j \boldsymbol{\eta}_j\|^2 + \sum_{j=1}^p \rho_{\lambda_n} \left(\frac{1}{\sqrt{n}} \|\Theta_j \boldsymbol{\eta}_j\| \right), \quad (10)$$

which is closely related to the group variable selection method (Yuan and Lin, 2007). Note that (10) is a $p \times q$ dimensional problem so is very challenging even if p is only of moderate size.

2.2 FAR Algorithm

A distinct advantage of the FAR criterion is that, when using the group Lasso penalty, $\rho_{\lambda_n}(t) = \lambda_n t$, there is a closed form solution for minimizing (10) over $\boldsymbol{\eta}_j$.

Proposition 1. If $\rho_{\lambda_n}(t) = \lambda_n t$, then the solution to (10) satisfies $\widehat{\mathbf{f}}_j = \Theta_j \widehat{\boldsymbol{\eta}}_j$ where

$$\widehat{\boldsymbol{\eta}}_j = \left(1 - \frac{\sqrt{n}\lambda_n}{\|S_j \mathbf{R}_j\|}\right)_+ (\Theta_j^T \Theta_j)^{-1} \Theta_j^T \mathbf{R}_j,$$

$S_j = \Theta_j (\Theta_j^T \Theta_j)^{-1} \Theta_j^T$, $\mathbf{R}_j = \mathbf{Y} - \sum_{k \neq j} \Theta_k \boldsymbol{\eta}_k$, and $z_+ = \max(0, z)$ represents the positive part of z .

The derivation of Proposition 1 involves simple algebra and a similar result is proved in Ravikumar *et al.* (2009) so we do not provide the proof here.

Recent research has shown that coordinate descent algorithms can be an extremely efficient approach for solving high dimensional sparse regression problems. These algorithms work by cycling through all the predictors, at each step optimizing one parameter while holding all the other terms fixed. Proposition 1 suggests a simple, but very efficient, coordinate descent algorithm for minimizing (10) for $\rho_{\lambda_n}(t) = \lambda_n t$.

Linear FAR Algorithm

0. Initialize $\widehat{\boldsymbol{\eta}}_j = \mathbf{0}$ and the projection matrices, $S_j = \Theta_j (\Theta_j^T \Theta_j)^{-1} \Theta_j^T$, for $j \in \{1, \dots, p\}$
1. Fix all $\widehat{\mathbf{f}}_k$ for $k \neq j$. Compute the residual vector $\mathbf{R}_j = \mathbf{Y} - \sum_{k \neq j} \widehat{\mathbf{f}}_k$.
2. Let $\widehat{\mathbf{P}}_j = S_j \mathbf{R}_j$ represent the unshrunk estimate for \mathbf{f}_j .
3. Let $\widehat{\mathbf{f}}_j = \alpha_j \widehat{\mathbf{P}}_j$ where $\alpha_j = \left(1 - \lambda_n \sqrt{n} / \|\widehat{\mathbf{P}}_j\|\right)_+$ is a shrinkage parameter.
4. Center $\widehat{\mathbf{f}}_j \leftarrow \widehat{\mathbf{f}}_j - \text{mean}(\widehat{\mathbf{f}}_j)$.
5. Repeat steps 1 through 4 for $j = 1, 2, \dots, p$ and iterate until convergence.

We repeat this algorithm over a grid of values for λ , using the previous values for the $\widehat{\boldsymbol{\eta}}_j$'s to initialize the parameters for the new λ . Since the parameters change very little for a small change in λ , the algorithm generally converges very rapidly. Note that the S_j 's only need to be computed once for all values of λ so the computation at each step of the algorithm is extremely fast. In addition it is clear from Proposition 1 that (10) will decrease at each step. Our FAR algorithm has similarities to the algorithm used to fit the SpAM approach of Ravikumar *et al.* (2009). In particular both methods have the advantage of decomposing the estimation of $\widehat{\mathbf{f}}_j$ into two simple, and separate,

steps. First, compute the unshrunk estimate, $\widehat{\mathbf{P}}_j$, and then apply the shrinkage factor, α_j . When $\alpha_j = 0$ then the j th predictor is absent from the model.

For a general penalty function, $\rho_{\lambda_n}(t)$, we use the local linear approximation method proposed in Zou and Li (2008) to solve (10). The penalty function can be approximated as $\rho_{\lambda_n}(\|\mathbf{f}\|/\sqrt{n}) \approx \rho'_{\lambda_n}(\|\mathbf{f}^*\|/\sqrt{n})\|\mathbf{f}\|/\sqrt{n} + C$, where \mathbf{f}^* is some vector that is close to \mathbf{f} and $C = \rho_{\lambda_n}(\|\mathbf{f}^*\|/\sqrt{n}) - \rho'_{\lambda_n}(\|\mathbf{f}^*\|/\sqrt{n})\|\mathbf{f}^*\|/\sqrt{n}$ is a constant. Hence the only required change to the FAR algorithm for optimizing over general penalty functions is to replace the calculation of α_j in Step 3. by,

$$\alpha_j = \left(1 - \rho'_{\lambda_n}\left(\frac{1}{\sqrt{n}}\|\widehat{\mathbf{f}}_j\|\right)\sqrt{n}/\|\widehat{\mathbf{P}}_j\|\right)_+,$$

where $\widehat{\mathbf{f}}_j$ represents the most recent estimate for \mathbf{f}_j . The initial estimate of $\widehat{\mathbf{f}}_j$ can be obtained by using the group Lasso penalty. This simple approximation allows the FAR algorithm to be easily applied to a wide range of penalty functions.

3 Theory

Denote by $\mathfrak{M}_0 = \{j : \beta_j(t) \neq 0, 1 \leq j \leq p\}$ the set of true functional predictors and let s_n represent the cardinality of \mathfrak{M}_0 . By minimizing the FAR criterion (10), we aim to identify the set \mathfrak{M}_0 and accurately estimate functions $\beta_j(t)$ for $j \in \mathfrak{M}_0$. In this section we discuss the theoretical properties of FAR in the setting where the f_j 's are linear functions, that is, $f_j(X_{ij}) = \int_0^1 X_{ij}(t)\beta_j(t)dt$. In particular we present two theorems. Theorem 1 concerns FAR's model selection properties. We show that, with probability tending to one, FAR will remove all noise predictors from the fitted model. Theorem 1 also places a bound on the L_2 error in the estimate for the f_j 's, where $j \in \mathfrak{M}_0$. Our second result, Theorem 2 shows the asymptotic normality of the estimator. Proofs of all the results in this section are provided in Appendix A.

In order to prove these results we make two sets of assumptions. Condition 1 relates to the level of accuracy in our basis approximations of $X_j(t)$ and $\beta_j(t)$. Alternatively, Condition 2 concerns the shape of the penalty function, the strength of the signal and the correlation structure of the predictors. All results are conditional on the observed predictors, $X_{ij}(t)$, $i = 1, \dots, n$, $j = 1, \dots, p$.

Recall that we have assumed there exists a q -dimensional orthogonal basis function

$\mathbf{b}(t)$ such that $X_{ij}(t) = \mathbf{b}(t)^T \boldsymbol{\theta}_{ij} + r_{ij}(t)$ and $\beta_j(t) = \mathbf{b}(t)^T \boldsymbol{\eta}_j + \tilde{r}_j(t)$. Note that when $j \in \mathfrak{M}_0$, $\beta_j(t) = 0$ and $\tilde{r}_j(t) = 0$, so the approximation error e_{ij} in (8) disappears. It seems reasonable to assume that as q grows the functions $X_{ij}(t)$ and $\beta_j(t)$ with $j \in \mathfrak{M}_0$ can be better approximated. In particular we make the following assumptions on the approximation accuracy.

Condition 1. *Suppose that provided $q = o(n/s_n)$, then $\max_{j \in \mathfrak{M}_0} \max_i \int_0^1 r_{ij}^2(t) dt = o(n^{-4\delta})$ and $\max_{j \in \mathfrak{M}_0} \int_0^1 \tilde{r}_j^2(t) dt = o(n^{-4\delta})$ for some $\delta > \frac{1}{4}$. Further, assume that $s_n = o(n^{2\delta - \frac{1}{2}})$, $\max_{j \in \mathfrak{M}_0} \max_i \int_0^1 X_{ij}^2(t) dt \leq M$ and $\max_{j \in \mathfrak{M}_0} \int_0^1 \beta_j^2(t) dt \leq M$ for some generic constant $M > 0$.*

Condition 1 puts restrictions on the number of nonzero parameters $s_n \times q$, that is, $s_n \times q = o(n)$ and $s_n = o(n^{2\delta - \frac{1}{2}})$ for some $\delta > \frac{1}{4}$. This condition is imposed because, in order to ensure that the approximation holds uniformly, the number of true predictors, s_n , cannot grow too fast, and to avoid over-fitting, the number of basis functions, q , must also be constrained. It is easy to derive from Condition 1 that e_{ij} , defined in (8), satisfies $\max_{i,j \in \mathfrak{M}_0} |e_{ij}| = o(n^{-2\delta})$ and $e_{ij} = 0$ for $j \in \mathfrak{M}_0^c$, since it is assumed that $s_n = o(n^{2\delta - \frac{1}{2}})$, the error term in (9) has mean $E[\varepsilon_i^* | \mathbf{X}] = o(n^{-1/2})$. If $\mathbf{b}(t)$ is the true basis from which $X_j(t)$ and $\beta_j(t)$ are generated then the approximation errors are 0 and Condition 1 is not needed. However, in general assuming a perfect representation seems to be a strong assumption.

Our second set of conditions concern the shape of the penalty function, the strength of the signal and the correlation structure of the predictors. Let $\boldsymbol{\eta}_0 = (\boldsymbol{\eta}_{0,1}^T, \dots, \boldsymbol{\eta}_{0,p}^T)^T \in R^{pq}$ be the true coefficients vector. For any index set $S \subset \{1, \dots, p\}$, we use $\boldsymbol{\eta}_S$ to denote the vector formed by stacking vectors $\boldsymbol{\eta}_j$, $j \in S$ one underneath each other, and Θ_S to denote the matrix formed by stacking the matrices Θ_j , $j \in S$ one after another.

Condition 2.

- (A) *For any fixed $\lambda > 0$, assume that $\rho_\lambda(t)$ is concave and non-decreasing in $[0, \infty)$, and has non-increasing first derivative $\rho'_\lambda(t)$. Further, assume that $\rho'_\lambda(0+) > 0$.*
- (B) *Let $a_n = \min_{j \in \mathfrak{M}_0} \|\Theta_j \boldsymbol{\eta}_{0,j}\| / \sqrt{n}$. Assume that $\frac{1}{\log n} a_n n^\alpha \rightarrow \infty$ with $0 < \alpha < \frac{1}{2}$.*
- (C) *Assume that $\rho'_{\lambda_n}(a_n/2) = O(n^{-1/2} \sqrt{\log n})$ and $\sup_{t \geq \frac{a_n}{2}} \rho''_{\lambda_n}(t) = O(n^{-1/2} \sqrt{\log n})$.*

(D) Assume that $\|\Theta_{\mathfrak{M}_0}^T\|_\infty = O(n)$, $\|\Theta_{\mathfrak{M}_0^c}^T\|_\infty = O(n)$ and,

$$0 < c_0 \leq \min_{j \in \mathfrak{M}_0} \Lambda_{\min}(\frac{1}{n} \Theta_j^T \Theta_j) < \Lambda_{\max}(\frac{1}{n} \Theta_{\mathfrak{M}_0}^T \Theta_{\mathfrak{M}_0}) \leq c_0^{-1},$$

where Λ_{\min} and Λ_{\max} are the smallest and largest eigenvalues of a matrix respectively and c_0 is a positive constant. Further, assume that

$$\max_{j \in \mathfrak{M}_0^c} \|\Theta_j (\Theta_j^T \Theta_j)^{-1} \Theta_j^T \Theta_{\mathfrak{M}_0} (\Theta_{\mathfrak{M}_0}^T \Theta_{\mathfrak{M}_0})^{-1}\|_{\infty,2} < \min \left(\frac{\sqrt{c_0} \rho'_{\lambda_n}(0+)}{2\sqrt{n} \rho'_{\lambda_n}(a_n/2)}, \frac{\sqrt{c_0} \rho'_{\lambda_n}(0+)}{8\sigma \sqrt{\log n}} \right), \quad (11)$$

$$\max_{j \in \mathfrak{M}_0^c} \|\Theta_j (\Theta_j^T \Theta_j)^{-1}\|_{\infty,2} < \frac{\rho'_{\lambda_n}(0+)}{8\sigma u_n}, \quad (12)$$

where u_n is a sequence satisfying $u_n \rightarrow \infty$ as $n \rightarrow \infty$, and $\|\mathbf{B}\|_{\infty,2} = \max_{\|\mathbf{x}\|_\infty=1} \|\mathbf{B}\mathbf{x}\|$.

Condition 2(A) requires that the penalty functional, $\rho_\lambda(t)$, is concave and singular at 0. From (8), we see that Condition 2(B) places a lower bound on the signal strength of the true predictors $j \in \mathfrak{M}_0$. In particular, it assumes that the weakest signal, a_n , can decay with sample size but the decay rate cannot be faster than $n^{-\alpha} \log(n)$. Although Condition 2(C) assumes the existence of the second derivative for $\rho_{\lambda_n}(t)$, it can be relaxed to the existence of the first order derivative by using the local concavity definition in Lv and Fan (2009). Condition 2(D) relates to the design matrix for the signal predictors, $\Theta_{\mathfrak{M}_0}$. We assume that the eigenvalues for the design matrix corresponding to each true predictor are bounded from below and above. The upper bound in condition (11) depends on the penalty function through the ratio $\rho'_{\lambda_n}(0+)/\rho'_{\lambda_n}(a_n/2)$, which is larger than 1 for concave penalties and equal to 1 for the group Lasso penalty, $\rho_{\lambda_n}(t) = \lambda_n t$.

Under Conditions 1 and 2, Theorem 1 shows that FAR possesses the oracle property for model selection.

Theorem 1. Assume that $qs_n = o(n^{\frac{1}{2}-\alpha} \sqrt{\log n})$ with α defined in Condition 2(B), $u_n/\sqrt{\log n} \rightarrow \infty$, and $\log(q(p - s_n)) = o(u_n)$. Then under Conditions 1 and 2, with probability tending to 1 as $n \rightarrow \infty$, there exists a local minimizer $\hat{\boldsymbol{\eta}}$ of $Q(\boldsymbol{\eta})$ such that

$$1) \hat{\boldsymbol{\eta}}_{\mathfrak{M}_0^c} = 0$$

$$2) \max_{j \in \mathfrak{M}_0} \frac{1}{\sqrt{n}} \|\Theta_j(\hat{\boldsymbol{\eta}}_j - \boldsymbol{\eta}_{0,j})\| \leq n^{-\alpha} \sqrt{\log n}.$$

Part 2 of Theorem 1 concerns the approximation accuracy of the basis coefficients rather than the functions themselves. However, the result extends naturally. Denote by $\hat{\mathbf{f}}_j = \Theta_j \hat{\boldsymbol{\eta}}_j$ and $\mathbf{f}_{0j} = (f_j(X_{j1}), \dots, f_j(X_{jn}))^T$, respectively the estimated and true values of the j -th functional component, both evaluated at the n training data points. Then we have the following corollary.

Corollary 1. *Suppose the conditions in Theorem 1 are satisfied. Then with probability tending to 1 as $n \rightarrow \infty$, there exists a FAR estimate such that*

$$\max_{j \in \mathfrak{M}_0} \|\hat{f}_j - f_j\|_{L_2(\hat{F}_j)} = \max_{j \in \mathfrak{M}_0} \frac{1}{\sqrt{n}} \|\hat{\mathbf{f}}_j - \mathbf{f}_{0j}\| \leq n^{-\alpha} \sqrt{\log n},$$

where $\hat{f}_j(x) = \int_0^1 x(t) \hat{\beta}_j(t) dt$ with $\hat{\beta}_j(t) = \mathbf{b}^T(t) \hat{\boldsymbol{\eta}}_j$, $x(t)$ is a given value of the j -th functional predictor, and \hat{F}_j is the corresponding empirical distribution of $x(t)$.

Theorem 2 shows the asymptotic normality of the FAR estimators that correspond to signal variables. As with Theorem 1 we first provide the result for the $\hat{\boldsymbol{\eta}}_j$'s and then extend to the functions.

Theorem 2. *Assume that the conditions in Theorem 1 hold and in addition, $\rho'_{\lambda_n}(a_n/2) = o(a_n n^{\alpha-1/2} s_n^{-1/2} (\log n)^{-1})$ and $s_n = o(n^{2\alpha} (\log n)^{-3})$. Then with probability tending to 1 as $n \rightarrow \infty$, there exists a strict local minimizer $\hat{\boldsymbol{\eta}}$ of $Q(\boldsymbol{\eta})$ such that $\hat{\boldsymbol{\eta}}_{\mathfrak{M}_0^c} = 0$ and*

$$\mathbf{c}^T [(\Theta_{\mathfrak{M}_0} \Theta_{\mathfrak{M}_0}^T)^{1/2} (\hat{\boldsymbol{\eta}}_{\mathfrak{M}_0} - \boldsymbol{\eta}_{0, \mathfrak{M}_0}) + n(\Theta_{\mathfrak{M}_0} \Theta_{\mathfrak{M}_0}^T)^{-1/2} \mathbf{v}_{0, \mathfrak{M}_0}] \xrightarrow{\mathcal{D}} N(0, \sigma^2),$$

where $\mathbf{c} \in \mathbf{R}^{qs_n}$ satisfies $\mathbf{c}^T \mathbf{c} = 1$ and $\mathbf{v}_{0, \mathfrak{M}_0}$ is a vector formed by stacking the vectors $\mathbf{v}_{0, k} = \rho'_{\lambda_n} \left(\frac{1}{\sqrt{n}} \|\Theta_k^T \boldsymbol{\eta}_{0, k}\| \right) \frac{1}{\sqrt{n}} \frac{\Theta_k \Theta_k^T \boldsymbol{\eta}_{0, k}}{\|\Theta_k^T \boldsymbol{\eta}_{0, k}\|}$, $k \in \mathfrak{M}_0$ underneath each other.

Let $f_{0j}^* = \boldsymbol{\theta}_j^{*T} \boldsymbol{\eta}_{0j}$ and $\hat{f}_j^* = \boldsymbol{\theta}_j^{*T} \hat{\boldsymbol{\eta}}_j$, with $\boldsymbol{\theta}_j^* \in R^q$ the coefficient vector when projecting a given new observation, $X_j^*(t)$, onto the basis function, $\mathbf{b}(t)$. Then as q increases, f_{0j}^* better approximates $f_j(X_j^*)$ for each fixed $j = 1, \dots, p$. Define $\mathbf{f}_0^* = (f_{01}^*, \dots, f_{0p}^*)^T$ and $\hat{\mathbf{f}}^* = (\hat{f}_1^*, \dots, \hat{f}_p^*)^T$. Taking $\mathbf{c} = (\Theta_{\mathfrak{M}_0} \Theta_{\mathfrak{M}_0}^T)^{-1/2} \Theta^* \tilde{\mathbf{c}}_0$ with $\Theta^* = \text{diag}(\boldsymbol{\theta}_1^*, \dots, \boldsymbol{\theta}_{s_n}^*) \in R^{(qs_n) \times s_n}$ in Theorem 2, we have the following asymptotic normality of \mathbf{f}_0^* .

Corollary 2. *Assume that the conditions in Theorem 2 hold. Then with probability tending to 1 as $n \rightarrow \infty$, there exists a FAR estimate such that $\hat{\mathbf{f}}_{\mathfrak{M}_0^c}^* = 0$. Moreover,*

$$\tilde{\mathbf{c}}_0^T [\hat{\mathbf{f}}_{\mathfrak{M}_0}^* - \mathbf{f}_{0, \mathfrak{M}_0}^* + n\Theta^* (\Theta_{\mathfrak{M}_0} \Theta_{\mathfrak{M}_0}^T)^{-1} \mathbf{v}_{0, \mathfrak{M}_0}] \xrightarrow{\mathcal{D}} N(0, \sigma^2 \tilde{\mathbf{c}}_0^T \Theta^* (\Theta_{\mathfrak{M}_0}^T \Theta_{\mathfrak{M}_0})^{-1} \Theta^{*T} \tilde{\mathbf{c}}_0),$$

where $\tilde{\mathbf{c}}_0$ is a constant vector in R^{s_n} satisfying $\tilde{\mathbf{c}}_0^T (\Theta^*)^T (\Theta_{\mathfrak{M}_0} \Theta_{\mathfrak{M}_0}^T)^{-1} \Theta^* \tilde{\mathbf{c}}_0 = 1$, and \mathbf{v}_0 is defined in Theorem 2.

4 Non-linear Functional Additive Regression

A limitation of linear functional regression models is that they can perform poorly when there is a non-linear relationship between $X(t)$ and Y . However, the infinite dimensional nature of $X(t)$ makes it challenging to model a non-linear relationship between the predictor and response. As a result relatively few papers have investigated this extension. Most methods focus on approximating $X(t)$ using its first few functional principal components and then implementing non-linear fits using the principal component scores as predictors (Muller and Yao, 2008). However, this unsupervised approach has the usual limitation; the directions which explain $X(t)$ best may not be the most appropriate for predicting the response.

In the multivariate setting, index models are commonly used for providing non-linear fits to high dimensional data. For a centered response, the standard single index model can be expressed in the form,

$$Y = g(\boldsymbol{\beta}^T \mathbf{X}) + \varepsilon,$$

where g is a general non-linear function and $\boldsymbol{\beta}$ is a norm one vector representing the best single direction to project the predictors into. A key advantage of the index model formulation is that $\boldsymbol{\beta}$ is chosen in a supervised fashion, incorporating both the response and predictors; potentially providing more accurate fits.

Index models can be naturally extended to functional predictors using the formulation, $f_j(X_{ij}) = g_j\left(\int \beta_j(t) X_{ij}(t) dt\right)$, where g_j and β_j are both non-parametric smooth functions (James and Silverman, 2005). Since β_j corresponds to a direction that we project $X_{ij}(t)$ into we impose the constraint $\|\beta_j\| = 1$. Using this non-linear representation the response centered FAR model, (3), can be expressed as,

$$Y_i = \sum_{j=1}^p g_j\left(\int \beta_j(t) X_{ij}(t) dt\right) + \varepsilon_i, \quad (13)$$

and the general FAR optimization criterion, (6), becomes,

$$\frac{1}{2n} \left\| \mathbf{Y} - \sum_{j=1}^p g_j \left(\int \beta_j(t) \mathbf{X}_j(t) dt \right) \right\|^2 + \sum_{j=1}^p \rho_{\lambda_n} \left(\frac{1}{\sqrt{n}} \|\mathbf{f}_j\| \right). \quad (14)$$

As in Section 2, we assume that $\beta_j(t)$ and $X_{ij}(t)$ are well approximated by an orthogonal q -dimensional basis, $\mathbf{b}(t)$, such that $\beta_j(t) \approx \mathbf{b}(t)^T \boldsymbol{\eta}_j$ and $X_{ij} \approx \mathbf{b}(t)^T \boldsymbol{\theta}_{ij}$. We further assume that $g_j(t)$ can be well approximated by a d -dimensional basis, $\mathbf{h}(t)$, such that $g_j(t) \approx \mathbf{h}(t)^T \boldsymbol{\xi}_j$. Using this basis representation (6) can be expressed as,

$$l_\lambda(\boldsymbol{\xi}|\boldsymbol{\eta}) = \frac{1}{2n} \left\| \mathbf{Y} - \sum_{j=1}^p H_j \boldsymbol{\xi}_j \right\|^2 + \sum_{j=1}^p \rho_{\lambda_n} \left(\frac{1}{\sqrt{n}} \|H_j \boldsymbol{\xi}_j\| \right), \quad (15)$$

where H_j is an n by d matrix whose i th row is given by $\mathbf{h}(\boldsymbol{\theta}_{ij}^T \boldsymbol{\eta}_j)^T$.

4.1 Non-linear FAR Algorithm

There are two obvious methods to fit the non-linear FAR model, both of which turn out to have significant problems in practice. The simplest approach is to first use an unsupervised procedure, such as functional PCA, to estimate the $\boldsymbol{\eta}_j$'s. Then conditional on the $\boldsymbol{\eta}_j$'s use a coordinate descent algorithm to optimize (15). In the $p = 1$ setting, similar unsupervised approaches have often been used (Muller and Yao, 2008). However, as discussed previously, since the choice of $\boldsymbol{\eta}_j$ is made in an unsupervised fashion, the direction it suggests will generally not be the most informative in terms of predicting Y .

Alternatively, one could use a coordinate descent algorithm to minimize $l_\lambda(\boldsymbol{\xi}|\boldsymbol{\eta})$ jointly over $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$. However, this also turns out to be a poor approach for several reasons. First, for an arbitrary basis function, \mathbf{h} , there is no guarantee that $l_\lambda(\boldsymbol{\xi}|\boldsymbol{\eta})$ will be convex in $\boldsymbol{\eta}$ so a coordinate descent algorithm may not produce a global minimum. Second, the penalty function in (15) means that minimizing $l_\lambda(\boldsymbol{\xi}|\boldsymbol{\eta})$ over $\boldsymbol{\eta}_j$ is difficult even if all the other terms are fixed. Third, depending on the value of λ , there will often be some estimates of $\boldsymbol{\xi}_j$ that are close to, but not exactly equal to, zero. This makes the corresponding estimates for $\boldsymbol{\eta}_j$ extremely unstable.

FAR adopts an approach between the simple, but potentially inaccurate, unsupervised method and the appealing, but infeasible, joint optimization procedure. We

use a two stage algorithm, analogous to a profile likelihood approach, where we first estimate the $\boldsymbol{\eta}_j$'s, in a supervised fashion, and then optimize (15) conditional on $\boldsymbol{\eta}_j$. This approach has the advantage of providing a more accurate estimate for the $\boldsymbol{\eta}_j$'s while avoiding the computational and practical difficulties of the joint optimization method.

Algorithm

A. Given initial values for the $\boldsymbol{\xi}_j$'s, compute the $\boldsymbol{\eta}_j$'s as the values minimizing

$$Q = \sum_{i=1}^n \left(Y_i - \sum_{j=1}^p \mathbf{h}(\boldsymbol{\theta}_{ij}^T \boldsymbol{\eta}_j)^T \boldsymbol{\xi}_j \right)^2. \quad (16)$$

Our approach for minimizing (16) is provided in Appendix B.

B. Let $\widehat{S}_j = \widehat{H}_j(\widehat{H}_j^T \widehat{H}_j)^{-1} \widehat{H}_j^T$ where the i th row of \widehat{H}_j is given by $\mathbf{h}(\boldsymbol{\theta}_{ij}^T \widehat{\boldsymbol{\eta}}_j)^T$. Conditional on $\widehat{\boldsymbol{\eta}}_1, \dots, \widehat{\boldsymbol{\eta}}_p$ from Step A., minimize $l_\lambda(\boldsymbol{\xi}|\widehat{\boldsymbol{\eta}})$ over $\boldsymbol{\xi}_1, \dots, \boldsymbol{\xi}_p$ using the following coordinate descent algorithm.

For each $j \in \{1, \dots, p\}$,

1. Fix all $\widehat{\mathbf{f}}_k$ for $k \neq j$. Compute the residual vector $\mathbf{R}_j = \mathbf{Y} - \sum_{k \neq j} \widehat{\mathbf{f}}_k(X_k)$.
2. Let $\widehat{\mathbf{P}}_j = \widehat{S}_j \mathbf{R}_j$ and $\alpha_j = \left(1 - \rho'_{\lambda_n} \left(\frac{1}{\sqrt{n}} \|\widehat{\mathbf{f}}_j\| \right) \sqrt{n} / \|\widehat{\mathbf{P}}_j\| \right)_+$ where $\widehat{\mathbf{f}}_j$ represents the most recent estimate for \mathbf{f}_j .
3. Let $\widehat{\mathbf{f}}_j = \alpha_j \widehat{\mathbf{P}}_j$.
4. Center $\widehat{\mathbf{f}}_j \leftarrow \widehat{\mathbf{f}}_j - \text{mean}(\widehat{\mathbf{f}}_j)$.

Repeat Steps 1. through 4. until convergence.

In practice, this algorithm is implemented over a grid of tuning parameters, $\lambda_1, \dots, \lambda_T$. Hence, for a given λ_t , the values for the $\boldsymbol{\xi}_j$'s in Step A. are obtained as the final Step B. estimates from the previous iteration using λ_{t-1} . We discuss our approach for selecting initial values of the parameters for $\lambda = \lambda_1$ in Appendix C.

4.2 Selecting Tuning Parameters

Both the linear and non-linear versions of FAR require choosing the tuning parameter, λ . As with all penalized regression methods, there are several possible approaches one

could adopt. Popular approaches include, BIC, AIC or cross-validation. The BIC and AIC methods require the calculation of the effective degrees of freedom. For the Lasso it has been shown that an unbiased estimate for this quantity is the number of non-zero coefficients (Zou, 2007). One could potentially use the same value for FAR. However, given FAR’s more complicated structure it is not clear that this is still an appropriate estimate. Computing the effective degrees of freedom for FAR is a topic for future research. Instead we selected λ using cross-validation for our real data examples and a separate validation data set for our simulations. This approach seemed to work well on the problems that we examined.

5 Simulations

In this section we compare the performance of FAR to several alternative linear and non-linear functional approaches in a series of simulation studies. We consider the linear setting in Section 5.1, while Section 5.2 contains our non-linear results.

5.1 Linear Additive Models

In this subsection we generated data from the linear model (2) with $\beta_0 = 0$ and $\varepsilon_i \sim_{i.i.d.} N(0, \sigma^2)$. The functional predictors, $X_{ij}(t)$, were simulated from a B-spline basis with two internal knots plus an error term,

$$X_{ij}(t_k) = \mathbf{b}(t_k)^T \boldsymbol{\theta}_{ij} + w_{ijk}, \quad w_{ijk} \sim N(0, \sigma_x^2), \quad \boldsymbol{\theta}_{ij} \sim N(0, \Theta),$$

where $\sigma_x = 0.5$ and each predictor was observed at 100 equally spaced time points, $0 = t_1, t_2, \dots, t_{100} = 1$. The basis coefficients, $\boldsymbol{\theta}_{ij}$, and the error terms, w_{ijk} , were all sampled independently from each other. The first two coefficient functions, $\beta_1(t)$ and $\beta_2(t)$, were also generated, from the same basis function, $\beta_j(t) = \mathbf{b}(t)^T \boldsymbol{\eta}_j$, while the remaining $p - 2$ predictors were noise variables with $\beta_j(t) = 0$.

Most functional regression methods utilize a functional principal components analysis (FPCA) decomposition of the predictors to form a low dimensional representation of the $X(t)$ ’s. The resulting PCA scores are then used as the predictors in the final regression model; the functional analogue of traditional principal components regression. In order to compare FAR to the FPCA approach we generated a range of

		FAR.S*	FAR.S	FAR.L*	FPCA.S*	FPCA.S	FPCA.L*
$\gamma = 69\%$ $\sigma = 1$ $p = 200$	FNR	0.015	0.000	0.010	0.030	0.015	0.035
	FPR	0.042	0.067	0.141	0.008	0.049	0.011
	Mean(PE)	1.254	1.155	1.219	1.328	1.337	1.329
	SE(PE)	0.022	0.011	0.015	0.011	0.008	0.011
$\gamma = 90\%$ $\sigma = 1$ $p = 50$	FNR	0.000	0.000	0.000	0.000	0.000	0.000
	FPR	0.005	0.084	0.002	0.010	0.060	0.007
	Mean(PE)	1.139	1.119	1.168	1.270	1.262	1.264
	SE(PE)	0.009	0.006	0.011	0.005	0.003	0.004
$\gamma = 90\%$ $\sigma = 2$ $p = 50$	FNR	0.025	0.000	0.035	0.000	0.000	0.000
	FPR	0.038	0.224	0.072	0.021	0.156	0.010
	Mean(PE)	2.407	2.350	2.378	2.206	2.225	2.193
	SE(PE)	0.032	0.023	0.028	0.011	0.010	0.009
$\gamma = 90\%$ $\sigma = 1$ $p = 200$	FNR	0.000	0.000	0.000	0.000	0.000	0.000
	FPR	0.006	0.032	0.038	0.008	0.025	0.006
	Mean(PE)	1.151	1.124	1.176	1.279	1.254	1.256
	SE(PE)	0.013	0.011	0.012	0.010	0.004	0.006
$\gamma = 99\%$ $\sigma = 1$ $p = 200$	FNR	0.000	0.000	0.000	0.000	0.000	0.000
	FPR	0.000	0.018	0.024	0.004	0.019	0.004
	Mean(PE)	1.114	1.121	1.236	1.121	1.110	1.124
	SE(PE)	0.009	0.008	0.018	0.007	0.004	0.007

Table 1: Comparison of FAR to FPCA based methods in five linear simulation settings. We use * to denote unshrunk estimators.

situations where the first principal component of $X(t)$ had varying predictive ability. In particular, let γ_j represent the proportion of variation in $\int_0^1 X_{ij}(t)\beta_j(t)dt$ that is explained by the first principal component of $X_{ij}(t)$. Then, in order to facilitate comparisons with the FPCA approach we choose Θ and the $\boldsymbol{\eta}_j$'s in such a way that 90% of the variance in $X_{ij}(t)$ was captured by the first FPC while $\gamma = \gamma_1 = \gamma_2$ ranged from approximately 69% to 99%, depending on the simulation setup. In the $\gamma \approx 99\%$ situation almost all the information about the response was contained in the first principal component of $X(t)$; an extremely favorable situation for the FPCA based methods. We also considered a setup with $\gamma \approx 69\%$, a more challenging situation where the first FPC explains most of the variance in the predictor but only slightly over 2/3rds of the variance in the response. The majority of our simulations corresponded to the intermediate situation, $\gamma \approx 90\%$, where most but not all of the relevant information was captured by the first principal component.

We implemented the linear version of FAR using two different penalty functions, the SCAD penalty (FAR.S) and the Lasso penalty (FAR.L). As a comparison we also

implemented two FPCA based methods by decomposing the predictors into functional principal components, selecting the components that explained at least 80% of the variance in each of the predictors, and finally using the resulting PCA scores to fit linear regression models to the responses. Since only two of the predictor functions were associated with the response we fit the linear regressions to the FPCA scores using both the SCAD (FPCA.S) and Lasso (FPCA.L) penalty functions to produce sparse fits. The FPCA estimates were produced using the standard smoothing approach of, first individually estimating the $\boldsymbol{\theta}_{ij}$'s using the least squares fit, $\hat{\boldsymbol{\theta}}_{ij} = (B^T B)^{-1} B^T \mathbf{X}_{ij}$, where B is the basis matrix associated with the spline basis, $\mathbf{b}(t)$, and then applying standard PCA to the resulting $\hat{\boldsymbol{\theta}}_{ij}$'s. Finally, we also generated unshrunk versions of each method by first selecting a subset of predictors using the regular FAR or FPCA methodologies, and then fitting this subset of variables using unpenalized versions of FAR or FPCA. This approach can reduce the overshrinkage often exhibited by the Lasso penalty and is analogous to the LARS/OLS hybrid discussed in Efron *et al.* (2004).

We used $n = 60$ training observations for each data set and tested a total of five linear settings corresponding to different numbers of predictors, $p = 50$ and 200 , different noise levels, $\sigma = 1$ and 2 , and different proportions of variance explained by the PCA scores, $\gamma = 69\%$, 90% and 99% . The tuning parameters for the FAR and FPCA methods were chosen by minimizing prediction error on a separately generated validation data set with similar characteristics to the training data but only $n = 50$ observations. For each simulation setting we fit each method to 100 different training sets and recorded the false positive rate (FPR), false negative rate (FNR), average prediction error on a separate test data set (Mean PE) and the standard error of the mean PE (SE PE). The FPR records the fraction of noise predictors incorrectly included in the model while the FNR corresponds to the fraction of signal variables incorrectly excluded. The simulation results are summarized graphically in Figure 1 and numerically in Table 1. Prediction errors that were either the best or were not statistically worse than the best result are shown in bold font.

We found that the shrunk version of the Lasso penalty performed uniformly worse on both the FAR and FPCA methods so we have only reported the unshrunk results for this penalty. In terms of prediction error, FAR was significantly superior to FPCA in three of the five simulation settings. Interestingly, even in the most favorable FPCA

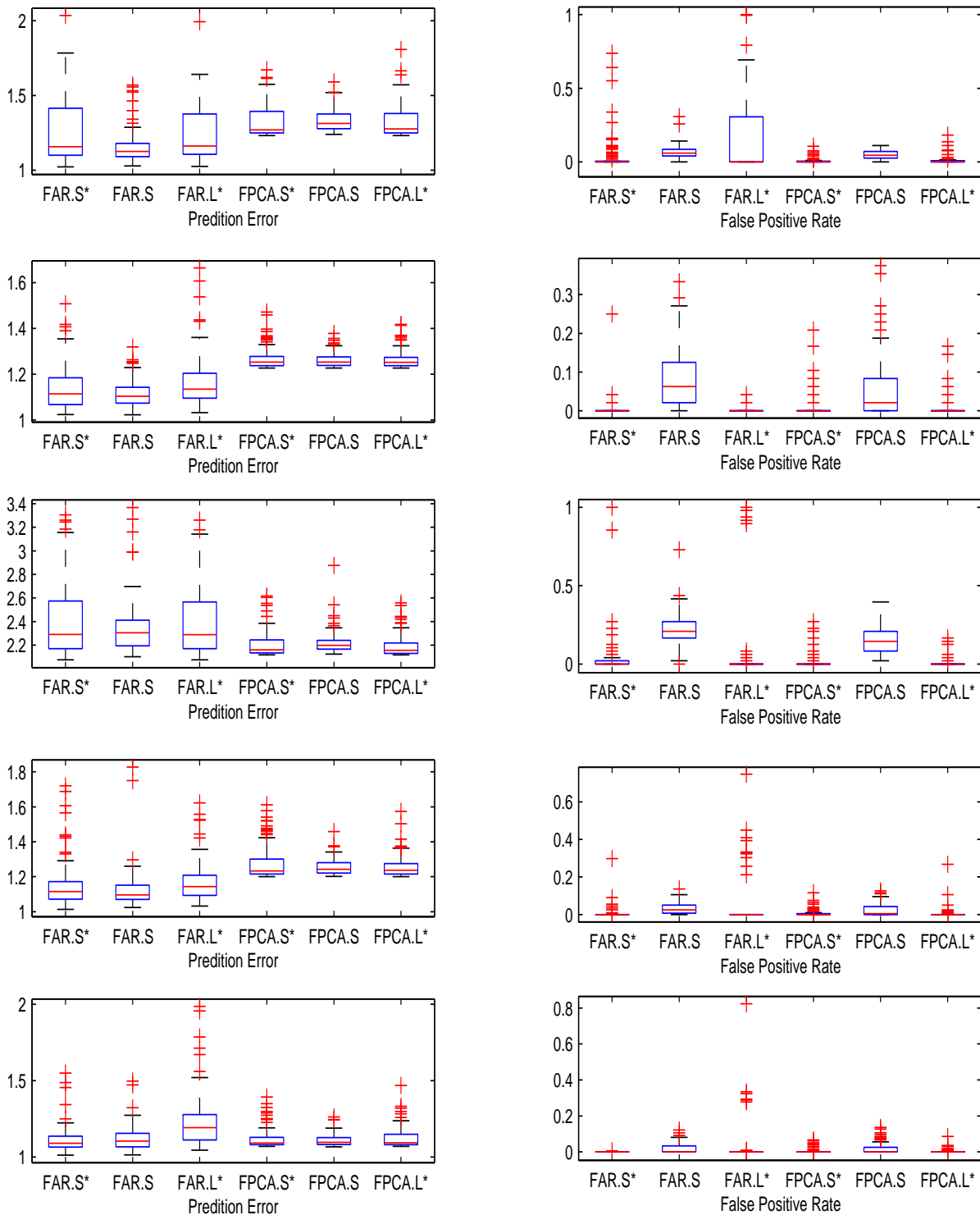


Figure 1: *Boxplots of Prediction Errors and False Positive Rates for the FAR and FPCA methods over 100 linear simulation runs. Each row provides results using the simulation settings from the corresponding row in Table 1. False Negative Rates were generally low so we have not plotted them here. We use * to denote unshrunk estimators.*

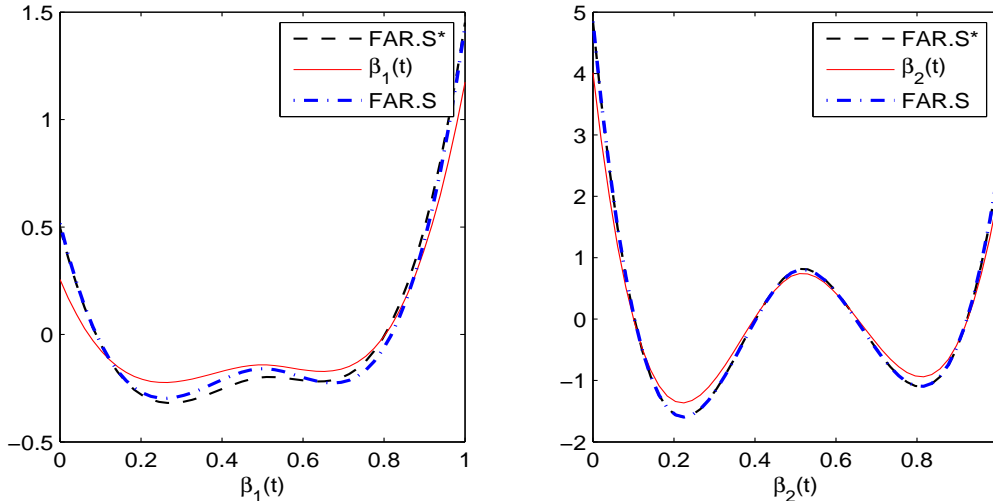


Figure 2: The $\beta_j(t)$ curves (solid red) for $j = 1, 2$ along with the average FAR estimates (black and blue dashed) in the $\gamma = 90\%$, $p = 200$ setting.

setting, where $\gamma = 99\%$, FAR was not statistically worse in terms of prediction error and was superior in terms of the false positive rate. The only setting where FPCA dominated was the situation where $\sigma = 2$ and $\gamma = 90\%$. This simulation had both high γ and high noise; a low bias, high variance situation that advantaged FPCA because it only had to estimate a single regression coefficient. The three different penalty functions gave similar results for the FPCA method. However, SCAD clearly dominated over the Lasso when using FAR. The unshrunk SCAD version of FAR generally gave the lowest prediction errors but the shrunk version was superior in terms of variable selection. Figure 2 shows the $\beta_j(t)$ curves for $j = 1, 2$ along with the average FAR estimates in the $\gamma = 90\%$, $p = 200$ setting. The fits in the other settings showed similar levels of accuracy.

5.2 Non-linear Models

We also examined five different non-linear simulation settings. With the exception that σ_x was set to 0.1, the predictors, $X_{ij}(t)$, and coefficient curves, $\beta_j(t)$, were all produced in an identical fashion to the linear setting. In particular, the $\beta_j(t)$'s were chosen so that the proportion of variation in $\int_0^1 X_{ij}(t)\beta_j(t)dt$ explained by the first principal component of $X_{ij}(t)$, varied between 69% and 99%. The key difference from the linear setting was that the responses were generated from model (13), where $\sigma = 0.1$ and the g_j 's were non-linear functions with p varying between 20 and 50

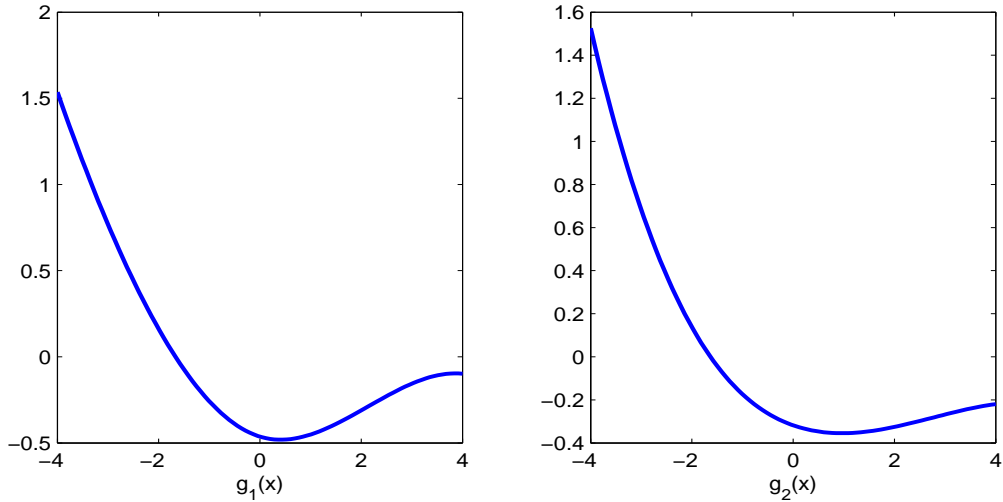


Figure 3: Plots of the non-linear functions, g_1 and g_2 , used to produce non-linear fits between the response and predictor.

depending on the simulation. To produce a sparse relationship between the predictors and the response we set $g_j = 0$ for $j = 3, 4, \dots, p$. The remaining two curves, g_1 and g_2 (illustrated in Figure 3), were generated from the B-spline basis, $\mathbf{b}(t)$, with the basis coefficients chosen to ensure a moderate level of non-linearity. To maintain comparable results we fixed g_1 and g_2 to be the same for all simulation settings.

We tested seven functional regression approaches. Since SCAD provided the best results in the linear setting we restricted attention to this penalty for the FAR and FPCA implementations in the non-linear simulations. We first fit two non-linear versions of FAR using the shrunk and unshrunk SCAD penalties, FAR and FAR*, respectively. Next, we fit the same linear FPCA methods, FPCA and FPCA*, from the linear simulation setup. To account for the non-linear relationships between the response and predictors, we also fit shrunk and unshrunk versions of the SpAM method (Ravikumar *et al.*, 2009), FPCA.NL and FPCA.NL*, to the principal component scores. SpAM essentially implements a penalized version of Generalized Additive Models (GAM), allowing for automatic variable selection in a non-linear but additive regression situation. Finally, we fit a standard linear regression, LS*, without any penalty function, to the p different principal component scores.

The five simulation settings corresponded to different combinations of $\gamma = 69\%$, 90% or 99% , $n = 100$ or 200 and $p = 20$ or 50 . In each setting we fit the methods to 100 separate data sets and, as in the linear setting, used a separate validation data set,

		FAR*	FAR	FPCA*	FPCA	FPCA.NL*	FPCA.NL	LS*
$\gamma = 69\%$ $n = 200$ $p = 20$	FNR	0.010	0.000	0.005	0.005	0.120	0.000	-
	FPR	0.042	0.105	0.021	0.212	0.003	0.044	-
	Mean(PE)	0.127	0.127	0.370	0.347	0.305	0.287	0.401
	SE(PE)	0.003	0.003	0.003	0.001	0.003	0.002	0.004
$\gamma = 90\%$ $n = 200$ $p = 20$	FNR	0.000	0.000	0.000	0.000	0.000	0.000	-
	FPR	0.057	0.090	0.012	0.335	0.003	0.020	-
	Mean(PE)	0.124	0.125	0.292	0.290	0.184	0.184	0.311
	SE(PE)	0.003	0.003	0.002	0.001	0.001	0.001	0.002
$\gamma = 90\%$ $n = 100$ $p = 50$	FNR	0.095	0.030	0.010	0.010	0.000	0.000	-
	FPR	0.016	0.103	0.022	0.374	0.007	0.102	-
	Mean(PE)	0.167	0.153	0.312	0.315	0.196	0.193	0.499
	SE(PE)	0.008	0.006	0.005	0.003	0.002	0.001	0.008
$\gamma = 90\%$ $n = 200$ $p = 50$	FNR	0.010	0.005	0.010	0.010	0.020	0.000	-
	FPR	0.005	0.061	0.004	0.132	0.000	0.004	-
	Mean(PE)	0.132	0.138	0.280	0.273	0.184	0.188	0.341
	SE(PE)	0.003	0.004	0.002	0.001	0.002	0.001	0.003
$\gamma = 99\%$ $n = 200$ $p = 20$	FNR	0.005	0.005	0.000	0.000	0.000	0.000	-
	FPR	0.082	0.216	0.056	0.521	0.001	0.026	-
	Mean(PE)	0.119	0.121	0.218	0.215	0.114	0.113	0.232
	SE(PE)	0.002	0.002	0.002	0.001	0.001	0.001	0.002

Table 2: Comparison of FAR to FPCA based methods in five non linear simulation settings. We use * to denote unshrunk estimators.

with identical characteristics to the training data, to select the tuning parameters. The simulation results are summarized graphically in Figure 4 and numerically in Table 2, with bold font indicating the statistically best prediction errors.

In terms of prediction error, both FAR methods dominated all the other approaches in all settings except the most favorable FPCA situation where $\gamma = 99\%$. In this setting the non-linear FPCA methods were superior, though the improvement over FAR was relatively small. The linear FPCA methods were inferior to both FAR and FPCA.NL but outperformed the unpenalized least squares approach. In comparing the shrunk and unshrunk versions of FAR, there was very little difference in the prediction errors. In general the unshrunk version resulted in lower false positive rates and, with the exception of the $n = 100$ setting, similar false negative rates.

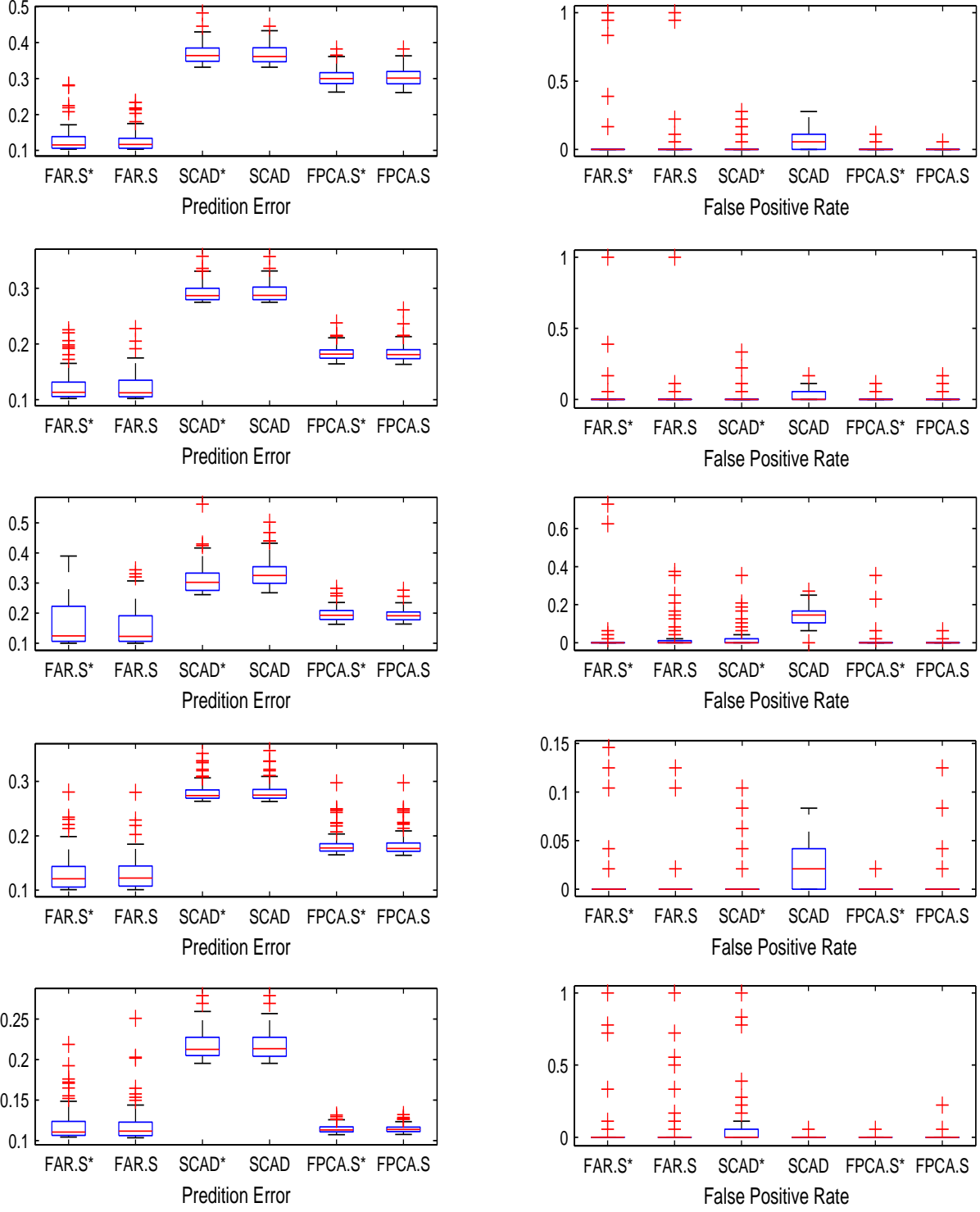


Figure 4: *Boxplots of Prediction Errors and False Positive Rates for the FAR and FPCA methods over 100 non-linear simulation runs. Each row provides results using the simulation settings from the corresponding row in Table 2. False Negative Rates were generally low so we have not plotted them here.*

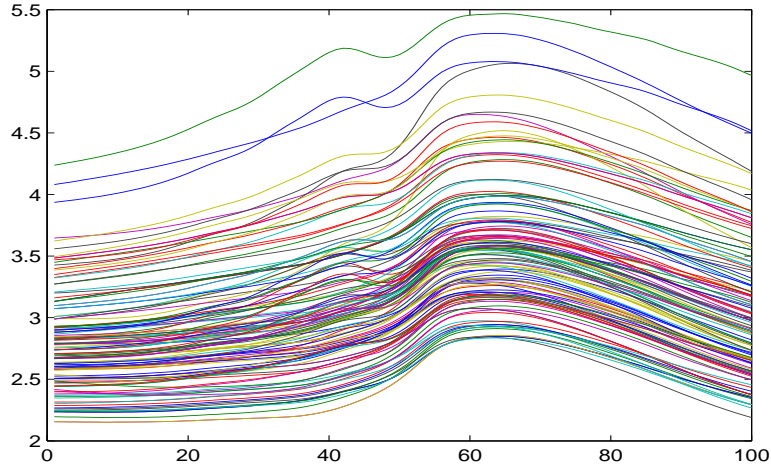


Figure 5: *Plots of the absorbance levels for the 215 meat samples over the wavelengths between 850 and 1050 nm.*

6 Real Data

In Sections 6.1 and 6.2 we respectively demonstrate FAR on low and high dimensional data sets.

6.1 Tecator Data

The Tecator data (available at StatLib) consists of 215 meat samples. For each sample we have 100 measurements of its absorbance, equally spaced over the wavelengths between 850 and 1050 nm. Thus we have a functional predictor, $X(t)$, for each of the 215 observations, as shown in Figure 5. We also include the first three derivatives of $X(t)$ and an additional 10 randomly created noise functions, for a total of 14 possible predictors. The purpose of the noise predictors is to judge how well the methods under consideration can identify the correct model. We also have three possible response variables; contents of moisture (water), fat and protein. We use the recommended decomposition for this data consisting of 129 observations for training, 43 for tuning and the remaining 43 as the test set.

For each of the three response variables we fit non-linear (NL-FAR) and linear (L-FAR) versions of FAR to the training data. We compared FAR to the FPCA.NL, FPCA and LS methods described in Section 5.2. We used the functional principal component scores that explained at least 85% of the variation in the curves as pre-

Response		NL-FAR	L-FAR	FPCA.NL	FPCA	LS	Mean
Moisture	PE	1.754	2.471	2.507	2.560	2.703	10.127
	$X(t)$	1	1	0	0	1	-
	Derivatives	0	0	2	2	3	-
	Noise	0	0	0	0	10	-
Fat	PE	1.793	2.614	2.597	3.126	3.198	13.122
	$X(t)$	1	1	0	0	1	-
	Derivatives	1	0	2	2	3	-
	Noise	0	0	0	0	10	-
Protein	PE	1.108	0.778	1.233	1.527	1.527	3.032
	$X(t)$	1	1	1	1	1	-
	Derivatives	0	1	1	3	3	-
	Noise	0	0	0	10	10	-

Table 3: Prediction errors and variables selected for five methods on each of the three response variables. The derivatives row indicates the number of derivatives of $X(t)$ that are selected in the final model.

dictors for the three comparison methods. The SCAD function was used for all of the penalty terms and an orthogonalized B-spline basis was used for all the methods. We found that, for non-linear FAR, the best results were obtained using relatively low dimensional bases; seven for $\beta(t)$ and five for $g(t)$. Linear FAR and the SpAM fit of FPCA.NL worked best with a higher, twelve-dimensional basis. The functional PCA fits used in FPCA.NL, FPCA and LS were also achieved by fitting the predictors using a twelve-dimensional basis.

The results of the analysis are presented in Table 3. The prediction error (PE) is computed based on the root mean squared error on the test data, using the unshrunk coefficient estimates. In addition to the five methods mentioned above we also computed the prediction error when using just the mean response from the training data (Mean). All methods show a significant improvement over that from using the training mean but, with one exception where Linear FAR is superior, the Non-Linear version of FAR dominates all the other methods for all responses. In terms of model accuracy FAR always includes $X(t)$ and either one or none of the derivatives. In all cases it correctly excludes all the noise variables. The competing FPCA methods often exclude $X(t)$, tend to include more of the derivatives, and in one case includes all the noise variables. There appear to be some clear non-linear relationships in the data with the non-linear versions of both FAR and FPCA generally outperforming their linear counterparts.

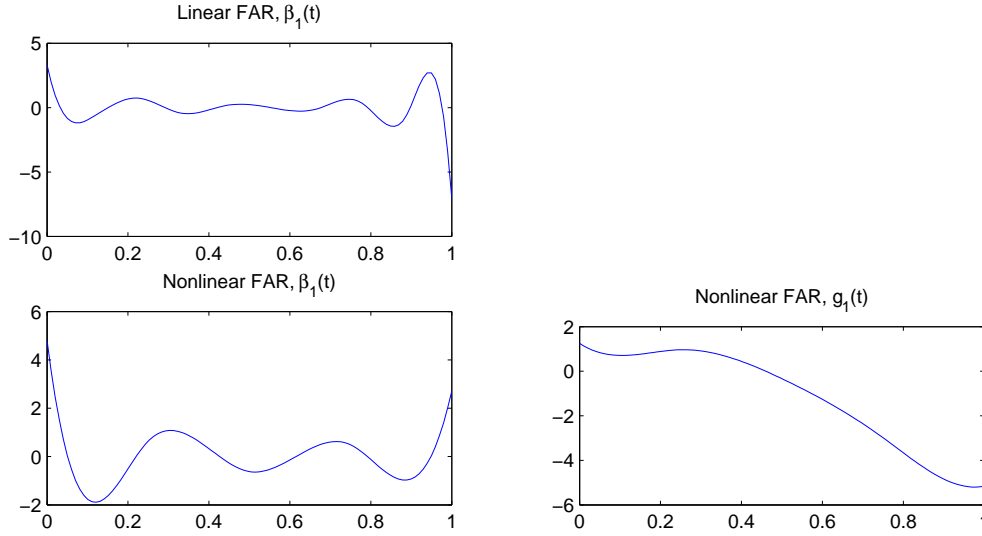


Figure 6: *Plots of $\beta(t)$ for linear FAR, and $\beta(t)$ and $g(t)$ for non-linear FAR on the moisture response.*

Figure 6 illustrates the estimated $\beta(t)$ and $g(t)$ curves from the linear and non-linear versions of FAR applied to the data with moisture as the response. The $\beta(t)$ curves for both versions of FAR have largest absolute values for the smallest and largest wavelengths and hence place most weight on these points. The $g(t)$ curve clearly illustrates the non-linear relationship with relatively little change predicted in moisture from around 0 to 0.3 and then a significant decline for larger values. Standard linear functional regression approaches would not be able to accurately model this relationship.

6.2 MEG Data

Our second data set consisted of Magnetoencephalography (MEG) recordings for 20 subjects conducted at the Center for Clinical Neurosciences, University of Texas Health Science Center at Houston. The MEG readings for each subject were recorded over 248 “channels” at 356 equally spaced time points. Each channel measured the intensity level of the magnetic field at a particular point on the brain. Multiple trials, consisting of reading a patient a word and measuring the MEG over time, were recorded for each patient. We averaged the trials for each channel and patient to produce 248 functional predictors. The response of interest was whether the patient was

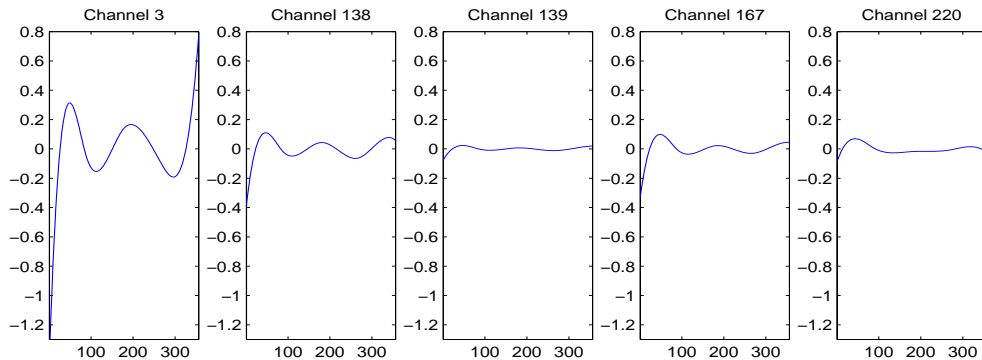


Figure 7: *Plots of $\beta(t)$ for linear FAR on the MEG data.*

left (14 subjects) or right (6 subjects) brain dominated. We coded $Y = 1$ and $Y = -1$ respectively for left and right brained subjects. Some channels were missing for some patients and were removed from the study, leaving a total of $p = 199$ predictors.

This was a very challenging data set because the ratio of predictors to observations was 10:1. We first fit the linear version of FAR to the full data set using a five dimensional basis. The tuning parameter was chosen as the point which minimized the classification error using 20-fold cross-validation. In this setting FAR selected only a five variable model (Channels 3, 138, 139, 167 and 220), which corresponded to a 20% cross-validated error rate. Figure 7 displays the $\beta(t)$ curves for each selected channel. All five channels put the bulk of their weight on the early time points. Channel 3 appears to provide the majority of the predictive power with smaller contributions from Channels 138 and 167.

We also fit the non-linear version of FAR. Given the small number of observations and the extra demands of fitting a non-linear regression method we felt it was prudent to first perform a marginal pre-screening to select a smaller subset of predictors for the final analysis. The marginal screening was performed by running non-linear FAR, using a 7-dimensional basis function, separately on each of the 194 predictors that linear FAR did not choose and selecting the 45 best predictors in terms of marginal prediction accuracy. Non-linear FAR was then run on the 50 predictors, including the 5 selected by linear FAR. 20-fold cross validation was again used to select the tuning parameter, resulting in five channels being selected. The channels were not the same as those selected by linear FAR. The cross-validated error rate was 25%, suggesting

that linear FAR may have a slight advantage on this data.

7 Discussion

FAR extends the recent linear penalized regression literature by incorporating functional predictors and modeling general non-linear relationships. It has several advantages over current functional regression methods. First, the penalized approach automatically deals with high dimensional data using an efficient coordinate descent algorithm. Second, the single index formulation provides a non-linear supervised method for projecting the predictors into a lower dimensional space, providing more accurate results than the traditional linear unsupervised PCA approach. Third, our theoretical results suggest that FAR should provide accurate model predictions and the simulation results show that FAR outperforms traditional approaches.

There are two obvious possible extensions for FAR. The first is to incorporate FAR into the generalized linear models setting. Conceptually, such an extension could be achieved by replacing the sum of squares term in (14) with the log likelihood and then using a modified version of the coordinate descent algorithm to maximize the criterion. The second possible extension would be to replace the single index model with a multiple index model of the form, $f_j(X_{ij}) = \sum_{k=1}^K g_{jk} \left(\int \beta_{jk}(t) X_{ij}(t) dt \right)$. This would increase the flexibility of FAR to model more general non-linear relationships.

A Proofs of Theoretical Results

Let $\boldsymbol{\eta} = (\boldsymbol{\eta}_1^T, \dots, \boldsymbol{\eta}_p^T)^T$ be a (pq) -vector and $\Theta = (\Theta_1, \dots, \Theta_p)$ be an $n \times (pq)$ matrix. With matrix notation, the regularization problem in (10) can be rewritten as

$$Q(\boldsymbol{\eta}) = \frac{1}{2n} \|Y - \Theta \boldsymbol{\eta}\|^2 + \sum_{j=1}^p \rho_{\lambda_n} \left(\frac{1}{\sqrt{n}} \|\Theta_j \boldsymbol{\eta}_j\| \right). \quad (17)$$

Since $\Theta_{\mathfrak{M}_0}^T \Theta_{\mathfrak{M}_0} = \sum_{j \in \mathfrak{M}_0} \Theta_j^T \Theta_j$, by Condition 2 it is easy to derive that $nc_0 \leq \Lambda_{\min}(\Theta_j^T \Theta_j) \leq \Lambda_{\min}(\Theta_{\mathfrak{M}_0}^T \Theta_{\mathfrak{M}_0}) \leq \Lambda_{\max}(\Theta_j^T \Theta_j) \leq \Lambda_{\max}(\Theta_{\mathfrak{M}_0}^T \Theta_{\mathfrak{M}_0}) \leq \frac{n}{c_0}$. Define the (qs_n) -dimensional hypercube

$$\mathcal{N} = \{ \boldsymbol{\eta} \in R^{pq} : \boldsymbol{\eta}_{\mathfrak{M}_0^c} = 0, \max_{k \in \mathfrak{M}_0} \|\Theta_k(\boldsymbol{\eta}_k - \boldsymbol{\eta}_{0,k})\| \leq n^{1/2-\alpha} \sqrt{\log n} \}. \quad (18)$$

Lemma 1. Define the event $\mathcal{E}_1 = \{\|\Theta_{\mathfrak{M}_0}^T \boldsymbol{\varepsilon}^*\|_\infty \leq 2c_0^{-1/2} \sigma \sqrt{n \log n}\}$. Then under Condition 2 and conditional on event \mathcal{E}_1 , there exists a vector $\boldsymbol{\eta} \in \mathcal{N}$ such that $\boldsymbol{\eta}_{\mathfrak{M}_0}$ is a solution to the following nonlinear equations

$$-\frac{1}{n} \Theta_{\mathfrak{M}_0}^T (\mathbf{Y} - \Theta_{\mathfrak{M}_0} \boldsymbol{\eta}_{\mathfrak{M}_0}) + \mathbf{v}_{\mathfrak{M}_0} = 0, \quad (19)$$

where $\mathbf{v}_{\mathfrak{M}_0}$ is a vector obtained by stacking the vectors $\mathbf{v}_k = \rho'_{\lambda_n}(\frac{1}{\sqrt{n}} \|\Theta_k \boldsymbol{\eta}_k\|) \frac{1}{\sqrt{n}} \frac{\Theta_k^T \Theta_k \boldsymbol{\eta}_k}{\|\Theta_k \boldsymbol{\eta}_k\|}$, $k \in \mathfrak{M}_0$ one underneath another.

Proof. For any $\tilde{\boldsymbol{\eta}} = (\tilde{\boldsymbol{\eta}}_1^T, \tilde{\boldsymbol{\eta}}_2^T, \dots, \tilde{\boldsymbol{\eta}}_m^T)^T \in \mathcal{N}$ and $k \in \mathfrak{M}_0$, by condition 2(D) and the definition of \mathcal{N} we have

$$\|\tilde{\boldsymbol{\eta}}_{\mathfrak{M}_0} - \boldsymbol{\eta}_{0, \mathfrak{M}_0}\|_\infty \leq \max_{k \in \mathfrak{M}_0} \|\tilde{\boldsymbol{\eta}}_k - \boldsymbol{\eta}_{0, k}\| \leq \max_{k \in \mathfrak{M}_0} \frac{\|\Theta_k(\tilde{\boldsymbol{\eta}}_k - \boldsymbol{\eta}_{0, k})\|}{(\Lambda_{\min}(\Theta_k^T \Theta_k))^{1/2}} < \frac{\log n}{\sqrt{c_0} n^\alpha}. \quad (20)$$

On the other hand, when n is large enough $\|\Theta_k \tilde{\boldsymbol{\eta}}_k\| \geq \|\Theta_k \boldsymbol{\eta}_{0, k}\| - \|\Theta_k(\tilde{\boldsymbol{\eta}}_k - \boldsymbol{\eta}_{0, k})\| > \sqrt{n} a_n / 2$. By Condition 2(A), $\rho'_{\lambda_n}(\cdot)$ is a decreasing function, thus for any $k \in \mathfrak{M}_0$, $\rho'_{\lambda_n}(\frac{1}{\sqrt{n}} \|\Theta_k \tilde{\boldsymbol{\eta}}_k\|) \leq \rho'_{\lambda_n}(a_n / 2)$. It follows from the definition of \mathbf{v} that

$$\|\mathbf{v}_{\mathfrak{M}_0}\|_\infty \leq \max_{k \in \mathfrak{M}_0} \rho'_{\lambda_n}(\frac{1}{\sqrt{n}} \|\Theta_k \tilde{\boldsymbol{\eta}}_k\|) \max_{k \in \mathfrak{M}_0} \frac{1}{\sqrt{n}} \frac{\|\Theta_k^T \Theta_k \tilde{\boldsymbol{\eta}}_k\|}{\|\Theta_k \tilde{\boldsymbol{\eta}}_k\|} \leq \frac{\rho'_{\lambda_n}(a_n / 2)}{\sqrt{c_0}}. \quad (21)$$

Therefore, $\|(\Theta_{\mathfrak{M}_0}^T \Theta_{\mathfrak{M}_0})^{-1} \mathbf{v}_{\mathfrak{M}_0}\|_\infty \leq c_0^{-1/2} \rho'_{\lambda_n}(a_n / 2) \|(\Theta_{\mathfrak{M}_0}^T \Theta_{\mathfrak{M}_0})^{-1}\|_\infty$. In addition, on the event \mathcal{E}_1 we have $\|(\Theta_{\mathfrak{M}_0}^T \Theta_{\mathfrak{M}_0})^{-1} \Theta_{\mathfrak{M}_0}^T \boldsymbol{\varepsilon}^*\|_\infty \leq 2\sigma c_0^{-1/2} \sqrt{n \log n} \|(\Theta_{\mathfrak{M}_0}^T \Theta_{\mathfrak{M}_0})^{-1}\|_\infty$. Since $\frac{1}{n} \Theta_{\mathfrak{M}_0}^T \Theta_{\mathfrak{M}_0}$ has bounded eigenvalues and $q s_n = o(n^{\frac{1}{2}-\alpha} \sqrt{\log n})$, it follows that $\|(\Theta_{\mathfrak{M}_0}^T \Theta_{\mathfrak{M}_0})^{-1}\|_\infty = O(\frac{1}{n} s_n q) = o(n^{-\frac{1}{2}-\alpha} \sqrt{\log n})$. Combing these and by Condition 2(C) we obtain

$$\begin{aligned} \|(\Theta_{\mathfrak{M}_0}^T \Theta_{\mathfrak{M}_0})^{-1} (n \mathbf{v}_{\mathfrak{M}_0} - \Theta_{\mathfrak{M}_0}^T \boldsymbol{\varepsilon})\|_\infty &\leq \|(\Theta_{\mathfrak{M}_0}^T \Theta_{\mathfrak{M}_0})^{-1} \mathbf{v}_{\mathfrak{M}_0}\|_\infty \\ &+ \|(\Theta_{\mathfrak{M}_0}^T \Theta_{\mathfrak{M}_0})^{-1} \Theta_{\mathfrak{M}_0}^T \boldsymbol{\varepsilon}^*\|_\infty \leq \frac{\log n}{\sqrt{c_0} n^\alpha} \end{aligned} \quad (22)$$

In view of (20) and (22) and by the continuity of the vector-valued function $\mathbf{g}(\boldsymbol{\eta}_{\mathfrak{M}_0}) = \boldsymbol{\eta}_{\mathfrak{M}_0} - \boldsymbol{\eta}_{0, \mathfrak{M}_0} - (\Theta_{\mathfrak{M}_0}^T \Theta_{\mathfrak{M}_0})^{-1} (n \mathbf{v}_{\mathfrak{M}_0} - \Theta_{\mathfrak{M}_0}^T \boldsymbol{\varepsilon})$, an application of Miranda's existence theorem (Vrahatis, 1989) shows that (19) indeed has a solution $\hat{\boldsymbol{\eta}}_{\mathfrak{M}_0}$ in \mathcal{N} . \square

Lemma 2. Define $\mathcal{E}_2 = \{\|\Theta_{\mathfrak{M}_0}^T \boldsymbol{\varepsilon}^*\|_\infty \leq 2\sigma n^{1/2} u_n\}$. Then under Conditions 2 and

conditional on the event $\mathcal{E}_1 \cap \mathcal{E}_2$, there exists a local minimizer of $Q(\boldsymbol{\eta})$ (17) in \mathcal{N} defined in (18).

Proof. By Lemma 1, we know that under Condition 2 there exists a vector $\hat{\boldsymbol{\eta}} \in \mathcal{N}$ such that $\hat{\boldsymbol{\eta}}_{\mathfrak{M}_0}$ is a solution to (18). We next show that under some additional conditions, $\hat{\boldsymbol{\eta}}$ is a local minimizer of $Q(\boldsymbol{\eta})$ in the original R^{pq} space.

We first constraint the objective function $Q(\boldsymbol{\eta})$ on the (qs_n) -dimensional subspace \mathcal{N} defined in (18). We will show that under Condition 2 and conditional on $\mathcal{E}_1 \cap \mathcal{E}_2$, $Q(\boldsymbol{\eta})$ is strictly convex around $\hat{\boldsymbol{\eta}}$. To this end, define $h(\boldsymbol{\eta}) = \sum_{j=1}^q \rho_{\lambda_n}(\frac{1}{\sqrt{n}} \|\Theta_j \boldsymbol{\eta}_j\|)$, which is a function in \mathbf{R}^{pq} . Note that

$$\begin{aligned} \frac{\partial^2}{\partial \boldsymbol{\eta}_k^2} h(\hat{\boldsymbol{\eta}}) &= \Theta_k^T \Theta_k \frac{\rho'_{\lambda_n}(\frac{1}{\sqrt{n}} \|\Theta_k \hat{\boldsymbol{\eta}}_k\|)}{\sqrt{n} \|\Theta_k \hat{\boldsymbol{\eta}}_k\|} \\ &+ \Theta_k^T \Theta_k \hat{\boldsymbol{\eta}}_k \hat{\boldsymbol{\eta}}_k^T \Theta_k^T \Theta_k \left(\frac{\rho''_{\lambda_n}(\frac{1}{\sqrt{n}} \|\Theta_k \hat{\boldsymbol{\eta}}_k\|)}{n \|\Theta_k \hat{\boldsymbol{\eta}}_k\|^2} - \frac{\rho'_{\lambda_n}(\frac{1}{\sqrt{n}} \|\Theta_k \hat{\boldsymbol{\eta}}_k\|)}{\sqrt{n} \|\Theta_k \hat{\boldsymbol{\eta}}_k\|^3} \right). \end{aligned} \quad (23)$$

It follows that $\Lambda_{\min}(\frac{\partial^2}{\partial \boldsymbol{\eta}_k^2} h(\hat{\boldsymbol{\eta}})) \geq \Lambda_{\max}(\Theta_k^T \Theta_k) \left(\frac{\rho''_{\lambda_n}(\frac{1}{\sqrt{n}} \|\Theta_k \hat{\boldsymbol{\eta}}_k\|)}{n} - \frac{\rho'_{\lambda_n}(\frac{1}{\sqrt{n}} \|\Theta_k \hat{\boldsymbol{\eta}}_k\|)}{\sqrt{n} \|\Theta_k \hat{\boldsymbol{\eta}}_k\|} \right)$. By the definition of \mathcal{N} , for any $k \in \mathfrak{M}_0$ and large enough n , $\|\Theta_k \hat{\boldsymbol{\eta}}_k\| \geq \|\Theta_k \boldsymbol{\eta}_{k,0}\| - \|\Theta_k(\hat{\boldsymbol{\eta}}_k - \boldsymbol{\eta}_{k,0})\| > \sqrt{n} a_n/2$. Thus by Condition 2 (A) and (C), $0 < \frac{\rho'_{\lambda_n}(\frac{1}{\sqrt{n}} \|\Theta_k \hat{\boldsymbol{\eta}}_k\|)}{\|\Theta_k \hat{\boldsymbol{\eta}}_k\|/\sqrt{n}} \leq \frac{\rho'_{\lambda_n}(a_n/2)}{a_n/2}$ and $0 > \rho''_{\lambda_n}(\frac{1}{\sqrt{n}} \|\Theta_k \hat{\boldsymbol{\eta}}_k\|) = O(n^{-1/2} \sqrt{\log n})$. Since $\Lambda_{\max}(\Theta_k^T \Theta_k) \leq n/c_0$, it is seen that

$$\Lambda_{\min}(\frac{\partial^2}{\partial \boldsymbol{\eta}_k^2} h(\hat{\boldsymbol{\eta}})) \geq -c_0^{-1} \left(\frac{\rho'_{\lambda_n}(a_n/2)}{a_n/2} + O(n^{-1/2} \sqrt{\log n}) \right).$$

Let H be a block diagonal matrix with block matrices $\frac{\partial^2}{\partial \boldsymbol{\eta}_k^2} h(\hat{\boldsymbol{\eta}})$, $k \in \mathfrak{M}_0$. Then it is easy to see that the Hessian matrix $\frac{\partial^2}{\partial \boldsymbol{\eta}_{\mathfrak{M}_0}^2} Q(\hat{\boldsymbol{\eta}}) = n^{-1} \Theta_{\mathfrak{M}_0}^T \Theta_{\mathfrak{M}_0} + H$. Since $\rho'_{\lambda_n}(a_n/2)/a_n = o(n^{\alpha-\frac{1}{2}} \sqrt{\log n}) = o(1)$, it follows that

$$\Lambda_{\min}(\frac{\partial^2}{\partial \boldsymbol{\eta}_{\mathfrak{M}_0}^2} Q(\hat{\boldsymbol{\eta}})) \geq \frac{1}{n} \Lambda_{\min}(\Theta_{\mathfrak{M}_0}^T \Theta_{\mathfrak{M}_0}) + \min_{k \in \mathfrak{M}_0} \Lambda_{\min}(\frac{\partial^2}{\partial \boldsymbol{\eta}_k^2} h(\hat{\boldsymbol{\eta}})) = c_0 - o(1). \quad (24)$$

Thus, for large enough n , (24) can be further bounded from below by $c_0/2$. Therefore, restricted in the space \mathcal{N} , the Hessian matrix $\frac{\partial^2}{\partial \boldsymbol{\eta}_{\mathfrak{M}_0}^2} Q(\boldsymbol{\eta})$ is strictly convex around $\hat{\boldsymbol{\eta}}$ and thus has a unique minimizer in a ball $\mathcal{N}_1 \subset \mathcal{N}$ centered at $\hat{\boldsymbol{\eta}}$. Since by Lemma 1 $\hat{\boldsymbol{\eta}}$ is a critical point, $\hat{\boldsymbol{\eta}}$ is this strict local minimizer in \mathcal{N}_1 .

We next show that $\hat{\boldsymbol{\eta}}$ is also a local minimizer in the original R^{pq} -dimensional

space. We will first show that for $\hat{\boldsymbol{\eta}}_{\mathfrak{M}_0}$ defined in Lemma 1, conditional on $\mathcal{E}_1 \cap \mathcal{E}_2$,

$$\max_{j \in \mathfrak{M}_0^c} \{\hat{\mathbf{v}}_j^T (\Theta_j^T \Theta_j)^{-1} \hat{\mathbf{v}}_j\}^{1/2} = \max_{j \in \mathfrak{M}_0^c} \|\Theta_j (\Theta_j^T \Theta_j)^{-1} \hat{\mathbf{v}}_j\| < n^{-1/2} \rho'_{\lambda_n}(0+), \forall j \in \mathfrak{M}_0^c, \quad (25)$$

where $\hat{\mathbf{v}}_j = n^{-1} \Theta_j^T (\mathbf{Y} - \Theta_{\mathfrak{M}_0} \hat{\boldsymbol{\eta}}_{\mathfrak{M}_0}) = n^{-1} \Theta_j^T \Theta_{\mathfrak{M}_0} (\boldsymbol{\eta}_{0, \mathfrak{M}_0} - \hat{\boldsymbol{\eta}}_{\mathfrak{M}_0}) + n^{-1} \Theta_j^T \boldsymbol{\varepsilon}^*$. By Lemma 1, we know that $\boldsymbol{\eta}_{0, \mathfrak{M}_0} - \hat{\boldsymbol{\eta}}_{\mathfrak{M}_0} = (\Theta_{\mathfrak{M}_0}^T \Theta_{\mathfrak{M}_0})^{-1} (n \mathbf{v}_{\mathfrak{M}_0} - \Theta_{\mathfrak{M}_0}^T \boldsymbol{\varepsilon}^*)$. Thus, for $j \in \mathfrak{M}_0^c$, $\hat{\mathbf{v}}_j = \Theta_j^T \Theta_{\mathfrak{M}_0} (\Theta_{\mathfrak{M}_0}^T \Theta_{\mathfrak{M}_0})^{-1} \mathbf{v}_{\mathfrak{M}_0} + n^{-1} [\Theta_j - \Theta_j^T \Theta_{\mathfrak{M}_0} (\Theta_{\mathfrak{M}_0}^T \Theta_{\mathfrak{M}_0})^{-1} \Theta_{\mathfrak{M}_0}^T] \boldsymbol{\varepsilon}^*$. Therefore,

$$\{\hat{\mathbf{v}}_j^T (\Theta_j^T \Theta_j)^{-1} \hat{\mathbf{v}}_j\}^{1/2} = \|\Theta_j (\Theta_j^T \Theta_j)^{-1} \hat{\mathbf{v}}_j\| \leq I_1 + I_2,$$

where $I_1 = \|\Theta_j (\Theta_j^T \Theta_j)^{-1} \Theta_j^T \Theta_{\mathfrak{M}_0} (\Theta_{\mathfrak{M}_0}^T \Theta_{\mathfrak{M}_0})^{-1} \mathbf{v}_{\mathfrak{M}_0}\|$, and $I_2 = n^{-1} \|\Theta_j (\Theta_j^T \Theta_j)^{-1} \Theta_j^T (\mathbf{I} - \Theta_{\mathfrak{M}_0} (\Theta_{\mathfrak{M}_0}^T \Theta_{\mathfrak{M}_0})^{-1} \Theta_{\mathfrak{M}_0}^T) \boldsymbol{\varepsilon}^*\|$. By (11), (12) and (21), conditional on $\mathcal{E}_1 \cap \mathcal{E}_2$,

$$\begin{aligned} I_1 &\leq \|\mathbf{v}_{\mathfrak{M}_0}\|_{\infty} \|\Theta_j (\Theta_j^T \Theta_j)^{-1} \Theta_j^T \Theta_{\mathfrak{M}_0} (\Theta_{\mathfrak{M}_0}^T \Theta_{\mathfrak{M}_0})^{-1}\|_{\infty, 2} < \frac{1}{2\sqrt{n}} \rho'_{\lambda_n}(0+), \\ I_2 &\leq n^{-1} \|\Theta_j (\Theta_j^T \Theta_j)^{-1} \Theta_j^T \Theta_{\mathfrak{M}_0} (\Theta_{\mathfrak{M}_0}^T \Theta_{\mathfrak{M}_0})^{-1}\|_{\infty, 2} \|\Theta_{\mathfrak{M}_0}^T \boldsymbol{\varepsilon}^*\|_{\infty} \\ &\quad + n^{-1} \|\Theta_j (\Theta_j^T \Theta_j)^{-1}\|_{\infty, 2} \|\Theta_j^T \boldsymbol{\varepsilon}^*\|_{\infty} < \frac{1}{2\sqrt{n}} \rho'_{\lambda_n}(0+). \end{aligned}$$

In summary, the above results on I_1 and I_2 show that the inequality (25) holds.

Let $\mathcal{B} = \{\boldsymbol{\eta} \in R^{pq} : \eta_{\mathfrak{M}_0^c} = 0\}$ be a subspace in R^{pq} . Take a sufficiently small ball \mathcal{N}_2 in R^{pq} centered at $\hat{\boldsymbol{\eta}}$ such that $\mathcal{N}_2 \cap \mathcal{B} \subset \mathcal{N}_1$. Since $\rho'_{\lambda_n}(t)$ is a continuous decreasing function and (25) holds for $\hat{\boldsymbol{\eta}} \in \mathcal{N}_2$, appropriately shrink the radius of the ball \mathcal{N}_2 gives that there exists a $\delta \in (0, \infty)$ such that for any $\boldsymbol{\eta} \in \mathcal{N}_2$,

$$\max_{j \in \mathfrak{M}_0^c} \|\Theta_j (\Theta_j^T \Theta_j)^{-1} \Theta_j^T (\mathbf{Y} - \Theta_{\mathfrak{M}_0} \boldsymbol{\eta}_{\mathfrak{M}_0})\| < n^{1/2} \rho'_{\lambda_n}(\delta). \quad (26)$$

Fix an arbitrary $\boldsymbol{\eta}_1 \in \mathcal{N}_2 / \mathcal{N}_1$, we will show that $Q(\boldsymbol{\eta}_1) > Q(\hat{\boldsymbol{\eta}})$. Let $\boldsymbol{\eta}_2$ be the projection of $\boldsymbol{\eta}_1$ onto \mathcal{B} . Then it follows from the definition of $\mathcal{N}_1, \mathcal{N}_2, \mathcal{B}$ and $\hat{\boldsymbol{\eta}}$ that $Q(\boldsymbol{\eta}_2) > Q(\hat{\boldsymbol{\eta}})$. Thus we only need to show that $Q(\boldsymbol{\eta}_1) \geq Q(\boldsymbol{\eta}_2)$.

Note that $Q(\boldsymbol{\eta}_1) - Q(\boldsymbol{\eta}_2) = \nabla Q(\boldsymbol{\eta}_3) (\boldsymbol{\eta}_1 - \boldsymbol{\eta}_2) = \sum_{j \in \mathfrak{M}_0^c} \boldsymbol{\eta}_{1j}^T \frac{\partial Q(\boldsymbol{\eta}_3)}{\partial \boldsymbol{\eta}_j}$, where $\boldsymbol{\eta}_3$ is a vector on the segment connecting $\boldsymbol{\eta}_1$ and $\boldsymbol{\eta}_2$. Since $\boldsymbol{\eta}_{2k} = 0$ for any $k \in \mathfrak{M}_0^c$, there exists a constant $0 < \gamma < 1$ such that $\boldsymbol{\eta}_{3k} = \gamma \boldsymbol{\eta}_{1k}$, $k \in \mathfrak{M}_0^c$. Then by the definitions of $\mathcal{B}, \mathcal{N}_1, \mathcal{N}_2$, we know that $\boldsymbol{\eta}_3 \in \mathcal{N}_2$. Shrink the ball \mathcal{N}_2 such that for any $\boldsymbol{\eta} \in \mathcal{N}_2$, $\|\Theta_k \boldsymbol{\eta}_k\| = \|\Theta_k (\boldsymbol{\eta}_k - \hat{\boldsymbol{\eta}}_k)\| \leq \sqrt{n} \delta$, $k \in \mathfrak{M}_0^c$. Since $\boldsymbol{\eta}_3 \in \mathcal{N}_2$, we have $\|\Theta_k \boldsymbol{\eta}_{3k}\| \leq \sqrt{n} \delta$

and thus $\rho'_{\lambda_n}(\frac{1}{\sqrt{n}}\|\Theta_k\boldsymbol{\eta}_{3k}\|) \geq \rho'_{\lambda_n}(\delta)$ for $k \in \mathfrak{M}_0^c$. Thus,

$$\begin{aligned} Q(\boldsymbol{\eta}_1) - Q(\boldsymbol{\eta}_2) &= \sum_{j \in \mathfrak{M}_0^c} \boldsymbol{\eta}_{1j}^T \left(-\frac{1}{n} \Theta_j^T (\mathbf{Y} - \Theta \boldsymbol{\eta}_3) + \frac{\rho'_{\lambda_n}(\frac{1}{\sqrt{n}}\|\Theta_j \boldsymbol{\eta}_{3j}\|)}{\sqrt{n}\|\Theta_j \boldsymbol{\eta}_{3j}\|} \Theta_j^T \Theta_j \boldsymbol{\eta}_{3j} \right) \\ &\geq -\frac{1}{n} \sum_{j \in \mathfrak{M}_0^c} \boldsymbol{\eta}_{1j}^T \Theta_j^T (\mathbf{Y} - \Theta \boldsymbol{\eta}_3) + \frac{1}{\sqrt{n}\gamma} \rho'_{\lambda_n}(\delta) \sum_{j \in \mathfrak{M}_0^c} \|\Theta_j \boldsymbol{\eta}_{3j}\| \equiv I_3 + I_4. \end{aligned}$$

Next note that by (26)

$$|I_3| \leq \frac{1}{n\gamma} \sum_{j \in \mathfrak{M}_0^c} \|\Theta_j \boldsymbol{\eta}_{3j}\| \|\Theta_j (\Theta_j^T \Theta_j)^{-1} \Theta_j^T (\mathbf{Y} - \Theta_{\mathfrak{M}_0} \boldsymbol{\eta}_{3, \mathfrak{M}_0})\| \leq \frac{1}{\sqrt{n}\gamma} \rho'_{\lambda_n}(\delta) \sum_{j \in \mathfrak{M}_0^c} \|\Theta_j \boldsymbol{\eta}_{3j}\|.$$

Thus, $Q(\boldsymbol{\eta}_1) - Q(\boldsymbol{\eta}_2) \geq 0$. This proves that $Q(\boldsymbol{\eta}_1) \geq Q(\hat{\boldsymbol{\eta}})$. Thus, $\hat{\boldsymbol{\eta}}$ is also a strict local minimizer in the original R^{pq} dimensional space. \square

Proof of Theorem 1

Proof. We only need to show that $P(\mathcal{E}_1 \cap \mathcal{E}_2) \rightarrow 1$. Then Theorem 1 follows easily from Lemmas 1 and 2. To this end, note that $P(\mathcal{E}_1 \cap \mathcal{E}_2) \geq 1 - P(\mathcal{E}_1^c) - P(\mathcal{E}_2^c)$. By Condition 1 and assumption that $s_n = O(n^{2\delta - \frac{1}{2}})$, it is easy to derive that $\|\mathbf{e}\|_\infty = O(n^{-1/2})$. Since $\|\Theta_{\mathfrak{M}_0}^T\|_\infty = O(n)$, we can derive that $\|\Theta_{\mathfrak{M}_0}^T \mathbf{e}\|_\infty \leq \|\Theta_{\mathfrak{M}_0}^T\|_\infty \|\mathbf{e}\|_\infty = O(\sqrt{n})$. Since $\boldsymbol{\varepsilon}^* = \boldsymbol{\varepsilon} + \mathbf{e}$ with $\mathbf{e} = (\sum_{j=1}^p e_{1j}, \dots, \sum_{j=1}^p e_{nj})^T$, we obtain that for large enough $n > 0$, $P(\mathcal{E}_1^c) \leq P(\|\Theta_{\mathfrak{M}_0} \boldsymbol{\varepsilon}\|_\infty > c_0^{-1/2} \sigma \sqrt{2n \log n})$. Let $\boldsymbol{\xi} = (\xi_1, \dots, \xi_{ps_n})^T = \Theta_{\mathfrak{M}_0} \boldsymbol{\varepsilon}$, then $\xi_i \sim N(0, n\sigma^2 d_i^2)$ with d_i^2 the i -th diagonal of matrix $n^{-1} \Theta_{\mathfrak{M}_0} \Theta_{\mathfrak{M}_0}^T$. Then $c_0 \leq d_i^2 \leq \frac{1}{c_0}$ for all i . Hence, if $\frac{qs_n}{n\sqrt{\log n}} \rightarrow 0$, we further obtain that

$$P(\mathcal{E}_1^c) \leq \sum_{i=1}^{qs_n} P(|\xi_i| > \sigma c_0^{-1/2} \sqrt{2n \log n}) \leq \frac{2qs_n}{n\sqrt{\log n}} \rightarrow 0.$$

Similarly, we can see that $\|\Theta_{\mathfrak{M}_0}^T \mathbf{e}\|_\infty \leq \|\Theta_{\mathfrak{M}_0}^T\|_\infty \|\mathbf{e}\|_\infty = O(\sqrt{n}) = o(u_n \sqrt{n})$. Moreover, since $\log(q(p - s_n)) = o(u_n)$,

$$P(\mathcal{E}_2^c) \leq P(\|\Theta_{\mathfrak{M}_0}^T \boldsymbol{\varepsilon}\|_\infty > \sigma \sqrt{n} u_n) \leq q(m - s_n) u_n^{-1} \exp(-\frac{u_n^2 c_0}{2}) \rightarrow 0.$$

Thus, Theorem 1 has been proved. \square

Proof of Theorem 2

Proof. Since $\hat{\eta}$ is a solution to (19), for any vector $\mathbf{c} \in \mathbf{R}^{s_n q}$ satisfying $\mathbf{c}^T \mathbf{c} = 1$, it can be written as

$$\begin{aligned} & \mathbf{c}^T [(\Theta_{\mathfrak{M}_0}^T \Theta_{\mathfrak{M}_0})^{1/2} (\hat{\boldsymbol{\eta}}_{\mathfrak{M}_0} - \boldsymbol{\eta}_{0, \mathfrak{M}_0}) + n(\Theta_{\mathfrak{M}_0}^T \Theta_{\mathfrak{M}_0})^{-1/2} \mathbf{v}_{0, \mathfrak{M}_0}] \\ &= \mathbf{c}^T (\Theta_{\mathfrak{M}_0}^T \Theta_{\mathfrak{M}_0})^{-1/2} \Theta_{\mathfrak{M}_0}^T \boldsymbol{\varepsilon} + \mathbf{c}^T (\Theta_{\mathfrak{M}_0}^T \Theta_{\mathfrak{M}_0})^{-1/2} \Theta_{\mathfrak{M}_0}^T \mathbf{e} + n \mathbf{c}^T (\Theta_{\mathfrak{M}_0}^T \Theta_{\mathfrak{M}_0})^{-1/2} (\hat{\mathbf{v}}_{\mathfrak{M}_0} - \mathbf{v}_{\mathfrak{M}_0}) \\ &\equiv I_1 + I_2 + I_3. \end{aligned} \quad (27)$$

It is easy to see that $I_1 \sim N(0, \sigma^2)$. As for I_2 , note that we have proved in Theorem 1 that $\|\mathbf{e}\|_\infty = o(n^{-1/2})$. Thus, $\|\mathbf{e}\| = o(1)$. So we can derive that

$$|I_2| \leq \|\mathbf{c}^T (\Theta_{\mathfrak{M}_0}^T \Theta_{\mathfrak{M}_0})^{-1/2} \Theta_{\mathfrak{M}_0}^T\| \|\mathbf{e}\| = \|\mathbf{e}\| = o(1). \quad (28)$$

Now let us consider I_3 . By Cauchy-Schwartz inequality we obtain that

$$|I_3| \leq \|\sqrt{n} \mathbf{c}^T (\Theta_{\mathfrak{M}_0}^T \Theta_{\mathfrak{M}_0})^{-1/2}\| \|\sqrt{n} (\hat{\mathbf{v}}_{\mathfrak{M}_0} - \mathbf{v}_{\mathfrak{M}_0})\| \leq c_0^{-1/2} \|\sqrt{n} (\hat{\mathbf{v}}_{\mathfrak{M}_0} - \mathbf{v}_{\mathfrak{M}_0})\|. \quad (29)$$

Note that $\hat{\mathbf{v}}_k - \mathbf{v}_{0, k} = g(\hat{\boldsymbol{\eta}}_k) - g(\boldsymbol{\eta}_{0, k}) = \frac{\partial}{\partial \boldsymbol{\eta}} g(\tilde{\boldsymbol{\eta}}_k) (\hat{\boldsymbol{\eta}}_k - \boldsymbol{\eta}_{0, k})$ with $\tilde{\boldsymbol{\eta}}_k$ lying on the segment connecting $\boldsymbol{\eta}_{0, k}$ and $\hat{\boldsymbol{\eta}}_k$. Define $g(\boldsymbol{\eta}_k) = \frac{1}{\sqrt{n}} \rho'_{\lambda_n} (\frac{1}{\sqrt{n}} \|\Theta_k \boldsymbol{\eta}_k\|) \frac{\Theta_k^T \Theta_k \boldsymbol{\eta}_k}{\|\Theta_k \boldsymbol{\eta}_k\|}$. Then

$$\frac{\partial}{\partial \boldsymbol{\eta}_k} g(\boldsymbol{\eta}_k) = \rho''_{\lambda_n} \left(\frac{1}{n} \|\Theta_k \boldsymbol{\eta}_k\| \right) \frac{\Theta_k^T \Theta_k \boldsymbol{\eta}_k \boldsymbol{\eta}_k^T \Theta_k^T \Theta_k}{n \|\Theta_k \boldsymbol{\eta}_k\|^2} + \frac{\rho'_{\lambda_n} (\frac{1}{n} \|\Theta_k \boldsymbol{\eta}_k\|)}{\sqrt{n}} \left\{ \frac{\Theta_k^T \Theta_k}{\|\Theta_k \boldsymbol{\eta}_k\|} - \frac{\Theta_k^T \Theta_k \boldsymbol{\eta}_k \boldsymbol{\eta}_k^T \Theta_k^T \Theta_k}{\|\Theta_k \boldsymbol{\eta}_k\|^3} \right\}.$$

By Condition 2(A), for any $k \in \mathfrak{M}_0$,

$$c_0^{-1} \left(-O\left(\frac{\sqrt{\log n}}{\sqrt{n}}\right) - \frac{2\rho'_{\lambda_n}(\frac{a_n}{2})}{a_n} \right) \leq \Lambda_{\min} \left(\frac{\partial}{\partial \boldsymbol{\eta}_k} g(\boldsymbol{\eta}_k) \right) \leq \Lambda_{\max} \left(\frac{\partial}{\partial \boldsymbol{\eta}_k} g(\boldsymbol{\eta}_k) \right) \leq c_0^{-1} \frac{2\rho'_{\lambda_n}(\frac{a_n}{2})}{a_n}.$$

This together with Theorem 1 ensures that

$$\begin{aligned} \|\hat{\mathbf{v}}_{\mathfrak{M}_0} - \mathbf{v}_{\mathfrak{M}_0}\| &= \left\{ \sum_{j \in \mathfrak{M}_0} \|\hat{\mathbf{v}}_j - \mathbf{v}_{0, j}\|^2 \right\}^{1/2} \leq c_0^{-1} \left(O\left(\frac{\sqrt{\log n}}{\sqrt{n}}\right) + \frac{2\rho'_{\lambda_n}(\frac{a_n}{2})}{a_n} \right) \left\{ \sum_{k \in \mathfrak{M}_0} \|\hat{\boldsymbol{\eta}}_k - \boldsymbol{\eta}_{0, k}\|^2 \right\}^{1/2} \\ &\leq c_0^{-3/2} \left(O\left(\frac{\sqrt{\log n}}{\sqrt{n}}\right) + o(n^{\alpha-1/2} s_n^{-1/2} (\log n)^{-1}) \right) O_p(s_n^{1/2} n^{-\alpha} \log n) = o_p(n^{-1/2}), \end{aligned}$$

where in the last two steps we have used the assumptions $s_n = o(n^{2\alpha} (\log n)^{-3})$ and $\rho'_{\lambda_n}(a_n/2) = o(a_n n^{\alpha-1/2} s_n^{-1/2} (\log n)^{-1})$. So it follows that $\sqrt{n} \|\hat{\mathbf{v}}_{\mathfrak{M}_0} - \mathbf{v}_{\mathfrak{M}_0}\| = o_p(1)$. Combining this with (29) yields $I_3 \xrightarrow{P} 0$. This together with (27) - (29) completes the

proof. □

B Step A of FAR Algorithm

To implement Step A. of the FAR algorithm, for each $j \in \{1, \dots, p\}$, we estimate $\boldsymbol{\eta}_j$ as the parameter that minimizes Q_j in (31). To implement the minimization of Q_j over $\boldsymbol{\eta}_j$ we observe that, with fixed $\boldsymbol{\xi}_j$, (31) is proportional to the log likelihood of a generalized linear model (GLM) with link function $g_j(x) = \mathbf{h}(x)^T \boldsymbol{\xi}_j$ and a Gaussian response distribution. The most common procedure for computing the MLE for a GLM is to use the Fisher scoring algorithm, $\boldsymbol{\eta}_j^{(l+1)} = \boldsymbol{\eta}_j^{(l)} - \left(E \frac{\partial^2 Q_j}{\partial \boldsymbol{\eta}_j^2} \right)^{-1} \frac{\partial Q_j}{\partial \boldsymbol{\eta}_j}$. Simple algebra shows that $\frac{\partial Q_j}{\partial \boldsymbol{\eta}_j} = - \sum_{i=1}^n \{R_{ij} - g_j(\boldsymbol{\theta}_{ij}^T \boldsymbol{\eta}_j)\} g'_j(\boldsymbol{\theta}_{ij}^T \boldsymbol{\eta}_j) \boldsymbol{\theta}_{ij}$ and $E \frac{\partial^2 Q_j}{\partial \boldsymbol{\eta}_j^2} = \sum_{i=1}^n g'_j(\boldsymbol{\theta}_{ij}^T \boldsymbol{\eta}_j) \boldsymbol{\theta}_{ij} \boldsymbol{\theta}_{ij}^T$. Hence (31) can be optimized by iteratively updating $\boldsymbol{\eta}_j$ using the following equation,

$$\boldsymbol{\eta}_j^{(l+1)} = \boldsymbol{\eta}_j^{(l)} + (\tilde{\Theta}_j^T \tilde{\Theta}_j)^{-1} \tilde{\Theta}_j^T \tilde{\mathbf{R}}_j, \quad (30)$$

where $\tilde{\Theta}_j$ is an n by q matrix with i th row given by $g'_j(\boldsymbol{\theta}_{ij}^T \boldsymbol{\eta}_j^{(l)}) \boldsymbol{\theta}_{ij}^T$ and $\tilde{\mathbf{R}}_j$ is a length n vector with i th component equal to $R_{ij} - g_j(\boldsymbol{\theta}_{ij}^T \boldsymbol{\eta}_j^{(l)})$. Equation 30 is a version of the iteratively reweighted least squares (IRLS) approach typically used to fit GLM's. The IRLS generally only requires a few iterations to fit a GLM and we have found (30) to be similarly efficient in the FAR setting. For certain values of $\lambda = \lambda_t$, some of the $\boldsymbol{\xi}_j$'s will be set to zero and hence it is not possible to compute the corresponding $\boldsymbol{\eta}_j$'s. In these cases we leave $\boldsymbol{\eta}_j$ at the value estimated using the previous $\lambda = \lambda_{t-1}$.

To fully minimize (16) requires a full coordinate descent algorithm where one iteratively reestimates $\boldsymbol{\eta}_j$ for each $j \in \{1, \dots, p\}$ until convergence. However, we found in our simulations that this step adds significant computational cost with very little change in the final solution. Hence, we only approximately minimize (16) by estimating $\boldsymbol{\eta}_j$ once for each j .

C Initial Parameters for the FAR Algorithm

To get initial values for the $\boldsymbol{\xi}_j$ parameters with $\lambda = \lambda_1$ we first set $\boldsymbol{\xi}_j = 0$ for all j . Then, for each $j \in \{1, \dots, p\}$:

1. Fix all $\hat{\mathbf{f}}_k$ for $k \neq j$ and compute the residual vector $\mathbf{R}_j = \mathbf{Y} - \sum_{k \neq j} \hat{\mathbf{f}}_k(X_k)$.

2. Use Sliced Inverse Regression (Li, 1991), with $\boldsymbol{\theta}_{ij}$ as the predictor, to estimate $\boldsymbol{\eta}_j$.
3. Given $\widehat{\boldsymbol{\eta}}_j$, estimate $\boldsymbol{\xi}_j$ using the usual least squares solution to

$$Q_j = \sum_{i=1}^n (R_{ij} - \mathbf{h}(\boldsymbol{\theta}_{ij}^T \widehat{\boldsymbol{\eta}}_j)^T \boldsymbol{\xi}_j)^2. \quad (31)$$

After Step 3. the FAR algorithm operates in an identical fashion for all values of λ .

References

- Alter, O., Brown, P. O., and Botstein, D. (2000). Singular value decomposition for genome-wide expression data processing and modeling. *Proc. Nat. Acad. Sci. USA* **97**, 10101–10106.
- Cardot, H., Ferraty, F., and Sarda, P. (2003). Spline estimators for the functional linear model. *Statistica Sinica* **13**, 571–591.
- Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression (with discussion). *The Annals of Statistics* **32**, 2, 407–451.
- Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* **96**, 1348–1360.
- Fan, J. and Lv, J. (2011). Non-concave penalized likelihood with np-dimensionality. *IEEE Transactions on Information Theory*, to appear.
- Ferraty, F. and Vieu, P. (2003). Curves discrimination: a nonparametric functional approach. *Computational Statistics and Data Analysis* **44**, 161–173.
- Hall, P., Poskitt, D. S., and Presnell, B. (2001). A functional data-analytic approach to signal discrimination. *Technometrics* **43**, 1–9.
- Hall, P., Reimann, J., and Rice, J. (2000). Nonparametric estimation of a periodic function. *Biometrika* **87**, 545–557.
- Hastie, T. and Mallows, C. (1993). Comment on “a statistical view of some chemometrics regression tools”. *Technometrics* **35**, 140–143.

- James, G. M. (2002). Generalized linear models with functional predictors. *Journal of the Royal Statistical Society, Series B* **64**, 411–432.
- James, G. M. and Silverman, B. W. (2005). Functional adaptive model estimation. *Journal of the American Statistical Association* **100**, 565–576.
- Li, K. (1991). Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association* **86**, 316–327.
- Lv, J. and Fan, Y. (2009). A unified approach to model selection and sparse recovery using regularized least squares. *Ann. Statist.* **37**, 3498–3528.
- Muller, H. G. and Stadtmuller, U. (2005). Generalized functional linear models. *Annals of Statistics* **33**, 774–805.
- Muller, H. G. and Yao, F. (2008). Functional additive models. *Journal of the American Statistical Association* **103**, 426–437.
- Ramsay, J. O. and Silverman, B. W. (2005). *Functional Data Analysis*. Springer, 2nd edn.
- Ravikumar, P., Lafferty, J., Liu, H., and Wasserman, L. (2009). Sparse additive models. *Journal of the Royal Statistical Society, B* **71**, 1009–1030.
- Storey, J. D., Xiao, W., Leek, J. T., Tompkins, R. G., and Davis, R. W. (2005). Significance analysis of time course microarray experiments. *PNAS* **102**, 12837–12842.
- Vrahatis, M. N. (1989). A short proof and a generalization of miranda’s existence theorem. *Proceedings of the American Mathematical Society* **107**, 701–703.
- Yuan, M. and Lin, Y. (2007). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B* **68**, 49–67.
- Zou, H. (2007). On the degrees of freedom of the lasso. *The Annals of Statistics* **35**, 2173–2192.
- Zou, H. and Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models (with discussion). *The Annals of Statistics* **36**, 1509–1566.