Functional Adaptive Model Estimation

GARETH M. JAMES* and BERNARD W. SILVERMAN†

Abstract

In this article we are interested in modeling the relationship between a scalar, Y, and a functional predictor, $X(t)$. We introduce a highly flexible approach called "Functional Adaptive Model Estimation" (FAME) which extends generalized linear models (GLM), generalized additive models (GAM) and projection pursuit regression (PPR) to handle functional predictors. The FAME approach can model any of the standard exponential family of response distributions that are assumed for GLM or GAM while maintaining the flexibility of PPR. For example standard linear or logistic regression with functional predictors, as well as far more complicated models, can easily be applied using this approach. A functional principal components decomposition of the predictor functions is used to aid visualization of the relationship between $X(t)$ and Y. We also show how the FAME procedure can be extended to deal with multiple functional and standard finite dimensional predictors, possibly with missing data. The FAME approach is illustrated on simulated data as well as on the prediction of five year survival for a patient given observations of blood chemistry levels over time and the prediction of arthritis based on bone shape. Asymptotic results are developed and used to provide tests for significance of the predictors. We end with a discussion of the relationships between standard regression approaches, their extensions to functional data and FAME.

Some key words: Functional predictor; Functional principal components; Generalized linear models; Generalized additive models; Projection pursuit regression.

1 Introduction

It is increasingly common to encounter regression problems where either the predictor, the response or both are functional in nature. A majority of the previous work in this area involves a functional response. For instance, Moyeed and Diggle (1994) and Zeger and Diggle (1994) model the relationship between response, $Y(t)$, and predictor, $X(t)$, both measured over time, using the equation,

$$Y(t) = \alpha_0(t) + \beta_0^T X(t) + \varepsilon(t) \tag{1}$$

where $\alpha_0(t)$ is a smooth function of t, β_0 is a fixed but unknown vector of regression coefficients and $\varepsilon(t)$ is a zero mean stationary Gaussian process. Hoover *et al.* (1998), Wu *et al.* (1998) and Lin and Ying (2001) use the varying-coefficient models proposed in Hastie and Tibshirani (1993) to extend (1) by allowing the regression coefficients to vary over time. Fahrmeir and Tutz (1994) and Liang and Zeger (1986) suggest

*Marshall School of Business, University of Southern California
†University of Oxford

ID	5 Year	Drug	Day	Bili	Alb	ID	5 Year	Drug	Day	Bili	Alb
1	No	Yes	0	14.5	2.60	2	Yes	Yes	2515	4.2	2.73
1	No	Yes	192	21.3	2.94	2	Yes	Yes	2882	3.6	2.80
2	Yes	Yes	0	1.1	4.14	2	Yes	Yes	3226	4.6	2.67
2	Yes	Yes	182	0.8	3.60	3	No	Yes	0	1.4	3.48
2	Yes	Yes	365	1.0	3.55	3	No	Yes	176	1.1	3.29
2	Yes	Yes	768	1.9	3.92	3	No	Yes	364	1.5	3.57
2	Yes	Yes	1790	2.6	3.32	3	No	Yes	743	1.8	3.25
2	Yes	Yes	2151	3.6	2.92	⋮	⋮	⋮	⋮	⋮	⋮

Table 1: *Subset of data, obtained from StatLib, of a Mayo Clinic trial in primary biliary cirrhosis (PBC) of the liver conducted between 1974 and 1984.*

an even more general framework where the response is modeled as a member of the exponential family of distributions.

We are interested in an alternative situation where the predictors are functional but the response is scalar. An example of such a situation is provided in Table 1. The data come from a randomized placebo controlled trial of the drug D-penicillamine on patients with primary biliary cirrhosis (PBC) of the liver conducted by the Mayo Clinic between 1974 and 1984 (Fleming and Harrington, 1991). For each patient we have a record of five year survival from enrollment in the study ("5 year"), whether they received the drug ("Drug"), day of each patient visit measured from registration ("Day"), serum bilirubin in mg/dl ("Bili") and albumin in gm/dl ("Alb"). Several other potential predictors were measured but we will restrict to these variables for illustrative purposes. Ultimately we wish to understand the relationship between five year survival and the predictors Bili, Alb and Drug. Thus we treat the data as consisting of one scalar and two functional predictors with a scalar response. Another example which we will refer to in this paper, involves two dimensional cross-sectional outlines of the knee end of the femur (the upper leg bone) and an indicator of arthritis (Ramsay and Silverman, 2002). The aim here is to use the functional image of the bone to predict the presence of arthritis.

This type of structure arises in numerous applications. Muller and Stadtmuller (2003) provide illustrations in astronomy (Hall *et al.*, 2000), DNA expression arrays with repeated measures (Alter *et al.*, 2000) and engineering (Hall *et al.*, 2001). However, there has been limited methodological work in this area. Hastie and Mallows (1993), Ramsay and Silverman (1997), Chapter 10 and Cardot *et al.* (2003b) discuss performing linear regression where the response is a scalar and the predictors functional. Ferraty and Vieu (2002) develop a nonparametric regression procedure. James and Hastie (2001) and Ferraty and Vieu (2003) use functional linear discriminant analysis models to perform classification for categorical responses with functional predictors. Marx and Eilers (1999), James (2002) and Muller and Stadtmuller (2003) suggest somewhat more general methods which provide extensions of generalized linear models (McCullagh and Nelder, 1989) to functional predictors. In this article we introduce a procedure that facilitates the modeling of highly non-linear response surfaces on general classes of response distributions using functional predictors. For standard p-dimensional predictors, non-linearity can be achieved through the use of procedures such as generalized additive models (Hastie and Tibshirani, 1990) or, if even more flexibility is required, through projection pursuit regression (Friedman and Stuetzle, 1981). Our approach, which we call "functional adaptive model estimation (FAME)", combines characteristics of projection pursuit regression with

generalized linear and additive models.

In Section 2 we present and motivate the FAME model for data with a single functional predictor as well as providing a fitting algorithm. Section 3 details the development of some asymptotic results for the FAME fit. These results are used to provide confidence intervals and significance tests for model parameters. We illustrate the FAME methodology on simulated data as well as the PBC data set in Section 4. Extensions to multiple functional and finite dimensional covariates are given in Section 5 and illustrated on the PBC and femur bone data. Section 6 presents a simulation study which compares the performance of the FAME approach with other possible methods. Finally, Section 7 provides a discussion of the relationship of the FAME methodology to other finite dimensional and functional approaches.

2 Functional adaptive model estimation

In order to motivate our approach we first briefly review generalized linear models (GLM), generalized additive models (GAM) and projection pursuit regression (PPR). Generalized linear models provide a flexible framework for regressing response variables from the exponential family of distributions. One models the relationship between predictors $\mathbf{X} = (X_1, X_2, \ldots, X_p)$ and response Y using the link function $g(\mu) = \beta_0 + \sum_{j=1}^{p} X_j \beta_j$ where $\mu = E(Y|\mathbf{X})$. While GLMs cover a wide class of response distributions they still assume a linear relationship between the predictors and $g(\mu)$. This linearity assumption is relaxed with generalized additive models using the link $g(\mu) = \beta_0 + \sum_{j=1}^{p} f_j(X_j)$ where f_j is a smooth function estimated as part of the fitting procedure. GAMs allow for non-linear but still additive relationships between the predictors and $g(\mu)$. The additivity of GAM has the advantage that it allows one to identify the effect of each predictor individually while holding all other predictors constant but it significantly restricts the range of functions that can be fit.

Projection pursuit regression removes the additivity constraint by modeling a Gaussian response using

$$Y = \beta_0 + \sum_{k=1}^{r} f_k(\mathbf{X}^T \beta_k) + \varepsilon$$

where r is arbitrary. PPR has several advantages over both GLM and GAM. First, it allows one to model a larger class of functions. For example, GAM can not model the simple interaction $g(\mu) = X_1 X_2$ while PPR can. In fact by setting r large enough one can model any continuous function. Second, by studying the β_k's one learns in which directions the variability of the predictors provide the most information about the response. However, because PPR does not utilize a link function it has less flexibility in terms of response distributions that can be modeled. Roosen and Hastie (1993) and more recently Lingjaerde and Liestol (1998) remove this constraint by adding a link of the form

$$g(\mu) = \beta_0 + \sum_{k=1}^{r} f_k(\mathbf{X}^T \beta_k) \tag{2}$$

where both f_k and β_k are estimated in the fitting procedure. This method is called generalized projection pursuit (GPP). The GLM and GAM link functions may both be considered special cases of (2).

The aim of this paper is to extend GPP to data with functional predictors using our "functional adaptive model estimation" procedure. FAME can model non-Gaussian responses with the ease of GLM and GAM, has the flexibility of PPR to fit non-linear response surfaces and can be applied to functional data. One

possible approach to fitting GPP to such data would be to sample the functional predictor, $X(t)$, over a fine grid of p time points to create a vector $\mathbf{X}$, thus removing the functional aspect of the problem. However, this approach has several potential problems. First, it necessitates modeling a very high-dimensional vector of coefficients, which may lead to an extremely unstable fit. Second, in many applications, such as the PBC study, there are only a few measurements taken on each function. Additionally, individuals may be measured at different sets of time points and/or have differing numbers of observations. For such data, it is not possible to create finite dimensional predictors by simple discretization and so (2) can not be directly applied. A more successful approach is to replace the summation $\mathbf{X}^T\beta_k$ with its functional analog, the integral

$$Z_{ik} = \int X_i(t)\beta_k(t)dt \tag{3}$$

where $\beta_k(t)$ is a coefficient function giving the weighting placed on $X(t)$ at each time. This method has a couple of advantages over the more ad hoc discretization approach. First, through the use of a smooth function to estimate $\beta(t)$, it properly utilizes the inherent correlation between nearby time points, effectively reducing the high dimensional nature of the data. Second, by utilizing smoothing techniques the integral can be calculated even on sparsely sampled predictors where the discretization approach would fail.

Combining (2) and (3) gives the FAME link

$$g(\mu_i) = \beta_0 + \sum_{k=1}^{r} f_k(Z_{ik}) = \beta_0 + \sum_{k=1}^{r} f_k\left(\int X_i(t)\beta_k(t)dt\right). \tag{4}$$

Equation 4 extends standard projection pursuit regression in two directions by introducing a link function to allow for non-Gaussian responses and replacing the summation $\mathbf{X}^T\beta_k$ with an integral over $X(t)\beta_k(t)$ to allow for functional predictors. Formally the FAME model can be written as

$$p(y_i;\theta_i,\phi) = \exp\left(\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i,\phi)\right), \tag{5}$$

$$g(\mu_i) = \beta_0 + \sum_{k=1}^{r} f_k(Z_{ik}), \tag{6}$$

$$Z_{ik} = \int X_i(t)\beta_k(t)dt, \quad i = 1,\ldots,N \tag{7}$$

where (5) is the response distribution, assumed to be a member of the exponential family with $\mu = E(Y_i|X_i)$ and the f_k's and β_k's are suitably smooth curves. The relationship between predictor and response is specified through the unobserved latent variables $Z_1,\ldots,Z_r$ which are linear functions of $X(t)$. Note that (5) and (6) are related through the standard exponential family identity $\mu = b'(\theta)$. As with standard PPR the FAME model can experience confounding of parameters. In particular, β_k and f_k are confounded because identical values of $f_k(Z_k)$ can be achieved by multiplying β_k by a constant and adjusting f_k accordingly. Hence we restrict

$$\int \beta_k(t)dt = 1 \quad k = 1,\ldots r. \tag{8}$$

Using (8) $\beta_k(t)$ can be interpreted as a weighting function on the predictor at any given time. In addition, for $r > 1$, f_k and f_j may be confounded. Thus we restrict $cor(Z_{ik},Z_{ij}) = 0$ for all $j \neq k$. Such a restriction should have the additional advantage of reducing collinearity between terms.

As specific examples of FAME we consider two of the most common situations. First, for a Gaussian

response with identity link the FAME model becomes

$$Y_i = \beta_0 + \sum_{k=1}^{r} f_k\left(\int X_i(t)\beta_k(t)dt\right) + \varepsilon_i, \quad \varepsilon_i \sim N(0,\sigma_y^2),$$

a functional analogue of projection pursuit regression. When the response is Bernoulli and a logistic link is used the FAME model reduces to

$$Y_i \sim Bern(p_i), \quad \log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \sum_{k=1}^{r} f_k\left(\int X_i(t)\beta_k(t)dt\right). \tag{9}$$

An alternative formulation of FAME can help facilitate interpretation. Consider the decomposition of the predictor function into a sum over its principal component curves,

$$X_i(t) = \bar{X}(t) + \sum_{m=1}^{\infty} \zeta_{im}\rho_m(t), \tag{10}$$

where $\rho_m(t)$ represents the mth principal component curve and ζ_{jm} the corresponding weighting for the ith individual. Principal component curves have similar interpretations to their finite dimensional counterparts with the mth component explaining the largest proportion of the variability in the predictors subject to being orthogonal to the first $m-1$ terms. Combining (7) and (10) we can reformulate Z_{ik} as

$$Z_{ik} = \alpha_k + \sum_{m=1}^{\infty} \zeta_{im}\beta^*_{km} \tag{11}$$

where $\alpha_k = \int \bar{X}(t)\beta_k(t)dt$ is the mean of Z_{ik} and $\beta^*_{km} = \int \beta_k(t)\rho_m(t)dt$. Using this parameterization, β^*_{km} gives the weight placed on the mth principal component curve in constructing Z_{ik}. For example, if $r = 1$ and $\beta^*_{km} = 0$ for all $m > 1$ then an individual's score on the first principal component would solely determine their value for Z_{ik} and hence μ_i. We explore these two different formulations of FAME further in Section 4.

2.1 FAME fitting procedure

In this section we present a fitting algorithm for FAME which is based on maximizing a penalized likelihood. In practice we only ever observe $X_i(t)$ at a finite set of time points so the predictors must be estimated using the observed values. Let $X_i(t) = \mathbf{B}(t)^T\gamma_i$,

$$\beta_k(t) = \mathbf{B}(t)^T\eta_k \quad \text{and} \quad f_k(t) = \mathbf{s}(t)^T\delta_k \tag{12}$$

where $\mathbf{B}(t)$ and $\mathbf{s}(t)$ are both orthogonal finite q-dimensional bases, chosen prior to fitting the model. We utilize cubic splines. If one assumes that the predictors have been measured without error, then the estimation can be achieved by interpolating the observations as nearly as possible, using $\mathbf{B}(t)$. In Section 2.2 we address the case in which the predictors are measured with error. For the FAME model given by (5)-(7) the log likelihood, up to additive constants, is

$$l(f_k,\beta_0,\beta_k,\phi) = \sum_{i=1}^{N}\left[\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i,\phi)\right] \tag{13}$$

subject to $g(\mu_i) = \beta_0 + \sum_{k=1}^{r} f_k(Z_{ik})$.

To initialize the FAME procedure we employ interpolating cubic splines with minimum integrated squared second derivative to estimate the X_i's. This is just one of many bases that could be used. An iterative approach is then used to maximize a penalized version of (13). We start by fitting the model with $r = 1$. At the first stage, β_0 and f_1 are held fixed and β_1 is estimated. The fit is achieved by maximizing (13) over η_1 subject to a penalty term $P(\beta)$ to ensure a smooth fit. There are several possible choices for $P(\beta)$. A common smoothness penalty involves using

$$P_1(\beta) = \lambda_\beta \int \beta_k''(t)^2 dt \tag{14}$$

which penalizes large second derivatives of β_k. However, in the original basis space the X_i's tend to vary little, if at all, in certain directions meaning that it is impossible to produce reasonable estimates of β_k in those dimensions. Hence it may be beneficial to penalize β_k away from these directions using

$$P_2(\beta) = \lambda_\beta \sum_{m=1}^{q} \int \left(\beta_k(t)\rho_m(t)/s_m\right)^2 dt \tag{15}$$

where ρ_m is the mth principal component function of X_i and s_m is the corresponding standard deviation of the principal component scores. $P_2(\beta)$ imposes a high penalty on β_k's that have significant variability in the directions of X_i with little variance. It is interesting to note that (15) has similarities to condition 1 in Cardot *et al.* (2003b). In this paper we explore both penalty approaches. The parameter λ_β can be selected using cross-validation and a standard non-linear optimization package used to maximize the penalized likelihood over η_k.

The second stage involves estimating β_0 and f_1 with all other parameters held constant. Notice however, that with β_1 fixed the Z_{i1}'s are also fixed and hence β_0 and f_1 can be estimated using any standard GAM package. The FAME procedure iterates through these two steps until the penalized likelihood converges. Then β_1, f_1 and the Z_{i1}'s are fixed and the process is repeated for the $r = 2$ model, producing estimates of β_2, f_2 and the Z_{i2}'s subject to zero correlation between Z_{i1} and Z_{i2}. This continues until r reaches the preset maximum value. This nested structure has the advantage that to reduce the number of components in the link function one simply eliminates the redundant values of f_k and β_k without needing to refit the model.

2.2 FAME with measurement error

In some circumstances it may be more reasonable to assume that the predictors, $X_i(t)$, have not been observed exactly. For example, one often has measurement error in medical experiments such as the PBC study. In this case if we denote the observed values by $X_i^{obs}(t)$ and the measurement error by $e_i(t)$ then

$$X_i^{obs}(t) = X_i(t) + e_i(t), \quad i = 1, \ldots, N. \tag{16}$$

We make the standard choice of modeling the error terms at each observed time as uncorrelated Gaussian random variables with variance σ_x^2. Hence if the ith individual is observed at times $t_{i1}, \ldots, t_{in_i}$ then from

(5)-(7) and (16) the log likelihood, up to additive constants, is

$$l(\sigma_x^2, f_k, \beta_0, \beta_k, \phi, X_i) = \sum_{i=1}^{N} \left[\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi) - \frac{1}{2} \sum_{l=1}^{n_i} \left[\log \sigma_x^2 + \frac{1}{\sigma_x^2} ||X_i^{obs}(t_{il}) - X_i(t_{il})||^2 \right] \right] \qquad (17)$$

subject to $g(\mu_i) = \beta_0 + \sum_{k=1}^{r} f_k(Z_{ik})$.

The FAME algorithm with measurement error in the predictors is fit in a similar manner to that outlined in Section 2.1 with the addition of one extra step in the iteration. Instead of initializing the procedure by fixing the X_i's one uses the current values of the other parameters as well as the observed measurements, $X_i^{obs}(t)$, and the responses, Y_i, to provide an updated estimate of the X_i's. It is an interesting feature of this problem that the responses provide additional information in the estimation of the X_i's. We again start with $r = 1$ but at each step use the current estimates of β_1, β_0 and f_1 along with the response to update the X_i's. The fit is obtained by maximizing (17) over the γ_{il}'s subject to the penalty term

$$\lambda_x \int X_i''(t)^2 dt \qquad (18)$$

which ensures smooth fits of the X_i's. To reduce computational burden λ_x is chosen prior to fitting the model using cross-validation on the predictors alone. The maximization of the penalized likelihood can be achieved relatively quickly using any standard non-linear optimization package because it is possible to calculate the derivatives analytically. An estimate of σ_x^2 is also produced using the maximum likelihood value

$$\sigma_x^2 = \frac{1}{\sum_{i=1}^{N} n_i} \sum_{i=1}^{N} \sum_{l=1}^{n_i} ||X_i^{obs}(t_{il}) - X_i(t_{il})||^2.$$

We then estimate β_0, f_1 and β_1 just as in the zero measurement error case and iterate until the penalized likelihood converges. At this point we fix β_1, f_1, σ_x^2, the X_i's and Z_{i1}'s and increase r by one. This process continues, with the X_i's now fixed, until the maximum value for r is reached.

3 Asymptotic Theory

In this section we derive asymptotic results for the FAME model under the assumptions of equation (12) i.e. that the β_k's and f_k's can be represented by finite dimensional bases and hence the FAME model is finite. Let $\xi^0 = (\beta_0, \eta_1, \ldots, \eta_r, \delta_1, \ldots, \delta_r)$ denote the true vector of parameters for the FAME model given by (5)-(7). For notational simplicity we will assume ϕ to be known. However, the theory can easily be extended to the case where ϕ is also estimated. We denote by $\hat{\xi}_N$ the corresponding estimators obtained from the penalized maximum likelihood fitting procedure. We show, under mild conditions, that $\hat{\xi}_N$ is a consistent estimator for ξ^0 and that $\sqrt{N}(\hat{\xi}_N - \xi^0)$ asymptotically has a Gaussian distribution. These results are then used to provide asymptotic confidence intervals for $\beta_k(t)$ and significance levels for f_k.

Let $l(\xi)$ be the likelihood function for the FAME model and

$$I_N = -E_{\xi^0}(l''(\xi^0)) = -E_{\xi^0}\left[\frac{\partial l^2}{\partial \xi \partial \xi} \right]$$

be the corresponding information matrix. In order to prove our results we make the following assumptions.

A-1 There exist functions M_i such that

$$\left| \frac{\partial^3}{\partial\xi_j\partial\xi_k\partial\xi_l} \left(\frac{Y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i,\phi) \right) \right| \leq M_i(Y_i)$$

where $P_{\xi^0}(\frac{1}{N}\sum_i M_i(Y_i) < m_1) \to 1$ for some $m_1 < \infty$ and all ξ.

A-2 $\lim_{N\to\infty} \frac{1}{N} E_{\xi^0}(-l''(\xi^0)) = \lim_{N\to\infty} I_N/N = \bar{I}$ where $\bar{I}$ is a positive definite matrix with finite components.

A-3 $\lim_{N\to\infty} \frac{1}{N^2} Var_{\xi^0}(l''(\xi^0)_{jk}) = 0$ for all j and k.

A-4 There exists an $\varepsilon > 0$ and $m_2 < \infty$ such that

$$E_{\xi^0}\left[\left| \frac{(Y_i - \mu_i)}{g'(\mu_i)Var(y_i)} \sum_{k=1}^{r} \frac{\partial f_k(Z_{ik})}{\partial\xi_j} \right|^{2+\varepsilon} \right] \leq m_2$$

for all i and j.

(A-1) and (A-3) place bounds on the third derivative and variance of the second derivative of the likelihood functions. For all common members of the exponential family, they will hold under very general conditions on f_k and $X_i(t)$. (A-2) requires that the information provided by Y_i and $X_i(t)$ approaches infinity. This assumption is standard in any linear models framework such as GLM. If, for example, the predictors converged to a constant, I_N/N would approach zero and a consistent estimator would not exist. Finally, (A-4) is required to ensure asymptotic normality of the estimators. Again for all standard members of the exponential family (A-4) will hold under general conditions on f_k and f_k'. Utilizing these assumptions Theorems 1 and 2 prove asymptotic consistency and normality of the solutions of the FAME likelihood equations.

Theorem 1 *Assuming (A-1) through (A-3) and (12) hold, asymptotically a sequence* $\{\hat{\xi}_N\}$ *of solutions of the FAME likelihood equations exists and is consistent for estimating* ξ^0.

Theorem 2 *Let* $\hat{\xi}_N$ *be a consistent solution of the FAME likelihood equations. Then assuming (A-1) through (A-4) and (12) hold,*

$$\sqrt{N}(\hat{\xi}_N - \xi^0) \Rightarrow N(0, \bar{I}^{-1}).$$

The proofs of these results utilize standard methods from maximum likelihood theory with the added complication that the observations are not identically distributed. In practice $\bar{I}$ will be approximated by

$I_N/N = \frac{1}{N}\left[\sum_{i=1}^{N} \frac{I_i^*}{Var(y_i)g'(\mu_i)^2} + D_\Omega\right]$ where

$$I_i^* = \begin{bmatrix} 1 & f_1'(Z_{i1})\gamma_i^{*T} & f_2'(Z_{i2})\gamma_i^{*T} & \cdots & \mathbf{s}^T(Z_{i1}) & \mathbf{s}^T(Z_{i2}) & \cdots \\ f_1'(Z_{i1})\gamma_i^* & f_1'(Z_{i1})^2\gamma_i^*\gamma_i^{*T} & f_1'(Z_{i1})f_2'(Z_{i2})\gamma_i^*\gamma_i^{*T} & \cdots & f_1'(Z_{i1})\gamma_i^*\mathbf{s}^T(Z_{i1}) & f_1'(Z_{i1})\gamma_i^*\mathbf{s}^T(Z_{i2}) & \cdots \\ f_2'(Z_{i1})\gamma_i^* & f_1'(Z_{i1})f_2'(Z_{i2})\gamma_i^*\gamma_i^{*T} & f_2'(Z_{i2})^2\gamma_i^*\gamma_i^{*T} & \cdots & f_2'(Z_{i2})\gamma_i^*\mathbf{s}^T(Z_{i1}) & f_2'(Z_{i2})\gamma_i^*\mathbf{s}^T(Z_{i2}) & \cdots \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots \\ \mathbf{s}(Z_{i1}) & f_1'(Z_{i1})\mathbf{s}(Z_{i1})\gamma_i^{*T} & f_2'(Z_{i2})\mathbf{s}(Z_{i1})\gamma_i^{*T} & \cdots & \mathbf{s}(Z_{i1})\mathbf{s}^T(Z_{i1}) & \mathbf{s}(Z_{i1})\mathbf{s}^T(Z_{i2}) & \cdots \\ \mathbf{s}(Z_{i2}) & f_1'(Z_{i1})\mathbf{s}(Z_{i2})\gamma_i^{*T} & f_2'(Z_{i2})\mathbf{s}(Z_{i2})\gamma_i^{*T} & \cdots & \mathbf{s}(Z_{i2})\mathbf{s}^T(Z_{i1}) & \mathbf{s}(Z_{i2})\mathbf{s}^T(Z_{i2}) & \cdots \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots \end{bmatrix} \tag{19}$$

and $\gamma_{ij}^* = \gamma_{ij} - \frac{\int B_j(t)dt}{\int B_q(t)dt}\gamma_{iq}$ for $1 \le j \le q-1$. D_Ω is a block diagonal matrix corresponding to the penalty terms on $\beta_k(t)$ and $f_k(t)$. For example when using $P_1(\beta)$, the penalty on η_k is $\Omega_{\eta_k} = \lambda_\beta \int \mathbf{B}''(t)\mathbf{B}''^T(t)dt$.

Theorem 2 suggests approaches for calculating pointwise confidence intervals on $\beta_k(t)$ and significance levels on f_k. We summarize these results in Corollaries 1 and 2.

Corollary 1 *Let $\hat{\beta}_k(t) = \mathbf{B}^T(t)\hat{\eta}_k$. Then under the assumptions given in Theorem 2, for any fixed t,*

$$P\left\{\hat{\beta}(t) - \Phi^{-1}_{1-\alpha/2}\sqrt{\mathbf{B}^T(t)\Sigma_{\hat{\eta}_k}\mathbf{B}(t)/N} \le \beta(t) \le \hat{\beta}(t) + \Phi^{-1}_{1-\alpha/2}\sqrt{\mathbf{B}^T(t)\Sigma_{\hat{\eta}_k}\mathbf{B}(t)/N}\right\} \to 1-\alpha$$

where $\Sigma_{\hat{\eta}_k}$ is equal to the block diagonal component of $\bar{I}^{-1}$ corresponding to η_k and Φ is the standard normal cdf.

Corollary 2 *Let $\Sigma_{\hat{\delta}_k}$ be the block diagonal component of $\bar{I}^{-1}_{(-\eta_k)}$ corresponding to δ_k where $\bar{I}^{-1}_{(-\eta_k)}$ is equal to $\bar{I}$ with all elements involving η_k removed. Then under (A-1), (A-3), (A-4), (12) and the null hypothesis of no relationship between Y and $X(t)$*

$$X_1^2 = N\hat{\delta}_1^T\Sigma^{-1}_{\hat{\delta}_1}\hat{\delta}_1 \Rightarrow \chi_q^2. \tag{20}$$

Under the null hypothesis that there are exactly r terms in the model

$$X_{r+1}^2 = \hat{\delta}_{r+1}^T\Sigma^{-1}_{\hat{\delta}_{r+1}}\hat{\delta}_{r+1} \Rightarrow \chi_q^2. \tag{21}$$

Notice that under the null hypothesis of no relationship between response and predictor $f_k = 0$ so that the information matrix given by (19) is non-singular which is a violation of (A-2). In fact it is easily seen that there is no consistent estimator of $\beta(t)$ in this case. However, a small modification of the proof of Theorem 2 shows that $\hat{\delta}_1$ will still converge to a normal distribution with the information matrix given by the terms in (19) that correspond to β_0 and δ_1 i.e.

$$I_i^* = \begin{bmatrix} 1 & \mathbf{s}^T(Z_{i1}) \\ \mathbf{s}(Z_{i1}) & \mathbf{s}(Z_{i1})\mathbf{s}^T(Z_{i1}) \end{bmatrix}.$$

The ability to remove the terms involving η_1 from the information matrix can significantly increase the power of the test. Corollary 2 suggests an iterative approach for choosing r. First fit FAME with $r = 1$ and calculate the significance of the first term using (20). Then proceed stepwise adding additional terms and testing significance using (21) until the $r+1$st term fails the test.

4 Applications

In this section we illustrate the FAME procedure on both a simulated data set and on the PBC data described in Section 1.

4.1 Simulated data

Figure 1 shows the key components of the FAME fit for a simulated data set. The model was run with $r = 1$, so we drop the subscript k in our discussion. To produce the simulated data we first generated β, f and 100 X_i's. These curves were all produced using third order polynomials with random Gaussian coefficients. The observed values of the predictors were obtained by sampling each X_i at 50 random time points and adding Gaussian noise. Finally, the responses were generated from a Gaussian distribution with mean $f(Z)$ where Z was given by (7). Since the predictors were sampled with uncertainty the FAME procedure with measurement error was employed. We used 15-dimensional cubic b-splines as the basis for β and the X_i's and $P_2(\beta)$ as the penalty term on β. Both λ_β and λ_x were chosen by cross-validation. The function f was fit using the GAM package in R. Figure 1(a) gives the true β curve, its estimate and 95% pointwise confidence intervals produced from Corollary 1. For this simulation we found that the best results for β were obtained using the penalty term $P_2(\beta)$ although a fairly similar fit was produced using $P_1(\beta)$. Figure 1(b) gives the observed responses and the mean response function together with its FAME fit. Notice that even though the responses have considerable noise and the Z_i's that generate the mean function are never observed it is possible to accurately recover all the components of the data. The β curve shows that individuals with low predictors at early and late times will have high Z scores and vice versa. The mean curve indicates that subjects with low Z scores will have high responses and vice versa. Thus individuals with high values of X_i at the early and late time periods will have high responses. Values of X_i in the middle time periods have comparatively less influence on Z_i and hence on the response. The fact that the response surface is clearly non-linear indicates that a standard functional GLM (James, 2002), which assumes linearity and a fixed link g, would not be adequate for this data. Another possible approach here would be to use a more advanced functional GLM approach which also estimates the link. Such a method could be expected to give similar results to FAME with r restricted to one.

In this example the first four principal component curves explain almost 100% of the variability in the X_i's. Hence from (11) we see that

$$Z_i \approx \alpha + \sum_{m=1}^{4} \zeta_{im}\beta^*_m \tag{22}$$

where ζ_{im} is the loading for the mth principal component on the ith predictor, X_i. Equation 22 provides an alternative method of presenting the FAME results, which is often more illuminating about the relationship between X_i and Z_i than the raw β curve. For this data $\beta^{*T} = (-20.296, -0.386, -0.006, 0.015)$ so in calculating Z_i almost all the weight is placed on the first component loading. Figures 1(c) and (d) provide plots of the first two principal component curves. The $+$ and $-$ curves correspond respectively to the mean function plus or minus three times the principal component. Hence an individual whose predictor curve looks like the $+$ curve for PC 1 would have $\zeta^T = (3,0,0,0)$ and their Z would be 20.296×3 below the average. The effect of this value of Z on the response can in turn be seen in Figure 1(b). Similar observations can be made for the other principal component curves. However, since the other components have much smaller β^* coefficients they have a comparatively low effect on the Z score and hence the response. A formulation such

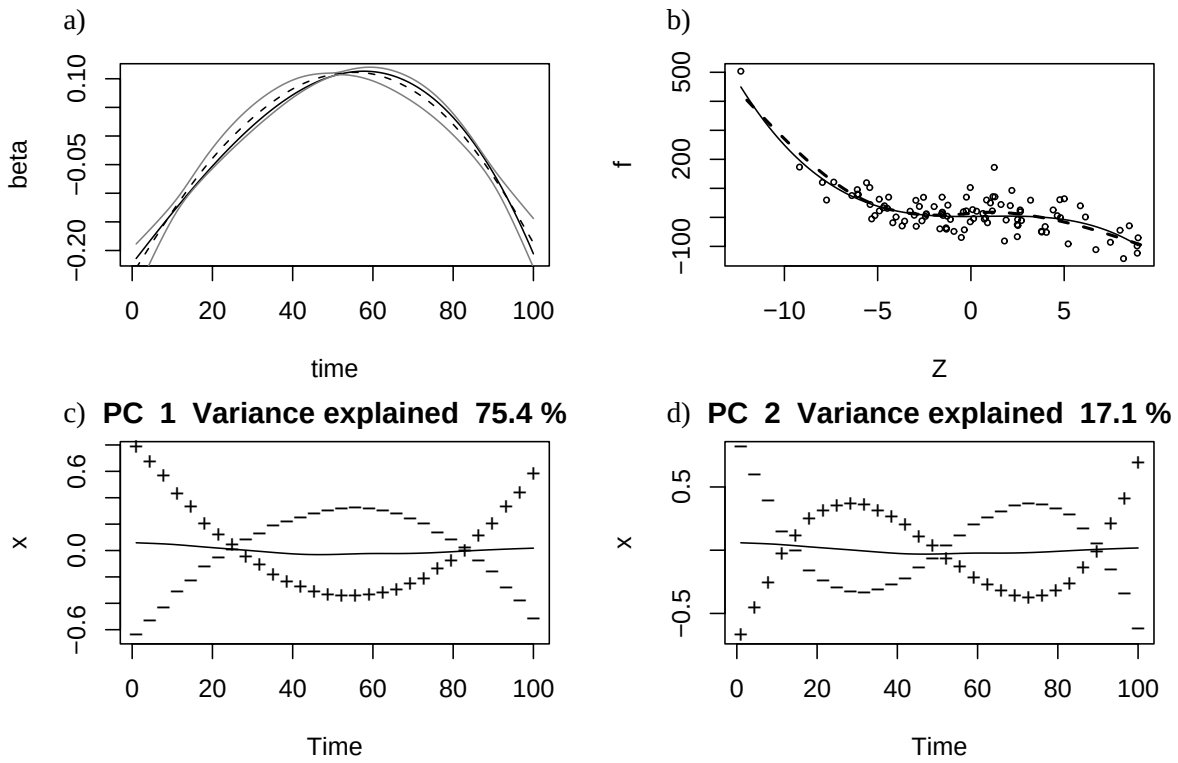

Figure 1: *(a) and (b) Results from the FAME fit to a simulated data set. Solid black curves indicate the truth, dashed lines give estimates and grey lines represent* 95% *confidence intervals. (c) and (d) The first two principal component curves of the predictors on the simulated data set.*

as (22) allows one to easily asses the types of variation in the predictors that have the greatest impact on the response. Notice that in this example β had a similar shape to the first principal component curve. This helps explain the superior performance of $P_2(\beta)$ which shrinks towards the dominant directions of variability.

4.2 PBC data

In the PBC data set the response is a binary variable indicating whether or not the patient survived five years. For now we use bilirubin levels as the sole predictor. In Section 5 we will extend the analysis to include multiple predictors. To ensure that no bias is introduced by the fact that patients with more observations are likely to have survived longer we choose to predict five year survival based only on measurements of bilirubin level up to 800 days from registration and use only patients with at least four observations. The average measurement time of these observations was essentially the same for patients that lived and those that did not. For each individual the number of observations varied from 1 to 16. After removing all those patients with fewer than four observations 169 remained of whom 147 lived at least five years and 19 died before five years. The remaining three patients lived at least four years but had no measurements after the five year point. To avoid censoring complications it was assumed that these patients lived for the full five year period. An alternative approach would be to remove them from the analysis but with such a small data set this was considered undesirable.

We applied the FAME model with Bernoulli responses given by (9). Since the predictors contained

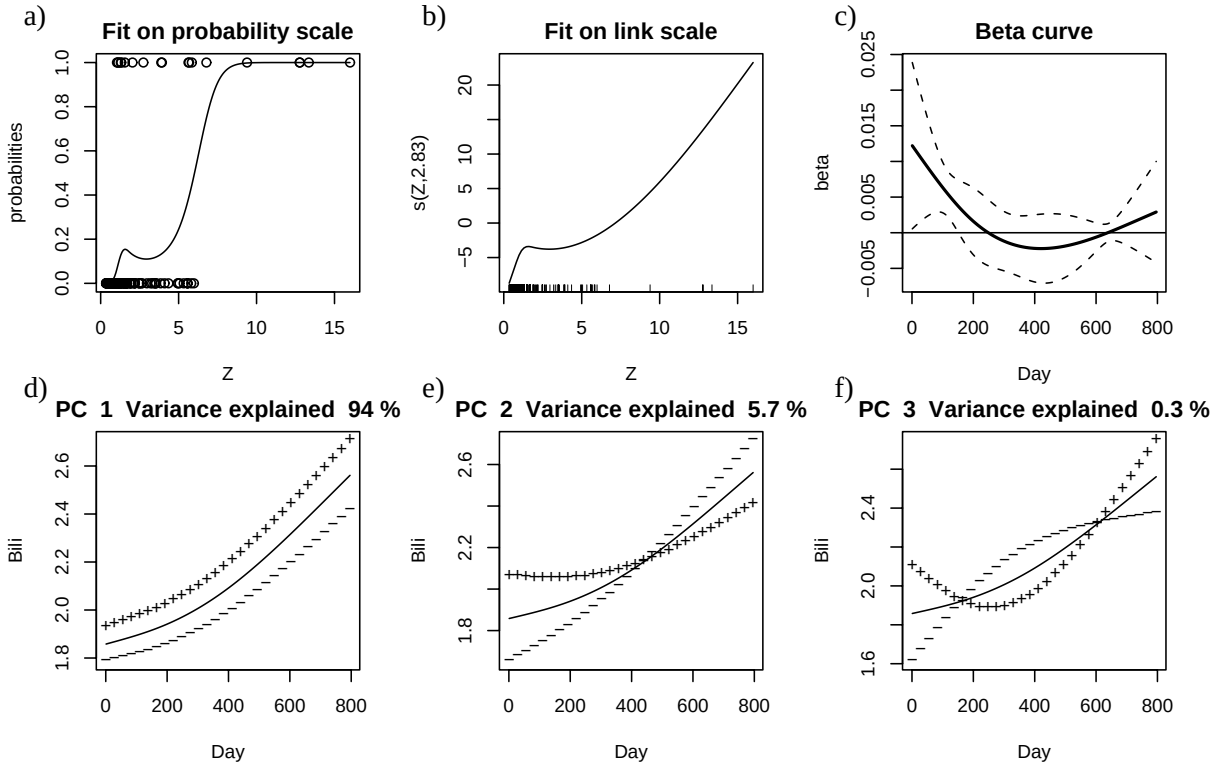

Figure 2: *Results from the FAME fit with $r = 1$ to the PBC data.*

a large degree of noise the most reasonable results were obtained using FAME with measurement error. Again, we utilized 15-dimensional cubic b-spline bases. Almost identical fits were produced using the penalties $P_1(\beta)$ and $P_2(\beta)$. The results from our first fit with r set to 1 are presented in Figure 2. Figure 2(a) gives the estimated probability of an individual failing to live five years based on their Z score. Notice that unlike a standard GLM fit in which the curve is restricted to be S-shaped the probability shows a sharp rise at the beginning and then levels off before increasing to 1. The reason for this shape is seen in Figure 2(b) which gives a plot of f_1 on the logit scale. With a standard GLM fit, with a fixed link g, this curve would be linear but here we notice an approximately piecewise linear shape with a steep slope for low values of Z followed by a flat segment and finally a positive slope for higher values of Z. The estimated degrees of freedom (dof) associated with this term was 2.8, significantly larger than the 1 dof of a linear fit. The difference in deviance between this non-linear fit and the corresponding linear one was highly significant, providing strong evidence of a non-linear relationship. The β curve, which places the highest weights on early time periods and little weight on later ones, is given in Figure 2(c). The dashed lines are approximate 95% pointwise confidence intervals calculated using Corollary 1. They show that only the early time periods have weights significantly different from zero. Thus the Z's can be interpreted as a measure of a patient's level of bilirubin early in the study. The first three principal component curves, which account for almost all the variability in the predictors, are provided in the remaining plots. The corresponding values for $\hat{\beta}_1^*$ were $(0.025, 0.065, 0.084)$ with all other coefficients negligible. At first glance this would imply that the third component is the most important in calculating Z_i. However, utilizing Theorem 2, the corresponding standard errors are $(0.001, 0.008, 0.051)$ so in fact only the first two components are significantly different

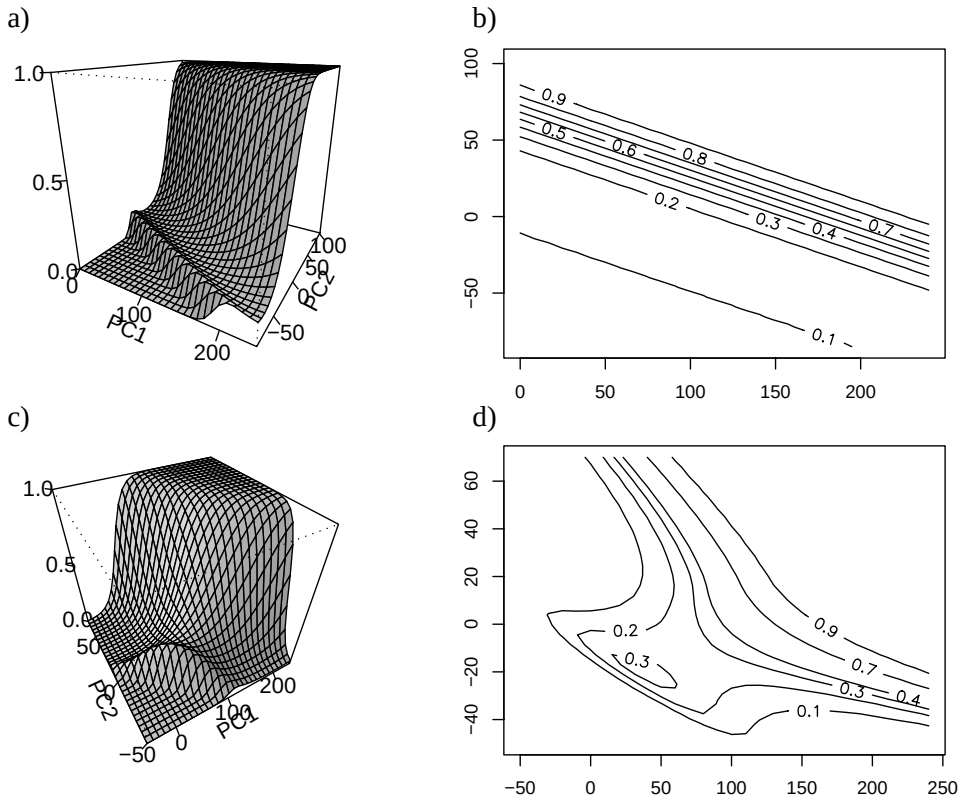

Figure 3: *Estimated probability of failing to survive five years based on a patient's score on the first two principal component directions. (a) and (b) respectively correspond to three dimensional and contour plots on a FAME fit with* $r = 1$ *while (c) and (d) represent the* $r = 2$ *fit.*

from zero. The first component corresponds to high values over all time periods while the second relates to high values at the early times. Since β_1^* and β_2^* are both positive the alternative formulation also suggests that values of bilirubin in the early time periods are most important in determining a patient's value for Z_i. These results are clinically sensible because liver failure is generally associated with high bilirubin levels. Since Z_i, and hence probability of failing to survive five years, appear to be primarily functions of the first two principal components we can plot these probabilities versus a patient's score on PCs one and two. Figures 2(a) and (b) provide these plots. Notice that for a fixed level of PC 1, i.e. overall average bilirubin, probabilities increase with PC 2 which suggests that high levels at early times are more dangerous than later times. Finally, using Corollary 2 we tested for a significant relationship between bilirubin levels over time and five year survival. The test produced a chi square statistic of 12.6 and an associated p-value of 0.013 providing strong evidence that bilirubin levels are a significant predictor.

Next, we fit FAME with $r = 2$. Again we plot probability of failing to survive as a function of PCs 1 and 2 in Figures 2(c) and (d). With $r = 2$ a more flexible surface is produced. With this fit PC 2 is not important for patient's with very high or low average levels of bilirubin. However, for patient's with moderate overall bilirubin levels it appears that those with very high or low values of PC 2 have the greatest probability of failing to survive. Since PC 2 forms a contrast between early and late times this suggests that given two patients with the same average bilirubin level the one with a stable pattern over time will have a higher

survival chance than one that has very high levels at some times and lower levels at others. In other words, variability in bilirubin is a risk factor in addition to the overall average level. The second term was not statistically significant but its reasonable clinical interpretation suggests that it may be important.

5 Extensions : Multivariate data

The FAME model presented in Section 2 and illustrated in Section 4 was for a single predictor. However, extending the model to multiple functional and finite dimensional covariates is straightforward. Suppose that for each individual we observe measurements from the predictor functions $X_{i1},\ldots,X_{ip}$ and a vector of standard covariates $\omega_i^T = (\omega_{i(p+1)},\ldots,\omega_{i(p+s)})$. The FAME model can be augmented in one of two ways. The first approach, most directly analogous to PPR, uses the same link function as standard FAME, (6), but replaces (7) by

$$Z_{ik} = \sum_{j=1}^{p} \int X_{ij}(t)\beta_{kj}(t)dt + \omega_i^T \beta_{k\omega}. \tag{23}$$

Equation 23 models Z_{ik} as a linear combination of all the predictors. In all other respects the FAME model remains identical. The second approach, more closely akin to GAM, fits a separate smooth function for each predictor. In this case (6) becomes

$$g(\mu_i) = \beta_0 + \sum_{j=1}^{p} f_j(Z_{ij}) + \sum_{j=p+1}^{p+s} f_j(\omega_{ij}) \tag{24}$$

where $Z_{ij} = \int X_{ij}(t)\beta_j(t)dt$. The first approach includes the second as a special case and has the advantage of providing a more flexible fit. However, it becomes very difficult to separate out the individual effects of each predictor using (23) while this is still possible with (24). Hence, as a general rule one should utilize the first approach when the ultimate goal is prediction of the response and the second if inference about the individual predictors is desired. We illustrate the two techniques on both the femur bone and PBC data sets.

5.1 Femur bone data

The femur bone data consists of two-dimensional functions describing cross-sectional images of bones from 96 individuals. The data were preprocessed to produce a matrix of 50 two-dimensional points, equally spaced by arc length, giving the outline of a specific section of each individual's femur bone. An image for a typical subject is provided in Figure 4(a). Full details of the preprocessing are given in Ramsay and Silverman (2002). By indexing the observations from 1 to 50 moving in a clockwise direction we can plot the data using two curves for each individual, one each for the x and y directions. Figures 4(b) and (c) show the x and y curves that correspond to Figure 4(a). For each person the data also include an indicator of arthritic bone change. We wish to use the x and y curves as predictors of arthritis. Since the curves here are really just two dimensions of a single function we are not primarily interested in the individual effect of x and y so it is natural to apply the multivariate version of FAME using (23).

Unlike the PBC study, in which we had only a few noisy observations of each curve, for this data we essentially have measurements of the entire function with no noise. Thus we fit the no measurement error version of FAME using a variety of values for r, 15-dimensional cubic b-spline bases and both $P_1(\beta)$ and $P_2(\beta)$. The fit using $P_2(\beta)$ and $r = 1$ is given in Figure 4. The β curves, along with 95% confidence intervals,

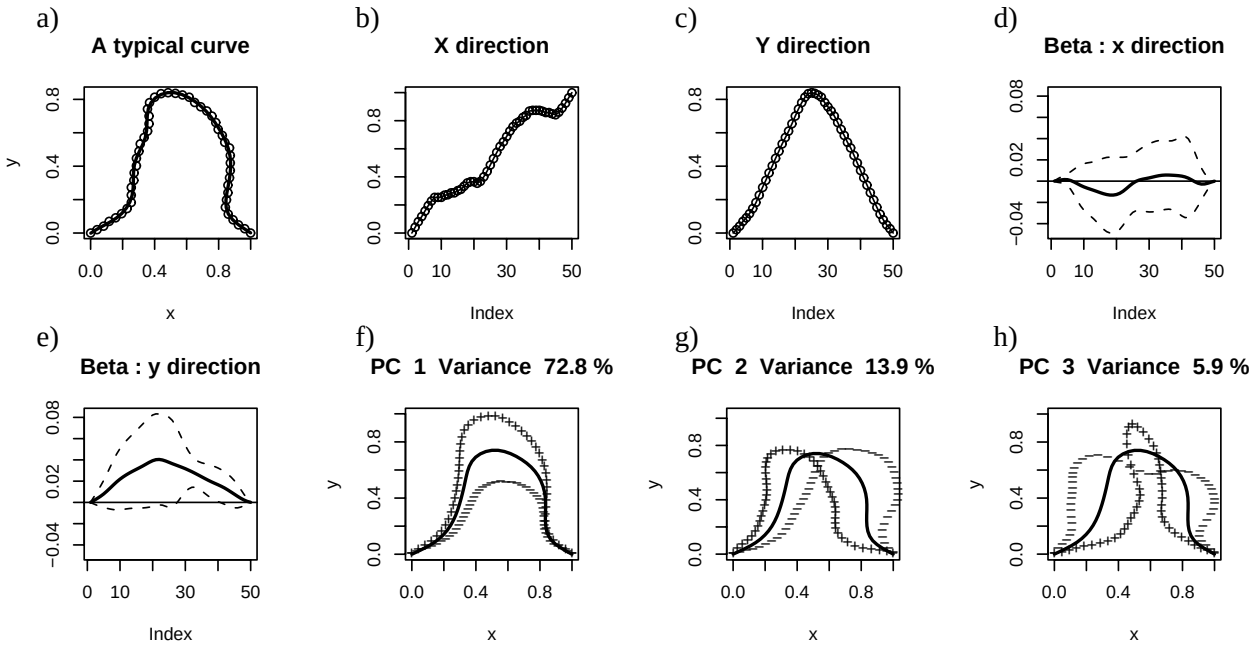

Figure 4: *Plots for the femur bone data showing a typical curve (a)-(c),* β *curves (solid) with confidence intervals (dashed) for the FAME fit (d) and (e) and the first three principal component curves (f)-(h).*

for the *x* and *y* directions are provided in Figures 4(d) and (e) respectively. The confidence intervals suggest no clear trend in the *x* direction and positive, but decreasing, weight on the second half of observations in the *y* direction. Unfortunately, the two-dimensional nature of the data makes the β curves more difficult to interpret. However, as in the one-dimensional case, one can analyze the data by decomposing the predictor functions into their first few principal component curves and examining the corresponding values of β^*. The first three principal component curves are given in Figures 4(f) through (h) and β^* is provided in Table 2. The first component accounts for a high proportion of all variability and primarily corresponds to variation in the *y* direction. The next two components relate more strongly to variability in the *x* direction. By examining β^* it is clear that most of the weight in calculating Z_i is placed on the first component indicating that this type of variation in the *y* direction is the most important predictor of arthritis. In fact, judging from the confidence intervals for β^* provided in parentheses the first component is the only significant term. The variability in the other components explains the wide confidence intervals in the first half of Figure 4(d). The last three columns of Table 2 provide the estimated degrees of freedom as well as significance values for the smooth term, $f_1(Z_i)$. The edf indicates that a linear fit is optimal and the p-value suggests that the bone images are highly significant predictors of arthritis. The slope of the smooth term was negative which, when combined with the positive value of β_1^*, indicates that individuals with shrunken bones in the *y* direction are at greater risk for developing arthritis. We next repeated the procedure with $r = 2$. The second term placed most of its weight on the second principal component but was not statistically significant.

5.2 PBC with multiple predictors

In addition to measurements of bilirubin levels, the PBC data set also contains observations of albumin levels and an indicator for the drug D-penicillamine. One of the original aims of the study was to test

	β^*			f_1		
	PC 1	PC 2	PC 3	EDF	χ^2	p-value
Term 1	0.181	0.005	-0.032	1	7.5	0.0062
	$(0.024, 0.338)$	$(-0.145, 0.155)$	$(-0.152, 0.088)$			

Table 2: *A table for the femur bone data. The first three columns contain the weights of* β *on the corresponding principal components of the predictors. Confidence intervals are provided in parentheses. The remaining columns give the estimated degrees of freedom and significance values for* f_1.

	β^*			Smooth terms		
	PC 1	PC 2	PC 3	EDF	χ^2	p-value
Bilirubin	0.035	0.007	0.001	1.33	13.7	0.008
	$(0.025, 0.045)$	$(-0.047, 0.061)$	$(-0.245, 0.247)$			
Albumin	0.034	0.010	-0.004	1	3.5	0.0614
	$(0.020, 0.048)$	$(-0.064, 0.084)$	$(-0.154, 0.146)$			
Drug	$\beta_{\text{drug}} = 0.438(-0.855, 1.731)$			1	0.5	0.480

Table 3: *A table for the PBC data using bilirubin, albumin and drug as predictors.* 95% *Confidence intervals for all* β *parameters are given in parentheses.*

the effectiveness of this treatment. Since we are potentially interested in the individual effects of all three predictors on five year survival rate we fit the multivariate FAME procedure using (24) and the smoothing parameters from the original analysis of Section 4.2. The results are shown in Table 3. For both bilirubin and albumin, almost all the weight is placed on the first principal component curve in calculating Z. The first principal component curve for bilirubin is shown in Figure 2(d) while the corresponding curve for albumin is shown in Figure 5(a). In both cases the first PC curve indicates a fairly constant positive difference from the mean curve over time. Bilirubin is found to be a highly significant predictor with a non-linear fit while albumin is only significant at the 10% level and provides no evidence of non-linearity. The D-penicillamine drug actually has a positive coefficient indicating lower survival rates for those on the medication but the result is not statistically significant. Since both Z_{Bili} and Z_{Alb} are primarily functions of their respective first principal component curves we can plot the probability of a patient failing to survive five years based on their scores on these two components. Figure 5(b) provides such a plot. Note that the probability of failing to survive five years rises as bilirubin levels increase and as albumin levels fall. A healthy liver secretes higher levels of albumin so these are clinically reasonable results.

6 Simulation study

6.1 FAME predictive performance

This section provides the results from a simulation study designed to test the performance of FAME in comparison with other approaches. We compared six different procedures on four test distributions. The first two methods were FAME with penalties $P_1(\beta)$ and $P_2(\beta)$ and 15-dimensional cubic b-spline bases. The third procedure, Functional Generalized Linear Models "FGLM" (James, 2002) provides a GLM fit to functional predictors. It is essentially identical to FAME with $r = 1$, f_1 restricted to be linear and the link g taken to be fixed. With the fourth approach, "S. Spline", we fit a cubic smoothing spline to each

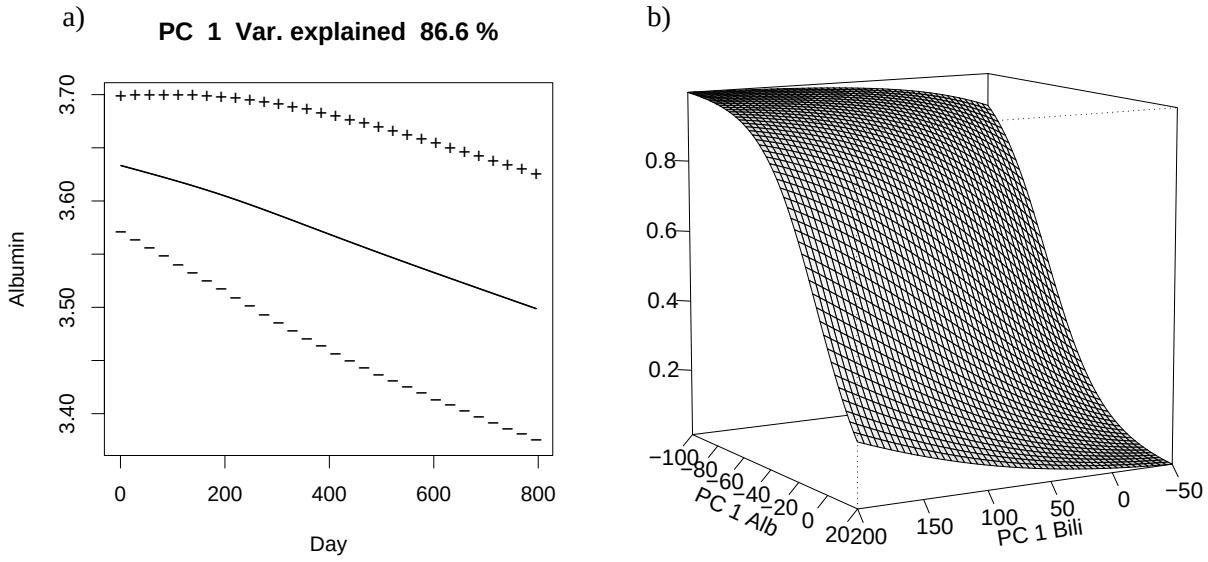

Figure 5: *The first principal component curve for albumin (a) and a plot of the probability of failing to survive five years as a function of Z_{Bili} and Z_{Alb} (b).*

individual predictor curve, produced estimates of the curve at each of ten equally spaced time points and used these ten observations as predictors in a standard linear regression. For the fifth method, "All points", the original measurements for each curve were sorted by time of observation and then used as predictors in a linear regression. This approach is only feasible if all the subjects have the same number of observations and may perform poorly if individual curves are observed at very different time points. The final procedure, "Average", just involved taking the mean of the existing observations for each curve and using this value as the predictor in a simple linear regression.

For each of the four simulations a test data set of 1000 observations was drawn from a given distribution. Each observation consisted of measurements along a predictor curve and a corresponding scalar response. In addition 100 training data sets were produced from the same distribution and fit with each of the six procedures. The goal was to use the training data and the predictors from the test data to provide as accurate predictions as possible for the 1000 test responses. The results for all four simulations are shown in Table 4 with standard errors over the 100 training data sets in parentheses. All results are shown as a percentage of the mean squared error produced by simply using the average of the training responses to predict the test responses. For example, on the first simulation the predictions from FAME produced mean squared deviations from the actual test responses that were only 3% of those obtained using the average of the training data. Taking the complement of this number gives the percentage of test sample error explained by using the predictor curves and is analogous to R^2. For instance FAME explained 97% of all the variability in the test responses in simulation 1.

The first simulation, intended to illustrate a situation where many simple approaches may work, involved producing responses that were a linear functional of the predictor curves. For each observation two predictor curves, X_{i1} and X_{i2}, were produced and each curve was sampled at ten random time points without measurement error. The curves were generated using cubic functions with randomly chosen Gaussian coefficients. Each response was then produced by taking a linear combination of the coefficients for each of the

Method	Simulation			
	1	2	3	4
FAME $P_1(\beta)$	3(0.1)	29(0.8)	38(1.5)	33(1.3)
FAME $P_2(\beta)$	3(0.1)	26(0.5)	33(0.4)	29(0.6)
FGLM	3(0.1)	60(0.3)	62(0.6)	63(0.6)
S. Spline	4(0.2)	66(0.5)	73(1.4)	67(1.0)
All points	60(1.6)	110(1.8)	137(4.3)	113(2.6)
Average	106(0.7)	101(0.3)	103(0.6)	102(0.4)

Table 4: *Results from the four simulation studies with standard errors in parentheses. Results are shown as a percentage of the mean squared error produced by simply using the average of the training responses to predict the test responses.*

two predictor curves and adding a small amount of random noise. A total of 50 observations were produced for each training data set. Since the data had little measurement error and did not involve any non-linear transformation of the predictors one would expect FAME to lose much of its advantage over other methods. Table 4 shows that FAME, using either type of penalty function, produced considerably improved results over those from using the training response mean. As one would expect given the linearity of the data, FGLM produced almost identical results. In problems involving linear data with more measurement error one may even expect FGLM to slightly outperform FAME because it should produce less variable results. The smoothing spline fit, S. Spline, gave similar, though slightly inferior results. The other two methods, All points and Average, both resulted in far inferior fits with the latter actually producing worse results than simply using the mean of the training response.

The second simulation used the distribution of the data from Section 4.1 with $\sigma_x = 0.1$ and $\sigma_y = 50$. These data had a non-linear relationship between the predictors and response, a situation where FAME might be expected to provide significant improvements over other approaches. In fact both FAME methods produced considerably superior results over the other, linear, methods. FGLM was the best of the linear approaches but still produced error rates approximately twice those of FAME. More complicated versions of FGLM exist in which the link function is also estimated. It is likely such an approach would produce results more similar to FAME with $r = 1$. S. Spline gave similar fits to FGLM but the other two methods failed completely. In this example the penalty term $P_2(\beta)$ gave slightly better results but it is possible that further fine tuning of the smoothing parameter λ_β would reduce the difference in performance.

The final two simulations were designed to test the robustness of FAME to violations of the Gaussian error assumptions for both the predictors and response. In Simulation 3 we replicated the data from the previous simulation but used noise from a t-distribution with three degrees of freedom, appropriately scaled to maintain the original standard deviations. This change in the error distribution produced only a minor deterioration in the performance of FAME. While a t-distribution is heavier tailed than the Gaussian it still has a symmetrical bell shape. For the final simulation we utilized errors from the exponential distribution, standardized to have mean zero and the correct standard deviations. Again there were only minor deteriorations noted in the FAME fit suggesting that the procedure is fairly robust to violations of the model assumptions.

6.2 Coverage, significance and power

We also performed simulations to test the true pointwise coverage of the confidence intervals on $\beta(t)$ and the type one error probability and power of hypothesis tests for a relationship between Y and $X(t)$. The results of these simulations are summarized in Figure 6. Figure 6(a) gives the true coverage levels of 90%, 95% and 99% confidence intervals from FAME fits to 100 simulated data sets for various values of the smoothing parameter λ_β. The data was simulated from essentially the same distribution as the second simulation of the previous section with $\sigma_x = 0$, as is assumed for the asymptotic results, and the sample of $X(t)$'s fixed across data sets to maintain comparability. For all reasonable values of the smoothing parameter the coverage levels are generally very close to, and in some cases even above, those predicted. When the parameter is set too high the coverage is reduced but one would expect the effect of the smoothing term to diminish with larger sample sizes. We also performed simulations with measurement error in the predictors and non Gaussian error terms. We found little change in the coverage for small amounts of measurement error and only an average reduction of 1% in the coverage for the *t* and exponential error distributions explored in the previous section.

To test significance levels and power we produced 200 data sets with predictors generated in an identical fashion to the previous simulation but with categorical $(0,1)$ responses. The log odds for the *i*th response was modeled using $\beta_0 + \beta_1 Z_i$. For $\beta_1 = 0$ we estimated the probability of a type one error for a particular nominal significance level, α, by calculating the fraction of p-values less than α. We used an identical approach to calculate power for various values of $\beta_1 > 0$. The results are summarized in Figure 6(b). With $\beta_1 = 0$ the observed and nominal significance levels are all very close. For comparison we also calculated the type one error probabilities when using p-values from the final GAM fit which treats the latent Z_i's as fixed. These errors where much higher. For example with $\alpha = 0.1$ the type one error probability was actually 0.24. This illustrates the importance of incorporating the variability of the latent variables in the analysis. As β_1 increases from zero the power increases in an approximately logistic fashion. These results were with $\lambda_\beta = 200$. We found the observed significance levels reduced even further if less flexibility was allowed in $\beta(t)$ and were higher for more flexible fits. Finally, we tested the power for a fit with $r = 2$ using log odds equal to $\beta_0 + \beta_1 Z_1 + \beta_2 Z_2^2$ where Z_1 and Z_2 represented two different linear combinations of the predictors. The powers for detecting significant effects for the first term, f_1, and the second term, f_2, are shown respectively in Figures 6(c) and (d). Both figures are plotted as a function of β_1. The power levels for f_1 are all high. The power for f_2 with $\alpha = 0.01$ is relatively low while for $\alpha = 0.05$ and 0.1 the power is moderate and increasing with β_1. In general, power will decrease as *r* increases because more flexible fits are produced. As one might expect, it is only possible to detect multiple f_k's provided the sample size is relatively large or the signal is clear.

7 Discussion

In this paper we have suggested a general methodology for fitting a flexible class of models to data consisting of functional predictors and scalar responses. Figure 7, which provides a summary of methods for modeling predictor-response data, indicates the relationship between FAME and other standard approaches. The six procedures in the upper boxes can all be used on data sets with standard *p*-dimensional predictors. Models range from least to most flexible moving from left to right. The top row corresponds to methods assuming a Gaussian response. Linear regression provides the simplest approach. Additive models give extra flexibility

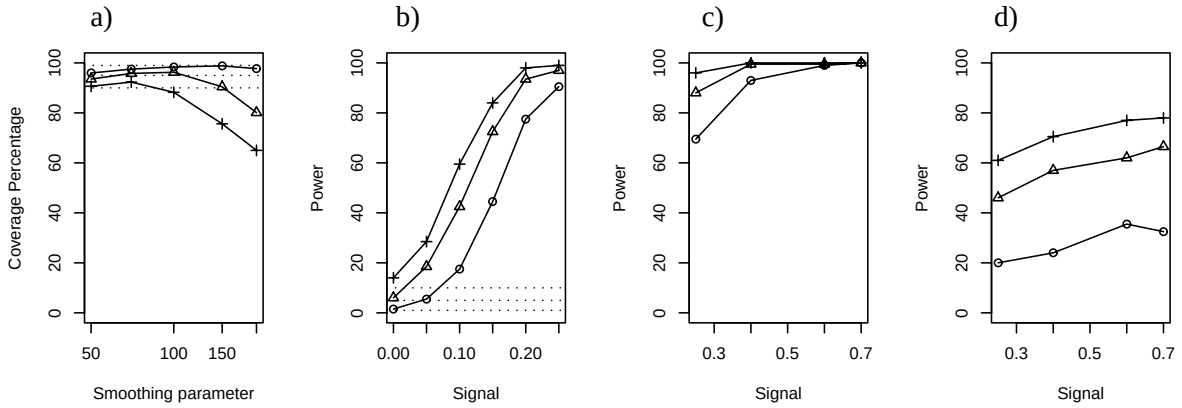

Figure 6: *(a) Coverage levels with pluses, triangles and circles respectively indicating* 90%, 95% *and* 99% *confidence intervals with dotted lines corresponding to theoretical coverage. (b) Significance levels and power for various values of* β_1 *with pluses, triangles and circles respectively corresponding to* $\alpha = 10\%, 5\%$ *and* 1% *and dotted lines indicating the correct significance levels. Note the* Z_i*'s ranged fairly uniformly between* -10 *and* 10 *so, for example, at* $\beta_1 = .1$ *the range of the log odds was* 2*. (c) Power for* f_1*. (d) Power for* f_2*.*

by permitting non-linear fits for each predictor. Finally, projection pursuit regression allows an almost unlimited range of possible relationships. Neural nets (Hastie *et al.*, 2001, Chapter 11) and boosting (Freund and Schapire, 1997) methods provide similar highly flexible fits and can be placed in the same category as PPR. The second row of Figure 7 gives extensions of these three methods to non-Gaussian responses through the use of a link between the mean of the response and the predictors.

All of the first six approaches require adaptation before they can be used for data with functional predictors. The bottom two rows of Figure 7 correspond to the functional extensions. Some of these methods have been previously explored but most have not. Functional linear regression is discussed in Ramsay and Silverman (1997) and functional GLM techniques are developed in Marx and Eilers (1999), James (2002) and Muller and Stadtmuller (2003). However, we are not aware of any previous work on the other four functional modeling types. FAME, which corresponds to the bottom right box, provides an extension of generalized projection pursuit to functional data of which the other methods can all be seen as special cases. Since neural networks with one hidden layer are a special case of projection pursuit regression, we note that FAME also provides a natural method for fitting neural networks to functional data.

The FAME methodology suggests a number of interesting areas for future work. First, the asymptotic hypothesis tests of Section 3 are only one of several possible approaches that might be taken. For example, Cardot *et al.* (2003a) develop hypothesis tests for functional linear models. Second, in implementing FAME we utilize high dimensional bases for the β_k's and X_i's. The exact choice of a basis and its dimension are not critical because of the use of penalty terms to regularize the fits. However, the simulations of Section 6 suggest that there is some sensitivity to a reasonable choice for the penalty coefficient λ_β. In this paper we utilized standard cross-validation but one might also use less computational approaches such as generalized cross-validation or possibly BIC or AIC type criterion. Third, the asymptotic theory might also be extended to arbitrary smooth functional data. In Section 3 we gave results for a finite dimensional FAME model but in principle it should be possible to derive similar results without this restriction. Fourth, for very sparse data, fitting each $X_i(t)$ individually could provide inaccurate estimates. It may be possible

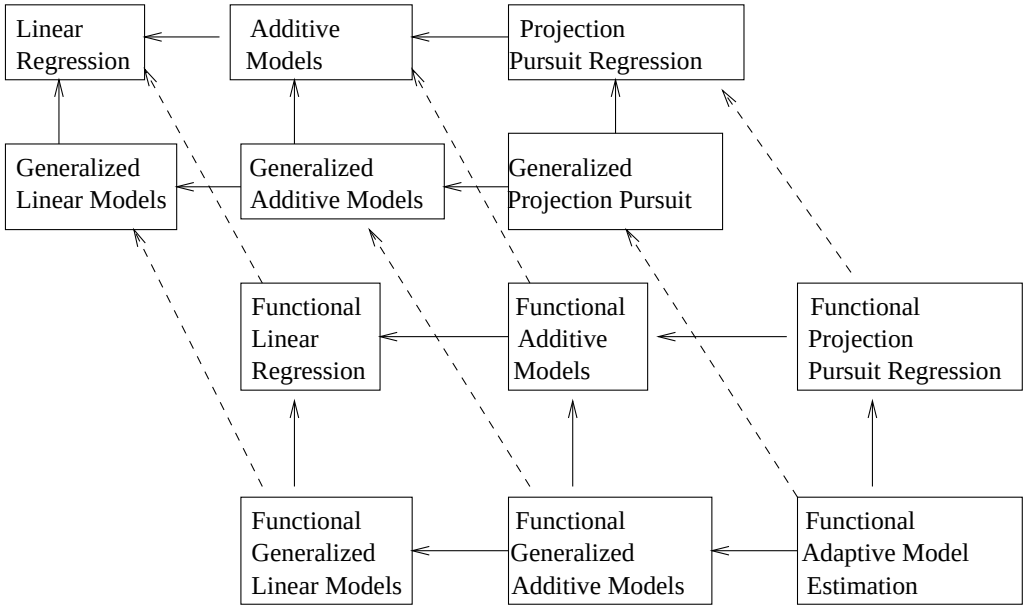

Figure 7: *A chart indicating the relationships among the various standard regression methodologies and their functional extensions. Arrows point from more to less general models.*

to produce better answers by building strength across the predictors by assuming common covariances. This approach is taken for functional data in James (2002) and James and Sugar (2003). Fifth, in practice we have found that the constraints placed on the β_k's and the Z_k's seem to produce identifiable parameter estimates. However, it is unclear what theoretical conditions need to be placed on the predictors to ensure identifiability. Finally, another interesting problem is the development of a functional generalized additive models procedure. Linear regression, generalized linear models and projection pursuit all naturally extend to functional data because they involve first taking a linear function of the predictors. In these cases the summation over $X_j\beta_j$ can be replaced by an integral over $X(t)\beta(t)$. However, no such linear function of the predictors is employed in additive models, making it unclear how best to proceed. One possibility would be to assume that the predictor functions lie approximately in a finite dimensional space by, for example, taking the first K principal component curves. An additive model could then be fit to the K weights for each curve and the results interpreted by examining the form of the principal component curves.

A Proofs

Proof of Theorem 1

Proofs of this result under fairly general conditions exit for iid data in many standard texts. We use a similar approach to that of Lehmann (1991) p 430 except that our problem is complicated by the fact that the observations are independent but not identically distributed.

First consider the FAME model given by (5) - (7) with the addition of the priors

$$\eta_k \sim N(0, \Omega_{\eta_k}^{-1}), \quad \delta_k \sim N(0, \Omega_{\delta_k}^{-1}) \tag{25}$$

where Ω_{η_k} and Ω_{δ_k} correspond to the penalty matrices for β_k and f_k. Up to additive constants, the log

likelihood function for this model is given by

$$l(\beta_0,\eta_k,\delta_k,\phi)=\sum_{i=1}^{N}\left(\frac{y_i\theta_i-b(\theta_i)}{a(\phi)}+c(y_i,\phi)\right)-\frac{1}{2}\sum_{k=1}^{r}\eta_k^T\Omega_{\eta_k}\eta_k-\frac{1}{2}\sum_{k=1}^{r}\delta_k^T\Omega_{\delta_k}\delta_k. \tag{26}$$

Equation 26 is identical to the penalized likelihood that is fit using the FAME algorithm. Hence the FAME parameter estimates are maximum likelihood estimates for the augmented model. So we need only prove that these mles have the desired asymptotic properties. However, the last two terms of (26) involving Ω_η and Ω_δ become negligible compared to the first as N approaches infinity so without loss of generality we may assume the likelihood is given by (13). In other words as N gets large the smoothness penalties have no effect on the fit. The functional nature of $X_i(t)$ posses no additional complications because for an orthogonal basis $\mathbf{B}(t)$ it is a simple matter to show that $Z_{ik}=\int X_i(t)\beta_k(t)dt=\gamma_i^T\eta_k$ so that the estimation problem involves fitting the standard finite dimensional variables η_k as well as β_0 and δ_k. We now show that the mles for (13) are consistent.

Consider a sphere Q_a centered at the true parameter ξ^0 with radius a. We first show that for small enough a with probability tending to 1, $l(\xi)<l(\xi^0)$ for all points ξ on the surface of Q_a. This will also show that $l(\xi)$ has a local maximum in the interior of Q_a. Expanding l about ξ^0 using Taylor's theorem gives

$$\begin{aligned}\frac{1}{N}l(\xi)-\frac{1}{N}l(\xi^0) &= \frac{1}{N}\sum_j l_j'(\xi^0)(\xi_j-\xi_j^0)\\ &+ \frac{1}{2N}\sum_j\sum_k l_{jk}''(\xi^0)(\xi_j-\xi_j^0)(\xi_k-\xi_k^0)\\ &+ \frac{1}{6N}\sum_j\sum_k\sum_l(\xi_j-\xi_j^0)(\xi_k-\xi_k^0)(\xi_l-\xi_l^0)\sum_{i=1}^{N}\alpha_{jkl}(y_i)M(y_i)\\ &= S_1+S_2+S_3\end{aligned}$$

where by (A-1) $0\le|\alpha_{jkl}|\le 1$. We will show that for large enough N and small enough a, S_2 is negative and S_1 and S_3 are small relative to S_2. First note that $E(l_j'(\xi^0))=0$ and hence by Chebychev's theorem

$$P(\frac{1}{N}|l_j'(\xi^0)|\ge a^2)\le\frac{I_{Njj}}{N^2a^2}. \tag{27}$$

where I_{Njj} is the (j,j)th component of I_N. By (A-2) the right hand side of this equation converges to zero. In addition, also by Chebychev's theorem,

$$P\left(\frac{1}{N}|l_{jk}''(\xi^0)-E(l_{jk}''(\xi^0))|>\varepsilon\right)<\frac{Var(l_{jk}''(\xi^0))}{N^2\varepsilon^2}.$$

By (A-2) the left hand side converges to $P\left(\frac{1}{N}|l_{jk}''(\xi^0)-(-\bar{I}_{jk})|>\varepsilon\right)$ while by (A-3) the right hand side converges to zero. Hence

$$\frac{1}{N}l_{jk}''(\xi^0)\to-\bar{I}_{jk}\quad\text{in probability}. \tag{28}$$

Let $p=2rq+1$ represent the total number of parameters in ξ. For ξ on Q_a, $|S_1|\le a\sum_j|l_j'(\xi^0)|/N$ and from (27) with probability tending to 1, $\sum_j|l_j'(\xi^0)|/N<pa^2$. Hence $|S_1|<pa^3$. Also with probability

tending to 1, $\sum_{i=1}^{N}|\alpha_{jkl}(y_i)|M(y_i)/N < 6m_1$ and hence $|S_3| < p^3a^3m_1$. Finally note that

$$2S_2 = \sum\sum[-\bar{I}_{jk}(\xi^0)(\xi_j-\xi_j^0)(\xi_k-\xi_k^0)] + \sum\sum\left\{\frac{1}{n}l''_{jk}(\xi^0) - [-\bar{I}_{jk}]\right\}(\xi_j-\xi_j^0)(\xi_k-\xi_k^0)$$

From (28) we see that the second term will have absolute value less than p^2a^3 with probability tending to 1. Also since $\bar{I}$ is positive definite for ξ on Q_a the first term is less than $-\lambda a^2$ for some $\lambda > 0$. Hence, combining the two terms, there exists $c > 0$ such that for small enough a, $S_2 < -ca^2$ with probability tending to 1. Therefore

$$\max(S_1+S_2+S_3) < -ca^2 + (p^3m_1+s)a^3$$

which is less than zero if $a < c/(p^3m_1+s)$.

As a result, with probability tending to 1, for any a sufficiently small there exists a value $\hat{\xi}_N \in Q_a$ at which $l(\xi)$ is a local maximum so that $l'(\hat{\xi}_N) = 0$. Therefore there exists a sequence $\hat{\xi}_N(a)$ of roots such that

$$P_{\xi^0}(||\hat{\xi}_N(a) - \xi^0|| < a) \to 1.$$

To complete the proof we need to show that we can determine a sequence which does not depend on a. Let ξ_N^* be the root closest to ξ^0. Then clearly $P_{\xi^0}(||\hat{\xi}_N^* - \xi^0|| < a) \to 1$ which completes our proof.

Proof of Theorem 2

We first state two lemmas.

Lemma 1 *Let $\mathbf{T}_N$ be a sequence of random vectors such that $\mathbf{T}_N \Rightarrow \mathbf{T}$. Let A_N be a sequence of random matrices such that each element converges in probability to the corresponding element of the constant non-singular matrix A. Then the solution $\mathbf{Y}_N$ of*

$$A_N\mathbf{Y}_N = \mathbf{T}_N \tag{29}$$

converges in distribution to

$$\mathbf{Y} = A^{-1}\mathbf{T}.$$

See Lehmann (1991) p 433 for a proof.

Lemma 2 *For random vectors $\mathbf{X}_N$ and $\mathbf{Y}$, a necessary and sufficient condition for $\mathbf{X}_N \Rightarrow \mathbf{Y}$ is that $\mathbf{X}_N^T\mathbf{t} \Rightarrow \mathbf{Y}^T\mathbf{t}$ for any real valued vector $\mathbf{t}$.*

See Billingsley (1986) p 397 for a proof.

To prove Theorem 2 we start by expanding $l'(\xi)$ about ξ^0 using Taylor's theorem,

$$l'_j(\xi) = l'_j(\xi^0) + \sum_k(\xi_k-\xi_k^0)l''_{jk}(\xi^0) + \frac{1}{2}\sum_k\sum_l(\xi_k-\xi_k^0)(\xi_l-\xi_l^0)l'''_{jkl}(\xi^*)$$

where ξ^* is a point on the line segment between ξ and ξ^0. By Theorem 1 a consistent solution of the likelihood equations $\hat{\xi}_N$ exists with probability tending to 1. Hence we replace ξ by $\hat{\xi}_N$ giving

$$-\frac{1}{\sqrt{N}} l'_j(\xi^0) = \sqrt{N} \sum_k (\hat{\xi}_{kN} - \xi^0_k) \left[\frac{1}{N} l''_{jk}(\xi^0) + \frac{1}{2N} \sum_l (\hat{\xi}_{lN} - \xi^0_l) l'''_{jkl}(\xi^*) \right]$$

This equation has the form of (29) with $\mathbf{Y}_N = \sqrt{N}(\hat{\xi}_N - \xi^0)$, $A_{jkN} = \frac{1}{N} l''_{jk}(\xi^0) + \frac{1}{2N} \sum (\hat{\xi}_{ln} - \xi^0_l) l'''_{jkl}(\xi^*)$ and $\mathbf{T}_N = -\frac{1}{\sqrt{N}} l'(\xi^0)$. Therefore by Lemma 1 we need only prove that $A_N \to \bar{I}$ in probability and $\mathbf{T}_N$ converges in distribution to a multivariate normal with mean zero and covariance $\bar{I}$.

As with the proof of Theorem 1 by (A-2) and (A-3) $\frac{1}{N} l''(\xi^0) \to \bar{I}$. Also by (A-1)

$$\frac{1}{N} |l'''_{jkl}| < \frac{1}{N} \sum M_i(Y_i) < m_1$$

with probability tending to 1. Hence, since $\hat{\xi}_N - \xi^0 \to 0$ in probability, $A_N \to \bar{I}$ in probability. Next consider

$$-\frac{1}{\sqrt{N}} l'(\xi^0)^T \mathbf{t} = \frac{-l'(\xi^0)^T \mathbf{t}}{\sqrt{\mathbf{t}^T I_N \mathbf{t}}} \sqrt{\mathbf{t}^T I_N \mathbf{t}/N}$$

where $\mathbf{t}$ is an arbitrary real valued vector. Then by Lindeberg's Theorem (Billingsley, 1986, p 369)

$$\frac{-l'(\xi^0)^T \mathbf{t}}{\sqrt{\mathbf{t}^T I_N \mathbf{t}}} \Rightarrow N(0,1)$$

provided Lyapounov's condition (Billingsley, 1986, p 371) holds. But by (A-4)

$$\lim_{N\to\infty} \frac{1}{(\mathbf{t}^T I_N \mathbf{t})^{1+\varepsilon/2}} \sum_{i=1}^{N} E \left[\left| \frac{(y_i - \mu_i)}{g'(\mu_i) Var(y_i)} \sum_{k=1}^{r} \left(\frac{\partial f(Z_{ik})}{\partial \xi} \right)^T \mathbf{t} \right|^{2+\varepsilon} \right] \leq \lim_{N\to\infty} \frac{m_2 ||\mathbf{t}||^{2+\varepsilon}}{(\mathbf{t}^T I_N / N \mathbf{t})^{1+\varepsilon/2} N^{\varepsilon/2}}.$$

This limit is equal to 0 because by (A-2) I_N converges to a positive definite matrix so Lyapounov's condition is satisfied. Similarly by (A-2) $\sqrt{\mathbf{t}^T I_N \mathbf{t}/N}$ converges to $\sqrt{\mathbf{t}^T \bar{I} \mathbf{t}}$ so for any $\mathbf{t}$

$$-\frac{1}{\sqrt{N}} l'(\xi^0)^T \mathbf{t} \Rightarrow N(0, \mathbf{t}^T \bar{I} \mathbf{t})$$

and by Lemma 2

$$-\frac{1}{\sqrt{N}} l'(\xi^0) \Rightarrow N(0, \bar{I}).$$

This completes the proof.

Proof of Corollary 1

By Theorem 2 $\sqrt{N}(\hat{\eta}_k - \eta_k) \Rightarrow N(0, \Sigma_{\hat{\eta}_k})$. Since $\hat{\beta}_k(t)$ is simply a linear combination of $\hat{\eta}_k$ this implies

$$\sqrt{N}(\hat{\beta}_k(t) - \beta_k(t)) \Rightarrow N(0, \mathbf{B}^T(t) \Sigma_{\hat{\eta}_k} \mathbf{B}(t))$$

or equivalently

$$P\left\{\sqrt{\mathbf{B}^T(t)\Sigma_{\hat{\eta}_k}\mathbf{B}(t)}\Phi^{-1}_{\alpha/2} \leq \sqrt{N}(\hat{\beta}_k(t)-\beta_k(t)) \leq \sqrt{\mathbf{B}^T(t)\Sigma_{\hat{\eta}_k}\mathbf{B}(t)}\Phi^{-1}_{1-\alpha/2}\right\} \rightarrow 1-\alpha.$$

Rearranging terms gives the desired result.

Proof of Corollary 2

Suppose that

$$\sqrt{N}\hat{\delta}_{r+1} \Rightarrow N(0,\Sigma_{\hat{\delta}_{r+1}}). \tag{30}$$

Then it is easily seen that

$$N\hat{\delta}^T_{r+1}\Sigma^{-1}_{\hat{\delta}_{r+1}}\hat{\delta}_{r+1} \Rightarrow \chi^2$$

where the degrees of freedom are equal to the number of components of δ_{+1} i.e. q. Therefore we need only show that (30) is true. If there are only r terms in the model then $f_{r+1}(Z_k)=0$ so $\delta_{r+1}=0$ which would suggest (30) via Theorem 2. But in this case $f'_{r+1}=0$ so it is clear from (19) that $\bar{I}$ is singular which violates (A-2). However, note that $\frac{\partial l^2}{\partial\eta_{r+1}\partial\eta_{r+1}}=\frac{\partial l}{\partial\eta_{r+1}}=0$. Hence, assuming that the limiting information matrix with components corresponding to η_{r+1} removed is non-singular, minor alternations of Theorems 1 and 2 show that (30) still holds given (A-1), (A-3) and (A-4).

B Derivation of I_N

First note that the constraint $\int\beta_k(t)dt=1$ implies

$$\eta_{kq}=1-\sum_{j=1}^{q-1}\frac{\int B_j(t)dt}{\int B_q(t)dt}\eta_{kj} \tag{31}$$

where $B_j(t)$ and η_{kj} are the jth components of $\mathbf{B}(t)$ and η_k. As a result of (31), $\frac{\partial Z_{ik}}{\partial\eta_{kj}}=\gamma^*_{ij}$ where $\gamma^*_{ij}=\gamma_{ij}-\frac{\int B_j(t)dt}{\int B_q(t)dt}\gamma_{iq}$. To derive I_N we need to calculate $\frac{\partial^2 l}{\partial\xi_k\partial\xi_j}$ which involves the terms $\frac{\partial^2 l}{\partial\eta_k\partial\eta_j}$, $\frac{\partial^2 l}{\partial\delta_k\partial\delta_j}$, $\frac{\partial^2 l}{\partial\eta_k\partial\delta_j}$, $\frac{\partial^2 l}{\partial\beta_0\partial\beta_0}$, $\frac{\partial^2 l}{\partial\eta_k\partial\beta_0}$ and $\frac{\partial^2 l}{\partial\delta_k\partial\beta_0}$. We illustrate the approach on the first three of these derivatives. The calculations are similar to those for GLMs.

$$\begin{aligned}
\frac{\partial l}{\partial\eta_k} &= \sum_{i=1}^N\left(\frac{y_i-b'(\theta_i)}{a(\phi)}\right)\frac{\partial\theta_i}{\partial\eta_k}-\Omega_\eta\eta_k \quad \left(\frac{\partial\theta_i}{\partial\eta_k}=\frac{a(\phi_i)f'_k(Z_{ik})\gamma^*_i}{g'(\mu_i)Var(y_i)}\right)\\
&= \sum_{i=1}^N\left(\frac{y_i-\mu_i}{Var(y_i)}\right)\frac{f'_k(Z_{ik})\gamma^*_i}{g'(\mu_i)}-\Omega_\eta\eta_k\\
\Rightarrow\frac{\partial^2 l}{\partial\eta_k\partial\eta_j} &= \sum_{i=1}^N\frac{1}{a(\phi_i)}\left[-b''(\theta_i)\frac{\partial\theta_i}{\partial\eta_k}\frac{\partial\theta_i}{\partial\eta_j}^T+\{y_i-b'(\theta_i)\}\frac{\partial^2\theta_i}{\partial\eta_k\partial\eta_j}\right]-\Omega_\eta I(j=k)\\
\Rightarrow -E\left[\frac{\partial^2 l}{\partial\eta_k\partial\eta_j}\right] &= \sum_{i=1}^N\frac{Var(y_i)}{a(\phi_i)^2}\frac{\partial\theta_i}{\partial\eta_k}\frac{\partial\theta_i}{\partial\eta_j}^T+\Omega_\eta I(j=k)=\sum_{i=1}^N\frac{f'_k(Z_{ik})f'_j(Z_{ij})\gamma^*_i\gamma^{*T}_i}{Var(y_i)g'(\mu_i)^2}+\Omega_\eta I(j=k)
\end{aligned}$$

The last line follows from the fact that $E\{y_i - b'(\theta_i)\} = 0$.

$$\begin{aligned}
\frac{\partial l}{\partial \delta_k} &= \sum_{i=1}^{N} \left(\frac{y_i - b'(\theta_i)}{a(\phi)} \right) \frac{\partial \theta_i}{\partial \delta_k} - \Omega_\delta \delta_k \quad \left(\frac{\partial \theta_i}{\partial \delta_k} = \frac{a(\phi_i)\mathbf{s}(Z_{ik})}{g'(\mu_i)Var(y_i)} \right) \\
&= \sum_{i=1}^{N} \left(\frac{y_i - \mu_i}{Var(y_i)} \right) \frac{\mathbf{s}(Z_{ik})}{g'(\mu_i)} - \Omega_\delta \delta_k \\
\Rightarrow \frac{\partial^2 l}{\partial \delta_k \partial \delta_j} &= \sum_{i=1}^{N} \frac{1}{a(\phi_i)} \left[-b''(\theta_i) \frac{\partial \theta_i}{\partial \delta_k} \frac{\partial \theta_i}{\partial \delta_j}^T + \{y_i - b'(\theta_i)\} \frac{\partial^2 \theta_i}{\partial \delta_k \partial \delta_j} \right] - \Omega_\delta I(j=k) \\
\Rightarrow -E\left[\frac{\partial^2 l}{\partial \delta_k \partial \delta_j} \right] &= \sum_{i=1}^{N} \frac{Var(y_i)}{a(\phi_i)^2} \frac{\partial \theta_i}{\partial \delta_k} \frac{\partial \theta_i}{\partial \delta_j}^T + \Omega_\delta I(j=k) = \sum_{i=1}^{N} \frac{\mathbf{s}(Z_{ik})\mathbf{s}^T(Z_{ij})}{Var(y_i)g'(\mu_i)^2} + \Omega_\delta I(j=k)
\end{aligned}$$

Finally

$$\begin{aligned}
\Rightarrow \frac{\partial^2 l}{\partial \eta_k \partial \delta_j} &= \sum_{i=1}^{N} \frac{1}{a(\phi_i)} \left[-b''(\theta_i) \frac{\partial \theta_i}{\partial \eta_k} \frac{\partial \theta_i}{\partial \delta_j}^T + \{y_i - b'(\theta_i)\} \frac{\partial^2 \theta_i}{\partial \eta_k \partial \delta_j} \right] \\
\Rightarrow -E\left[\frac{\partial^2 l}{\partial \eta_k \partial \delta_j} \right] &= \sum_{i=1}^{N} \frac{Var(y_i)}{a(\phi_i)^2} \frac{\partial \theta_i}{\partial \eta_k} \frac{\partial \theta_i}{\partial \delta_j}^T = \sum_{i=1}^{N} \frac{f_k'(Z_{ik})\gamma_i^* \mathbf{s}^T(Z_{ij})}{Var(y_i)g'(\mu_i)^2}
\end{aligned}$$

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.

Billingsley, P. (1986). *Probability and Measure*. John Wiley and Sons, 2nd edn.

Cardot, H., Ferraty, F., Mas, A., and Sarda, P. (2003a). Testing hypotheses in the functional linear model. *Scandinavian Journal of Statistics* **30**, 241–255.

Cardot, H., Ferraty, F., and Sarda, P. (2003b). Spline estimators for the functional linear model. *Statistica Sinica* **13**, 571–591.

Fahrmeir, L. and Tutz, G. (1994). *Multivariate Statistical Modeling Based on Generalized Linear Models*. Springer-Verlag.

Ferraty, F. and Vieu, P. (2002). The functional nonparametric model and applications to spectrometric data. *Computational Statistics* **17**, 545–564.

Ferraty, F. and Vieu, P. (2003). Curves discrimination: a nonparametric functional approach. *Computational Statistics and Data Analysis* **44**, 161–173.

Fleming and Harrington (1991). *Counting Processes and Survival Analysis*. Wiley.

Freund, Y. and Schapire, R. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences* **55**, 119–139.

Friedman, J. H. and Stuetzle, W. (1981). Projection pursuit regression. *Journal of the American Statistical Association* **76**, 817–823.

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.

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized Additive Models*. Chapman and Hall.

Hastie, T. J. and Tibshirani, R. J. (1993). Varying-coefficient models (with discussion). *Journal of the Royal Statistical Society, Series B* **55**, 757–796.

Hastie, T. J., Tibshirani, R. J., and Friedman, J. (2001). *The Elements of Statistical Learning*. Springer.

Hoover, D. R., Rice, J. A., Wu, C. O., and Yang, L. P. (1998). Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. *Biometrika* **85**, 809–822.

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 Hastie, T. J. (2001). Functional linear discriminant analysis for irregularly sampled curves. *Journal of the Royal Statistical Society, Series B* **63**, 533–550.

James, G. M. and Sugar, C. A. (2003). Clustering for sparsely sampled functional data. *Journal of the American Statistical Association* **98**, 397–408.

Lehmann, E. L. (1991). *Theory of Point Estimation*. Wadsworth and Brooks, Pacific Grove, California, 1st edn.

Liang, K. Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13–22.

Lin, D. Y. and Ying, Z. (2001). Semiparametric and nonparametric regression analysis of longitudinal data. *Journal of the American Statistical Association* **96**, 103–113.

Lingjaerde, O. C. and Liestol, K. (1998). Generalized projection pursuit regression. *SIAM Journal of Scientific Computing* **20**, 844–857.

Marx, B. D. and Eilers, P. H. (1999). Generalized linear regression on sampled signals and curves: A P-spline approach. *Technometrics* **41**, 1–13.

McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models*. Chapman and Hall, London, 2nd edn.

Moyeed, R. A. and Diggle, P. J. (1994). Rates of convergence in semi-parametric modeling of longitudinal data. *Australian Journal of Statistics* **36**, 75–93.

Muller, H. G. and Stadtmuller, U. (2003). Generalized functional linear models. *Unpublished* .

Ramsay, J. O. and Silverman, B. W. (1997). *Functional Data Analysis*. Springer.

Ramsay, J. O. and Silverman, B. W. (2002). *Applied Functional Data Analysis*. Springer.

Roosen, C. B. and Hastie, T. J. (1993). Logisitc response projection pursuit. *AT&T Bell Laboratories, Doc. BL011214-930806-09TM, Murray Hill, NJ* .

Wu, C. O., Chiang, C. T., and Hoover, D. R. (1998). Asymptotic confidence regions for kernel smoothing of a varying-coefficient model with longitudinal data. *Journal of the American Statistical Association* **93**, 1388–1402.

Zeger, S. L. and Diggle, P. J. (1994). Semiparametric models for longitudinal date with applications to CD4 cell numbers in HIV seroconverters. *Biometrics* **50**, 689–699.