Generalized Linear Models with Functional Predictors

GARETH M. JAMES
Marshall School of Business, University of Southern California

Abstract

In this paper we present a technique for extending generalized linear models (GLM) to the situation where some of the predictor variables are observations from a curve or function. The technique is particularly useful when only fragments of each curve have been observed. We demonstrate, on both simulated and real world data sets, how this approach can be used to perform linear, logistic and censored regression with functional predictors. In addition, we show how functional principal components can be used to gain insight into the relationship between the response and functional predictors. Finally, we extend the methodology to apply GLM and principal components to standard missing data problems.

Some key words: Censored regression; Functional data analysis; Functional principal components; Generalized linear models; Logistic regression.

1 Introduction

Generalized linear models provide a framework for relating response and predictor variables (McCullagh and Nelder, 1989). For a random variable Y with distribution,

$$p(y;\eta,\phi) = \exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right)$$

we model the relationship between predictor $\mathbf{X}$ and response Y as

$$g(\mu) = \beta_0 + \beta_1^T \mathbf{X} \tag{1}$$

where $\mu = E(Y;\theta,\phi) = b'(\theta)$ and g is referred to as the link function. Common examples include the identity link used for normal response data and the logistic link used for binary response data. Generalized linear models provide a very flexible class of procedures. However, they assume that the predictor has a finite dimension. In this paper we extend GLM to handle functional predictors which may be measured at different times and with different numbers of observations for each individual.

One of the difficulties with these sorts of data sets is that, when predictors are functional, observations from the same individual will generally be correlated. A great deal of research has been conducted on data with correlated outcomes. Situations where such data arise include twin studies (Cessie and Houwelingen, 1994), two-period cross-over designs (Jones and Kenward, 1989), ophthalmological studies (Gao *et al.*, 2001) and longitudinal data (Diggle *et al.*, 1994). Numerous models have been proposed for the response variable. 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,

$$Y(t) = \alpha_0(t) + \beta_0^T X(t) + \varepsilon(t) \tag{2}$$

ID	End	Outcome	Drug	Day	Bili	Alb	ID	End	Outcome	Drug	Day	Bili	Alb
1	400	Dead	Yes	0	14.5	2.60	2	5169	Alive	Yes	2515	4.2	2.73
1	400	Dead	Yes	192	21.3	2.94	2	5169	Alive	Yes	2882	3.6	2.80
2	5169	Alive	Yes	0	1.1	4.14	2	5169	Alive	Yes	3226	4.6	2.67
2	5169	Alive	Yes	182	0.8	3.60	3	1012	Dead	Yes	0	1.4	3.48
2	5169	Alive	Yes	365	1.0	3.55	3	1012	Dead	Yes	176	1.1	3.29
2	5169	Alive	Yes	768	1.9	3.92	3	1012	Dead	Yes	364	1.5	3.57
2	5169	Alive	Yes	1790	2.6	3.32	3	1012	Dead	Yes	743	1.8	3.25
2	5169	Alive	Yes	2151	3.6	2.92	⋮	⋮	⋮	⋮	⋮	⋮	⋮

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

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 (2) by allowing the regression coefficients to vary over time. Alternatively, Gao *et al.* (2001) model categorical responses using a combined smoothing spline analysis of variance and log-linear model approach, while James and Hastie (2001) use a functional linear discriminant analysis model. Fahrmeir and Tutz (1994) and Liang and Zeger (1986) suggest an even more general framework where the response is modeled as a member of the exponential family of distributions.

This work has tended to focus on the situation where the predictor and response are observed together at varying times. However, in many cases, one wishes to model the relationship between a single, time independent, response and a functional predictor. For example, one might wish to predict whether an individual possesses a genetic disorder based on various predictors measured over time. Alternatively, one may wish to calculate the probability of a successful transplant operation based on historical measurements of a patient. In both these situations a single response is observed but the predictors are functional because they are measured over time. Most of the methods listed above cannot easily be applied to such problems because they assume a separate response at each time that a predictor is observed. Hastie and Mallows (1993) and Ramsay and Silverman (1997) (Chapter 10) discuss performing regression where the response is a scalar and the predictors functional but they primarily address the situation where the predictors are all measured at the same time points.

Table 1 provides a typical example of a functional data set with unequally spaced observations. These data were obtained from StatLib and 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 Harrrington, 1991). For each patient we have a record of the time, in days, between the earlier of death or end of study ("End"), alive or dead ("Outcome"), 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 for illustrative purposes we will restrict to these variables. There are two response variables of interest. The first is survival time, a right censored variable, and the second is the five year survival outcome. For both situations, each patient has multiple measurements of both bilirubin and albumin but only one, time independent, response so a model such as (2) cannot be applied. Furthermore, there are different numbers of measurements for each patient and they are taken at different times so it is not possible to use a standard multiple regression model.

One possible solution would be to ignore time trends and to use either the first measurement or the average

over all observations for each person. However, if there is a time trend, both of these methods will make inefficient use of the available information. A superior approach might be to fit a smooth parametric curve, such as a natural cubic spline, to each individual's observations and use the resulting coefficient vector as a predictor. This method has the advantage that it accounts for any time trends in the data. Unfortunately, it has several drawbacks. First, many of the individuals only have a very small number of observations so it may not be possible to fit curves for each of them. Second, even if the curves can all be fit, it is not obvious how to adjust for the varying levels of accuracy in the coefficients caused by differences in the number and spacing of observations.

In this paper we present an approach, which we call functional generalized linear models (FGLM), that directly models the relationship between a single response, from any member of the exponential family of distributions, and a functional predictor. The predictors are modeled as cubic splines and it is assumed that the spline coefficients for all individuals have a common mean and variance for which both the response and predictors are used to fit. The predicted coefficients for each individual can then be used in the linear portion of the link function to relate the predictors to the response. We have successfully applied FGLM to situations in which each subject has observations at differing time points. Furthermore, the method works well on sparse data sets such as the PBC data, since it does not rely on fitting a separate curve to each person. A large range of possible distributions can be assumed for the response variable, allowing the modeling of both continuous and categorical data. In addition, the relationship between functional predictor and scalar response can be assessed through the use of functional principal components. Tests are also developed for a relationship between predictor and response.

In section 2 we outline and motivate the general modeling procedure. Functional linear, logistic and censored regression are developed as special cases of this model. Section 3 provides details of an EM algorithm that works well for fitting the functional generalized linear model. Examples on simulated and real data sets are given in section 4. Functional principal components ideas are used in section 5 to provide a better understanding of the exact form of the relationship between predictor and response. In practice one may wish to incorporate multiple functional and finite dimensional predictors into the model. This extension is developed in section 6. Finally, extensions to missing data problems are provided in section 7.

2 The functional generalized linear model

In this section we develop the functional generalized linear model. We then illustrate three particular examples, linear, censored and logistic regression.

2.1 The general model

When the predictor $X(t)$ is functional, the link function given by (1) cannot be directly applied. However, a natural generalization is to replace the summation over the finite dimensional space with an integral over the infinite dimensional one,

$$g(\mu) = \beta_0 + \int \omega_1(t)X(t)dt, \tag{3}$$

where $\omega_1(t)$ is the functional analogue of β_1. Unfortunately, in practice $X(t)$ is only ever observed at a finite set of time points. One might imagine simply replacing the integral with a summation over the observed times. However, this approach has several potential problems. First, it may necessitate fitting an extremely high dimensional vector of coefficients, resulting in large or infinite variance terms. Second, it is not clear how to handle individuals with observations that are measured at different sets of time points or individuals with differing numbers of observations. Both these problems are related to the fact that this procedure fails

to make use of the intrinsic relationship between points observed at similar times. Instead we assume that each predictor can be modeled as a smooth curve from a given functional family. We choose to make use of natural cubic splines (Silverman 1985; Green and Silverman 1994). The resulting parameterization is

$$X(t) = \mathbf{s}(t)^T\gamma, \qquad \gamma \sim N(\mu_\gamma, \Gamma) \tag{4}$$

where $\mathbf{s}(t)$ represents the q-dimensional spline basis at time t, γ the q-dimensional spline coefficients for the predictor and μ_γ and Γ the mean and variance of the γ's. A q-dimensional natural cubic spline will have $q-2$ knots. Combining (3) and (4) we arrive at the final link function,

$$\begin{aligned} g(\mu_i) &= \beta_0 + \int \omega_1(t)\mathbf{s}(t)^T\gamma_i dt \\ &= \beta_0 + \beta_1^T\gamma_i, \end{aligned} \tag{5}$$

where $\beta_1 = \int \omega_1(t)\mathbf{s}(t)dt$. Further, we assume that, at any given time t, instead of $X(t)$, one observes $x(t)$ where,

$$x(t) = X(t) + e(t).$$

We model $e(t)$ as a zero-mean stationary Gaussian process. This term represents the deviations of observations from the spline fit due to measurement error or other factors. Let $\mathbf{x}_i$ and $\mathbf{e}_i$ be the vectors of observations and measurement errors for individual i at times $t_{i1}, \ldots, t_{in_i}$ and let $S_i = (\mathbf{s}(t_{i1}), \ldots, \mathbf{s}(t_{in_i}))^T$ be the corresponding spline basis matrix. Then the functional generalized linear model can be written as

$$\begin{aligned} p(y_i;\theta_i,\phi) &= \exp\left(\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i,\phi)\right), \\ g(\mu_i) &= \beta_0 + \beta_1^T\gamma_i, \quad \gamma_i \sim N(\mu_\gamma, \Gamma), \\ \mathbf{x}_i &= S_i\gamma_i + \mathbf{e}_i, \quad \mathbf{e}_i \sim N(0, \sigma_x^2 I), \quad i = 1, \ldots, N, \end{aligned}$$

where N represents the number of observed response-predictor pairs. We use spline bases because they allow one to fit a large variety of functional forms. Using a basis with a large number of knots models very flexible curves while restricting the number of knots forces less flexible, possibly more interpretable, curves. However, the above model can be fit with equal ease using Fourier transforms, orthogonal polynomial bases or any other finite dimensional basis.

2.2 Specific Models

The FGLM model from the previous section can be used with a large number of response variable distributions. In this section we give details for three important specific examples.

2.2.1 Functional Linear Regression

The best known special case of GLM is linear regression, in which the response is assumed to be normally distributed and g is taken to be the identity function. Under these conditions the FGLM model of Section 2.1 reduces to

$$\begin{aligned} Y_i &= \beta_0 + \beta_1^T\gamma_i + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma_y^2), \quad \gamma_i \sim N(\mu_\gamma, \Gamma) \\ \mathbf{x}_i &= S_i\gamma_i + \mathbf{e}_i, \quad \mathbf{e}_i \sim N(0, \sigma_x^2 I). \end{aligned}$$

An algorithm for fitting this model is presented in the appendix A.1 and examples of its application are given in Section 4.1. As with standard linear regression the predicted response will be $E(Y|\mathbf{x})$. It can be shown that,

$$\gamma|\mathbf{x} \sim N\left(\left[\sigma_x^2\Gamma^{-1}+S^TS\right]^{-1}\left[\sigma_x^2\Gamma^{-1}\mu_\gamma+S^T\mathbf{x}\right],\left[\Gamma^{-1}+S^TS/\sigma_x^2\right]^{-1}\right). \tag{6}$$

Hence $E(Y|\mathbf{x})$ can be easily computed using

$$Y|\mathbf{x} \sim N\left(\beta_0+\beta_1^T(\sigma_x^2\Gamma^{-1}+S^TS)^{-1}(\sigma_x^2\Gamma^{-1}\mu_\gamma+S^T\mathbf{x}),\beta_1^T(\Gamma^{-1}+S^TS/\sigma_x^2)^{-1}\beta_1+\sigma_y^2\right). \tag{7}$$

2.2.2 Functional Censored Regression

When using life expectancy as the response variable right censoring is a common problem. The functional linear regression model can be extended to the case where right censoring exists in the response. In this situation we assume that Y_i is observed for $i = 1,\ldots,m$ but for $i = m+1,\ldots,N$ we observe only that $Y_i > c_i$ where c_i is a known constant. In all other respects the censored and standard linear regression models are identical. We present an algorithm for fitting this model in appendix A.2 and an example using the PBC data in section 4.2. Predictions for new or uncensored responses are given by $E(Y|\mathbf{x})$ calculated according to (7). However, predictions for a censored response are given by

$$E(Y_i|Y_i > c_i,\mathbf{x}_i) = \mu_{Y|x}+\sigma_{Y|x}\frac{\phi\left(\frac{c_i-\mu_{Y|x}}{\sigma_{Y|x}}\right)}{1-\Phi\left(\frac{c_i-\mu_{Y|x}}{\sigma_{Y|x}}\right)}, \tag{8}$$

where $\mu_{Y|x}$ and $\sigma^2_{Y|x}$ are the mean and variance of $Y|\mathbf{x}$ from (7) and ϕ and Φ are the standard normal density and cumulative distribution functions.

2.2.3 Functional Logistic Regression

Finally we illustrate the case in which the response is a Bernoulli variable. The PBC data using five year survival as the response is a typical example. In this case $E(Y|\mathbf{x}) = P(Y = 1|\mathbf{x})$ represents the survival rate. This probability can be modeled using several possible link functions including the probit or complementary log-log. We will demonstrate the technique on the canonical and most commonly applied link function, the logistic, under which the functional generalized linear model becomes

$$Y_i = \begin{cases} 1 & \text{with probability } (1+\exp(-\beta_0-\beta_1^T\gamma_i))^{-1}, \\ 0 & \text{with probability } (1+\exp(\beta_0+\beta_1^T\gamma_i))^{-1}, \end{cases} \tag{9}$$

$$\mathbf{x}_i = S_i\gamma_i+\mathbf{e}_i, \quad \mathbf{e}_i \sim N(0,\sigma_x^2 I), \quad \gamma_i \sim N(\mu_\gamma,\Gamma). \tag{10}$$

An algorithm for fitting this model and examples of its application are presented respectively in appendix A.3 and section 4.3. In general, a new response is predicted as 1 if $E(Y|\mathbf{x}) = P(Y = 1|\mathbf{x}) > 0.5$ and 0 otherwise. Thus we predict $Y = 1$ if

$$\left(1+\exp(-\beta_0-\beta_1^T\mu_{\gamma|x})\right)^{-1} > 0.5$$

where $\mu_{\gamma|x}$ is computed using (6). It should be noted that, while

$$P(Y = 1|\mathbf{x}) > 0.5 \text{ iff } \left(1+\exp(-\beta_0-\beta_1^T\mu_{\gamma|x})\right)^{-1} > 0.5,$$

in general $\left(1+\exp(-\beta_0-\beta_1^T\mu_{\gamma|x})\right)^{-1}$ will provide a biased estimate of $P(Y=1|\mathbf{x})$. A method to obtain unbiased probability estimates is presented in section 4.3.

3 The fitting procedure

In this section we outline a general approach to fitting the functional generalized linear model. The observed data likelihood for this model is extremely difficult to optimize directly because the γ_i's are unobserved. However, if the γ_i's are treated as missing data then $\mathbf{x}_i$ and Y_i are conditionally independent and the full data log likelihood factors, up to additive constants, into three distinct parts:

$$l(\mu_\gamma,\Gamma,\sigma_x^2,\beta_0,\beta_1,\phi) = \sum_{i=1}^{N}\left[\frac{y_i\theta_i-b(\theta_i)}{a(\phi)}+c(y_i,\phi)\right] \quad (11)$$

$$- \sum_{i=1}^{N}\left[\frac{n_i}{2}\log\sigma_x^2+\frac{1}{2\sigma_x^2}\left(\mathbf{x}_i-S_i\gamma_i\right)^T\left(\mathbf{x}_i-S_i\gamma_i\right)\right] \quad (12)$$

$$- \sum_{i=1}^{N}\left[\frac{1}{2}\log|\Gamma|+\frac{1}{2}(\gamma_i-\mu_\gamma)^T\Gamma^{-1}(\gamma_i-\mu_\gamma)\right]. \quad (13)$$

We use the EM algorithm (Dempster *et al.*, 1977; Laird and Ware, 1982) which iterates between a maximization and an expectation step to optimize the observed likelihood. Since θ_i is a function of β_0 and β_1 the M-step involves maximizing the expected values of (11) with respect to β_0,β_1 and ϕ, (12) with respect to σ_x^2 and (13) with respect to μ_γ and Γ. The three maximizations all involve separate parameters so the M-step can be performed in three parts. Maximizing the expected value of (12) and (13) involves setting

$$\sigma_x^2 = \frac{1}{\sum n_i}\sum_{i=1}^{N}\left[\left(\mathbf{x}_i-S_i\hat{\gamma}_i\right)^T\left(\mathbf{x}_i-S_i\hat{\gamma}_i\right)+\text{trace}\left(S_iV_{\gamma_i}S_i^T\right)\right], \quad (14)$$

$$\mu_\gamma = \frac{1}{N}\sum_{i=1}^{N}\hat{\gamma}_i, \quad (15)$$

$$\Gamma = \frac{1}{N}\sum_{i=1}^{N}\left[V_{\gamma_i}+(\hat{\gamma}_i-\mu_\gamma)(\hat{\gamma}_i-\mu_\gamma)^T\right], \quad (16)$$

where,

$$\hat{\gamma}_i = E(\gamma_i|\mathbf{x}_i,Y_i), \quad V_{\gamma_i} = Var(\gamma_i|\mathbf{x}_i,Y_i). \quad (17)$$

The maximization procedure for (11) depends on the distribution of Y_i.

The E-step consists of calculating the expected value and variance of the γ_i's given $\mathbf{x}_i$, Y_i and the current estimates of the parameters. The procedure for calculating these values also depends on the distribution Y_i. For certain distributions there is a closed form equation while for others a Monte Carlo approach must be adopted.

The fitting procedure iterates through these steps until the parameters have converged. Details of the fitting procedures for the linear, censored and logistic regression models are presented in the appendixes.

Method	Simulation					
	1	2	3	4	5	6
Simple Regression	97.2	97.2	97.2	78.6	78.6	100.2
Mean Regression	92.3	92.3	92.3	32.7	32.7	100.5
Multiple Regression	91.3	91.3	91.3	9.6	9.6	103.3
Filtering Regression	166.5	98.8	NA	182.0	NA	72.1
Functional Regression	23.6	88.3	37.9	2.7	2.1	13.8
Optimal Regression	22.7	22.7	22.7	1.2	1.2	13.7

Table 2: Results from the simulation study of section 4.1. The FGLM method produces close to optimal results for all but the second simulation.

4 Examples

In this section we present several applications of the functional generalized linear models approach. In section 4.1 we use a simulation study to illustrate some situations in which the functional linear regression approach can be expected to provide advantages over other methods. In sections 4.2 and 4.3 the functional censored and functional logistic regression methods are applied to the primary biliary cirrhosis (PBC) data set.

4.1 Functional Linear Regression Simulation Study

We performed six simulations whose purpose was to predict a time independent real valued response based on longitudinal observations. Four different "straw men" were compared to the functional regression approach of this paper. The results are summarized in table 2. The first approach, *simple regression*, used the first observation for each individual as the predictor in a standard linear regression. The second, *mean regression*, used the average of the observations for each individual as the sole predictor. In the third approach, *multiple regression* was performed on the time-ordered observations. This method could only be adopted because every individual had observations at the same number of time points. The final method, *filtering regression*, was a less simplistic approach in which a natural cubic spline was fit to each individual's observations and the spline coefficients were used as the predictors in a multiple regression.

Simulation 1 involved producing data from the functional linear regression model of section 2.2.1. First 100 random γ's were generated from a six-dimensional normal distribution with an identity covariance matrix. Next, 100 curves were created by using each of the γ's as the coefficients of a natural cubic spline with four equally spaced knots. The curve fragments to be used as the predictors, were then generated by sampling at six random locations from each of the curves and adding a small amount of random normal noise. Finally, the response variables were generated by multiplying each of the γ's by a fixed six-dimensional vector and adding a small amount of random normal noise. All methods were fit using this training data. A separate test set of size 1000 was generated using the same distribution. The average squared deviations of the predictions of each method from the observed test set responses are recorded in the first column of table 2. The deviations have been standardized by dividing by the mean squared deviation achieved by simply using the mean response from the training data. For example the mean squared deviation of the simple regression method was 97.2% of that achieved by using the mean of the training responses. The numbers are analogous to $1-R^2$ but calculated on the new test data. The final row of table 2 lists results for the optimal regression which uses $E(Y|\mathbf{x})$ calculated from the true simulation distribution and hence has the lowest possible squared deviation. Naturally the optimal regression approach could not be used on a real data set where the distribution is unknown. The functional regression method performed favorably with deviations of 23.6% compared

to 22.7% for the optimal regression. The straw man methods all performed poorly. The simple, mean and multiple regression methods worked slightly better than the sample mean i.e. had results slightly less than 100%. However, the filtering regression approach produced considerably worse results.

The filtering and functional regression methods both require the choice of a spline basis or, equivalently, knot locations. In simulation 1 the correct knot locations were used. In simulations 2 and 3 we illustrate the effect of incorrect knot selection. Simulation 2 fitted the functional and filtering regression methods using a two knot spline while simulation 3 fitted a six knot spline. Both simulations used the test and training data from simulation 1. In simulation 2 the filtering method improved to a level comparable to the other straw men and the performance of the functional regression method declined considerably, although it still produced results superior to those of the other approaches. In the third simulation the filtering regression method could not be used because at least eight observations were required to fit a natural cubic spline with six knots and each curve had only six measurements. However, there was no difficulty in fitting the FGLM model and its predictions were again considerably better than those of the other approaches. In general it has been our experience that the FGLM method suffers less from an overly flexible spline basis than it does from an overly rigid basis.

Simulations 4 and 5 explored the effect of violations of several of the FGLM model assumptions. Instead of the assumed spline basis, the training and test data were generated according to a fourth degree polynomial with no knots. In addition, the response and predictor error terms and the γ's were produced using a t distribution with 10 degrees of freedom. In simulation 4 the FGLM and filtering regression methods were fit using a three knot spline while a six knot spline was used in simulation 5. The optimal regression results were again obtained using the correct model specification. Despite the incorrect modeling assumptions the FGLM approach still produced results close to optimal, with the six knot spline giving a slightly better fit than the three knot spline. Of the straw men the multiple regression method worked best but was still clearly inferior to FGLM. The filtering regression method again performed poorly with no fit possible for the six knot spline. In general, in sparse data situations, the filtering approach works best when using low dimensional bases. For example, when using a two knot spline the standardized mean squared error for the filtering regression reduced to 17.1%.

While this article has concentrated on spline bases, any finite dimensional basis can be equally easily applied. To illustrate the improvements that are possible when using FGLM, even for relatively simple curves, the final simulation made use of orthogonal polynomials. The training and test data were produced in a similar fashion to those of the first simulation except that the curves were generated from a standard cubic with no constant term. The filtering and functional regression procedures were then fit using a three dimensional orthogonal polynomial basis instead of a spline basis. This situation is more favorable to the filtering approach because only three parameters need be estimated for each curve as opposed to six in the first simulation. However, while the filtering method does perform considerably better than the other straw-men, the functional regression approach still provides a significantly superior fit.

These simulations concentrate on situations in which the predictors have been sparsely sampled. In these circumstances the FGLM approach generally provides considerable improvement over the filtering method. In data sets with a large number of observations per predictor the two methods will give more similar results.

4.2 Functional Censoring Example

In this section we illustrate the FGLM approach on the primary biliary cirrhosis (PBC) data set. We choose to predict life expectancy, from date of enrollment in the study, based on the first four measurements of bilirubin level. For each individual the number of observations varied from one to sixteen. After removing all those patients with fewer than four observations 169 remained of whom 65 died prior to the end of the study. Since 104 of the responses were right censored, we fit the functional censored regression model of section 2.2.2.

Coefficient	Estimate	Std. Error	t	P-value
β_0	5268.66	355.2	14.83	< 0.001
β_{11}	7.60	2.2	3.39	0.001
β_{12}	-27.85	9.8	-2.85	0.004
β_{13}	-12.69	7.0	-1.80	0.071
β_{14}	-1.62	5.4	-0.30	0.765
β_{15}	36.73	8.8	4.18	< 0.001
β_{16}	27.89	18.7	1.49	0.136

Table 3: Estimated coefficients and significance levels for the Mayo Clinic trial using life expectancy as the response. Several of the coefficients are highly significant so there appears to be a relationship between predictor and response.

A natural cubic spline with four equally spaced knots, or equivalently six dimensions, was used to create the basis matrix. The average time of observation for patients who died was only seven days lower than for those who survived, indicating that there was no bias from differences in measurement times between censored and uncensored people.

The estimated coefficients are shown in table 3. β_{1j} is the jth element of β_1 and gives the weight applied to the corresponding element of γ in calculating $g(\mu)$. The standard error estimates were obtained using

$$Var\begin{bmatrix}\beta_0\\ \beta_1\end{bmatrix} \approx Var\left\{\left[E(A^TA)\right]^{-1}E(A^T)Y|\mathbf{x}\right\} = \left[E(A^TA)\right]^{-1}E(A^T)Var\{Y|\mathbf{x}\}E(A)\left[E(A^TA)\right]^{-1}, \qquad (18)$$

where A is an m by $q+1$ matrix whose ith row is $(1,\gamma_i^T)$. The various components of (18) are estimated during the fitting procedure so no extra computations are required. Equation (18) only makes use of the uncensored responses $Y_1,\ldots,Y_m$, and ignores the partial information provided by the censored responses. As a result it may tend to somewhat overestimate the standard errors. However, despite this fact, when the asymptotic normality of the maximum likelihood estimates is used to calculate the p-values, there are several highly significant coefficients. In addition to p-values for individual coefficients it is possible to estimate the overall significance level of the fit using the fact that

$$\beta_1^T Var(\beta_1)^{-1}\beta_1 \qquad (19)$$

will asymptotically have a χ^2_q distribution under the hypothesis of no relationship between $\mathbf{x}$ and Y. Note that $Var(\beta_1)$ can be calculated from (18). For this data set (19) gives a value of 68.75 with a corresponding p-value of less than 0.001. There is strong evidence of a relationship between bilirubin levels and life expectancy.

Generalized linear models can be used to either gain insights into the relationship between predictor and response or to estimate the response given values of a predictor. Unfortunately, there is no simple way to relate the coefficients in table 3 to the predictor, bilirubin. The best we can say is that a relationship between $\mathbf{x}$ and Y exists. We address this problem in detail in section 5. However, using the FGLM method to make predictions is straight forward. Censored responses are estimated by using (8) and uncensored responses by using $E(Y|\mathbf{x})$ from (7).

This data set was collected to study the effectiveness of the drug D-penicillamine. To test whether there was any effect from the drug on life expectancy we fit the functional censoring model separately to the control and treatment groups and estimated life expectancy on the censored observations. There was no apparent improvement for those on the medication. In fact there was some evidence that the treatment group may be performing worse than the control group with a mean life expectancy of 4237 days for the former and 4618

for the latter. However, the mean survival times are not significantly different at the 5% level. In section 6 we directly examine the drug effect by extending the model to include multiple predictors.

4.3 Functional Logistic Regression Example

In this section we illustrate the binary response version of FGLM to model five year survival on the PBC data set. The framework is identical to that of section 4.2 except that now Y equals one or zero depending on whether the patient did or did not survive five years from date of registration. Three individuals who were studied for less than five years were removed from the data set. As with regular logistic regression the standard errors of β_0 and β_1 can be estimated using the diagonal elements of the inverse Hessian,

$$Var\begin{bmatrix}\beta_0\\ \beta_1\end{bmatrix} \approx \left(\sum_i E\begin{bmatrix}\pi_i(1-\pi_i) & \pi_i(1-\pi_i)\gamma_i^T\\ \pi_i(1-\pi_i)\gamma_i & \pi_i(1-\pi_i)\gamma_i\gamma_i^T\end{bmatrix}\right)^{-1}, \tag{20}$$

where $\pi_i = P(Y_i = 1|\mathbf{x}_i)$. The Hessian is computed during the fitting process so no extra calculations are required. It is possible to estimate the overall significance level of the fit using (19) where $Var(\beta_1)$ is calculated from (20). For this response the chi-square statistic is 22.54 with a corresponding p-value of 0.002. Hence there is strong evidence of a relationship between bilirubin observations and the probability of an individual surviving five years.

Once again, interpretation of the β_1 coefficients poses a problem which we address in Section 5. However, prediction of responses can still be achieved with relative ease. Generally one predicts $Y = 1$ iff

$$\pi = P_{Y,\gamma}(Y = 1) = E_\gamma(1+\exp(-\beta_0 - \beta_1^T\gamma))^{-1} > 0.5.$$

It is possible to estimate π in a variety of ways. The most accurate approach is to simulate $\gamma_1^*, \ldots, \gamma_n^*$ from (6) and produce the Monte Carlo estimate,

$$\pi_{MC}^n = \frac{1}{n}\sum_{j=1}^{n}(1+\exp(-\beta_0 - \beta_1^T\gamma_j^*))^{-1}. \tag{21}$$

Equation (21) can be calculated quickly for moderate n. However, it may be computationally expensive if a large number of estimates are required. A simple alternative is to use the biased "plug in" approximation,

$$\pi_{Plug} = (1+\exp(-\beta_0 - \beta_1^T E(\gamma|\mathbf{x})))^{-1},$$

where $E(\gamma|\mathbf{x})$ is given by (6). It can be shown that $\pi_{Plug} > 0.5$ iff $\pi_{MC}^{\infty} > 0.5$. Hence, if prediction of Y is the sole objective and computational speed is an issue π_{Plug} may be used. However, if accurate probability estimates are required the more precise Monte Carlo procedure is preferable. The functional logistic regression approach produces fairly accurate predictions for the PBC data with only 15 out of 166 patients misclassified.

5 Assessing the relationship between predictor and response

Although the results from sections 4.2 and 4.3 clearly showed a relationship between bilirubin levels and both life expectancy and five year survival, they gave no simple way of understanding the form of the relationship. For instance, in standard linear regression a positive sign on the β coefficient indicates a positive relationship between $\mathbf{x}$ and Y, but for FGLM there is no such simple explanation. In this section we develop two methods to improve the interpretability of the FGLM results. In section 5.1 we use a functional principal components

decomposition of the predictor curves and in section 5.2 we show how to compute the weight function $\omega_1(t)$ from (3). The function $\omega_1(t)$ is the analogue of the coefficients in a standard linear regression.

5.1 Functional principal components

We can decompose the covariance matrix of the γ_i's into

$$\Gamma = \Delta D \Delta^T,$$

where Δ is the q by q matrix of eigenvectors of Γ and D is the diagonal matrix of eigenvalues. Then an alternative parameterization of γ is

$$\gamma = \Delta\alpha,$$

where $Var(\alpha) = D$. Under this formulation the functional generalized linear model of section 2.1 can be rewritten as

$$\begin{aligned} p(y_i;\theta_i,\phi) &= \exp\left(\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i,\phi)\right) \\ g(\mu_i) &= \beta_0' + {\beta_1'}^T\alpha_i, \quad \alpha_i \sim N(0,D) \\ \mathbf{x}_i|\alpha_i &= S_i(\mu_\gamma + \delta_1\alpha_{i1} + \cdots + \delta_q\alpha_{iq}) + \mathbf{e}_i, \quad \mathbf{e}_i \sim N(0,\sigma_x^2 I) \end{aligned} \tag{22}$$

where

$$\beta_0' = \beta_0 + \beta_1^T\mu_\gamma, \quad \text{and} \quad \beta_1' = \Delta^T\beta_1. \tag{23}$$

Note that (22) provides a functional principal components decomposition of $X(t)$ (Shi *et al.*, 1996; Ramsay and Silverman, 1997; James *et al.*, 2000). This model is easily fit by implementing the EM algorithm of section 3, performing an eigenvalue decomposition of the fitted Γ to produce Δ and D, and using (23) to obtain the transformed coefficients. This model's principal advantage is that each β coefficient has a natural interpretation in terms of the functional principal component curves of the predictor. For example, if the identity link is used, β_{11}' gives the average increase in Y_i for a one unit increase in α_{i1}. From (22) we note that a one unit increase in α_{i1} will cause $X_i(t)$ to increase by $\mathbf{s}(t)^T\delta_1$ above the population average. Hence, β_{11}' can be interpreted as the average increase in Y_i when $X_i(t)$ increases by $\mathbf{s}(t)^T\delta_1$.

Figure 1 illustrates the principal components approach on the fits of sections 4.2 and 4.3 where bilirubin was used as a predictor of life expectancy and five year survival. The plot gives the mean and first three principal component curves, for bilirubin levels, with life expectancy as the response. The curves when using five year survival as the response where almost identical. The dotted lines correspond to 90% pointwise confidence intervals. They were produced by generating bootstrapped γ's from a normal distribution with mean μ_γ and variance Γ and then using these γ's to estimate a new variance matrix Γ^* and hence new principal component curves. This procedure was repeated 100 times and at each time point the 5th and 95th largest values for the curves were taken as the upper and lower confidence bounds. Upon examining the mean curve it appears that the average bilirubin level is increasing over time. Liver failure is generally associated with increasing bilirubin levels so, as a whole, the population is getting sicker.

The first principal component curve models individuals with bilirubin levels starting above the average and continuing to increase. Thus an individual with average bilirubin levels will have an α_{i1} of zero, one with bilirubin levels above average and continuing to climb will have a positive α_{i1} and one with bilirubin levels below average and declining will have a negative α_{i1}. Table 4 gives the transformed coefficients using both life expectancy and five year survival as the response. Notice that β_{11}' is negative for both responses (-10.41 and -0.016) indicating that, for example, an individual with bilirubin levels initially 0.2 mg/dl above the

	Life Expectancy				Five Year Survival			
Coef.	Estimate	Std. Error	t	P-value	Estimate	Std. Error	t	P-value
β'_0	4440.2	268.0	16.57	< 0.001	2.88	0.417	6.91	< 0.001
β'_{11}	-10.41	1.9	-5.52	< 0.001	-0.016	0.004	-4.60	< 0.001
β'_{12}	0.11	6.1	0.02	0.986	0.021	0.011	2.03	0.043
β'_{13}	1.54	10.5	0.15	0.883	-0.020	0.019	-1.04	0.297
β'_{14}	-16.53	9.3	-1.77	0.077	-0.079	0.046	-1.73	0.083
β'_{15}	-50.98	14.6	-3.50	0.001	-0.192	0.100	-1.91	0.056
β'_{16}	11.93	12.5	0.95	0.340	-0.007	0.370	-0.02	0.985

Table 4: Transformed coefficients and significance levels for the Mayo Clinic trial using life expectancy and five year survival as the response. The first and fifth principal component curves are significant for life expectancy while the first and second are significant for five year survival.

mean and increasing by day 800 to 0.35 mg/dl above average can be expected to live 104 days below the average and has an odds ratio of $e^{-0.16} = 0.85$ of surviving five years relative to an average person. Since increasing bilirubin levels are generally associated with liver failure this is a clinically sensible result. The first principal component explains 91% of the variance in the γ's and is highly significant for both the life expectancy and five year survival responses. Clearly this is an important predictor.

The second principal component curve models individuals with below average bilirubin levels at the start that, relative to the mean, increase and then begin to decline again. This component explains about 6% of the variability in the γ's and is marginally significant for five year survival but not life expectancy. An individual with bilirubin levels initially 0.2 mg/dl below the mean increasing to 0.4 mg/dl above the mean by day 600 and then declining again will have an odds ratio of $e^{0.21} = 1.23$ of surviving five years relative to an average person. The third component models a cubic pattern. It only explains about 2% of the variability in the γ's and is not a significant predictor.

The other three components explain only a small fraction of the variability in the γ's. Interestingly the fifth principal component is significant for life expectancy. This suggests that we may be underestimating the standard errors for the higher level principal components. As one would expect, given that the γ's exhibit less variability in these directions, the higher level curves generally have greater standard errors. Notice also that the life expectancy coefficient $\beta'_0 = 4440$ can now be interpreted as the expected survival time of an individual with bilirubin levels equal to the population average. Also the five year survival coefficient $\beta'_0 = 2.88$ gives the probability of an average individual surviving five years as $1/(1+e^{-2.88}) = 0.95$.

5.2 Weight function

An alternative method for visualizing the relationship between predictor and response is to plot the weight function, $\omega_1(t)$. This function gives the weight placed on the predictor $X(t)$ at each time in determining the value of the response. High absolute values of the curve indicate times with a large influence while small values correspond to periods with little influence. We can model $\omega_1(t)$ using the same spline basis as for the predictors $X_i(t)$ i.e. $\omega_1(t) = \eta_1^T \mathbf{s}(t)$. In this case, provided $\mathbf{s}(t)$ is an orthogonal basis,

$$\int \omega_1(t) X_i(t) dt = \eta_1^T \int \mathbf{s}(t)\mathbf{s}(t)^T dt \gamma_i = \eta_1^T \gamma_i.$$

Hence, an alternative formulation of (5) would be

$$g(\mu_i) = \beta_0 + \eta_1^T \gamma_i.$$

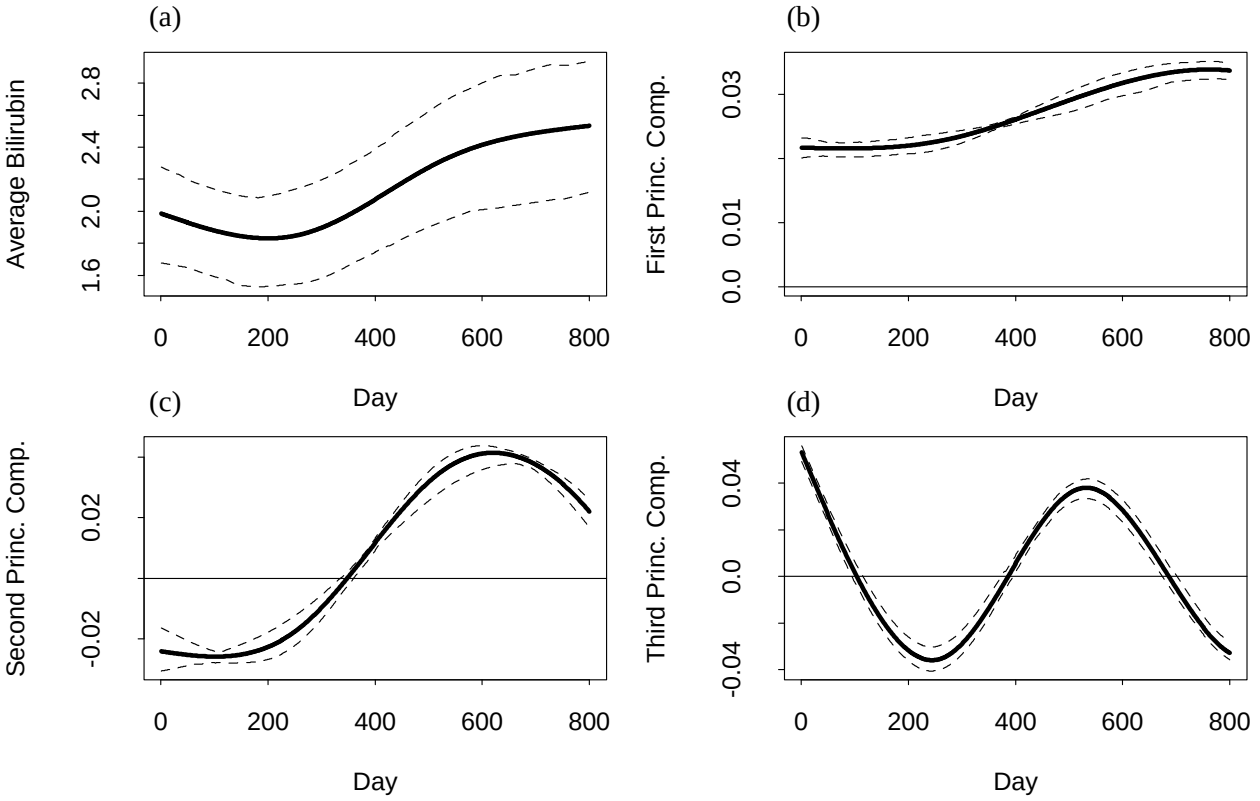

Figure 1: *Mean and first three principal component curves for bilirubin. Solid lines are the estimates while the dotted lines are* 90% *pointwise confidence intervals.*

As a result η_1 and β_1 are identical and $\omega_1(t)$ can be estimated by using $\beta_1^T \mathbf{s}(t)$. Figure 2 shows a plot of $\omega_1(t)$ for the PBC data using life expectancy as the response. The weights are negative in the early and late time periods indicating that people with high bilirubin levels in these stages have lower life expectancies. This confirms the results from the previous section. Interestingly, the middle time periods have slightly positive coefficients indicating higher life expectancy for high bilirubin levels between days 200 to 600. However, this result must be interpreted carefully because patients with high levels in this time period will likely have high bilirubin levels at the early and late time periods also.

6 Multiple predictor variables

The functional generalized linear model of section 2.1 assumes only one functional predictor variable. In practice one may wish to perform a regression with multiple functional and/or finite dimensional predictors. For the *i*th person we denote the observations from the jth functional predictor by $\mathbf{x}_{ij}$, the corresponding spline basis matrix by S_{ij} and the vector of fixed dimensional predictors by $\mathbf{z}_i$. Then, in analogy with generalized linear models, we assume a linear relationship between predictors. When fitting p functional predictors

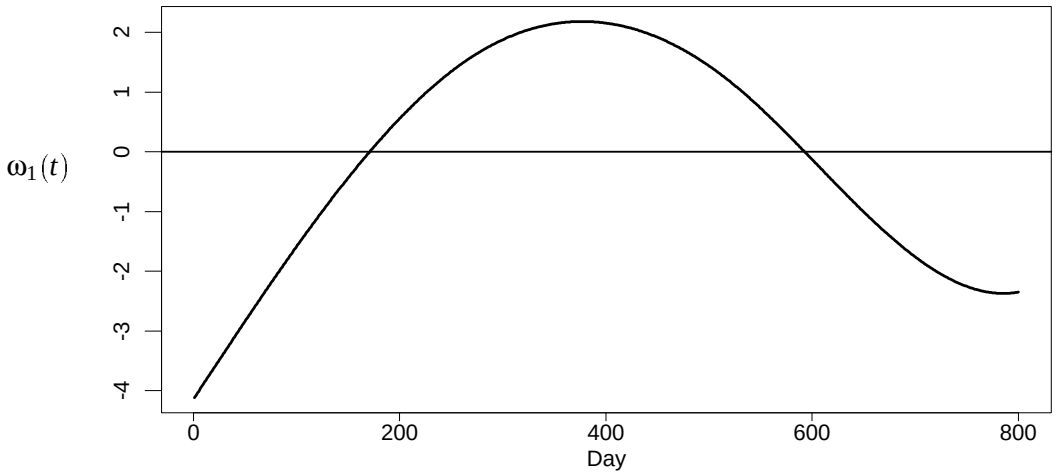

Figure 2: *A plot of* $\omega_1(t)$*. This curve gives the weight placed on the bilirubin level for an individual at any given time.*

the FGLM model becomes

$$\begin{aligned} p(y_i;\theta_i,\phi) &= \exp\left(\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i,\phi)\right), \\ g(\mu_i) &= \beta_0 + \beta_Z^T \mathbf{z}_i + \sum_{j=1}^{p} \beta_j^T \gamma_{ij}, \quad \gamma_{ij} \sim N(\mu_j, \Gamma_j), \\ \mathbf{x}_{ij} &= S_{ij}\gamma_{ij} + \mathbf{e}_i, \quad \mathbf{e}_i \sim N(0, \sigma_{xj}^2 I). \end{aligned}$$

Here β_j denotes the coefficient vector for the jth functional predictor and β_Z the coefficient vector for the finite dimensional predictors. The mean and variance of γ_{ij} for the jth predictor are given by μ_j and Γ_j and the γ_{ij}'s are assumed independent. As with the model of section 2.1, the extended FGLM model can be fit to a large variety of response variables. In particular it provides generalizations of multiple linear and logistic regression as well as censored regression with multiple predictors.

The functional multiple generalized linear model is fit using an EM procedure similar to that of section 3. Details of the functional multiple linear regression (FMLR) fitting algorithm are provided in appendix A.4. To aid interpretation one can decompose each of the functional predictors into their respective principal component curves and transform the coefficients according to

$$\beta_0' = \beta_0 + \sum_{j=1}^{p} \beta_j^T \mu_j, \quad \text{and} \quad \beta_j' = \Delta_j^T \beta_j, \tag{24}$$

where Δ_j is the matrix of eigenvectors of Γ_j.

To illustrate this approach we used two functional predictors, bilirubin and albumin levels, as well as an indicator of D-penicillamine use, to predict life expectancy. The results are shown in figure 3 and table 5. Figure 3 illustrates the mean curve and first two principal component curves for both bilirubin and albumin. The bilirubin curves are extremely similar to those of section 5. The mean curve for albumin indicates decreasing levels. A healthy liver secretes albumin so this is further indication of a sickening population. The first principal component models individuals with above average albumin levels with a slight increase and then decrease relative to the mean while the second component models a cubic relationship.

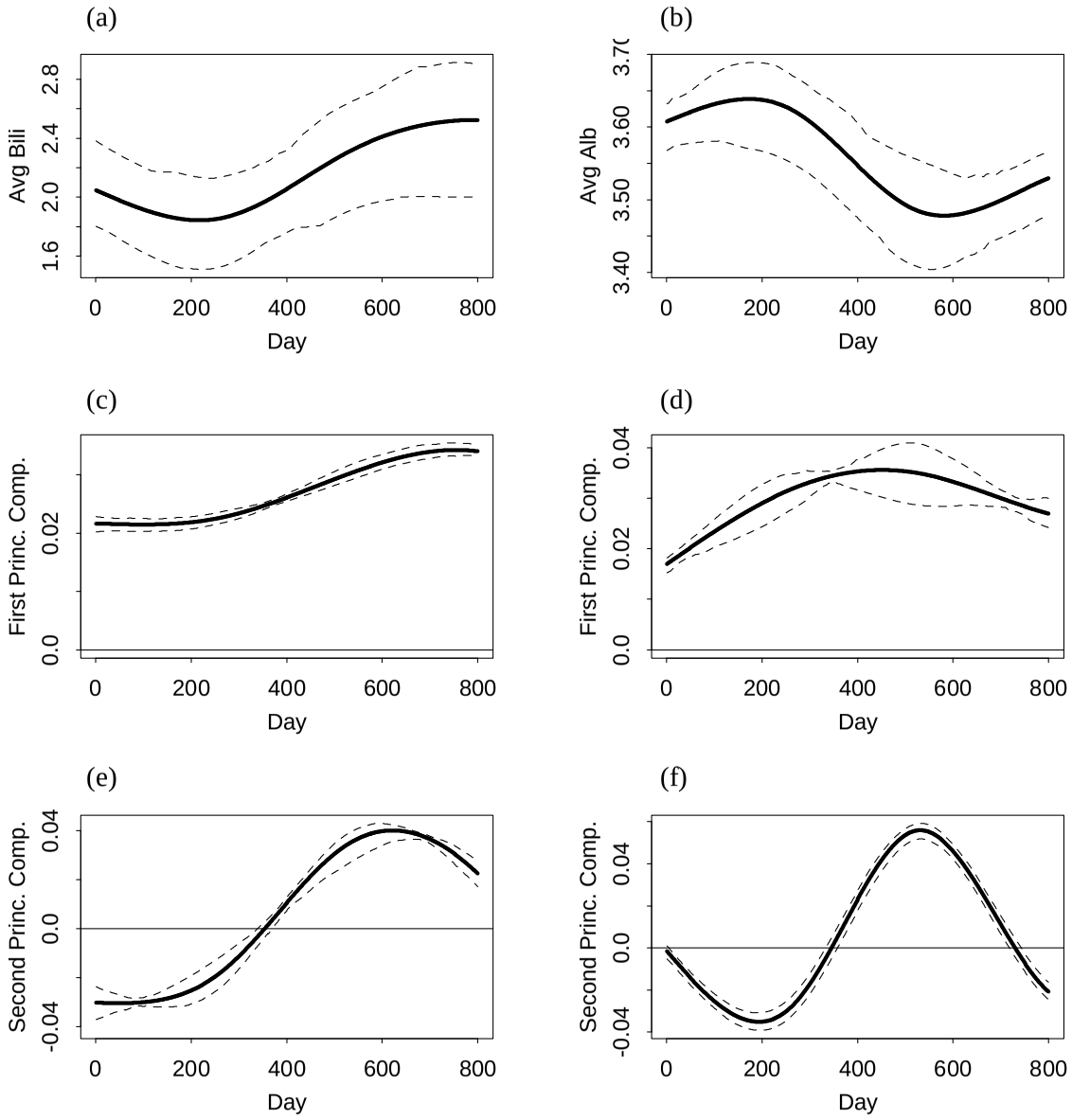

Figure 3: *Mean and first two principal component curves for bilirubin (a, c and e) and albumin (b, d and f).*

Coef.		Est.	Sd Err	t	P-value	Coef.		Est.	Sd. Err	t	P-value
	β'_0	4399.2	392.2	11.22	< 0.001		β_Z	-207.5	520.1	-0.40	0.690
	β'_{11}	-11.2	2.9	-3.90	< 0.001		β'_{21}	75.9	27.7	2.73	0.006
	β'_{12}	4.1	9.3	0.44	0.657		β'_{22}	64.4	36.6	1.76	0.078
Bili-	β'_{13}	-6.1	13.2	-0.46	0.646	Alb-	β'_{23}	42.8	46.8	0.92	0.360
rubin	β'_{14}	45.4	15.8	2.88	0.004	umin	β'_{24}	-100.7	61.0	-1.65	0.099
	β'_{15}	-2.6	16.6	-0.16	0.874		β'_{25}	-312.6	70.4	-4.44	< 0.001
	β'_{16}	6.3	6.6	0.96	0.336		β'_{26}	498.9	80.9	6.17	< 0.001

Table 5: Transformed coefficients and significance levels for the PBC data using life expectancy as the response and bilirubin, albumin and drug as the predictors. The first and fourth principal component curves are significant for bilirubin while the first, fifth and sixth are significant for albumin. Drug is not a significant predictor.

The standard errors in table 5 are calculated using (18) where the ith row of A is now $(1, \gamma_{i1}, \ldots, \gamma_{ip}, \mathbf{z}_i)$. We note that for both bilirubin and albumin the first principal component is highly significant. The negative coefficient for β'_{11} and positive coefficient for β'_{21} imply increased life expectancy for individuals with lower bilirubin and higher albumin levels, both clinically sensible results. There is strong evidence that albumin and bilirubin are both significant predictors for life expectancy. One may also directly gauge the effect of the drug D-penicillamine on life expectancy by examining β_Z. The estimated coefficient is negative suggesting that the drug may actually reduce life expectancy, although the result is not statistically significant. Finally, notice that β_{14}, β_{25} and β_{26} are all significant. Since the first three principal components explain almost all the variability in bilirubin and albumin levels it seems likely that the model has underestimated the standard errors of these coefficients.

7 Missing data

This paper has focused on the situation where the predictors have a functional form so that a spline basis matrix S_i is appropriate. However, through the use of an alternative basis matrix, this methodology may be extended immediately to GLM with finite dimensional predictors that suffer from missing data. For finite, q-dimensional predictors, there will in general be no reason to assume a particular functional relationship between observations from adjacent dimensions so a spline basis matrix would be inappropriate. For an observation $\mathbf{x}_i$, with the first m dimensions observed and the remaining $q - m$ missing, a natural choice would be to replace S_i by an m by q matrix with the first m columns consisting of the identity matrix and remaining entries zero. This parameterization allows the data to directly model any relationship between different dimensions through the covariance matrix Γ. Using this basis matrix, the model of section 2.1 provides a natural representation of the relationship between a predictor, containing missing data, and a response which means that the FGLM fitting procedure can be implemented immediately in this setting. The FGLM approach provides three useful tools for working with missing data. First, it enables GLMs to be fit when missing data is present in the predictors. Second, the estimated γ's can be used to impute the missing observations. Finally, the eigen-decomposition of Γ provides a method for performing principal components on missing data.

A simulation study was conducted to illustrate the improvements that are possible when using FGLM to perform linear regression with missing predictor values. Six simulations were conducted. For each a training data set was created by randomly sampling 50 γ's from a mean zero five-dimensional multivariate normal distribution. The first three data sets had correlations between each dimension of 0.5 while the other three

Method	$\rho = .5$			$\rho = .9$		
	5% miss.	10% miss.	20% miss.	5% miss.	10% miss.	20% miss.
Substitution	2.02	3.92	8.08	1.40	2.89	7.51
FGLM	1.64	2.28	5.20	0.32	0.61	0.92
Optimal	1.40	2.21	4.79	0.30	0.52	0.75

Table 6: Results from missing data simulations. The FGLM approach achieves squared deviations very close to that of the optimal fit while the substitution method has consistently higher deviations.

had correlations of 0.9. The responses were produced as a linear combination of the γ's plus a small amount of random normal noise. Finally, the predictors were created by randomly removing a fixed percentage of the elements from each γ, to simulate the missing data, and adding random normal measurement errors to the remaining elements. Three procedures were fit to these data sets. The first, a simple approach commonly applied in practice, substituted the mean of each dimension for any missing observations before applying a standard linear regression fit. We call this method *substitution regression*. The second was the FGLM approach using the basis matrix described earlier. The final method was the optimal fit where the response is predicted using the true distribution from which the data were simulated. This procedure gives the best possible estimates but cannot be used on real problems. We compared the fitted models for each of the procedures using a separate test set of 1000 observations drawn from the same distribution as the training data. The mean squared deviations between predictions and actual responses, standardized in the same manner as table 2, are shown in table 6. The first three columns show results for a correlation of 0.5 and data missing at random with probabilities 5%, 10% and 20% while the remaining three columns correspond to a correlation of 0.9. Notice that the FGLM approach is producing deviations very close to the lowest possible. The substitution regression method produces consistently higher deviations. As we would expect the FGLM approach provides the greatest advantages over the substitution method when there are larger correlations between dimensions and higher percentages of missing data.

Acknowledgments

I would like to thank the Trevor Hastie and Catherine Sugar as well as the referees and editor for numerous useful suggestions and comments.

A Appendix

A.1 Functional Simple Linear Regression Algorithm

When Y is assumed to have a normal distribution with the identity link function then the distribution of γ conditional on $\mathbf{x}$ and Y is normal with

$$\begin{aligned} Var(\gamma|\mathbf{x},Y) &= \left[\Gamma^{-1} + S^T S/\sigma_x^2 + \beta_1\beta_1^T/\sigma_y^2\right]^{-1}, \\ E(\gamma|\mathbf{x},Y) &= Var(\gamma|\mathbf{x},Y)\left[\Gamma^{-1}\mu_\gamma + S^T\mathbf{x}/\sigma_x^2 + \beta_1(y-\beta_0)/\sigma_y^2\right]. \end{aligned}$$

Hence

$$V_{\gamma_i} = \left[\Gamma^{-1} + S_i^T S_i/\sigma_x^2 + \beta_1\beta_1^T/\sigma_y^2\right]^{-1}, \tag{25}$$

$$\hat{\gamma}_i = V_{\gamma_i}\left[\Gamma^{-1}\mu_\gamma + S_i^T \mathbf{x}_i/\sigma_x^2 + \beta_1(y_i - \beta_0)/\sigma_y^2\right] \tag{26}$$

Furthermore, (11) reduces to

$$-\frac{1}{2}\sum_{i=1}^{N}\left[\log\sigma_y^2 + (y_i - \beta_0 - \beta_1^T\gamma_i)^2/\sigma_y^2\right]. \tag{27}$$

Hence, using standard least squares, the maximization of the expected value of (11) is achieved using (28) and (29),

$$\begin{bmatrix}\beta_0\\ \beta_1\end{bmatrix} = \begin{bmatrix} N & \sum_i \hat{\gamma}_i^T \\ \sum_i \hat{\gamma}_i & \sum_i (V_{\gamma_i} + \hat{\gamma}_i\hat{\gamma}_i^T)\end{bmatrix}^{-1}\begin{bmatrix}\sum_i y_i \\ \sum_i \hat{\gamma}_i y_i\end{bmatrix}, \tag{28}$$

$$\sigma_y^2 = \frac{1}{N}\sum_{i=1}^{N}\left[(Y_i - \beta_0 - \beta_1^T\hat{\gamma}_i)^2 + \beta_1^T V_{\gamma_i}\beta_1\right]. \tag{29}$$

Thus the **functional linear regression algorithm** iterates through a two step procedure until the parameters have converged.

1. In the E-step $\hat{\gamma}_i$ and V_{γ_i} are calculated according to (26) and (25) using the current estimates of the parameters.
2. Then in the M-step the parameters $\sigma_x^2, \mu_\gamma, \Gamma, \beta_0, \beta_1$ and σ_y^2 are calculated using (14), (15), (16), (28) and (29) respectively.
3. Return to 1. unless the parameters have converged.

A.2 Functional Censored Regression Algorithm

The functional censored regression model is almost identical to that of linear regression with the exception that a portion of the responses are assumed censored and hence missing. The EM algorithm has been commonly applied to this problem when the predictor has finite dimension (Schmee and Hahn, 1979). One still estimates σ_x^2, μ_γ and Γ using (14), (15) and (16). Likewise, for responses that have been observed $\hat{\gamma}_i$ and V_{γ_i} are still calculated using (26) and (25). There is no simple closed form solution for $\hat{\gamma}_i$ and V_{γ_i} when the response is censored but they may still be computed using a simple form of Monte Carlo estimation. For the *i*th censored response, generate a sample, $\gamma_1^*, \ldots, \gamma_n^*$, according to the distribution given by (6). Then an unbiased estimate for $\hat{\gamma}_i$ is,

$$\hat{\gamma}_i = \frac{\sum_{j=1}^{n}\gamma_j^* P(Y_i > c_i|\gamma_j^*)}{\sum_{j=1}^{n} P(Y_i > c_i|\gamma_j^*)}, \tag{30}$$

where $P(Y_i > c_i|\gamma^*) = 1 - \Phi\left(\frac{c_i - \beta_0 - \beta_1^T\gamma^*}{\sigma_y}\right)$. Similarly V_{γ_i} can be estimated using

$$V_{\gamma_i} = \frac{\sum_{j=1}^n \gamma_j^* \gamma_j^{*T} P(Y_i > c_i|\gamma_j^*)}{\sum_{j=1}^n P(Y_i > c_i|\gamma_j^*)} - \hat{\gamma}_i \hat{\gamma}_i^T. \tag{31}$$

β_0 and β_1 are estimated using a similar equation to (28),

$$\begin{bmatrix}\beta_0 \\ \beta_1\end{bmatrix} = \begin{bmatrix} N & \sum_i \hat{\gamma}_i^T \\ \sum_i \hat{\gamma}_i & \sum_i (V_{\gamma_i} + \hat{\gamma}_i \hat{\gamma}_i^T)\end{bmatrix}^{-1} \begin{bmatrix} \sum_{i=1}^m y_i + \sum_{i=m+1}^N E(Y_i|Y_i > c_i, \mathbf{x}_i) \\ \sum_{i=1}^m \hat{\gamma}_i y_i + \sum_{i=m+1}^N E(\gamma_i Y_i|Y_i > c_i, \mathbf{x}_i)\end{bmatrix}, \tag{32}$$

where $E(Y_i|Y_i > c_i, \mathbf{x}_i)$ is given by (8) and $E(\gamma_i Y_i|Y_i > c_i, \mathbf{x}_i)$ can be estimated using the above mentioned Monte Carlo approach with

$$E(\gamma_i Y_i|Y_i > c_i, \mathbf{x}_i) = \frac{\sum_{j=1}^n \gamma_j^* E(Y_i|Y_i > c_i, \gamma_j^*) P(Y_i > c_i|\gamma_j^*)}{\sum_{j=1}^n P(Y_i > c_i|\gamma_j^*)}, \tag{33}$$

and

$$E(Y_i|Y_i > c_i, \gamma_j^*) = \beta_0 + \beta_1^T \gamma_j^* + \frac{\sigma_y \phi\left((c_i - \beta_0 - \beta_1 \gamma_j^*)/\sigma_y\right)}{1 - \Phi\left((c_i - \beta_0 - \beta_1 \gamma_j^*)/\sigma_y\right)}.$$

Finally σ_y^2 is estimated using

$$\sigma_y^2 = \frac{1}{N}\sum_{i=1}^N E(\varepsilon_i^2|Y_i, \mathbf{x}_i) = \frac{1}{N}\left[\sum_{i=1}^m \left((Y_i - \beta_0 - \beta_1^T \hat{\gamma}_i)^2 + \beta_1^T V_{\gamma_i} \beta_1\right) + \sum_{i=m+1}^N E(\varepsilon_i^2|Y_i > c_i, \mathbf{x}_i)\right] \tag{34}$$

where an unbiased estimate of $E(\varepsilon_i^2|Y_i > c_i, \mathbf{x}_i)$ is obtained using the Monte Carlo approach with

$$E(\varepsilon_i^2|Y_i > c_i, \mathbf{x}_i) = \frac{\sum_{j=1}^n \left[\sigma_y^2 P(Y_i > c_i|\gamma_j^*) + \sigma_y (c_i - \beta_0 - \beta_1^T \gamma_j^*) \phi\left(\frac{c_i - \beta_0 - \beta_1^T \gamma_j^*}{\sigma_y}\right)\right]}{\sum_{j=1}^n P(Y_i > c_i|\gamma_j^*)}. \tag{35}$$

Thus the **functional censored regression algorithm** iterates through a two step procedure until the parameters have converged.

1. In the E-step we calculate the expected values and variances of the γ_i's using (26), (30), (25) and (31) and the conditional expectations of Y_i, $Y_i\gamma_i$ and ε_i^2 given $Y_i > c_i$ using (8), (33) and (35)
2. Then in the M-step the parameters, $\sigma_x^2, \mu_\gamma, \Gamma, \beta_0, \beta_1$ and σ_y^2, are estimated using (14), (15), (16), (32) and (34) respectively.
3. Return to 1. unless the parameters have converged.

This procedure is an example of Monte Carlo EM (Tanner, 1994).

A.3 Functional Logistic Regression Algorithm

As with the previous two algorithms, the logistic regression fitting procedure estimates σ_x^2, μ_γ and Γ using (14), (15) and (16). There is no closed form solution for $\hat{\gamma}_i$ or V_{γ_i} so we again make use of Monte Carlo.

For the ith response, generate a sample, $\gamma_1^*, \ldots, \gamma_n^*$, according to the distribution given by (6). Then unbiased estimates for $\hat{\gamma}_i$ and V_{γ_i} are given by,

$$\hat{\gamma}_i = \frac{\sum_{j=1}^n \gamma_j^* P(Y = y_i|\gamma_j^*)}{\sum_{j=1}^n P(Y = y_i|\gamma_j^*)} \quad \text{and} \quad V_{\gamma_i} = \frac{\sum_{j=1}^n \gamma_j^* \gamma_j^{*T} P(Y = y_i|\gamma_j^*)}{\sum_{j=1}^n P(Y = y_i|\gamma_j^*)} - \hat{\gamma}_i \hat{\gamma}_i^T, \tag{36}$$

where $P(Y = y_i|\gamma_j^*)$ is calculated from (9).

In analogy with the iteratively re-weighted least squares approach used to fit standard logistic regression problems, the expected value of (11) can be maximized by choosing β_0 and β_1 such that

$$E\left(A^T W A\right) \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = E\left(A^T W Z\right) \tag{37}$$

where A is an N by $q+1$ matrix whose ith row is $(1, \gamma_i^T)$, W is a diagonal matrix with the ith diagonal $\pi_i(1-\pi_i)$, Z is a vector with ith element $\beta_0 + \gamma_i^T \beta_1 + (y_i - \pi_i)/\pi_i(1-\pi_i)$ and $\pi_i = P(Y_i = 1|\mathbf{x}_i)$. Since π_i depends on the current parameter estimate an iterative approach must be taken to solve this equation where β_0 and β_1 are updated by incrementing them by,

$$\left(E(A^T W A)\right)^{-1} E\left(A^T W Z\right) = \left(\sum_i E \begin{bmatrix} \pi_i(1-\pi_i) & \pi_i(1-\pi_i)\gamma_i^T \\ \pi_i(1-\pi_i)\gamma_i & \pi_i(1-\pi_i)\gamma_i\gamma_i^T \end{bmatrix} \right)^{-1} E \begin{bmatrix} \sum_i (y_i - \pi_i) \\ \sum_i (y_i - \pi_i)\gamma_i \end{bmatrix}. \tag{38}$$

Implementing (37) requires calculating $E(\pi_i(1-\pi_i)), E(\pi_i(1-\pi_i)\gamma_i), E(\pi_i(1-\pi_i)\gamma_i\gamma_i^T), E(y_i - \pi_i)$ and $E((y_i - \pi_i)\gamma_i)$. These quantities are estimated using

$$\frac{\sum_{j=1}^n U_j P(Y = y_i|\gamma_j^*)}{\sum_{j=1}^n P(Y = y_i|\gamma_j^*)} \tag{39}$$

where

$$U_j = \begin{cases} P(Y=1|\gamma_j^*)(1 - P(Y=1|\gamma_j^*)) & \text{for } E(\pi_i(1-\pi_i)) \\ \gamma_j^* P(Y=1|\gamma_j^*)(1 - P(Y=1|\gamma_j^*)) & \text{for } E(\pi_i(1-\pi_i)\gamma_i) \\ \gamma_j^* \gamma_j^{*T} P(Y=1|\gamma_j^*)(1 - P(Y=1|\gamma_j^*)) & \text{for } E(\pi_i(1-\pi_i)\gamma_i\gamma_i^T) \\ (y_i - P(Y=1|\gamma_j^*)) & \text{for } E(y_i - \pi_i) \\ \gamma_j^*(y_i - P(Y=1|\gamma_j^*)) & \text{for } E((y_i - \pi_i)\gamma_i). \end{cases}$$

Thus the **functional logistic regression algorithm** iterates through a two step procedure until the parameters have converged.

1. In the E-step the expected value and variance of the γ_i's are calculated using (36) and the expected values of $\pi_i(1-\pi_i), \pi_i(1-\pi_i)\gamma_i, \pi_i(1-\pi_i)\gamma_i\gamma_i^T, y_i - \pi_i$ and $(y_i - \pi_i)\gamma_i$ are calculated using (39).

2. In the M-step the parameters $\sigma_x^2, \mu_\gamma, \Gamma, \beta_0$ and β_1 are estimated using equations (14), (15), (16) and (37).

3. Return to 1. unless the parameters have converged.

In practice we have found that one step of (38) provides a reasonable estimate of the coefficients. This saves a great deal of computation as the various expected values only have to be calculated once per M-step.

A.4 Functional Multiple Linear Regression Algorithm

Let $\gamma_i'^T = (\gamma_{i1}^T, \ldots, \gamma_{ip}^T)$. When Y is assumed to have a normal distribution with the identity link function then the distribution of γ_i', conditional on $\mathbf{x}_{ij}$ and Y_i, is normal with

$$E(\gamma_i'|\mathbf{x}_{ij}, Y_i) = Var(\gamma_i'|\mathbf{x}_{ij}, Y_i)^{-1} \begin{bmatrix} \Gamma_1^{-1}\mu_1 + S_{i1}^T\mathbf{x}_{i1}/\sigma_{x1}^2 + (Y_i - \beta_0 - \beta_Z^T\mathbf{z}_i)\beta_1/\sigma_y^2) \\ \vdots \\ \Gamma_p^{-1}b\mu_p + S_{ip}^T\mathbf{x}_{ip}/\sigma_{xp}^2 + (Y_i - \beta_0 - \beta_Z^T\mathbf{z}_i)\beta_p/\sigma_y^2) \end{bmatrix}. \tag{40}$$

and

$$Var(\gamma_i'|\mathbf{x}_{ij}, Y_i) = \begin{bmatrix} \Gamma_1^{-1} + S_{i1}^T S_{i1}/\sigma_{x1}^2 + \beta_1\beta_1^T/\sigma_y^2 & \beta_1\beta_2^T/\sigma_y^2 & \cdots & \beta_1\beta_p^T/\sigma_y^2 \\ \vdots & \vdots & \ddots & \vdots \\ \beta_p\beta_1^T/\sigma_y^2 & \beta_p\beta_2^T/\sigma_y^2 & \cdots & \Gamma_p^{-1} + S_{ip}^T S_{ip}/\sigma_{xp}^2 + \beta_p\beta_p^T/\sigma_y^2 \end{bmatrix}^{-1} \tag{41}$$

Therefore $\hat{\gamma}_{ij} = E(\gamma_{ij}|\mathbf{x}_{ij}, Y_i)$ and $V_{\gamma_{ij}} = Var(\gamma_{ij}|\mathbf{x}_{ij}, Y_i)$ can be computed using (40) and (41). This in turn allows σ_{xj}^2, μ_j and Γ_j to be calculated using

$$\sigma_{xj}^2 = \frac{1}{\sum n_{ij}} \sum_{i=1}^{N} \left[(\mathbf{x}_{ij} - S_{ij}\hat{\gamma}_{ij})^T (\mathbf{x}_{ij} - S_{ij}\hat{\gamma}_{ij}) + \text{trace}\left(S_{ij} V_{\gamma_{ij}} S_{ij}^T \right) \right], \tag{42}$$

$$\mu_j = \frac{1}{N} \sum_{i=1}^{N} \hat{\gamma}_{ij}, \tag{43}$$

$$\Gamma_j = \frac{1}{N} \sum_{i=1}^{N} \left(V_{\gamma_{ij}} + (\hat{\gamma}_{ij} - \mu_j)(\hat{\gamma}_{ij} - \mu_j)^T \right). \tag{44}$$

Finally $\beta_0, \beta_1, \ldots, \beta_p, \beta_Z$ and σ_y^2 are estimated using,

$$\begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \\ \beta_Z \end{bmatrix} = \left[\sum_i E \begin{pmatrix} 1 & \gamma_i'^T & Z_i^T \\ \gamma_i' & \gamma_i'\gamma_i'^T & \gamma_i'\mathbf{z}_i^T \\ \mathbf{z}_i & \mathbf{z}_i\gamma_i'^T & \mathbf{z}_i\mathbf{z}_i^T \end{pmatrix} \right]^{-1} \begin{bmatrix} \sum_i y_i \\ \sum_i y_i E\gamma_i' \\ \sum_i y_i \mathbf{z}_i \end{bmatrix}, \tag{45}$$

and

$$\sigma_y^2 = \frac{1}{N} \sum_{i=1}^{N} \left((Y_i - \beta_0 - \beta_Z^T\mathbf{z}_i - \sum_{j=1}^{p} \beta_j^T\hat{\gamma}_{ij})^2 + \sum_{j=1}^{p} \beta_j^T V_{\gamma_{ij}} \beta_j \right). \tag{46}$$

Thus the **functional multiple linear regression algorithm** iterates through a two step procedure until the parameters have converged.

1. In the E-step the expected value and variance of the γ_i's are calculated using (40) and (41).
2. In the M-step the parameters σ_{xj}^2, μ_j and Γ_j are estimated using equations (42), (43) and (44) and $\beta_0, \beta_1, \ldots, \beta_p, \beta_Z$ are estimated using (45). Finally, σ_y^2 is estimated using (46).
3. Return to 1. unless the parameters have converged.

References

Cessie, S. L. and Houwelingen, J. C. V. (1994). Logistic regression for correlated binary data. *Applied Statistics* **43**, 95–108.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Ser. B* **39**, 1–22.

Diggle, P. J., Liand, K. Y., and Zeger, S. L. (1994). *Analysis of Longitudinal Data*. New York: Oxford University Press.

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

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

Gao, F., Wahba, G., Klein, R., and Klein, B. (2001). Smoothing spline ANOVA for multivariate bernoulli observations with applications to opthalmology data. *Journal of the American Statistical Association* **96**, 127–147.

Green, P. J. and Silverman, B. W. (1994). *Nonparametric Regression and Generalized Linear Models : A roughness penalty approach*. Chapman and Hall: London.

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

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

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., Hastie, T. J., and Sugar, C. A. (2000). Principal component models for sparse functional data. *Biometrika* **87**, 587–602.

Jones, B. and Kenward, M. G. (1989). *Design and Analysis of Cross-Over Trials*. London: Chapman Hall.

Laird, N. M. and Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics* **38**, 963–974.

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.

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.

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

Schmee, J. and Hahn, G. (1979). A simple method for regression analysis with censored data. *Technometrics* **21**, 417–432.

Shi, M., Weiss, R., and Taylor, J. (1996). An analysis of paediatric cd4 counts for acquired immune deficiency syndrome using flexible random curves. *Applied Statistics* **45**, 2, 151–164.

Silverman, B. W. (1985). Some aspects of the spline smoothing approach to non-parametric regression curve fitting (with discussion). *R. Statist. Soc. B* **47**, 1–52.

Tanner, M. (1994). *Tools for Statistical Inference*. Springer, 2nd edn.

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.