IMS Lecture Notes–Monograph Series
Optimality: The Third Erich L. Lehmann Symposium
Vol. 57 (2009) 173–189

DOI: 10.1214/09-LNMS5711

A Functional Generalized Linear Model with Curve Selection in Cervical Pre-cancer Diagnosis Using Fluorescence Spectroscopy

Hongxiao Zhu[1] **and Dennis D. Cox**[2]

Rice University

Abstract: A functional generalized linear model is applied to spectroscopic data to discriminate disease from non-disease in the diagnosis of cervical pre-cancer. For each observation, multiple functional covariates are available, and it is of interest to select a few of them for efficient classification. In addition to multiple functional covariates, some non-functional covariates are also used to account for systematic differences caused by these covariates. Functional principal components are used to reduce the model to multivariate logistic regression and a grouped Lasso penalty is applied to the reduced model to select useful functional covariates among multiple curves.

Contents

1. Introduction

Classification with functional data is a challenging problem due to the high dimensionality of the observation space. One solution is to reduce the dimension and use the reduced features for classification, such as the work of Hall et al. [6], Zhao et al. [19] and Ferré and Villa [5]. Another way is to use generalized linear regression by treating the class labels as responses and functional observations as predictors, which was proposed by James [8] and Müller and Stadtmüller [11]. Ratcliffe et al.

[1]Department of Statistics, Rice University, 6100 Main St. MS-138, Houston, Texas 77005, U.S.A., e-mail: hxzhu@stat.rice.edu
[2]Department of Statistics, Rice University, 6100 Main St. MS-138, Houston, Texas 77005, U.S.A., e-mail: dcox@stat.rice.edu
AMS 2000 subject classifications: 60K35; secondary 60K37.
Keywords and phrases: functional generalized linear model, curve selection, grouped lasso, fluorescence spectroscopy, cervical cancer.

[14] and Leng and Müller [9] applied this type of modeling to medical and gene expression data, respectively. Our basic concern in this study is the case when there are multiple functions per observation in a classification problem, and we wish to perform a curve selection to select few important curves and perform classification based on the selected curves.

The example that motivated our work is fluorescence spectroscopy data being investigated for cervical pre-cancer diagnosis. Fluorescence spectroscopy is an optical technique proposed for cervical pre-cancer screening. As a non-invasive, low-cost diagnosis tool, it provides a promising alternative to the existing methods for early-stage cancer diagnosis. One important step in this type of diagnosis is to discriminate the diseased observations from normal based on the high dimensional functional data — the fluorescence spectral measurements. In many clinical studies, several different spectra can be produced and used simultaneously for diagnosis ([12]), which makes the classification difficult since introducing more spectra not only provides more information but also more noise. Among these multiple spectral curves, it is suspected that some spectral curves contain more disease related information and hence are more "important" than others (see [3]). Furthermore, in order to produce an inexpensive commercial device, we would like to measure as few spectra as is necessary. This makes it beneficial to use statistical analysis to find out those curves that are good enough for diagnosis and remove the unnecessary ones, which can improve the diagnostic accuracy and reduce the cost.

The data studied in this paper are from a clinical study in which multiple fluorescence spectra were measured at the same sites where biopsies were taken for pathological diagnosis. Each observation consists of several spectral curves measured in the following way: an excitation light at a certain fixed excitation wavelength is produced to illuminate the cervical tissue. The excitation light is absorbed by various endogenous fluorescent molecules in tissue, resulting in emission of fluorescent light. The emitted fluorescent light is measured by an optical detector and the spectrum is obtained as one smooth curve. The excitation light is varied at several different wavelengths and gives multiple spectral curves for each measurement. The left panel of Figure 1 shows the plot of all spectral curves from one measurement. Each measurement contains 16 spectral curves measured at excitation wavelengths ranging from 330 nm to 480 nm with increments of 10 nm. Each spectral curve contains fluorescence intensities recorded on a range of emission wavelengths between 385nm and 700nm. If we use a color plot to represent the intensities, we can stack all the 16 spectra and obtain an image as shown in the right panel of Figure 1. We call such fluorescence spectroscopy measurements excitation-emission matrices (EEMs).

This study aims to select a subset of spectral curves from the 16 available curves for the purpose of classification. We will look at the problem from the functional data analysis ([13]) point of view and propose a functional generalized linear model, which will select among multiple functional predictors and perform binary classification. The proposed model allows both functional predictors and non-functional predictors. The non-functional predictors are variables associated with the measurements which may cause systematic difference in spectra, such as tissue type of the measurement site, or the menopausal status of patients.

The structure of this paper is as follows: Section 2 introduces the functional generalized linear model with curve selection and Section 3 provides a simulation study. The real data application to the fluorescence spectroscopy data is presented in Section 4, and details on determining related parameters are discussed in Section 5. A more general discussion is given in Section 6.

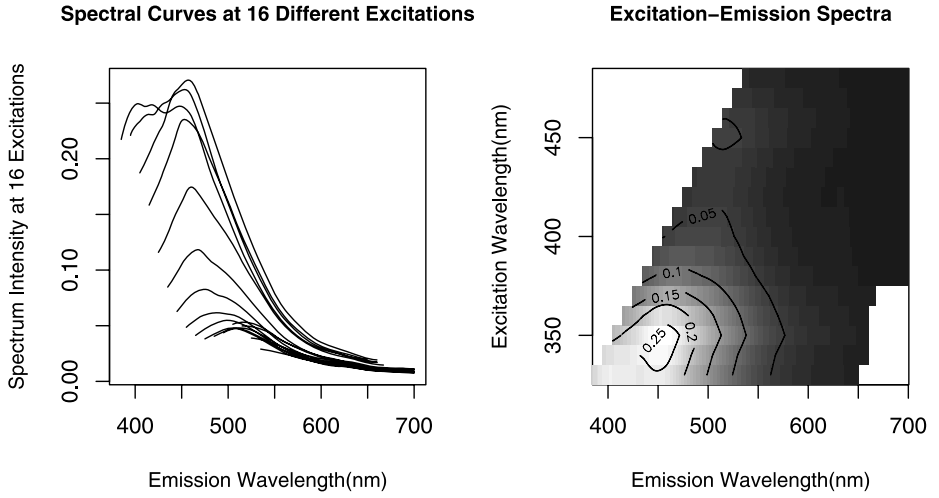

FIG 1. *Left Panel: Fluorescence spectral curves at different excitation wavelengths. Right Panel: The image plot of fluorescence spectroscopy data (EEM).*

2. Functional Generalized Linear Model with Curve Selection

Consider n i.i.d. observations where each observation contains J functions. For $i = 1, \ldots, n$ and $j = 1, \ldots, J$, let $x_{ij}(t)$ denote the jth function observed from the ith observation, where $E[x_{ij}(t)] = \mu_j(t)$. Note that the J functions within each observation can be rather arbitrary hence we assume different mean function $\mu_j(t)$ for each $x_{ij}(t)$. In addition to functional data, we assume there is a non-functional vector z_i associated with each observation. Suppose the responses we observed are binary variables y_i. Similarly to James [8] and Müller and Stadtmüller [11], we propose a functional generalized linear model to connect the binary responses with the predictors. Let $p_i = Pr\,(y_i = 1|z_i, x_{ij}(t), j = 1, \ldots, J)$ and

$$p_i = g^{-1}(\eta_i), \tag{2.1}$$

$$\eta_i = \alpha_0 + z_i^T \alpha + \sum_{j=1}^{J} \int_{T_j} \beta_j(t)(x_{ij}(t) - \mu_j(t))dt, \tag{2.2}$$

where T_j is the domain of $x_{ij}(t)$, α_0 is the univariate intercept, α is a vector of coefficients for the non-functional predictors, and the $\beta_j(t)$'s are the functional regression coefficients. For convenience, we center $x_{ij}(t)$ at its mean in the integrand. Here the link function $g(\cdot)$ is a one-to-one continuous function. To perform curve selection, we propose the following constraint on the functional regression coefficients:

$$\sum_{j=1}^{J} ||\beta_j||_{L^2} < s, \tag{2.3}$$

where $||f||_{L^2} = (\int f^2(t)dt)^{1/2}$, s is a pre-defined constant. Note that (2.3) is a combined constraint of L^2 norm and l^1 norm. This is an extension of the group-wise variable selection in the multivariate setting proposed by Yuan and Li [18]. Because of the properties of this combined constraint, we expect $\beta_j \equiv 0$ for a number of j's, depending on the value of s.

Due to the infinite dimensionality of functional data, multivariate methods can not be used directly for solving the above proposed model. One can discretize $x_{ij}(t)$ on a finite grid and transform the problem to a multivariate regression model, but the number of grid points is an issue and there will be high correlation between contiguous grid points because of the "functional" properties of $x_{ij}(t)$. A natural choice is to apply standard functional dimension reduction methods to reduce the dimension first and solve the problem on the reduced space. If we assume $\forall j, x_{ij}(t) \in \mathcal{H}_j$ for some separable Hilbert space $\mathcal{H}_j$, and $E[x_{ij}(t)] = \mu_j(t)$, we can expand $x_{ij}(t) - \mu_j(t)$ on a set of orthonormal basis $\{\phi_k^j\}_{k=1}^{\infty}$

$$x_{ij}(t) - \mu_j(t) = \sum_{k=1}^{\infty} c_{ijk}\phi_k^j(t) \tag{2.4}$$

and a truncated version of (2.4) can be used to approximate $x_{ij}(t)$ since $\sum_{k=1}^{\infty} |c_{ijk}|^2 < \infty$. And similarly, we assume $\beta_j(t) \in \mathcal{H}_j, \forall j$, and this gives

$$\beta_j(t) = \sum_{k=1}^{\infty} b_{jk}\phi_k^j(t). \tag{2.5}$$

Note that the orthonormal basis $\{\phi_k^j\}_{k=1}^{\infty}$ can be chosen to be a known basis such as a Fourier basis or a wavelet basis. If in addition, we assume $x_{ij}(t) \in L_2[\Omega \times T_j]$ for the domain T_j and the underlying sample space Ω, i.e., $\int_{T_j} E[x_{ij}(t)^2]dt < \infty, \forall j$, Mercer's theorem and Karhunen-Loève theorem ([2]) suggest taking the orthonormal basis to be the eigenfunctions of the covariance operator K, where K is defined by

$$Kx(t) = \int x(s)k(s,t)ds, \qquad k(s,t) = Cov(x(s), x(t)). \tag{2.6}$$

In this case, the coefficients $\{c_{ijk}, k = 1, \ldots, \infty\}$ are called functional principal component scores of the functional data. Using the functional principal component method is different from using a known basis in that the eigenbasis functions need to be estimated. Various estimating methods are proposed as in Ramsay and Silverman [13], and in Hall, Müller and Wang [7].

Once the functional principal component scores or the orthonormal basis coefficients have been estimated, we can reduce equation (2.2) to

$$\eta_i = \alpha_0 + z_i^T\alpha + \sum_{j=1}^{J}\sum_{k=1}^{\delta_j} c_{ijk}b_{jk}, \tag{2.7}$$

where δ_j is the truncation parameter for the jth functional predictor. We thus transfer the functional regression to multivariate regression. The constraint condition (2.3) will be reduced to

$$\sum_{j=1}^{J} ||b_j||_2 < t, \tag{2.8}$$

where $b_j = (b_{j1}, \ldots, b_{j\delta_j})$ and $||\cdot||_2$ stands for the Euclidean norm. Curve selection can thus be performed through selecting variables in (2.7) using the grouped Lasso type constraint (2.8), i.e., if one curve $x_j(t)$ is selected, then the coefficients

$b_{jk}, k = 1, \ldots, \delta_j$, will all be non-zero. The Lasso (Least Absolute Shrinkage and Selection Operator) was first proposed by Tibshirani [16] for model selection in linear regression models. The basic idea was to find a subset of the covariates with non-zero coefficients by applying an l_1 constraint to the regression coefficients based on the ordinary least square estimation. Yuan and Lin [18] extended the regular Lasso to cases where the covariates can be grouped, such as multi-factor ANOVA. They combine the l_1 and l_2 constraints so that the resulting model selects variables at the group level and is invariant under group-wise orthogonal transformation. To solve our problem based on the reduced model (2.7) and (2.8), we borrow the algorithm proposed by Meier et al. [10], where they extend the group-wise lasso regression of Yuan and Lin [18] to a logistic regression setup. Assume the link function in (2.1) is a logit link, i.e.,

$$\log\left(\frac{p_i}{1-p_i}\right) = \eta_i. \tag{2.9}$$

The estimate can be obtained by minimizing the convex function

$$Q_\lambda(\theta) = -l(\theta) + \lambda \sum_{j=1}^{J} s(\delta_j)||b_j||_2, \tag{2.10}$$

where $\theta = \{\alpha_0, \alpha, b_j, j = 1, \ldots, J\}$, and $l(\cdot)$ is the log-likelihood function:

$$l(\theta) = \sum_{i=1}^{n} \{y_i \eta_i - \log(1 + \exp(\eta_i))\}. \tag{2.11}$$

Here $s(\delta_j)$ is used to rescale the penalty with respect to the dimensionality of b_j, usually taken to be $\sqrt{\delta_j}$, and $\lambda > 0$ is the tuning parameter to control the amount of penalty. Note that in the model of Meier et al. [10], they only allow one unpenalized term, i.e., only the intercept term is unpenalized. In our proposed model, in addition to the intercept α_0, we allow the coefficients α of nonfunctional predictors to be unpenalized. Meier et al. stated the attainability of the minimum of the optimization problem in their paper and provided a proof. Actually, some conditions must be satisfied for the attainability to hold. Here we provide a general sufficient condition for the minimum of equation (2.10) to be attained.

Proposition 1. *For* $0 < \sum_{i=1}^{n} y_i < n, \lambda > 0, s(\delta_j) > 0, \forall j$, *assume the design matrix* X *formed by*

$$X = \begin{pmatrix} 1 & z_1^T & c_{111} & \ldots & c_{11\delta_1} & \ldots & \ldots & c_{1J1} & \ldots & c_{1J\delta_J} \\ \vdots & & & & & & & & & \\ 1 & z_n^T & c_{n11} & \ldots & c_{n1\delta_1} & \ldots & \ldots & c_{nJ1} & \ldots & c_{nJ\delta_J} \end{pmatrix}$$

is an n *by* m *matrix of rank* m, $n \geq m$. *Assume the maximum likelihood estimator for the logistic regression (with log-likelihood in equation (2.11)) exists. Then the equation (2.10) has an unique minimizer* θ^*.

The proof for Proposition 1 is in the Appendix. Meier et al. [10] proposed a Block Coordinate Gradient Descent algorithm to solve the group lasso logistic regression and provided an R package called *grplasso*. We will use this package to perform curve selection based on reduced model in equations (2.7) and (2.8). The initiation of the algorithm is the same as in *grplasso*.

3. Simulation Study

To verify the performance of the proposed method in classification problems with multiple functional covariates, we generate $n = 1000$ i.i.d. observations. Each observation contains one non-functional covariate and three functional covariates. The non-functional covariate is generated from uniform $(0, 1)$ distribution. And the three functional covariates are generated using the first 4 cosine basis functions on the domain $[0, 1]$, i.e., using basis $\phi_0(t) = 1, \phi_k(t) = \sqrt{2}\cos(k\pi t), k = 1, \ldots, 3$. For each functional covariate, the 4 coefficients of the cosine basis are generated independently from a normal distribution with some fixed mean and variance 0.5. We set the coefficient functions for the first and third functional covariates to be zero and set the coefficient function for the second to be non-zero. Figure 2 shows the plot of both non-functional covariates and functional covariates for the first 50 observations. The binary responses y_i are generated by sampling from a Bernoulli distribution with success probability $p_i = (1 + \exp(-\eta_i))^{-1}$, where η_i is computed from equation (2.2) using numerical integration. The proportion of 1's among the

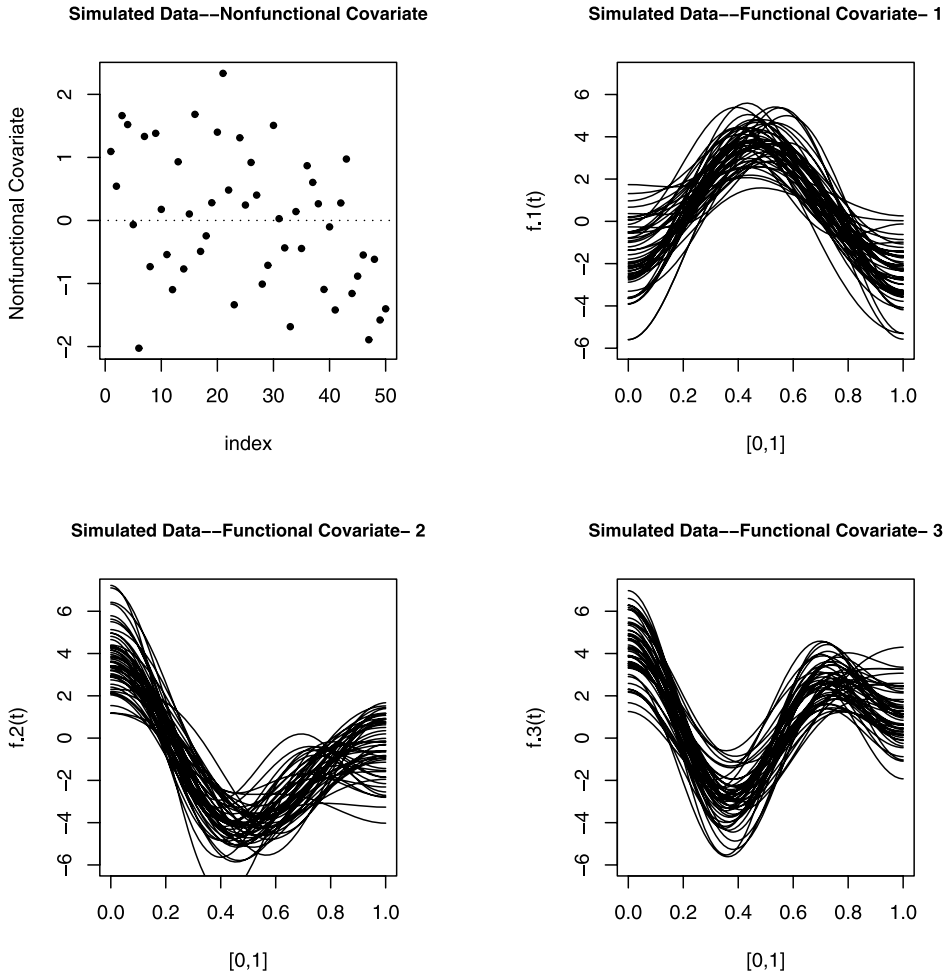

FIG 2. *Data plot of both non-functional covariates and functional covariates for the first* 50 *observations used in simulation.*

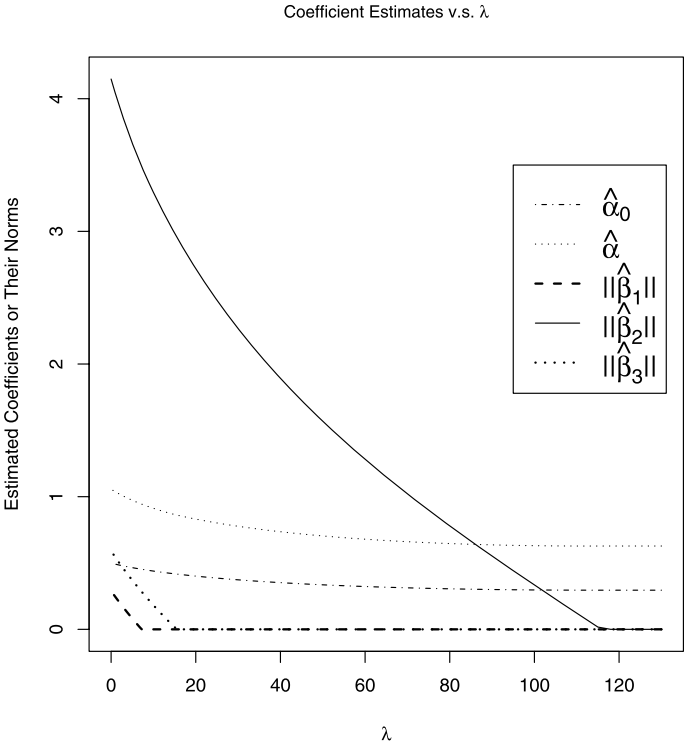

FIG 3. *Estimated paths of coefficient vector at different λ values.*

binary y_i's is 57.3%. The data are randomly split into a training set and a test set with 800 observations in the training set and 200 observations in the test set.

To apply the proposed model to these data, one can choose different types of orthonormal basis for dimension reduction. Since the data are generated using cosine basis, we will show the results of using the cosine basis so that the estimated coefficients can be compared with their known true values. We have also tried using functional principal components, and the curve selection and prediction results are very similar to that of using cosine basis.

For the choice of cosine basis, we reduce the dimension of the functional predictors using the first 4 cosine basis functions. The group-wise lasso regression algorithm of Meier et al. [10] is then applied to the reduced scores. Figure 3 shows the estimation paths for the regression coefficients as a function of λ. Note that for the estimated coefficient function $\hat{\beta}_j$, we plotted their L^2 norm, i.e., $||\hat{\beta}_j|| = \sqrt{\int_{T_j} \hat{\beta}_j(t)^2 dt}$, where the function $\hat{\beta}_j$ are obtained through inverse transform of the estimated coefficients $\hat{b}_j$. From Figure 3, we see that for a large range of λ, i.e., $15.7 < \lambda < 115$, the method correctly picked out the non-zero coefficient function $\hat{\beta}_2$. The values of $\hat{\beta}_2(t)$ at 6 selected λ's is plotted in Figure 4 in comparison with the true $\beta_2(t)$. Table 1 shows the estimated coefficients(in form of the cosine basis scores $\hat{b}_j$) compared with the true values under the 6 selected λ's. From Table 1, we see that as the penalty parameter λ increases, the magnitudes of the estimated coefficients shrink toward 0. When $\lambda = 0$, the estimates are equal to the maximum likelihood estimates, which gives nonzero estimates to all coefficients. When λ ranges from 22.4 to 89.6, the coefficients corresponding to the first and third curve

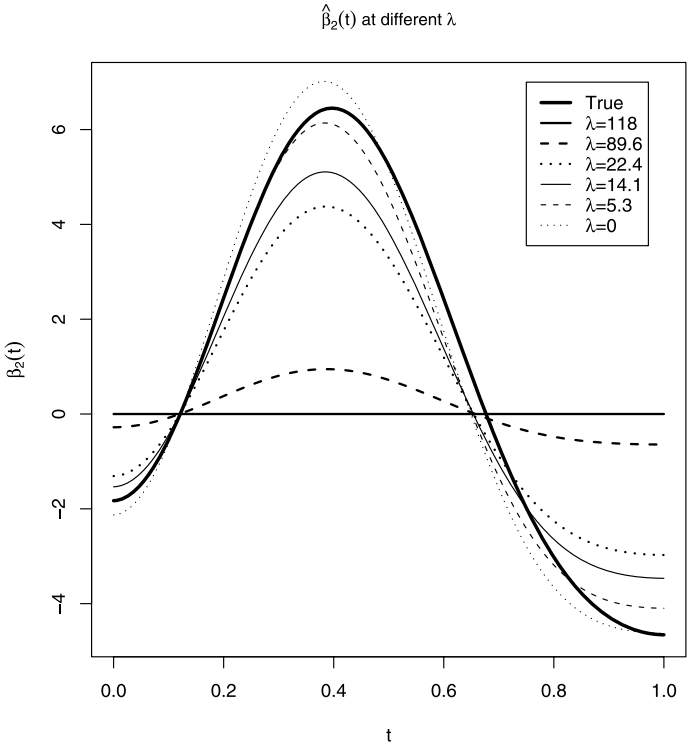

FIG 4. *Estimated coefficient function* $\hat{\beta}_2(t)$ *at* 6 *selected* λ *values and the true* $\beta_2(t)$.

are exactly 0, and the coefficients corresponding to the second curve are nonzero. For $\lambda > 14.1$, the estimates are (almost all) closer to 0 than the true values. We believe that these shrinkage effects are caused by the continuous-shrinkage property of Ridge and Lasso penalty (see Tibshirani [16]). It has been suggested that there may be large bias in the estimators related to the inconsistency of the original Lasso

TABLE 1
The estimated coefficient values compared with the true values at different λ*s*

		Estimated coefficients at different λ values					
Coef	*True Values*	$\lambda = 118$	$\lambda = 89.6$	$\lambda = 22.4$	$\lambda = 14.1$	$\lambda = 5.3$	$\lambda = 0$
α_0	0.5	0.3	0.3	0.39	0.42	0.46	0.5
α	1	0.63	0.64	0.82	0.87	0.97	1.06
b_{11}	0	0	0	0	0	0.03	0.15
b_{12}	0	0	0	0	0	−0.04	−0.17
b_{13}	0	0	0	0	0	0.04	0.18
b_{14}	0	0	0	0	0	0	−0.01
b_{21}	1	0	0.13	0.58	0.67	0.79	0.9
b_{22}	2	0	0.31	1.43	1.67	2.01	2.29
b_{23}	−3	0	−0.42	−1.92	−2.24	−2.66	−3.02
b_{24}	−1	0	−0.18	−0.84	−0.99	−1.21	−1.41
b_{31}	0	0	0	0	0	0.02	0.03
b_{32}	0	0	0	0	0.01	0.07	0.13
b_{33}	0	0	0	0	0.04	0.34	0.56
b_{34}	0	0	0	0	0.01	0.09	0.14

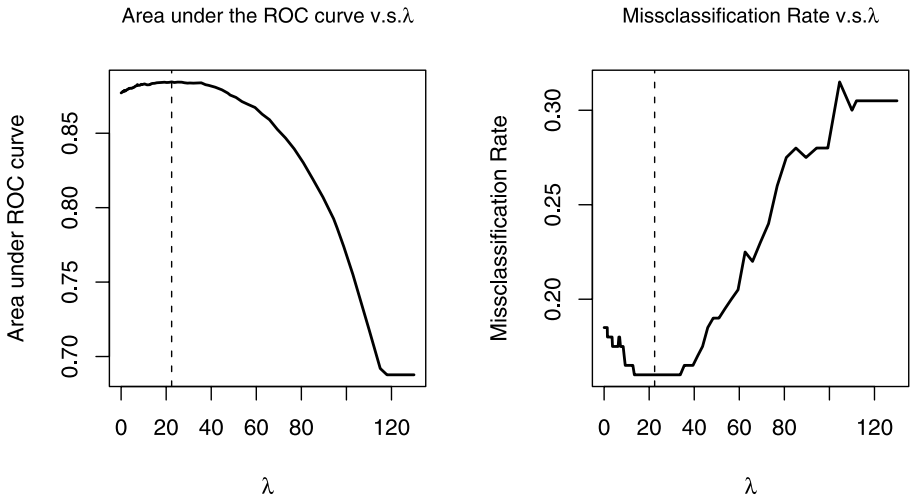

FIG 5. *Prediction results at different λ values.*

under certain conditions, i.e., that the Lasso does not satisfy the "oracle properties" (Fan and Li [4], Zhao and Yu [20]). Some modifications have been proposed to overcome the drawbacks of Lasso and make the estimators satisfy the oracle properties(see Zou [22]). In this study, we only focus on the curve selection and prediction, but more research can be done on the consistency of the grouped-Lasso regression under the functional data setup.

To perform prediction on test set, the estimated coefficient function $\hat{\beta}_j(t), j = 1, 2, 3$ are plugged into the test set using (2.2) and the estimated success probability $\hat{p}_i$ are computed for each observation, from which we can plot a ROC curve (see Zweig & Campbell [23]) for each λ. From each ROC curve, we pick a point that maximizes the sum of sensitivity and specificity, and this point will be used as the optimal classification point. The misclassification rate at the optimal point and the corresponding area under the ROC curves are computed at different values of λ and plotted in Figure 5. From Figure 5, we find that when λ is around 22.4, the prediction on the test set gives the best sensitivity(93%) and specificity(73%) and an fairly large area under ROC curve (0.88), and the corresponding misclassification rate is 16%.

4. Real Data Application–Fluorescence Spectral Curve Selection and Cervical Pre-Cancer Diagnosis

Totally 717 EEM measurements were made on 306 patients, and each measurement contains 16 spectral curves. Measurements were taken from different sites on the cervix and may include repeated measurements at the same site. All the measurements were made using the same instrument (called FastEEM3) in the same clinic (British Columbia Cancer Agency, Vancouver, CA). Data were split into a training set and a test set with 396 measurements in the training set and 321 in the test set. The proportions of diseased cases within each set are 0.21, 0.20, respectively. Two non-functional covariates are considered in this study: the colposcopic tissue type of the measurements, and the menopausal status. Colposcopic tissue type is a binary variable indicating two types of tissue — squamous and columnar, which is obtained

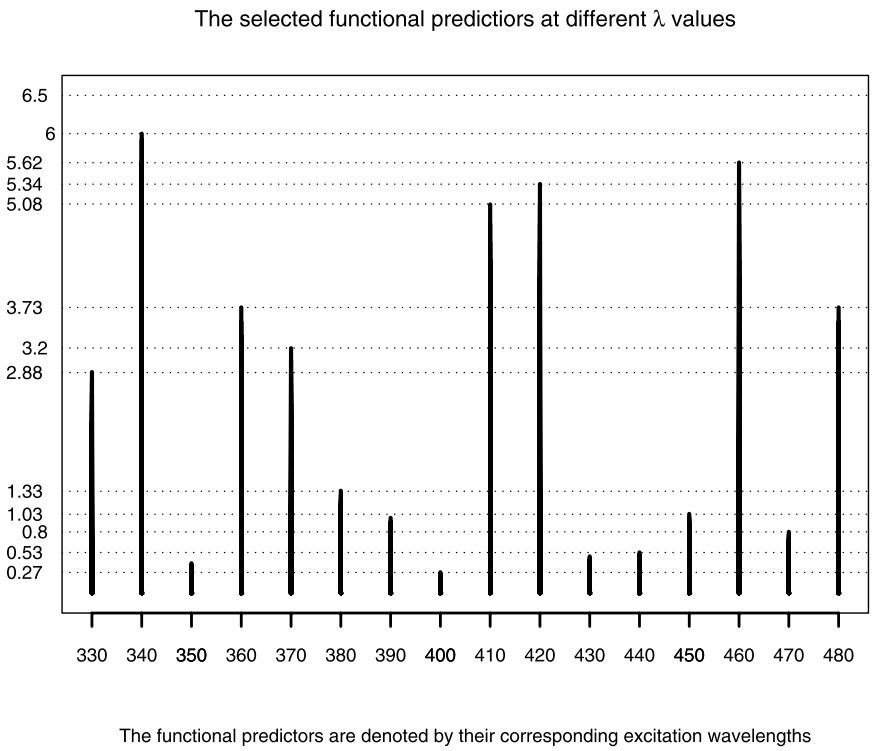

FIG 6. *The selected functional predictors (fluorescence spectral curves denoted by excitation wavelengths) at different λ values.*

prior the fluorescence spectroscopy measurements. Menopausal status of a patient is a categorical variable which has three levels: pre-, peri- and post-menopause. The first 5 functional principal components are chosen as the scores extracted from each functional predictor, which reduce the data to a total of 80 scores. To reduce bias, the test set scores (the scores of orthonormal basis) are computed based on only information obtained from the training set. For example, the eigenfunctions used for computing functional principal components scores of the test set are estimated from the training set.

The grouped lasso logistic regression is used to pick the excitation wavelengths as λ decreases from 6 to 0. Due to the large number of curves, the plot of coefficient path is hard to visualize. In Figure 6, we summarize the excitation spectral curves selected at different λ values. For example, from Figure 6 we find that when $3.73 < \lambda \leq 5.08$, the curves at excitations 340, 410, 420, 460 are selected. At larger values of λ, the penalty is heavier, and fewer curves are selected. When $\lambda = 0$, there is no penalty, and all curves are selected. The order of selection from larger λ values to smaller λ values suggests the importance of curves in the regression. For example, the excitation curves are ordered by $340 > 460 > 420 > 410$ according to the order of being selected. The estimated coefficients at different values of λ are used to predict in the test set, from where we can evaluate the performance of different λ values. Due to the fact that the total proportion of diseased cases is small, the misclassification rate is not an ideal criterion for evaluating the prediction result (see [21], page 22 for details). To reduce the risk of false negatives, we wish to keep the sensitivity high enough and sacrifice some specificity. Hence for each fixed λ,

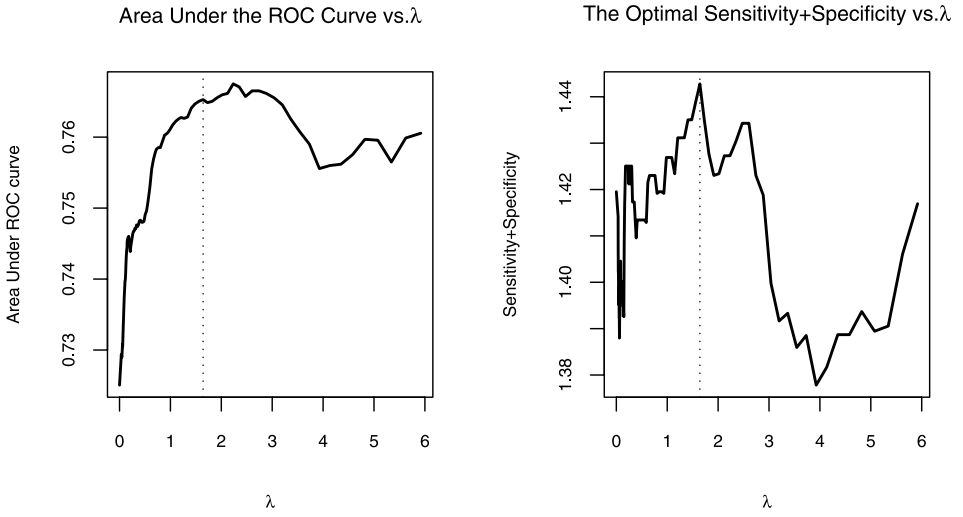

FIG 7. *Prediction results at different λ values.*

we pick a point from the empirical ROC curve using the criterion that the sum of the sensitivity and specificity is maximized. The Figure 7 shows the area under the curves and the optimal sum of the sensitivity and specificity at different values of λ. At $\lambda = 1.64$, the sum is maximized at 1.44 with sensitivity 87% and specificity 57%, and the corresponding area under ROC curve is 0.77, and misclassification rate is 38%.

Since the main purpose of the above analysis is for curve selection rather than classification, once the functional covariates are selected, different classifiers can be applied to perform classification based on the selected subset of curves. In addition to logistic regression, we also performed classification with 3 other classifiers using the selected curves. By choosing $\lambda = 1.64$, we selected function predictor curves at excitations: 330, 340, 360, 370, 410, 420, 460 and 480, and used the first 5 functional principal components to reduce the dimension. We refitted the logistic model without penalty and compared the prediction results on the test set with 3 other classifiers in Table 2. The corresponding ROC curves are plotted in Figure 8. From Figure 8, we find that logistic regression, k-nearest neighbor (KNN) and linear discriminant analysis(LDA) provide similar ROC curves. The highest sum of sensitivity and specificity is 1.43, obtained by KNN, which is only slightly smaller than the grouped lasso results at $\lambda = 1.64$. The LDA method provides the same specificity with logistic regression but higher sensitivity.

TABLE 2

The classification results using 4 different methods on the selected curves. Auc: Area under ROC curve. MisR: Misclassification rate. Sens: Sesitivity. Speci: Specificity. Sum: The sum of sensitivity and specificity. Logistic: logistic regression. KNN: k-nearest neighbor. LDA: linear discriminant analysis. SVM: support vector machine

Method	Auc	MisR	Sens	Speci	Sum
Logistic	0.76	31%	71%	68%	1.39
KNN	0.68	27%	68%	74%	1.43
LDA	0.75	31%	75%	68%	1.42
SVM	0.64	28%	48%	79%	1.26

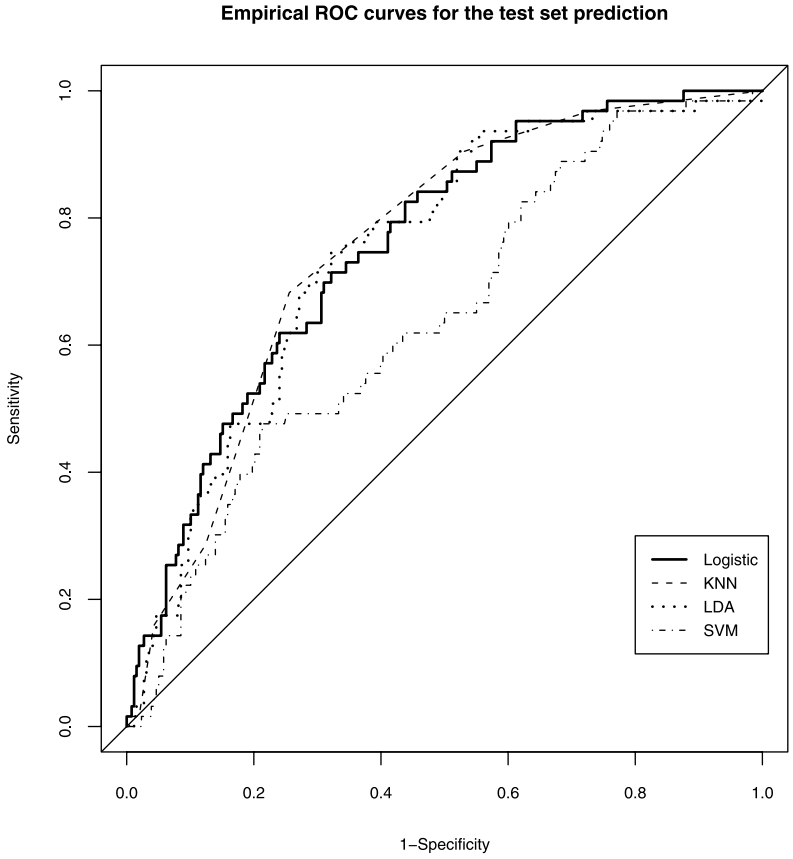

FIG 8. *ROC curves obtained when training 4 different classifiers based on selected curves and predicting on the test set.*

5. Determining the Related Parameters

In our proposed model, two types of parameters need to be determined: the tuning parameter λ and the truncation parameters δ_j, $j = 1, \ldots, J$. In this section, we discuss how to determine these parameters.

The choice of tuning parameter λ is important for prediction. In Meier et al. [10] and in our paper, a test set is used to choose the λ with the best prediction performance. However, there are also cases where only a small number of observations are available and splitting out a test set is not possible. In this case, we can adopt model selection criteria such as AIC, practical C_p or BIC. AIC tends to select a model with optimal prediction, while BIC tends to identify the true sparse model if the true model is included in the candidate set (see Yang [17]). In the grouped Lasso linear regression model, Yuan and Lin [18] proposed an approximation to the degrees of freedom and used a C_p criterion for selecting the tuning parameter λ. Whether this criterion can be extended to logistic regression case for selecting λ is an open question.

In addition to the tuning parameter λ, the truncation parameter δ_j in equation (2.7) is also one concern of the study. In the real application of Section 4, we let $\delta_j \equiv \delta$ and reported the curve selection and prediction results with $\delta = 5$. To find

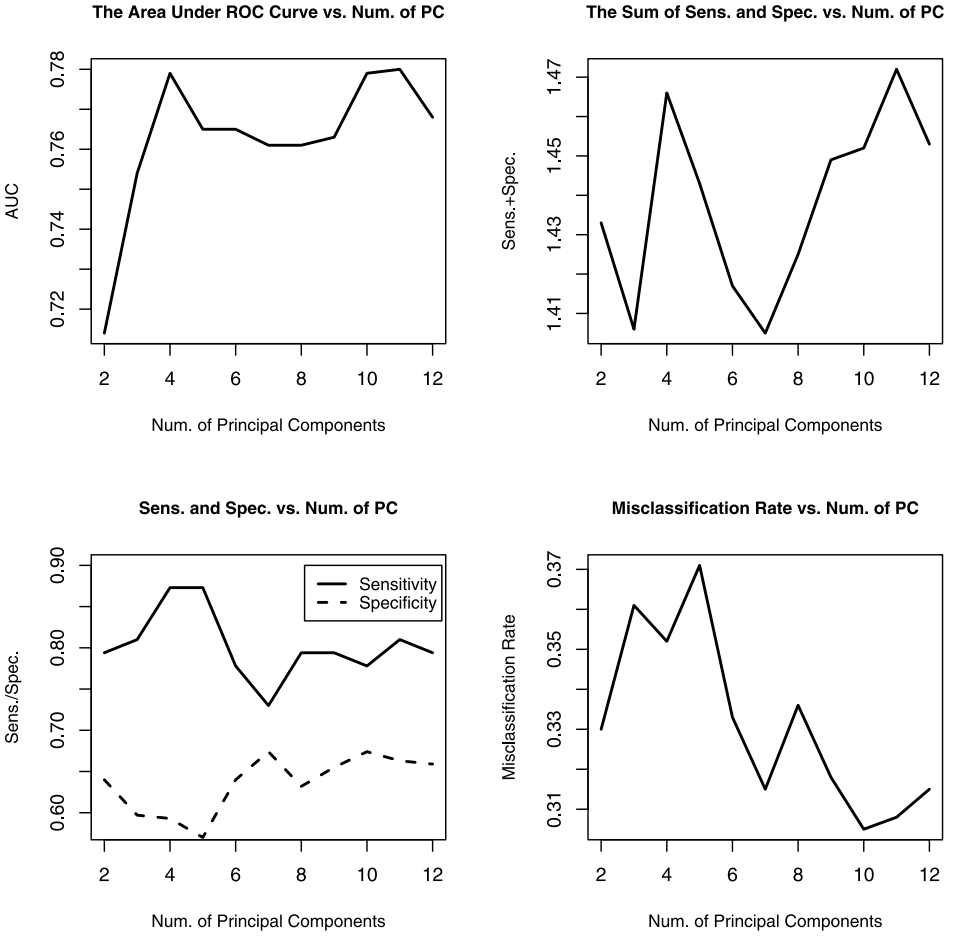

FIG 9. *Prediction results using different number of basis.*

out whether other choices of δ are better for prediction, we compute the prediction results for the test set at different number of δ but fixing $\lambda = 1.64$. The quantitative prediction results are plotted in Figure 9. From Figure 9, we can see that using 11 functional principal components, the area under the ROC curve are maximized at 0.780, and the sum of sensitivity and specificity are maximized at 1.47, with a relatively small misclassification rate 31%. The sensitivity and specificity reach 81%, 66%, respectively.

It is also suspected that the optimal λ may interact with δ so determining one by fixing the other may be suboptimal. In our study, we also have tried to determine both the parameters by training the model under different combinations of them, and predicting on the test set. It turns out that at around $\lambda = 1.64$ the prediction results of the model is better than other choices of λ, and this is quite stable across different choice of δ, especially for δ greater than 3. In Figure 10, We plot the area under the ROC curve for 11 different δ and for appropriately selected λ values across a meaningful range, i.e., $\lambda = (5, 3, 1.64, 1.5, 1, 0.27)$. It shows that the line with $\lambda = 1.64$ stays on the top for δ's larger than 3. The reason for the small interaction between λ's and δ's can be the following: the orthogonal basis approximation

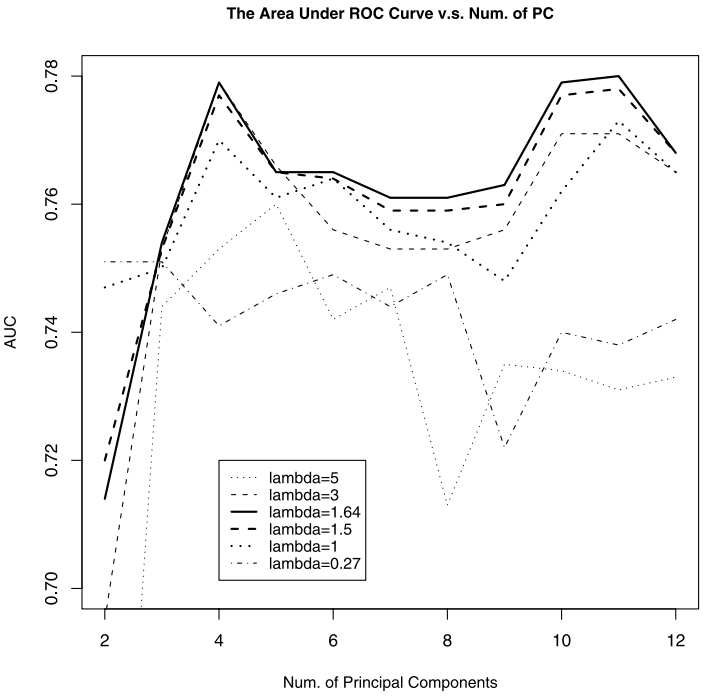

FIG 10. *The area under the ROC curves for 6 different* λ *values and* 11 *different choice of FPC's.*

tends to be accurate with only a few components. For example, in functional principle components, over 97% of the variability will be counted in the first priciple component score for all excitation curves. Later components only add details to the model but does not change the likelihood dramatically. Therefore the miminum of equation (2.10) as a function of λ does not change much when δ changes. But this is not true for non-orthogonal basis approximation methods such as B-spline.

Note that choosing $\delta_j \equiv \delta$ is just a convenient choice, which has the advantage that it leaves only two parameters to determine and cross validation is feasible for determining these parameters. However, it also brings in the risk of loosing information. In general, one may use different truncation parameters if there are large differences on the properties of the curves such as smoothness. If all curves are obtained through similar sources and are similar in shape and other above mentioned properties, it would be safe to choose a common δ. As an alternative, since the step of estimating $\{c_{ijk}, k = 1, \ldots, \delta_j, j = 1, \ldots, J\}$ can be independent of the group-wise Lasso step, one can use approximation criteria such as error sum of squares (SSE) to determine the truncation parameters for each curve. For example, if using functional principal component, we can choose a level of approximation (e.g., let the percentage of variabilities explained to be greater than 99%) and select the number of eigenfunctions to achieve this. However, better approximation does not necessarily give better prediction.

6. Discussion

We have proposed a functional logistic regression model to perform classification and curve selection. This model automatically selects among the functional co-

variates through the grouped Lasso variable selection. The proposed model gives information about which curves will be selected if we are willing to use a subset of the functional covariates for classification. For example, under penalty $\lambda = 5$, the best four functional predictors selected in our real data application are curves at excitations 340, 410, 420 and 460. The selected functional covariates can then be used with different classifiers for accurate classification.

There are several aspects that can be studied in more detail. Firstly, the basis expansion step can be combined more tightly with the grouped Lasso regression step using techniques similar to Müller and Stadtmüller [11]. It is necessary to investigate the consistency properties of the estimated coefficient function $\beta_j(t)$'s, such as the oracle property. The algorithm of Meier et al. [10] requires that the tuning parameter λ to be predefined on a grid of values, where they proposed a way to find the range of λ of interest. This method, although faster, makes it difficult to find a precise λ value that is optimal for prediction purposes. Efficient algorithms for searching λ is of great importance especially when functional data is involved.

Alternative methods for curve selection can be formulated through the Bayesian paradigm. Bayesian variable selection models can be derived for selecting variables at a group level and thus can be used for curve selection as well.

Appendix: Proof of Proposition 1

The proof of Proposition 1 uses a result stated in the following lemma.

Lemma 1. *Let $f : \mathbb{R}^n \mapsto \mathbb{R}$ be a strictly convex function with a minimizer $\tilde{x}$, and let $g : \mathbb{R}^n \mapsto [0, \infty)$ be a convex function. Then $f + g$ has a unique minimizer x^* in $\mathbb{R}^n$.*

Proof. Let $h(x) = f(x) + g(x)$. It is easy to show that $h(x)$ is strictly convex from the definition. We claim that the existence of a minimizer $\tilde{x}$ of f implies that h is coercive, which means $h(x) \to \infty$ as $||x|| \to \infty$. The coerciveness and strict convexity of h implies the existence of a unique minimizer x^*.

To show that h is coercive, it is sufficient to show that f is coercive (since $g \geq 0$). The minimizer $\tilde{x}$ of f is the unique minimizer of f by strict convexity. Also, f is convex hence is continuous on $\mathbb{R}^n$ (see [15], page 82). Thus $\forall\ r > 0, \forall\ x$ such that $||x - \tilde{x}|| > r$, we claim

$$f(x) > \frac{b}{r}||x - \tilde{x}|| + f(\tilde{x}),$$

where $b = \inf\{f(x) : ||x - \tilde{x}|| = r\} - f(\tilde{x})$. Note that b exists and $b > 0$ by continuity of f. To show this inequality, let $x_0 = r(x - \tilde{x})/(||x - \tilde{x}||) + \tilde{x}$, so that x_0 lies on the line formed by x and $\tilde{x}$, with $||x_0 - \tilde{x}|| = r$ and $||x - x_0|| = ||x - \tilde{x}|| - r$. Thus $f(x_0) - f(\tilde{x}) \geq b$ by the definition of b. Now let $\alpha = r/||x - \tilde{x}||$. We see that $x_0 = \alpha x + (1 - \alpha)\tilde{x}$. By strict convexity of f,

$$f(x_0) < \alpha f(x) + (1 - \alpha) f(\tilde{x}).$$

Thus

$$\begin{aligned}
\frac{b}{r}||x - \tilde{x}|| + f(\tilde{x}) &\leq (f(x_0) - f(\tilde{x}))\frac{||x - \tilde{x}||}{r} + f(\tilde{x}) \\
&< (\alpha f(x) + (1 - \alpha) f(\tilde{x}) - f(\tilde{x}))\frac{||x - \tilde{x}||}{r} + f(\tilde{x}) \\
&= f(x).
\end{aligned}$$

Since $||x - \tilde{x}|| \geq ||x|| - ||\tilde{x}||$, $||x|| \to \infty$ implies $||x - \tilde{x}|| \to \infty$, which implies $f(x) \to \infty$ by the above inequality and the facts that $b > 0, r > 0$, $f(\tilde{x})$ finite. Therefore, f is coercive, and so is h.

Since h is coercive, we have $h(x) \to \infty$ as $||x|| \to \infty$. Therefore, if we pick an arbitrary point $x_1 \in \mathbb{R}^n$, there exists a constant $\delta > 0$ such that $h(x) > h(x_1)$ for all $||x - x_1|| > \delta$. Since the domain $||x - x_1|| \leq \delta$ is compact and $h(x)$ is strictly convex on it, $h(x)$ has a unique minimizer in $||x - x_1|| \leq \delta$, which we denote as x^*. (A strictly convex real valued function defined on a compact domain has a unique minimum on its domain.) This x^* is also the global minimizer since $h(x) > h(x_1) \geq h(x^*)$ on $||x - x_1|| > \delta$. □

Proof of Proposition 1. Based on results in Lemma 1, we let f to be $-l(\theta)$ and g to be $\lambda \sum_{j=1}^{J} s(\delta_j)||b_j||_2$, therefore our objective function in equation (2.10) is the sum of f and g, where $\theta = \{\alpha_0, \alpha, b_j, j = 1, \ldots, J\}$, and $l(\theta) = \sum_{i=1}^{n} y_i \eta_i - \log(1 + \exp(\eta_i))$ with $\eta_i = \alpha_0 + z_i^T \alpha + \sum_{j=1}^{J} \sum_{k=1}^{\delta_j} c_{ijk} b_{jk}$.

Firstly, we show that $-l(\theta)$ is strictly convex. It is sufficient to show that its Hessian is positive definite. Since the Hessian takes the form

$$\bigtriangledown_\theta^2(-l(\theta)) = X^T D X,$$

where $D = \text{diag}\{\exp(\eta_i)/(1 + \exp(\eta_i))^2, i = 1, \ldots, n\}$. It is positive definite since X is of rank m (full rank). Secondly, since the maximum likelihood estimator exists, $-l(\theta)$ has an unique minimizer. The existence of maximum likelihood estimator for logistic regression requires some conditions for the design matrix X. Basically, the n rows of X can not be completely separated or quasi-completely separated in $\mathbb{R}^m$. See [1] for details. In practice, as long as we can find a numerical solution for the MLE at $\lambda = 0$, we would believe that the maximum likelihood estimator exists. Finally, let $g(b) = \lambda \sum_{j=1}^{J} s(\delta_j)||b_j||_2$, $b^T = (b_1^T, \ldots, b_J^T)$. It is easy to see that $g(b)$ is convex by the triangle inequality. Therefore by Lemma 1, $Q_\lambda(\theta)$ has a unique minimizer θ^*. □

Acknowledgements

This research was supported by the National Cancer Institute grant PO1-CA82710 and by the National Science Foundation grant DMS0505584. We thank the referees for constructive comments, and thank Dr. Wotao Yin for helpful discussions on the convex optimization.

References

[1] Albert, A. and Anderson, J. A. (1984). On the existence of maximum likelihood estimates in logistic regression models. *Biometrika* **71** 1–10.
[2] Ash, R. B. (1975). *Topics in Stochastic Processes.* Academic Press, New York.
[3] Chang, S. K., Follen, M., Malpica, A., Utzinger, U., Staerkel, G., Cox, D., Atkinson, E. N., MacAulay, C. and Richards-Kortum, R. (2002). Optimal excitation wavelengths for discrimination of cervical neoplasia. *IEEE Transactions on Biomedical Engineering* **49** 1102–1110.
[4] Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96** 1348–1360.
[5] Ferré, L. and Villa, N. (2006). Multilayer perceptron with functional inputs: An Inverse Regression Approach. *Scand. J. Statist.* **33** 807–823.

[6] Hall, P., Poskitt, D. S. and Presnell, B. (2001). A Functional data-analytic approach to signal discrimination. *Technometrics* **43** 157–214.
[7] Hall, P., Müller, H. and Wang, J. (2006). Properties of principal component methods for functional and longitudinal data analysis. *Ann. Statist.* **34** 1493–1517.
[8] James, G. M. (2002). Generalized linear models with functional predictors. *J. Roy. Statist. Soc. Ser. B* **64** 411–432.
[9] Leng, X. and Müller, H. (2005). Classification using functional data analysis for temporal gene expression data. *Bioinformatics* **22** 68–76.
[10] Meier, L., Geer, S. and Bühlmann, P. (2008). The group Lasso for logistic regression. *J. Roy. Statist. Soc. Ser. B* **70** 53–71.
[11] Müller, H. and Stadtmüller, U. (2005). Generalized functional linear models. *Ann. Statist.* **33** 774–805.
[12] Ramanujam, N., Mitchell, M. F., Mahadevan, A., Thomsen, S., Malpica, A., Wright, T., Atkinson, N. and Richards-Kortum, R. (1996). Spectorscopic diagnosis of cervical intraepithelial neoplasia(CIN) in vivo using laser induced fluorescence spectra at multiple excitation wavelengths. *Lasers Surg. Med.* **19** 63–67.
[13] Ramsay, J. and Silverman, B. (1997). *Functional Data Analysis*. Springer, New York.
[14] Ratcliffe, S. J., Heller, G. Z. and Leader, L. R. (2002). Functional data analysis with application to periodically stimulated foetal heart rate data. II: Functional logistic regression. *Statist. Med.* **21** 1115–1127.
[15] Rockafellar, R. T. (1970). *Convex Analysis*. Princeton Univ. Press.
[16] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288.
[17] Yang, Y. (2005). Can the strengths of AIC and BIC be shared? A conflict between model identification and regression estimation. *Biometrika* **92** 937–950.
[18] Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *J. Roy. Statist. Soc. Ser. B* **68** 49–67.
[19] Zhao, X., Marron, J. S. and Wells, M. T. (2004). The functional data analysis view of longitudinal data. *Statist. Sinica* **4** 789–808.
[20] Zhao, P. and Yu, B. (2006). On model selection consistency of Lasso. *J. Mach. Learn. Res.* **7** 2541–2563.
[21] Zhou, X. and Obuchowski, N. A. (2002). *Statistical Methods in Diagnostic Medicine*. Wiley, New York.
[22] Zou, H. (2006). The adaptive Lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101** 1418–1429.
[23] Zweig, M. H. and Campbell, G. (1993). Receiver-operating characteristic (ROC) plots: A fundamental evaluation tool in clinical medicine. *Clinical Chemistry* **39** 561–577.