ORİJİNAL ARAŞTIRMA ORIGINAL RESEARCH

Variable Selection for Functional Logistic Regression in fMRI Data Analysis

fMRI Veri Analizinde Fonksiyonel Lojistik Regresyon İçin Değişken Seçimi

Nedret BİLLOR,[a]
Jessica GODWIN[b]

[a]Department of Mathematics & Statistics, Auburn University, College of Sciences and Mathematics, Auburn, AL
[b]Department of Statistics, University of Washington, Seattle, WA, USA

Geliş Tarihi/*Received:* 16.01.2015
Kabul Tarihi/*Accepted:* 11.02.2015

Yazışma Adresi/*Correspondence:*
Nedret BİLLOR
Auburn University, College of Sciences and Mathematics, Department of Mathematics & Statistics, Auburn, AL, 36849
USA/A.B.D.
billone@auburn.edu

ABSTRACT This study was motivated by classification problem in Functional Magnetic Resonance Imaging (fMRI), a noninvasive imaging technique which allows an experimenter to take images of a subject's brain over time. As fMRI studies usually have a small number of subjects and we assume that there is a smooth, underlying curve describing the observations in fMRI data, this results in incredibly high-dimensional datasets that are functional in nature. High dimensionality is one of the biggest problems in statistical analysis of fMRI data. There is also a need for the development of better classification methods. One of the best things about fMRI technique is its noninvasiveness. If statistical classification methods are improved, it could aid the advancement of noninvasive diagnostic techniques for mental illness or even degenerative diseases such as Alzheimer's. In this paper, we develop a variable selection technique, which tackles high dimensionality and correlation problems in fMRI data, based on L_1 regularization-group lasso for the functional logistic regression model where the response is binary and represent two separate classes; the predictors are functional. We assess our method with a simulation study and an application to a real fMRI dataset.

Key Words: Functional data; logistic regression; LASSO; variable selection; fMRI data

ÖZET Bu çalışmada bir deneğin zamana dayalı olarak beyninin görüntülerinin alınmasını sağlayan noninvazif bir görüntüleme tekniği olan Fonksiyonel Manyetik Rezonans Görüntülemedeki (fMRI) sınıflandırma problemi incelenmiştir. fMRI çalışmaları az sayıda deneğe sahip ve fMRI verisindeki gözlemlerin pürüzsüz eğriler olarak varsayılması söz konusu olduğundan, doğası gereği fonksiyonel olan yüksek boyutlu veri kümelerinin ortaya çıkmasına neden olur. Yüksek boyutluluk; fMRI verisinin istatistiksel analizindeki en büyük problemlerinden biridir. Daha iyi sınıflama yöntemlerinin geliştirilmesine de ihtiyaç vardır. fMRI tekniğinin en iyi tarafı noninvazif olmasıdır. Eğer istatistiksel sınıflama yöntemleri geliştirilirse, akıl hastalığı ya da hatta Alzheimer gibi dejeneratif hastalıklar için noninvazif tanı yöntemlerinin gelişimine yardımcı olabilir. Bu makalede, fMRI verisindeki yüksek boyutluluk ve korelasyon problemlerini ele alan, yanıt değişkeninin iki ayrı sınıfı gösterdiği ve açıklayıcı değişkenlerin fonksiyonel olduğu lojistik regresyon modeli için lasso grubu L_1 düzenlemesine dayalı, bir değişken seçim tekniği geliştirdik. Yöntemimizin performansını simülasyon çalışması ve gerçek bir fMRI veri seti uygulaması ile saptadık.

Anahtar Kelimeler: Fonksiyonel veri; lojistik regresyon; LASSO; değişken seçimi; fMRI verisi

Turkiye Klinikleri J Biostat 2015;7(1):1-10

Functional data analysis (FDA) is a relatively new area within the discipline of statistics. Functional data are data that have been measured discretely over a continuum, usually time. Instead of treating the many discrete measurements as individual observations, one makes the assumption that these measurements represent a smooth, underlying

doi: 10.5336/biostatic.2015-43642

curve. This curve, then, is considered as one observation.

Much of this work was motivated by Functional Magnetic Resonance Imaging (fMRI), a noninvasive imaging technique which allows an experimenter to take images of a subject's brain over time. These images are taken while the subject performs a task such as finger tapping or correctly identifying images of human faces amidst a series of images containing both human faces and objects. fMRI is currently being used to assess which areas of the brain are activated while performing certain tasks. This is done by dividing the brain into voxels, the three-dimensional analog of a pixel, and measuring brain activation. Depending on how one defines a voxel, a typical fMRI image has over 1,000,000 voxels. As fMRI studies usually have a small number of subjects, this results in datasets that are incredibly highdimensional. High dimensionality is one of the biggest problems in statistical analysis of fMRI data.[1] Initial statistical analysis of fMRI data was univariate in nature.[2] This is obviously simplistic for data that take into account four dimensions: three spatial dimensions and one temporal dimension. The second decade of fMRI research focused on multivariate data analysis. Considering the fact that, the capturing of images happens over time, the next logical step in this progression is functional data analysis. In an fMRI session, images of a subject's brain are taken at discrete time points. However, one would expect the brain's response to a stimulus to be continuous in nature. This makes fMRI imaging data a great candidate for functional data analysis. In Viviani et al.[3] published a paper using functional principal component analysis in fMRI. They showed the results to be much more interpretable than multivariate PCA. Since then more statistical analysis of fMRI data has been functional in nature. There is a need to attack the problem of high dimensionality of brain imaging data. There is also a need for the development of better classification methods.[1] One of the best things about Functional Magnetic Resonance Imaging is its noninvasiveness. If statistical classification methods are improved, it could aid the advancement of noninvasive diagnostic techniques for mental illness or even degenerative diseases such as Alzheimer's. There is some research being done in this area, but there is room for more advancement.[1]

Section 2 contains overview of functional data analysis and steps to be taken in this type of data. Functional logistic regression and principal component functional logistic regression for multiple functional predictors are described in Section 3. Group Lasso for functional logistic regression is developed for classification in section 4. An application from fMRI data analysis and a simulation study are conducted to show the performance of the proposed methodology in section 5. We finally conclude the paper with a discussion and conclusion section.

FUNCTIONAL DATA ANALYSIS

Consider sample curves of the form $\{x_i(t),\ t \in T,\ i = 1, \ldots, n\}$, where T is an interval over which the observations were measured. The observations belong to the Hilbert space, $L_2(T)$, of square-integrable functions with the inner product

$$< f, g >= \int_T fg dt, \forall f, g \in L_2(T). \quad (1)$$

A vector $x_i = (x_{i1},...,x_{iN})$ represents the discrete measurements for the i^{th} subject of one variable, x, at N points in T. There are functional analogs of the traditional summary statistics. The functional sample mean, $\bar{\mathbf{x}}(t)$, is defined below:

$$\bar{x}(t) = n^{-1} \sum_{i=1}^{n} x_i(t). \quad (2)$$

The sample mean is computed point-wise at $t \in T$. Similarly, one can compute the covariance between measurements at two time points s and t

$$cov(s,t) = (n-1)^{-1} \sum_{i=1}^{n} (x_i(s) - \bar{x}_i(s))(x_i(t) - \bar{x}_i(t)). \quad (3)$$

BASIS EXPANSION OF FUNCTIONAL DATA

Let observations $\{x_i(t),\ t \in T,\ i = 1, \ldots, n\}$ belong a subspace of L_2 spanned by the p-dimensional basis system of independent functions, $\{\phi_1,\ldots, \phi_p\}$. The assumed smooth functional observation,

or linear expansion, $x_i(t)$, can be expressed in terms of the sum

$$x_i(t) = \sum_{k=1}^{p} a_k \phi_k. \quad (4)$$

Here, each a_k is called a basis coefficient. There are many different basis systems that can be used in a basis expansion. One of the most common systems is the Fourier basis system. A Fourier basis is defined by a Fourier series,

1, sin(ωt), cos(ωt), sin(2ωt), cos(2ωt), sin(3ωt), cos(ω3t), ...

Fourier bases are useful for data that are periodic, e.g. weather patterns.[4] Additionally, derivative estimation is easier with a Fourier basis as the derivatives of sin(t) and cos(t) are known and easy to compute.

For data that are not cyclical in nature, the most common choice is the B-spline basis system. They are computationally efficient and are flexible enough to approximate most non-periodic functions. In this paper, only B-spline bases will be used.

After a basis system has been chosen, the basis coefficients must be estimated. One method of basis coefficient estimation is that of least squares estimation. Denote the discrete observations of a functional dataset $y_j, j = 1,\ldots,N$, where N is the number of time points at which observations were taken. Assume independently and identically distributed measurement errors, ε_j, with E $[\varepsilon_j] = 0$ and Var $[\varepsilon_j] = \sigma^2$). However, least squares smoothing is inappropriate if the error assumptions (i.e. errors are independently and identically distributed measurement errors, ε_j, with E $[\varepsilon_j] = 0$ and Var $[\varepsilon_j] = \sigma^2$) are not true. In the case of functional data measured over time, observations at adjacent time points are likely correlated, violating the standard error assumptions.

Therefore a more common method, called spline smoothing by roughness penalty, is often used for estimating basis coefficients. This method is designed to estimate a curve that is rough enough to describe observed features of the data, but suppresses high-frequency features of the data, including noise. To find the coefficients of a smooth approximation of this type, the sum of squared errors is minimized with the added constraint of a roughness penalty. Roughness of a function is described by the curvature, or the squared second derivative. The quantity penalized is the integrated squared second derivative,

$$PEN_2(x) = \int [D^2 x(s)]^2 ds. \quad (5)$$

The corresponding sum of squared errors to be minimized is as follows:

$$PENSSE_\lambda(x|\mathbf{y}) = [\mathbf{y} - x(t)]'\mathbf{W}[\mathbf{y} - x(t)] + \lambda PEN_2(x), \quad (6)$$

where **W** is the matrix of weights describing the covariance structure of the errors. The smoothing parameter λ is chosen by the method of generalized cross validation developed by Craven and Wahba.[5] This is done by choosing λ such that it minimizes the following equation

$$GCV(\lambda) = \frac{n \times SSE}{(n - df(\lambda))^2}. \quad (7)$$

Here $df(\lambda)$ = trace ($S_{\phi\lambda}$), where

$$S_{\emptyset,\lambda} = \mathbf{\Phi}(\mathbf{\Phi}'\mathbf{W}\mathbf{\Phi})^{-1}\mathbf{\Phi}\mathbf{W} \quad (8)$$

is the hat matrix of the spline smoother.

Once you have smoothed functional observations, the statistical analysis can begin. As with traditional statistics, the beginning of FDA focused on univariate statistics. Our focus, however, is on a set of multiple functional predictors. In any functional data set with multiple functional predictors, and especially in fMRI, dimensionality is an issue. It may be important, for the sake of interpretation and computational expense, to select a smaller subset of important variables from the dataset. In any dataset, classification is often of interest. This particularly resonates within fMRI. Currently classification is used to explore brain functionality. For example, in a certain study subjects are given one of two stimuli. Does the brain behave differently in the presence of each stimulus so that one may be able to predict which stimulus a subject was presented with? If so, it would be important to identify

which areas of the brain are associated with this difference. As classification improves within the realm of fMRI, it could contribute to diagnosis of brain degeneration or mental illness in clinical settings. The rest of this paper focuses on a method of dimension reduction and variable selection in a dataset with multiple functional predictors and a binary response.

FUNCTIONAL PRINCIPAL COMPONENT LOGISTIC REGRESSION

Traditional principal component analysis (PCA) is a method of data reduction for multivariate datasets.[6] Consider a data matrix $X_{n\times m}$, where n is the sample size and m is the number of variables. Let $E[\mathbf{X}] = 0$ and $C = (\mathbf{X'X})^{-1}$ be the variance-covariance matrix. PCA is performed on the covariance matrix or correlation matrix of $\mathbf{X}$. It reduces dimension by finding a linear combination of the variables that has the maximum variance; this linear combination is the first principal component (PC). The next PC is found by finding a linear combination of the variables that is independent of the first PC and has the next largest variance. This goes on until $\min\{n-1, m\}$ PCs have been found.

FUNCTIONAL PRINCIPAL COMPONENT ANALYSIS

Consider a sample containing observations of one functional predictor. There is correlation between observations at adjacent points in T. In a linear model framework, this leads to the problem of high multicollinearity. Escabias et al. propose a method of functional principal component analysis (fPCA) for the logistic regression model with one functional predictor that alleviates this issue.[8] We will extend this method to a functional logistic regression model with multiple functional predictors.

As outlined by Ramsay and Silverman, fPCA is merely a functional analog of the traditional multivariate principal component analysis.[4] Assume functional observations of one variable $x_i(t) \in L_2$ where $i = 1,\ldots n$, with the usual functional sample mean, $\bar{x}(t)$, and sample covariance function, $\hat{C}(s,t), s,t \in T$. Without loss of generality, assume $\bar{x}(t) = 0$. The functional principal components, ξ_j, are found by solving the following functional eigenequation,

$$\int_T \hat{C}(s,t)f(s)ds = \lambda f(t). \quad (9)$$

The solutions to (9) are the eigenvalues, λ, and eigenfunctions, $f(t)$, of the covariance matrix $\mathbf{C}$. The number of eigenvalues is $n-1$. The i^{th} component of the j^{th} principal component, ξ_{ij}, is expressed as

$$\xi_{ij} = \int_T x_i(t)f_j(t)dt. \quad (10)$$

The solution of (9) cannot always be computed.

When the n sample functions of a functional predictor belong to the space $L_2(T)$ spanned by orthonormal bases $\{\phi_1,\ldots,\phi_p\}$, the functional PCs are equivalent to the multivariate PCs of the matrix $\mathbf{A\Psi}$.[8] Here, $\mathbf{A}$ is the $n\times p$ matrix of coefficients of the basis expansions and $\mathbf{\Psi}$ is a $p\times p$ matrix whose components are defined as

$$\psi_{ij} = \int_T \phi_i\phi_j dt. \quad (11)$$

FUNCTIONAL PRINCIPAL COMPONENT LOGISTIC REGRESSION FOR ONE FUNCTIONAL PREDICTOR

Escabias et al.[8] develop a principal component functional logistic regression model for one functional variable. We describe this method before extending it to the case with multiple functional predictors. Consider observations $\{(y_i, \mathbf{x_i}(t)), t\in T, i=1,\ldots,n\}$ where $\mathbf{x_i}(t)$ is a functional predictor. Each $x_{ij}(t)$ is the i^{th} observation at the j^{th} time point, and each $y_i \in \{0,1\}$. The conditional distribution of $Y_i|\mathbf{X_i}(t)$ is Bernoulli(π_i), with

$$\pi_i = E[Y_i|\mathbf{X_i}(t)] = \frac{exp\{\alpha + \int x_i(t)\beta(t)dt\}}{1+exp\{\alpha+\int x_i(t)\beta(t)dt\}} \; i = 1,\ldots,n, \quad (12)$$

where $\alpha \in \mathbb{R}$ and $\beta(t)$, the parameter, is a function. Making the logit transform, a generalized model is formed:[9]

$$l_i = \ln\left(\frac{\pi_i}{1-\pi_i}\right) = \alpha + \int_T x_i(t)\beta(t)dt \; i = 1,\ldots,n. \quad (13)$$

Under the assumption that $\beta(t)$ belongs to the same L_2 space spanned by $\{\phi_1,\ldots,\phi_p\}$,

$$\beta(t) = \sum_{k=1}^{p} \beta_k \phi_k. \quad (14)$$

The l_i's can be expressed in terms of the $\mathbf{A\Psi}$ matrix,

$$\mathbf{L} = \alpha 1_{n\times 1} + \mathbf{A\Psi}\beta, \quad (15)$$

where β is a $p \times 1$ dimensional vector containing the coefficients β_k for the basis expansion of $\beta(t)$. The basis coefficients β_k can be estimated using a Newton-Raphson algorithm maximizing the likelihood equation

$$Y'(X - \Pi) = 0. \quad (16)$$

In (16), $Y = (y_1,...y_n)$, $\prod = (\pi_1,...,\pi_n)$ and $X = (1|\mathbf{A\Psi})$.[8] Once the coefficients are estimated,

$$\hat{\beta}(t) = \sum_{k=1}^{p} \hat{\beta}_k \emptyset_k. \quad (17)$$

The vector $\mathbf{L}$ in (15) can be reexpressed in terms of the principal components of $\mathbf{A\Psi}$:

$$\mathbf{L} = \alpha 1_{nx1} + \mathbf{\Gamma V}'\beta = \mathbf{\Gamma}_\gamma \quad (18)$$

where $\mathbf{\Gamma}=(\xi_{ij})$ is the matrix of the p principal components and $\mathbf{V}$ is the matrix of the eigenvectors. This notation allows for estimation of the components of β,

$$\hat{\beta} = \mathbf{V}\ \hat{\gamma} \quad (19)$$

Reduction of the effects of multicollinearity occurs when a number of PCs, $s \leq p$, is chosen.

FUNCTIONAL PRINCIPAL COMPONENT LOGISTIC REGRESSION FOR MULTIPLE FUNCTIONAL PREDICTORS

The extension of PCA to the model with multiple functional predictors lies in the definition of the inner product.[6] Let $\{(y_i, \mathbf{x_i}^m(t)), t \in T, i = 1,...,n, m = 1,...,M\} \in L_2(T)$ spanned by $\{\phi_1,...,\phi_p\}$, where each $\mathbf{x_i}^m(t)$ is a functional predictor and each $y_i \in \{0,1\}$. The inner product, then, of two functional PCs can be defined as follows:

$$< \xi_i, \xi_j >= \sum_{m=1}^{M} \int_T \xi_i^m \xi_j^m dt. \quad (20)$$

The functional logistic regression model with multiple predictors is still defined by (13). The definition of π_i, however, changes with the change in inner product:

$$\pi_i = E[Y_i|\mathbf{X}_i(t)] = \frac{exp\{\alpha + \sum_{m=1}^{M} \int_T x_i^m(t)\beta^m(t)dt\}}{1 + exp\{\alpha + \sum_{m=1}^{M} \int_T x_i^m(t)\beta^m(t)dt\}}, i = 1,...,n. \quad (21)$$

Making the logit transform,

$$l_i = \alpha + \sum_{m=1}^{M} \int_T x_i^m(t)\beta^m(t)dt, \quad i = 1,...,n. \quad (22)$$

To perform the dimension reducing PCA, we redefine the design matrix $\mathbf{A\Psi}$. Here,

$$\mathbf{A}_{n\times(Mp)} = [\mathbf{A}^1|...|\mathbf{A}^M], \quad (23)$$

and

$$\mathbf{\Psi}_{(Mp)\times(Mp)} = \begin{vmatrix} \mathbf{\Psi^1} & \mathbf{0} & ... & ... \\ \mathbf{0} & \mathbf{\Psi^2} & \mathbf{0} & ... \\ ... & ... & ... & ... \\ \mathbf{0} & ... & ... & \mathbf{\Psi^M} \end{vmatrix}. \quad (24)$$

Each $\mathbf{A^m}$ is an $n \times p$ matrix of basis coefficients. $\mathbf{\Psi}$ is a diagonal block matrix with each $\mathbf{\Psi^m}$ having dimensions $p \times p$. Now, $\beta = (\beta_1', ..., \beta_M')'$. Redefining (15),

$$\mathbf{L} = \alpha 1_{nx1} + \mathbf{A\Psi}\beta = \alpha 1_{nx1} + \sum_{m=1}^{M} \mathbf{A}^m \mathbf{\Psi}^m \beta^m. \quad (25)$$

This definition depends on the assumption that each observation $x_i{}^m(t)$ and the respective parameter function, $\beta^m(t)$ can be defined by the same set of basis functions $\{\emptyset_1^m(t), ..., \emptyset_p^m(t)\}$. Below, we express $\mathbf{L}$ in terms of the principal components

$$\mathbf{L} = \alpha 1_{nx1} + \sum_{m=1}^{M} \mathbf{\Gamma}^m \mathbf{V}^{m\prime} \beta^m, \quad (26)$$

where $\mathbf{\Gamma}^m = (\xi_{ij}^m)_{nxp}$ are the principal components of the $\boldsymbol{A\Psi}^m$ and $\boldsymbol{V^m}$ is the matrix of eigenvectors. The number of PCs $s_m \leq p_m$, to be chosen for each of the M should be determined by cumulative variance. For simplicity, we chose the same number of PCs, s, for each of the M predictors. After the dimension has been reduced on the within-variable level, (21) becomes

$$\pi_{i(s)} = \frac{exp\{\alpha_{(s)} + \sum_{m=1}^{M} \sum_{j=1}^{s} \xi_{ij(s)}^m f_{ij(s)}^m \beta_{ij(s)}^m\}}{1 + exp\{\alpha_{(s)} + \sum_{m=1}^{M} \sum_{j=1}^{s} \xi_{ij(s)}^m f_{ij(s)}^m \beta_{ij(s)}^m\}} \ i = 1,...,n. \quad (27)$$

We can re-express (25) as

$$\mathbf{L} = \alpha_{(s)} 1_{nx1} + \sum_{m=1}^{M} \mathbf{\Gamma}_{(s)}^{m} \mathrm{V}_{(s)}^{m\prime} \beta_{(s).}^{m} \quad (28)$$

The vectors β^m can be estimated as in (19),

$$\hat{\beta}^m = \mathbf{V}^m \hat{\gamma}^m. \quad (29)$$

Although multicollinearity has been dealt with on a within variable basis, as M gets large there could be multiple predictors providing similar information. In the case of fMRI data, time series of adjacent voxels are expected to be similar. There is a need to reduce dimension on the multiple variable level by selecting only those functional predictors which are relevant to the response. This will alleviate another potential source of multicollinearity.

GROUP LASSO FOR FUNCTIONAL LOGISTIC REGRESSION MODEL

The principal component analysis allows for removal of redundant information on a within predictor basis. As stated before, this is necessary due to the autocorrelation of observations between time points. There is also a need to select only those predictors which provide relevant information to the model.

There are many methods of variable selection used in linear models and generalized linear models. Model selection techniques, such as stepwise and forward selection, cycle algorithmically through subsets of variables until certain criteria are met. The variables included in the various steps of the algorithm are determined by previously selected p-values. These methods are inherently subjective, as it is up to the person analyzing the model to choose the ”best” model based on a set of criteria. The criteria that determine the quality of the model are also chosen by the analyst.

LASSO

A more objective method of variable selection, the lasso, was introduced by Tibshirani in 1996. The lasso is a method that simultaneously performs model selection and parameter estimation. It is an L_1 regularization technique that performs this variable selection by shrinking certain β coefficients to exactly 0, excluding those predictors from the model. The other, non-zero, coefficients represent variables that are relevant to the model. This is done by solving the least squares estimation subject to a constraint on the β Assume a standard regression model with independent observations $\{(y_i, \mathbf{x}_i,\ i = 1,...,n\}$ where $\mathbf{x}_i = (\mathrm{x}_{i1},...,\mathrm{x}_{ip})$ The estimates of regression coefficients by the lasso method $(\hat{\alpha}, \hat{\beta})$ are

$$(\hat{\alpha}, \hat{\beta}) := arg\ min\left\{\sum_{i=1}^{n}(y_i - \alpha - \sum_{i=1}^{p}\beta_i x_{ij})^2\right\}, \quad (30)$$

under $\sum_j |\beta_j| \leq t$, where $t \geq 0$. Note that is not penalized.

Tibshirani also applied the lasso to the logistic regression model.[10] Consider independent observations $\{(y_i, \mathbf{x}_i,\ i = 1,...,n\}$ where $y_i \in \{0, 1\}$. The variable selection and model estimation are performed by maximizing the loglikelihood function,

$$l(\beta) = \sum_{i=1}^{n} y_i(\mathbf{x}_i'\beta) - \ln(1 + \exp\{\mathbf{x}_i'\beta\}, \quad (31)$$

under $\sum_j |\beta_j| \leq t$, where $t \geq 0$. An iterated reweighted least squares algorithm is used to compute $\hat{\beta}$ under these conditions.

GROUP LASSO

Consider a linear model with multiple predictors, some of them categorical. A categorical predictor with l levels will be represented in the model by $l - 1$ variables. The lasso only has the ability to shrink individual regression coefficients to zero. In the case of the categorical predictor, this has little interpretation. If the categorical predictor is not relevant to the response, all l–1 variables should be removed from the model.

Consider independent observations $\{(y_i, \mathbf{x}_i),\ i = 1,...,n\}$ where $\mathbf{x}_i = \mathrm{x}_{i1}', ..., \mathrm{x}_{iM}')'$. Each x_{im} represents a group of predictors. The linear regression model is defined as

$$\mathbf{Y} = \alpha + \sum_{m=1}^{M} \mathbf{x}_m \beta_m + \varepsilon, \quad (32)$$

where $\alpha \in \mathbb{R}$ is the intercept, each β_m is a vector whose components are the regression coefficients for the m^{th} group of predictors and $\mathbf{Y}_{nx1}$ is the vector of responses. Yuan and Lin[11] developed a method of variable selection called the group lasso considers each of the M groups of variables for inclusion or exclusion in the model. The coefficient estimates are defined as

$$\hat{\beta} = \arg\min\left(\|\mathbf{Y} - \mathbf{X}\beta\|_2^2 + \lambda \sum_{m=1}^{M} \|\beta_m\|_2 \right), \qquad (33)$$

where λ is a tuning parameter and $\beta = (\alpha, \beta_1', \ldots, \beta_M')'$. The penalty is a mixture of L_1 and L_2 regularization methods, the lasso and the ridge regression penalties.

GROUP LASSO FOR FUNCTIONAL LOGISTIC REGRESSION

Meier et al.[7] describe a method of group lasso for the multivariate logistic regression model. Consider independent observations $\{(y_i, \mathbf{x}_i),\ i = 1,\ldots,n\}$ where $y_i \in \{0,1\}$ and $\mathbf{x}_i = (x_{i1}', \ldots, x_{iM}')'$. Each x_i^m represents a group of predictors. Group lasso is performed by minimizing the following convex function:

$$S_\lambda(\beta) = -l(\beta) + \lambda \sum_{m=1}^{M} s(df_m)\|\beta^m\|_2, \qquad (34)$$

where df_m is the degrees of freedom of the m^{th} group of predictors. The use of $s(df_m) = df_m^{1/2}$ is suggested.[7] The solution to this equation is the logistic group lasso estimator, $\hat{\beta}_\lambda$. It is found using a block co-ordinate gradient descent minimization algorithm. The algorithm uses a second-order Taylor series expansion,

$$S_\lambda(\hat{\beta}^{(t)} + \mathbf{d}) \approx -\left\{ l(\hat{\beta}^{(t)}) + \mathbf{d}^T \nabla l(\hat{\beta}^{(t)}) + \frac{1}{2}\mathbf{d}^T H^{(t)} \mathbf{d} \right\} + \lambda \sum_{m=1}^{M} s(df_m) \left\| \hat{\beta}_m^{(t)} + \mathbf{d}_m \right\|_2 = M_\lambda^{(t)}(\mathbf{d}). \qquad (35)$$

The algorithm summarized in Figure 1 begins by assuming an initial parameter vector, $\beta^{(0)}$. For each of the M groups of variables, the algorithm finds $\mathbf{d}$ that minimizes $M_\lambda^t(\mathbf{d})$. If this $\mathbf{d}$ is not identically 0, the estimate of β is updated. The updated estimate $\hat{\beta}^{(t+1)}$ is the previous estimate, $\beta^{(t)}$ plus a scalar times $\mathbf{d}$. This algorithm proceeds for each group until some convergence criterion is met. The choice of λ is dependent upon n and the degrees of freedom of each of the M groups. In the multivariate model, group lasso performed on a dataset containing M groups of discrete predictors. A number of groups of predictors less than M is selected. This version of the group lasso is shown to be asymptotically consistent.[7] The minimization can be done in R, using the package grplasso written by Meier et al.[7]

Step	*Algorithm*
1	Let $\beta \in \mathbb{R}^{p+1}$ be an initial parameter vector
2	For $g = 0, \ldots, G$
	$H_{gg} \leftarrow h_g(\beta) I_{\mathrm{df}_g}$
	$\mathbf{d} \leftarrow \arg\min_{\mathbf{d}\mid \mathbf{d}_k = 0, k \neq g} \{M_\lambda(\mathbf{d})\}$
	if $\mathbf{d} \neq \mathbf{0}$
	$\alpha \leftarrow$ line search
	$\beta \leftarrow \beta + \alpha\mathbf{d}$
	end
	end
3	Repeat step 2 until some convergence criterion is met

FIGURE 1: Group Lasso Algorithm: Outline of the block coordinate gradient descent algorithm used to perform grouped variable selection in the logistic regression model.[7]

We apply this group lasso method to the functional logistic regression model with multiple functional predictors. Recall observations $(y_i, \mathbf{x}_i^m(t)), t \in T, i = 1, \ldots, n,\ m = 1, \ldots, M\} \in L_2^M(T)$ spanned by $\{\phi_1, \ldots, \phi_p\}$ where each $\mathbf{x}_i^m(t)$ is a functional predictor and each $y_i \in \{0,1\}$. Consider the model in (26), after principal components have been chosen. The loglikelihood function, $l(\beta)$, of the functional logistic regression model is

$$l(\beta) = \sum_{i=1}^{n} y_i(\alpha_{(s)} + \sum_{m=1}^{M} \int x_i^m(t)\beta_{(s)}^m(t)dt) - \ln(1 + \exp\{\alpha_{(s)} + \sum_{m=1}^{M} \int x_i^m(t)\beta_{(s)}^m(t)dt\}). \qquad (36)$$

Using the definition $l(\beta)$ in (36) we minimize the objective equation,

$$S_\lambda(\beta) = -l(\beta) + \lambda \sum_{m=1}^{M} s(df_m)\|\beta^m\|_2. \qquad (37)$$

In the case of our method of principal component logistic regression with multiple functional predictors, the degrees of freedom in (37) is equiv-

alent to the number of chosen PCs, *s*. In the functional case, each of the *M* functional predictors is defined by a group of *s* coefficients. When one entire group of coefficients is shrunk to 0, it excludes the corresponding single functional predictor from the model. For any set of coefficients that are not equal to zero, the corresponding functional predictor is included in the model. In essence, the group lasso performs single variable selection in the functional logistic regression model with multiple functional predictors.

NUMERICAL EXAMPLES

In a Functional Magnetic Resonance Imaging experiment, an experimenter aims to measure the amount of activation in each voxel of the brain. When a part of the brain is active, there is increased bloodflow to the area. fMRI measures the change in blood flow using the bloodoxygen-level-dependent (BOLD) contrast.[2] Assessing which parts of the brain are active during an fMRI experiment allows researchers to determine which parts of the brain respond to certain stimuli. There is a need for classification tools in the statistical analysis of fMRI. Logistic regression could be used to distinguish between brains at rest and those presented with stimuli. Another application would be to distinguish between subjects receiving one of two particular stimuli. For example, an experimenter may play pieces of music or speech to a subject.[12] Being able to classify which stimulus was presented allows one to learn more about the way the brain works. Classification also has an application in diagnosis of mental illness or degenerative disease.

SIMULATION STUDY

We assess our methodology using a simulation study. Using the R package neuRosim, we were able to simulated preprocessed fMRI data. The package can be used to simulated fMRI time series or complete 4D fMRI volumes. With neuRosim, one can define the onset and duration of stimuli. One can specify the effect size of the stimuli, TR and times of spatial and temporal noise.[13]

We simulated preprocessed four dimensional fMRI data for an area of 4000 voxels containing two non-overlapping regions of activation. We used neuRosim to create block designs of a stimulus followed by rest. Design 1 presented an effect size that was larger in Region 1 than in Region 2. Design 2 created activation that was larger in Region 2 than in Region 1. Half of the observations in each simulation were simulated under Design 1, the other half under Design 2. Our goal was to use the developed method of group lasso for functional logistic regression with multiple functional predictors to classify the validation set into the proper groups. We simulated 15 datasets each at two levels of signal-to-noise ratio (SNR), 0.75 and 3.87, and two levels of subject size, 30 and 50. An observation simulated under Design 1 was given a y value of 1, otherwise $y_i = 0$. Two-fold cross validation was then used to form training and validation sets. All analysis was performed in R; the package grplasso was used to perform the final variable selection.[7] Observations with $\hat{\pi}_i > 0.5$ were classified as $y_i = 1$, otherwise $y_i = 0$.

For each of the four sets, two-fold cross validation resulted in 30 models. We report the number of voxels selected out of the initial 4000 voxels. We also report sensitivity, defined as the ratio of true positives to true positives plus false negatives; false positive rate, the ration of false positives to false positives plus true negatives; and the accuracy, defined as the ratio of true positives plus true negatives to the number of observations in the validation set. These findings can be seen in Table 1. We did not report the number of principal components selected in the table. In the cases where $n = 50$, the original number of basis functions was 43. After PCA, 8, 9 or 10 principal components were selected every time. In the cases where $n = 30$, 7 or 8 PCs were chosen. The method classifies well, even after use of a small number of voxels. As expected, the cases with fewer subjects have lower sensitivity and accuracy and a higher false positive rate. In the two simulations with lower SNR classification appears to improve, which is surprising.

fMRI EXAMPLE

To test this methodology on a real dataset, we used fMRI data collected by Auburn University's MRI

TABLE 1: Simulation Results: The values reported are the means, followed by their standard deviations in parentheses.

	No Voxels Selected	Sensitivity	False Positive Rate	Accuracy
SNR = 3.87 n = 50	5.414 (1.842)	0.935 (0.096)	0.053 (0.079)	0.936 (0.060)
SNR = 0.75 n = 50	5.267 (1.617)	0.953 (0.099)	0.044 (0.082)	0.949 (0.670)
SNR = 3.87 n = 30	3.867 (1.332)	0.825 (0.188)	0.166 (0.197)	0.840 (0.158)
SNR = 0.75 n = 30	3.967 (1.449)	0.874 (0.153)	0.108 (0.159)	0.871 (0.106)

SNR: Signal-to-noise ratio.

Research Center. The data was collected from 6 subjects on a 7T MRI scanner, and each scanning session lasted 1000s. The data were preprocessed by the experimenters using SPM8. Slice timing correction was made. Spatial realignment, normalization, and smoothing were performed. And, finally, the data were detrended. The complete raw data voxel-wise time series were then extracted using MarsBaR.

The experimental design was a block design with two conditions.In one condition, subjects were asked to use four lines to connect all dots in Figure 2. In the other, subjects were asked to use five lines to connect all dots in Figure 2. Conditions were presented in a random sequence, and each condition was followed by a period of rest. All subjects were able to connect the nine dots using five lines, but only one (Subject 5) was able to solve the puzzle using four lines. Our aim was to use group lasso for functional logistic regression to classify the six subjects as having solved the four line puzzle, $y = 1$, or as being unable to solve the puzzle, $y = 0$.

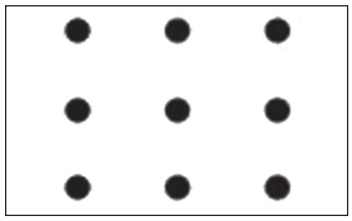

FIGURE 2: fMRI Nine Dot Experiment: Subjects were asked to use four and five lines to connect all nine dots in the figure above.

According to the experimenters, there were two important regions of interest (ROIs) in this study. These regions are the left and right anterior temporal lobes (ATL) which can be seen in Figure 3. They are associated with semantic memory, knowledge of objects and facts. From the right ATL, 7560 voxel time series were extracted. From the left ATL, 6584 voxel time series were extracted.

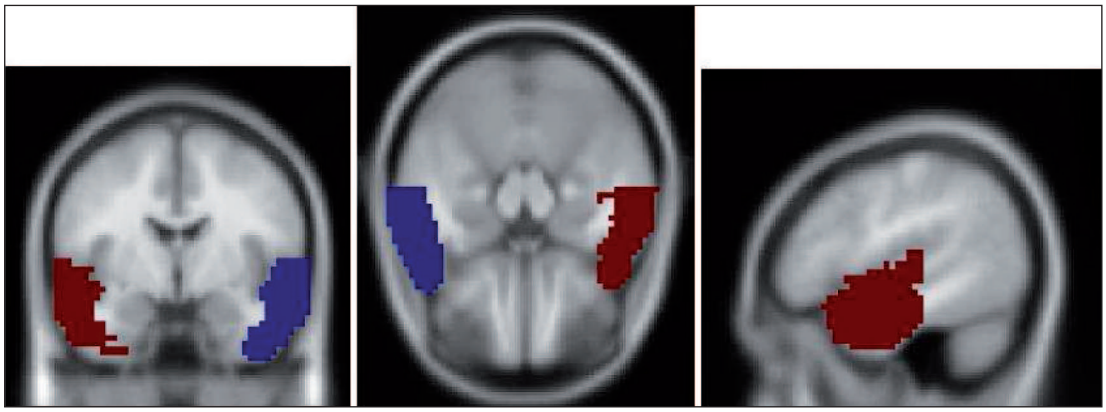

FIGURE 3: Voxel Mask of the Anterior Temporal Lobe: The right anterior temporal lobe is in red, the left in blue.

This led to a total of time series from 14144 voxels. To perform PCA on the 14144 $\mathbf{A}\Psi$ matrices of the spline smooths, the number of basis functions, p, must be less than the number of subjects. We chose to use 5 basis functions. After performing the principal component analysis, 3 PCs were chosen. From the 14144 voxels, the group lasso procedure selected 11 voxels. From these 11 voxel time series, the classification procedure correctly selected Subject 5 as having solved the puzzle.

CONCLUSION

We have developed a viable method of variable selection for functional logistic regression by employing the group lasso in an interesting way. There are obvious limitations with small sample sizes. Being limited to a number of spline basis functions that is less than the sample size could lead to poor estimation of the functional observations. The method employed is also computationally expensive for a large number of functional variables and subjects. Each spline smooth procedure must be performed for all variables and subjects. The application in the field of Functional Magnetic Resonance Imaging is exciting. In future studies on real data, it would be interesting to study the neurological significance of the voxels selected by the group lasso.

REFERENCES

1. Tian TS. Functional data analysis in brain ımaging studies. Front Psychol 2010;1:35.
2. Huettel SA, Song AW, McCarthy G. Functional Magnetic Resonance Imaging. 2nd ed. Sunderland, MA: Sinauer Associates; 2009. p.1-542.
3. Viviani R, Grön G, Spitzer M. Functional principal component analysis of fMRI data. Hum Brain Mapp 2005;24(2):109-29.
4. Ramsay JO, Silverman BW. Functional Data Analysis. 2nded. New York: Springer-Verlag; 2005. p.1-426.
5. Craven P, Wahba G. Smoothing noisy data with spline functions. Numer Math 1979;31: 377-403.
6. Joliffe IT. Principal Component Analysis. 2nd ed. New York: Springer-Verlag; 2004. p.1-487.
7. Meier L, van de Geer S, Bühlmann P. The group lasso for logistic regression. J R Stat Soc Series B 2008;70(1):53-71.
8. Escabias M, Aguilera AM, Valderrama MJ. Principal component estimation of functional logistic regression: discussion of two different approaches. J Nonparametr Stat 2004;16(3-4):365-84.
9. Müller H, Stadtmüller U. Generalized functional linear models. Ann Stat 2005;33(2):774-805.
10. Tibshirani R. Regression Shrinkage and Selection via the Lasso. J R Stat Soc Series B 1996;58(1):267-88.
11. Yuan M, Lin Y. Model selection and estimation in regression with grouped variables. J R Stat Soc Series B 2006;68(1):49-67.
12. Ryali S, Kaustubh S, Abrams D, Menon V. Sparse logistic regression for whole-brain classification of fMRI data. Neuroimage 2010;51(2):752-64.
13. Welvaert M, Durnez J, Moerkerke B, Verdoolaege G, Rosseel Y. neuRosim: An R Package for Generating fMRI Data. J Stat Softw 2011;40(10):1-18.