Biometrics 00, 1–25 DOI: 10.1111/j.1541-0420.2005.00454.x
July 2011

Robust Classification of Functional and Quantitative Image Data Using Functional Mixed Models

Hongxiao Zhu[1], Philip J. Brown[2], and Jeffrey S. Morris[3,*]

[1]Statistical and Applied Mathematical Sciences Institute

Research Triangle Park, NC 27709, U.S.A.

[2]School of Mathematics, Statistics and Actuarial Science, University of Kent, U.K.

[3]Department of Biostatistics, University of Texas MD Anderson Cancer Center,

Houston, TX 77230, U.S.A.

email: jefmorris@mdanderson.org

Summary: This paper describes how to perform classification of complex, high-dimensional functional data using the functional mixed model (FMM) framework. The FMM relates a functional response to a set of predictors through functional fixed and random effects, which allows it to account for various factors and between-function correlations. Classification is performed through training the model treating class as one of the fixed effects, and then predicting on the test data using posterior predictive probabilities of class. Through a Bayesian scheme, we are able to adjust for factors affecting both the functions and the class designations. While the method we present can be applied to any FMM-based method, we provide details for two specific Bayesian approaches: the Gaussian, wavelet-based functional mixed model (G-WFMM) and the robust, wavelet-based functional mixed model (R-WFMM). Both methods perform modeling in the wavelet space, which yields parsimonious representations for the functions, and can naturally adapt to local features and complex nonstationarities in the functions. The R-WFMM allows potentially heavier tails for features of the functions indexed by particular wavelet coefficients, leading to a down weighting of outliers that makes the method robust to outlying functions or regions of functions. The models are applied to a pancreatic cancer mass spectroscopy data set and compared with some other recently developed functional classification methods.

Key words: Bayesian Modeling; Classification; Discrimination; Functional data analysis; Image Analysis; Mixed models; Robust Regression; Wavelets.

1. Introduction

In recent years, an increasing number of application areas yield functional data, which consist of curves observed on some fine grid. The scope of functional data includes quantitative image data: images whose pixel intensities represent some quantitative measure that can be viewed as functions on a higher dimensional domain. While functional data can have many different characteristics, they are increasingly high-dimensional, with automated measurements taken on finer and finer grids, and also more complex, with many applications yielding functions that are highly structured and have many local features.

One important problem of interest in functional data analysis is classification, whereby one wishes to assign an individual to a predefined discrete class based on the observed functional or image data. Existing methods for functional data classification can be organized into the following categories: (1) **Density-based**. The functional data are first projected to some finite dimensional feature space (through functional principal component analysis (FPCA), splines, etc.) on which the densities of each class are estimated, either parametrically (e.g., linear discriminant analysis) or non-parametrically (e.g., kernel density estimation (KDE) or using Bayesian nonparametrics). Classification of new observations is performed based on the estimated densities (see Hall, Poskitt and Presnell (2001), James (2001), Ferraty and Vieu (2003), etc.). Alternatively, the joint distribution of class and function can be estimated and used to perform classification (Bigelow and Dunson (2009)). (2) **Regression-based**. A regression model is constructed linking categorical responses with functional predictors, frequently through generalized linear models. The model parameters are estimated and used for classification (e.g., James (2002), Müller and Stadtmüller (2005), Müller (2005), Leng and Müller (2005), Zhu, Vannucci and Cox (2010)). (3) **Algorithmic-based**. Dimension reduction is performed to transform to a multivariate problem, and then one of a variety of

nonparametric classification tools such as k-nearest neighbor or support vector machines are applied for classification (see Ramsay (2000), Li and Yu (2008)) .

While there are a large number of methods for functional data classification in the current literature, there are still important aspects that are not simultaneously handled by existing methods, including adjustment for covariates affecting either the function or the class, classification of subjects based on multiple observed functions, robustness to outliers in classification, and the ability to handle complex, extremely high dimensional functional and image data. Most density-based approaches in current literature assume i.i.d. functions, and so do not naturally provide a way to classify subjects based on multiple observed functions that are expected to be correlated; furthermore, the approaches do not adjust for factors affecting the functions nor for other predictors of class. Similarly, algorithmic-based approaches tend not to account for other factors influencing the class or the observed functions when performing classification and cannot easily handle multiple correlated functional predictors. Regression-based methods can naturally accommodate other predictors of class and can be robustified through introducing heavier-tailed link functions or error distributions, but they typically do not adjust for factors affecting the functions and typically cannot handle multiple correlated functional predictors. Furthermore, many current methods do not scale up to the setting of complex, high-dimensional data, either because their functional representations are not flexible enough to capture complex features of the functions or because they cannot be feasibly applied to extremely high-dimensional functions.

In this paper, we introduce a novel method for functional data classification using the functional mixed model (FMM) framework which, as we will demonstrate, is able to account for all of these factors simultaneously. The FMM relates a functional response to a set of predictors through functional fixed and random effects, and is typically used to estimate and perform inference on fixed effect functions characterizing, for example, the systematic

difference in the mean functions between groups. Here, we will show how to perform classification in this framework by first fitting an FMM to the training data with class as one of the fixed effect predictors and then performing classification of the test data using posterior predictive probabilities of class membership. This approach has numerous advantages, and if the particular FMM used is flexible enough to capture the relevant features in the observed functions, it has the potential to outperform other standard approaches.

The inclusion of general fixed and random effect covariates allows one to adjust for the effects of confounding factors on the function, which can be of high importance in many applications. For example, in mass spectrometry proteomics, it has been shown that observed functions collected in different time blocks can differ systematically. By modeling the block effects, the FMM can automatically adjust for these factors when building the classification model. Similarly, effects of factors such as gender, age, and hospital on the function can be taken into account. The inclusion of fixed and random effects makes it possible to model multiple functions from the same individual, and thus take into account the within-subject correlations among the functions when performing classification. In our proposed method, we not only model the covariates that affect the functional data, but also are able to hierarchically model the covariates that directly affect the class designation. A generalization of this approach can be used to integrate information across multiple predictors, both functional and scalar, in performing classification.

In contrast with most regression-based classification methods, the FMM treats functional observations as responses and class labels as predictors, and therefore can be considered a density-based method. Efron (1975) showed that in simple multivariate settings, classification based on normal likelihood (referred to as the *normal discriminant procedure*) is asymptotically more efficient than that based on regression (generalized linear models), but generally is not as robust since it may be more susceptible or sensitive to model misspecification. By

analogy, one might expect that, relative to functional regression approaches, density-based classification approaches modeling the function as response may improve efficiency in the functional classification setting as long as the model is flexible enough to capture the true features of the functional data. Thus, it is interesting to consider classification based on FMM-based methods with flexible representations for the functions.

While the proposed classification approach can be used with any FMM-based method, we will provide details using two specific FMM approaches: the Gaussian, wavelet-based functional mixed model (G-WFMM, Morris and Carroll (2006)) and the robust, wavelet-based functional mixed model (R-WFMM, Zhu, Brown and Morris (2011)). Both demonstrate outstanding flexibility and computational feasibility for modeling complex, high-dimensional functional data. These methods perform modeling in the wavelet space, which yields parsimonious representations for the functions, can naturally adapt to local features in the functions, and accommodates various nonstationarities in the within-function covariance surfaces, including different variances and varying smoothness at different parts of the functions or images. Further, both of these methods are computationally convenient, having been applied to extremely large functional and image data sets using available automated code in which the user simply provides the functional responses and covariate design matrices if they are satisfied with the supplied automated choices of wavelet basis and vague proper priors. The R-WFMM has the additional advantage of modeling with a more flexible class of likelihoods, allowing potentially heavier tails for features of the functions indexed by particular wavelet coefficients, as determined by the data, and leading to a down weighting of outliers that makes the method robust to outlying functions or regions of functions. Both of these approaches naturally extend to functions with higher dimensional domains (e.g., quantitative image data (see Morris, et al. (2011)), so the methods we describe here can also be applied to classify individuals based on image data.

The outline for the rest of the paper is as follows: Section 2 introduces the general FMM-based classification approach. Implementation details are discussed in Section 3. Two specific methods, the Gaussian WFMM and the Robust WFMM, are presented in Section 3.1 and Section 3.2, respectively. The method is applied to a pancreatic cancer mass spectrometry application and compared with some alternative methods in Section 4, with conclusions and a discussion following in Section 5. The online supplementary materials contain some derivations, further computational details, and further results.

2. FMM-based Classification

2.1 *Classification using FMM Framework*

Let $Y_i(t), i = 1, \ldots, n$ be functional observations on a compact set $\mathcal{T}$, and $c_i \in \{1, ..., q\}$ be the corresponding class labels. A functional data classification model aims to find a "rule" to assign new observation $Y^0(t)$ to one of the q classes. To adjust for commonly encountered issues in many applications, we include the possibility of two types of covariates: $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^T$ are covariates corresponding to factors that influence the functional observations, and $\mathbf{z}_i = (z_{i1}, \ldots, z_{im})^T$ are covariates indicating a possible clustering structure within the data induced by the experimental design. We treat $\mathbf{x}_i$ as covariates for fixed effects and $\mathbf{z}_i$ as covariates for random effects. Covariates for fixed effects are usually profile variables such as the age of patients, the types of tissue, etc. Random effect covariates are usually variables indicating subgroup designation, such as the family the observation belongs to, the hospital at which the measurement was made, block structure from the experimental design, or subject indicators when there are multiple functions per subject. We model the relationship between the functional observations, their class labels and all other covariates through the following functional mixed model:

$$Y_i(t) = \mathbf{v}_i^T \boldsymbol{G}(t) + \mathbf{x}_i^T \boldsymbol{B}(t) + \mathbf{z}_i^T \boldsymbol{U}(t) + E_i(t), \tag{1}$$

where $\mathbf{v}_i$ is a vector with the c_i^{th} component 1 and 0 elsewhere, and $\boldsymbol{G}(t) = (G_1(t), \ldots, G_q(t))^T$ denotes the functions of the group mean for the q classes. Here $\boldsymbol{B}(t) = (B_1(t), \ldots, B_p(t))^T$

and $\boldsymbol{U}(t) = (U_1(t), \ldots, U_m(t))^T$ are the coefficient functions of fixed effects and random effects, respectively. $E_i(t)$ is the residual error function. The random effect coefficients and the error terms are assumed to be mean zero with covariance functions $\text{Cov}\{U_j(t)\} = \Sigma_U(s,t)$ and $\text{Cov}\{E_i(t)\} = \Sigma_E(s,t)$, independently across j, i, respectively, and with $\boldsymbol{U}(t)$ and $\boldsymbol{E}(t) = (E_1(t), \ldots, E_n(t))^T$ independent of each other.

Denote $\mathbf{Y}(t) = (Y_1(t), \ldots, Y_n(t))^T, \mathbf{V} = (\mathbf{v}_1, \ldots, \mathbf{v_n})^T, \mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^T, \mathbf{Z} = (\mathbf{z}_1, \ldots, \mathbf{z}_n)^T$. In a typical classification problem, the data are split into a training set and a test set. In the training set, the class labels (therefore the covariates $\mathbf{V}$) are known, while in the test set $\mathbf{V}$ need to be predicted. In this paper, we aim to estimate the regression coefficients and covariance parameters, denoted as $\boldsymbol{\Theta} = \{\boldsymbol{G}(t), \boldsymbol{B}(t), \boldsymbol{U}(t), \Sigma_U(s,t), \Sigma_E(s,t)\}$ based on the training set, and predict the class labels c^0 (or $\mathbf{v}^0$) for future functional observation $Y^0(t)$ (test data). Note that because $\mathbf{V}$ are known in the training data, the coefficients $\boldsymbol{G}(t)$ and $\boldsymbol{B}(t)$ are both treated as fixed effects.

The approach for fitting model (1) depends on the specifics of the chosen FMM-based method, including how the random functions are represented. For now, assume we have some training process yielding estimates $\boldsymbol{\Theta}$, and that the prediction for new observations can be summarized in two cases:

(1) **The random effect for the new observation is available in the test data.** This happens when the new observation is drawn from a population from which all or part of the training data are drawn. In this case, we can compute the posterior odds of $c^0 = j$ versus $c^0 = 1$ as

$$\text{Odds}(j) = \frac{f(Y^0 \mid c^0 = j, \mathbf{x}^0, \mathbf{z}^0, \mathbf{Y}, \mathbf{V}, \mathbf{X}, \mathbf{Z})}{f(Y^0 \mid c^0 = 1, \mathbf{x}^0, \mathbf{z}^0, \mathbf{Y}, \mathbf{V}, \mathbf{X}, \mathbf{Z})} \cdot \frac{f(c^0 = j)}{f(c^0 = 1)} \tag{2}$$

for $j = 2, \ldots, q$. Here $f(c^0 = j)$ and $f(c^0 = 1)$ are the pre-specified prior probabilities for the class designation, $f(\cdot \mid c^0, \mathbf{x}^0, \mathbf{z}^0, \mathbf{Y}, \mathbf{V}, \mathbf{X}, \mathbf{Z})$ represents of the posterior predictive density of the new function Y^0, and $\{\mathbf{Y}, \mathbf{V}, \mathbf{X}, \mathbf{Z}\}$ represent the training data. The

posterior predictive density is obtained by

$$\int f(Y^0 \mid c^0, \mathbf{x}^0, \mathbf{z}^0, \boldsymbol{\Theta})\ f(\boldsymbol{\Theta}|\mathbf{Y}, \mathbf{V}, \mathbf{X}, \mathbf{Z})\ d\boldsymbol{\Theta}, \tag{3}$$

with $f(\boldsymbol{\Theta}|\mathbf{Y}, \mathbf{V}, \mathbf{X}, \mathbf{Z})$ being the posterior density of the parameters. Detailed computation of the density of a random process depends on how the functional data is modeled and how the model is estimated, which will be discussed in detail in Section 3.

The posterior predictive density estimated by (3), based on a Bayesian approach, integrates over the posterior uncertainty of the parameters. If a frequentist approach is used, then a point estimate of $\boldsymbol{\Theta}$ can be obtained (denoted as $\widehat{\boldsymbol{\Theta}}$), and the posterior predictive density in (2) can be replaced by the conditional predictive density $f(Y^0 \mid c^0, \mathbf{x}^0, \mathbf{z}^0, \widehat{\boldsymbol{\Theta}})$. Of course, an advantage of using the posterior predictive density is that it takes into account the variability of estimated parameters.

(2) **The random effect for the new observation is not available in the test data.** For instance, the new observation corresponds to a patient from a new hospital. In this case, we simply replace the first factor of the integrand in (3) with $f(Y^0 \mid c^0, \mathbf{x}^0, \widetilde{\boldsymbol{\Theta}})$, with $\widetilde{\boldsymbol{\Theta}}$ being a subset of $\boldsymbol{\Theta}$ with $\mathbf{U}(t)$ omitted. In other words, with $\mathbf{z}^0$ unknown, we work with the marginal likelihood of Y^0 with the random effects integrated out.

With the odds computed using equation (2), the posterior predictive probabilities for class designations can be computed straightforwardly using

$$\Pr(c^0 = j \mid Y^0, \mathbf{x}^0, \mathbf{z}^0, \mathbf{Y}, \mathbf{V}, \mathbf{X}, \mathbf{Z}) = \begin{cases} \frac{\text{Odds}_{(j)}}{1+\sum_j \text{Odds}_{(j)}} & \text{for } j = 2, \ldots, q. \\ \frac{1}{1+\sum_j \text{Odds}_{(j)}} & \text{for } j = 1. \end{cases}$$

2.2 *Prediction on Correlated Functions*

The previous section deals with prediction of class based on a single function. It applies when all new observations in the test set are independent (i.e., none share the same random effect). In other situations, multiple new observations may share the same random effect, for example, subjects from the same subgroup or settings with multiple functions per subject. This induces correlation among the test functions that one may want to take into account

when performing prediction. This can be done using the joint likelihood of the correlated functions when computing the posterior predictive odds ratio. Here we discuss two cases: one in which all individual units within the block share the same class label and one in which they do not.

(a) **The correlated observations all share the same class label.** This is this case, for example, in the setting in which we want to classify individuals based on replicate functions. We compute the joint likelihood of the multiple new observations, and the odds ratio can be computed as described above. In particular, assume there are L correlated observations from the test set, denoted as $\{Y^{0,l}, \mathbf{x}^{0,l}, \mathbf{z}^{0,l}\}, l = 1, \ldots, L$. Let c^0 be the common class label. The posterior odds can be computed by

$$\text{Odds}(j) = \frac{f(\{Y^{0,l}\}_l \mid c^0 \equiv j, \{\mathbf{x}^{0,l}, \mathbf{z}^{0,l}\}_l, \mathbf{Y}, \mathbf{V}, \mathbf{X}, \mathbf{Z})}{f(\{Y^{0,l}\}_l \mid c^0 \equiv 1, \{\mathbf{x}^{0,l}, \mathbf{z}^{0,l}\}_l, \mathbf{Y}, \mathbf{V}, \mathbf{X}, \mathbf{Z})} \cdot \frac{f(c^0 \equiv j)}{f(c^0 \equiv 1)}.$$

When $\mathbf{z}^{0,l}$ are available in the test data, the joint likelihood of $\{Y^{0,l}\}_l$ is conditionally independent. When $\mathbf{z}^{0,l}$ are not available, we need to compute the joint likelihood of the correlated observations by integrating out the random effects.

(b) **The correlated observations do not necessarily share the same class label.** In this case, the joint posterior predictive density will be conditional on a vector $\mathbf{c^0} = (c^{0,l}, \ldots, c^{0,L})^T$ of class designation. There will be q^L possible choices of $\mathbf{c^0}$, and therefore $q^L - 1$ posterior odds to compute.

2.3 *Incorporating Direct Covariates in FMM-based Classification*

In some applied settings, one may wish to account for covariates that directly affect the class designation but not necessarily the functional predictor itself. For example, in clinical applications one may wish to condition on known scalar clinical factors in addition to a functional response from a genomic or proteomic assay. Here we describe how to account for these when classifying using the FMM framework.

Suppose $\tilde{\mathbf{x}}_i$ are factors that directly affect class designation c_i through parameter vectors $\boldsymbol{\eta}$,

and that $\boldsymbol{\Theta}$ are the parameters in the FMM. Assuming that $f(\mathbf{Y}(t), c_i|\boldsymbol{\Theta}) \propto f(\mathbf{Y}(t)|c_i, \mathbf{X}, \mathbf{Z}, \boldsymbol{\Theta})$ $f(c_i|\tilde{\mathbf{x}}_i, \boldsymbol{\eta})$, we can handle the direct covariates $\tilde{\mathbf{x}}_i$ by first fitting a model for $f(c_i|\tilde{\mathbf{x}}_i, \boldsymbol{\eta})$ (e.g., using a generalized linear model) and then substituting the $f(c^0 = j|\tilde{\mathbf{x}}^0) = \Pr(c^0 = j \mid \tilde{\mathbf{x}}^0, \hat{\boldsymbol{\eta}})$ for $f(c^0 = j)$ in equation (2). If the model $\mathcal{M}$ is fitted using a Bayesian approach, its posterior predictive distribution can be easily computed by

$$f(c^0 = j \mid \tilde{\mathbf{x}}^0, \{c_i, \tilde{\mathbf{x}}_i\}_i) = \int f(c^0 = j \mid \tilde{\mathbf{x}}^0, \boldsymbol{\eta}) f(\boldsymbol{\eta}|\{c_i, \tilde{\mathbf{x}}_i\}_i) d\boldsymbol{\eta}. \tag{4}$$

Then the combined posterior predictive probability will become

$$\begin{aligned} & f(c^0 = j \mid Y^0, \mathbf{x}^0, \tilde{\mathbf{x}}^0, \mathbf{z}^0, \mathbf{Y}, \mathbf{X}, \mathbf{Z}, \{c_i, \tilde{\mathbf{x}}_i\}_i) \\ \propto \quad & f(Y^0 \mid c^0 = j, \mathbf{x}^0, \mathbf{z}^0, \mathbf{Y}(t), \mathbf{V}, \mathbf{X}, \mathbf{Z}) \cdot f(c^0 = j \mid \tilde{\mathbf{x}}^0, \{c_i, \tilde{\mathbf{x}}_i\}_i). \end{aligned}$$

Note that, using this approach, we can combine information across a series of different functional or scalar predictors to perform classification as long as we can write a series of conditionally independent models for each.

3. Specific Implementation Details for G-WFMM/R-WFMM

Note that model (1) is not completely specified since no distributional assumptions are made for $\boldsymbol{U}(t)$ and $\boldsymbol{E}(t)$, no structure has been assumed on the functional quantities, and no assumption has been made on the covariances $\Sigma_U(t, s)$ and $\Sigma_E(t, s)$, which in high-dimensional settings have too many parameters to estimate in an unstructured fashion. Different methods for specifying these details and fitting the FMM have been considered. Guo (2002) made Gaussian assumptions, represented the functions through smoothing splines and made simple, stationary covariance assumptions, and fit the model using Kalman filters. Morris and Carroll (2006) made Gaussian assumptions, represented the functions using wavelets, and used heteroscedastic diagonal covariances in the wavelet space to accommodate nonstationary covariance features, and fit the model using a fully Bayesian approach with shrinkage priors to induce adaptive smoothing of the fixed effect functions. Zhu, Brown and Morris (2011) also used Bayesian modeling and wavelet-space representations, but assumed

double exponential distributions in the wavelet space, which corresponds to mixtures of double exponentials in the data space, leading to robust estimates of fixed and random effect functions that naturally down weight outliers. Aston, Chiou and Evans (2010) made Gaussian assumptions, represented the data using a principal component (PC) decomposition of the marginal covariance function, used heteroscedastic diagonal covariances in the PC score space, and fit the model using restricted maximum likelihood. While the classification method described in this paper can be used with any FMM-based method, in this section we provide implementation details based on G-WFMM (Morris and Carroll 2006) and R-WFMM (Zhu, Brown, and Morris 2011).

3.1 *Classification using Gaussian Wavelet-based FMM (G-WFMM)*

Based on model (1), let the components of $\boldsymbol{Y}(t)$ take values on a common interval $\mathcal{T}$. The G-WFMM fits the model based on Gaussian assumptions for the random effects and errors. In particular, $\boldsymbol{U}(t)$ is assumed to be a mean zero multivariate Gaussian process with an $m \times m$ between-function covariance matrix $\mathbf{P}$ and a within-function covariance surface $Q(t_1, t_2) \in \mathcal{T} \times \mathcal{T}$, denoted as $\boldsymbol{U}(t) \sim \mathcal{N}(\mathbf{P}, Q)$. This implies that $\text{Cov}\{U_l(t_1), U_k(t_2)\} = \mathbf{P}_{lk} Q(t_1, t_2)$. The residual error is assumed to be $\boldsymbol{E}(t) \sim \mathcal{N}(\mathbf{R}, S)$ independent of $\boldsymbol{U}(t)$. A useful special case of this model is to let $\mathbf{P} = \mathbf{R} = \mathbf{I}$ (i.e., the components of $\boldsymbol{U}(t)$, respectively $\boldsymbol{E}(t)$, are independent). Note that, if desired, the covariance parameters $\mathbf{P}$, $\mathbf{R}$, Q, and S can all be allowed to vary by class c, for which we will discuss details in Section 3.3.

If the functional responses $Y_i(t)$ are all measured on the same equally-spaced fine grid of length T, a discretized version of model (1) on the grid can be represented in matrix form:

$$\mathbf{Y} = \mathbf{VG} + \mathbf{XB} + \mathbf{ZU} + \mathbf{E}, \tag{5}$$

with $\mathbf{Y}, \mathbf{G}, \mathbf{B}, \mathbf{U}$, and $\mathbf{E}$ each having T columns, and each column corresponding to one position on the grid. The random effects and error distributions become mean-zero normal random matrices: $\mathbf{U} \sim \mathcal{N}(\mathbf{P}, \mathbf{Q})$, $\mathbf{E} \sim \mathcal{N}(\mathbf{R}, \mathbf{S})$, with $\mathbf{Q}$ and $\mathbf{S}$ matrices of size $T \times T$.

The discrete wavelet transform (DWT) can then be applied to the rows of $\mathbf{Y}, \mathbf{G}, \mathbf{B}, \mathbf{U}$, and $\mathbf{E}$. Whereas in practice this is generally done using a fast recursive algorithm, for didactic purposes the DWT can be represented as a linear transformation by matrix $\mathbf{W}$, which for most choices of wavelets is either orthogonal or nearly orthogonal (i.e., $\mathbf{D} = \mathbf{YW}^T$, $\mathbf{G}^* = \mathbf{GW}^T$, $\mathbf{B}^* = \mathbf{BW}^T$, $\mathbf{U}^* = \mathbf{UW}^T$ and $\mathbf{E}^* = \mathbf{EW}^T$). This induces a wavelet-space functional mixed model :

$$\mathbf{D} = \mathbf{VG}^* + \mathbf{XB}^* + \mathbf{ZU}^* + \mathbf{E}^*, \tag{6}$$

where rows of $\mathbf{D}, \mathbf{G}^*, \mathbf{B}^*, \mathbf{U}^*$, and $\mathbf{E}^*$ correspond to the DWT of the rows of $\mathbf{Y}, \mathbf{G}, \mathbf{B}, \mathbf{U}$, and $\mathbf{E}$, respectively, and the columns correspond to the individual wavelet coefficients, double-indexed by their resolution levels $j = 1, \ldots, J$ and locations $k = 1, \ldots, K_j$. The induced distributional assumptions are $\mathbf{U}^* \sim \mathcal{N}(\mathbf{P}, \mathbf{Q}^*)$ and $\mathbf{E}^* \sim \mathcal{N}(\mathbf{R}, \mathbf{S}^*)$, with $\mathbf{Q}^* = \mathbf{WQW}^T$ and $\mathbf{S}^* = \mathbf{WSW}^T$. The whitening property of the wavelet transform (e.g., Vidakovic (1999), pages 10-13) tends to induce decorrelation of the model coefficients in the wavelet domain, so that one might make reasonable independence assumptions for the covariance matrices of $\mathbf{U}^*$ and $\mathbf{E}^*$ across their columns (i.e., $\mathbf{Q}^* = \text{diag}(\{q^*_{jk}\})$, $\mathbf{S}^* = \text{diag}(\{s^*_{jk}\})$). By indexing these wavelet-space variance components by both j and k, this model is parsimonious yet flexible enough to accommodate important types of nonstationarities in $\mathbf{Q}$ and $\mathbf{S}$.

To induce adaptive regularization of $G_c(t), c = 1, \ldots, q$ and $B_a(t), a = 1, \ldots, p$, spike-slab priors are assumed for the fixed effects in the wavelet space G^*_{cjk}, the c^{th} component in the $(j,k)^{th}$ column of $\mathbf{G}^*$, and B^*_{ajk}, the a^{th} component in the $(j,k)^{th}$ column of $\mathbf{B}^*$. That is, $G^*_{cjk} = \gamma^G_{cjk} N(0, \tau^G_{cj}) + (1 - \gamma^G_{cjk}) I_0$, and $\gamma^G_{cjk} \sim \text{Bernoulli}(\pi^G_{cj})$. Similarly, $B^*_{ajk} = \gamma^B_{ajk} N(0, \tau^B_{aj}) + (1 - \gamma^B_{ajk}) I_0$, and $\gamma^B_{ajk} \sim \text{Bernoulli}(\pi^B_{aj})$. Here $\pi^G_{cj}, \pi^B_{aj}, \tau^G_{cj}$, and τ^B_{aj} are regularization parameters that can be estimated using conditional maximum likelihood in an empirical Bayes approach, or given hyperpriors themselves (e.g., set $\pi^G_{cj} \sim \text{Beta}(a^G, b^G)$, $\pi^B_{aj} \sim \text{Beta}(a^B, b^B)$, $\tau^G_{cj} \sim \text{IG}(\alpha^G, \beta^G)$, and $\tau^B_{aj} \sim \text{IG}(\alpha^B, \beta^B)$). A Markov chain Monte Carlo (MCMC) scheme is used

to obtain posterior samples for the quantities in model (6), which are then projected back to the data space using the inverse discrete wavelet transform (IDWT) to yield posterior samples in model (1).

The following steps build the G-WFMM using training data and then use it to perform classification for test data.

Step 1. Applying G-WFMM to the training data, obtain M posterior samples for the model parameters, denote $\mathbf{\Theta} = \left\{\mathbf{G}^*, \mathbf{B}^*, \mathbf{U}^*, \{q_{jk}\}, \{s_{jk}\}, \{\gamma^G_{cjk}\}, \{\gamma^B_{ajk}\}, \{\tau^G_{cj}\}, \{\tau^B_{aj}\}, \{\pi^G_{cj}\}, \{\pi^B_{aj}\}\right\}$.

Step 2. Prediction on the test set. Assume that an observation from the test set has response Y^0 (with wavelet coefficients $\mathbf{d}^0$) and covariates $\mathbf{x}^0$, $\mathbf{z}^0$, and denote the unknown class designation vector as $\mathbf{v}^0$. Furthermore, denote c^0 as the class label for this observation. Then $\mathbf{v}^0 = \mathbf{e}_j$if and only if $c^0 = j$, where $\mathbf{e}_j$ is the unit vector of length q with the j^{th} component equal to 1 and 0 elsewhere. The posterior predictive density in equation (2) can be computed by

$$\begin{aligned} & f(Y^0 \mid c^0 = j, \mathbf{x}^0, \mathbf{z}^0, \mathbf{Y}, \mathbf{V}, \mathbf{X}, \mathbf{Z}) \\ = \;& f(\mathbf{d}^0 \mid c^0 = j, \mathbf{x}^0, \mathbf{z}^0, \mathbf{D}, \mathbf{V}, \mathbf{X}, \mathbf{Z}) \\ = \;& \int f(\mathbf{d}^0 \mid c^0 = j, \mathbf{x}^0, \mathbf{z}^0, \mathbf{\Theta}) f(\mathbf{\Theta}|\mathbf{D}, \mathbf{V}, \mathbf{X}, \mathbf{Z}) d\mathbf{\Theta} \\ = \;& \int \left[\prod_{j,k} f(d^0_{jk} \mid c^0 = j, \mathbf{x}^0, \mathbf{z}^0, \mathbf{\Theta}_{jk})\right] f(\mathbf{\Theta}|\mathbf{D}, \mathbf{V}, \mathbf{X}, \mathbf{Z}) d\mathbf{\Theta}, \end{aligned} \tag{7}$$

where $\mathbf{\Theta}_{jk}$ are the parameters in $\mathbf{\Theta}$ indexed by (j,k), and d^0_{jk} is the $(j,k)^{th}$ component of $\mathbf{d}^0$. The integration in equation (7) can be numerically approximated by averaging over the joint posterior samples from the MCMC,

$$\begin{aligned} & \int \left[\prod_{j,k} f(d^0_{jk} \mid c^0 = j, \mathbf{x}^0, \mathbf{z}^0, \mathbf{\Theta}_{jk})\right] f(\mathbf{\Theta}|\mathbf{D}, \mathbf{V}, \mathbf{X}, \mathbf{Z}) d\mathbf{\Theta} \\ \approx \;& \frac{1}{M} \sum_{t=1}^{M} \prod_{j,k} f(d^0_{jk} \mid c^0 = j, \mathbf{x}^0, \mathbf{z}^0, \mathbf{\Theta}^{(t)}_{jk}), \end{aligned} \tag{8}$$

where $\mathbf{\Theta}^{(t)}_{jk}$ denote the t^{th} posterior samples of $\mathbf{\Theta}_{jk}$ from Step 1. Typically, we would

compute these likelihoods by combining information across all wavelet coefficients (j,k). The sparsity property of the wavelet coefficients suggests most signals can be efficiently represented by a relatively small proportion of the total wavelet coefficients. Thus, if desired, one could perform classification using only the subset of most important wavelet coefficients. For example, Morris et al. (2011) describe a method to obtain the smallest subset of wavelet coefficients that simultaneously preserves at least $100(1-\alpha)\%$ of the total variation in each of the observed functions. By doing this, one can speed up calculations by a factor of 20 or more while retaining almost all information in the original functions and simultaneously performing an additional layer of denoising that could potentially improve classification performance.

When $\mathbf{z}^0$ is available, we use $\mathbf{U}^*$ when computing the densities, in particular,

$$f(d^0_{jk} \mid c^0 = j, \mathbf{x}^0, \mathbf{z}^0, \mathbf{\Theta}_{jk}) = \phi(d^0_{jk} |\ \mathbf{e}_j^T\mathbf{G}^*_{jk} + (\mathbf{x}^0)^T\mathbf{B}^*_{jk} + (\mathbf{z}^0)^T\mathbf{U}^*_{jk}, s^*_{jk}),$$

where $\phi(\cdot|\mu,\sigma^2)$ represents the density for normal distribution with mean μ and variance σ^2, and $\mathbf{G}^*_{jk}, \mathbf{B}^*_{jk}$, and $\mathbf{U}^*_{jk}$ are the $(j,k)^{th}$ column of $\mathbf{G}^*, \mathbf{B}^*$ and $\mathbf{U}^*$, respectively. Note that here $\mathbf{e}_j$ is a vector with the j^{th} component 1 and 0 elsewhere. When $\mathbf{z}^0$ is not available, we can compute the densities by integrating out the random effects:

$$f(d^0_{jk} \mid c^0 = j, \mathbf{x}^0, \mathbf{\Theta}_{jk}) = \phi(d^0_{jk} |\ \mathbf{e}_j^T\mathbf{G}^*_{jk} + (\mathbf{x}^0)^T\mathbf{B}^*_{jk}, q^*_{jk} + s^*_{jk}). \tag{9}$$

If classifying a block of correlated functions that all share the same class, equation (9) is replaced by the joint likelihood

$$f(\mathbf{d}^0_{jk} \mid c^0 = j, \mathbf{X}^0, \mathbf{\Theta}_{jk}) = \phi(\mathbf{d}^0_{jk} |\ \mathbf{E}_j^T\mathbf{G}^*_{jk} + \mathbf{X}^0\mathbf{B}^*_{jk}, q_{jk}\mathbf{J}_L + s_{jk}\mathbf{I}_L)$$

with $\mathbf{d}^0_{jk} = (d^{0,1}_{jk}, \ldots, d^{0,L}_{jk})^T$, $\mathbf{X}^0 = (\mathbf{x}^{0,1}, \ldots, \mathbf{x}^{0,L})^T$, $\mathbf{E}_j = (\mathbf{e}_j, \ldots, \mathbf{e}_j)^T$, $\mathbf{J}_L$ is an L by L matrix of ones and $\mathbf{I}_L$ is an L by L identity matrix.

3.2 *Classification using Robust Wavelet-based FMM (R-WFMM)*

While the G-WFMM is very flexible in many ways, one rigid aspect of the model is the Gaussian assumptions made on the residual errors, random effects, and slabs of the spike-

slab prior distributions to regularize the fixed effect functions. This leads to estimates of the fixed and random effect functions that are sensitive to outlying curves and regions of curves, which induce outlying wavelet coefficients that can exercise undue influence on the classification. Heavier-tailed distributions would better accommodate outliers in the data and would down weight the influence of outlying wavelet coefficients in the classification, potentially improving performance.

Zhu, Brown and Morris (2011) introduced R-WFMM, which models using heavier-tailed distributions in the wavelet space and thus yields robust estimation and inference of the fixed and random effect functions. Simulation studies revealed that the R-WFMM resulted in greatly improved fixed and random effect estimates in the presence of outlying curves and curve regions and did not give away much efficiency relative to the G-WFMM when no outliers were present. Further, this model led to more adaptive estimates of the fixed and random effect functions that attenuated spurious features of the functions while retaining true local features. We describe the model configuration here and provide details for how to use it to perform FMM-based classification.

Denote the $(j,k)^{th}$ column of model (6) as

$$\mathbf{d}_{jk} = \mathbf{V}\mathbf{g}^*_{jk} + \mathbf{X}\mathbf{b}^*_{jk} + \mathbf{Z}\mathbf{u}^*_{jk} + \mathbf{e}^*_{jk}, \tag{10}$$

where $\mathbf{d}_{jk} = \{D_{ijk}\}_{i=1}^n$, $\mathbf{g}^*_{jk} = \{G^*_{cjk}\}_{c=1}^q$, $\mathbf{b}^*_{jk} = \{B^*_{ajk}\}_{a=1}^p$, $\mathbf{u}^*_{jk} = \{U^*_{ljk}\}_{l=1}^m$, and $\mathbf{e}^*_{jk} = \{E^*_{ijk}\}_{i=1}^n$. We specify the following hierarchical model on these parameters:

$$\begin{aligned}
E^*_{ijk} &\sim N(0,\lambda_{ijk}),\ \ \lambda_{ijk} \sim g_1^E(\nu^E_{jk}),\ \ \nu^E_{jk} \sim g_2^E(\mathbf{\Theta}^E),\\
U^*_{ljk} &\sim N(0,\phi_{ljk}),\ \ \phi_{ljk} \sim g_1^U(\nu^U_{jk}),\ \ \nu^U_{jk} \sim g_2^U(\mathbf{\Theta}^U),\\
B^*_{ajk} &\sim \gamma^B_{ajk}N(0,\psi^B_{ajk}) + (1-\gamma^B_{ajk})\delta_0,\ \ \psi^B_{ajk} \sim g_1^B(\nu^B_{aj}),\ \nu^B_{aj} \sim g_2^B(\mathbf{\Theta}^B),\ \gamma^B_{ajk} \sim \text{Bernoulli}(\pi^B_{aj}),\\
G^*_{cjk} &\sim \gamma^G_{cjk}N(0,\psi^G_{cjk}) + (1-\gamma^G_{cjk})\delta_0,\ \ \psi^G_{cjk} \sim g_1^G(\nu^G_{cj}),\ \nu^G_{cj} \sim g_2^G(\mathbf{\Theta}^G),\ \gamma^G_{cjk} \sim \text{Bernoulli}(\pi^G_{cj}),
\end{aligned}$$

where δ_0 is a point mass at 0 and E^*_{ijk}, U^*_{ljk}, B^*_{ajk}, and G^*_{cjk} are mutually independent.

The *individual scale parameters* $\lambda_{ijk}, \phi_{ljk}, \psi^B_{ajk}, \psi^G_{cjk}$ are mutually independent with specified mixing distributions g_1^E, g_1^U, g_1^B, and g_1^G indexed by *population scale parameters* $\nu^U_{jk}, \nu^E_{jk}, \nu^B_{aj}$, and ν^G_{cj}, which are also mutually independent with prior distributions g_2^E, g_2^U, g_2^B, and g_2^G indexed by specified hyperparameter vectors $\mathbf{\Theta}^E, \mathbf{\Theta}^U$, $\mathbf{\Theta}^B$, and $\mathbf{\Theta}^G$, respectively. Note that the G-WFMM is a special case of this model, with a degenerate distribution for $g_1(\bullet)$, $\lambda_{ijk} \sim \delta_{s_{jk}}$, $\phi_{ljk} \sim \delta_{q_{jk}}$, and $\psi_{ajk} = \delta_{\tau_{aj}}$. The individual scale parameters λ_{ijk} serve as wavelet-space outlier weights. A relatively large λ_{ijk} (across i) suggests curve i is an outlier with respect to a feature of the curve corresponding to the wavelet basis function (j,k) and will result in a down weighting of observation D_{ijk} in estimating the corresponding fixed and random effects. Similarly, relatively large ϕ_{ljk} (across l) indicate random effect unit l is an outlier for feature (j,k) and will result in some downweighting of the D_{ijk} corresponding to random effect unit l, which are those with $Z_{il} \neq 0$.

While many different choices can be considered for $g_1(\bullet)$ and $g_2(\bullet)$, for our calculations we will assume $g_1(\nu_{jk}) = \text{Exp}(\nu^2_{jk}/2)$ for each model component E, U, B, and G, and choose $g_2(\bullet)$ to be such that $\{\nu^2_{jk}\}$ are Gamma distributions, with their parameters determined using the empirical Bayes approach. This leads to a model that behaves like a Laplace distribution across i and l, and through mixing over (j,k) in the second level behaves like a normal-exponential-gamma across j and k at the residual and random effect levels, which is a distribution that has been shown to have outstanding variable selection and shrinkage properties (NEG, Griffin and Brown, 2005). Based on the above model setup, an MCMC algorithm is proposed that yields posterior samples of all model parameters that can be used to perform Bayesian estimation and inference. The algorithm is efficient enough to handle large data sets as encountered in practice and is able to run automatically. Computational details can be found in the supplementary materials.

Classifications based on R-WFMM can be performed using a two-step procedure similar

to that described in Section 3.1. The MCMC algorithm on the training data leads to M posterior samples for the model parameters

$$\begin{aligned}\boldsymbol{\Theta} = \{\mathbf{G}^*, \mathbf{B}^*, \mathbf{U}^*, \{\gamma^G_{cjk}\}, \{\gamma^B_{ajk}\}, \{\lambda_{ijk}\}, \{\phi_{ljk}\}, \{\psi^B_{ajk}\}, \{\psi^G_{cjk}\}, \\ \{(\nu^E_{jk})^2\}, \{(\nu^U_{jk})^2\}, \{(\nu^G_{cj})^2\}, \{(\nu^B_{aj})^2\}, \{\pi^G_{cj}\}, \{\pi^B_{aj}\}\}.\end{aligned} \tag{11}$$

Equations (7) and (8) still describe the posterior predictive densities, except the densities are not Gaussian. When $U^0(t)$ is estimable in the training set, the densities are DE:

$$f(d^0_{jk} \mid c^0 = j, \mathbf{x}^0, \mathbf{z}^0, \boldsymbol{\Theta}_{jk}) = \mathrm{DE}(d^0_{jk} |\ \mathbf{e}_j^T\mathbf{G}^*_{jk} + (\mathbf{x}^0)^T\mathbf{B}^*_{jk} + (\mathbf{z}^0)^T\mathbf{U}^*_{jk}, 1/\nu^E_{jk}),$$

where $\mathrm{DE}(\cdot|\mu, b)$ represents the density for a DE distribution with mean μ and scale parameter b. When $\mathbf{z}^0$ is not available, we can compute the densities by integrating out the random effects, which gives

$$f(d^0_{jk} \mid c^0 = j,\ \mathbf{x}^0, \boldsymbol{\Theta}_{jk}) = \frac{\nu^E_{jk}\nu^U_{jk}}{2\left((\nu^U_{jk})^2 - (\nu^E_{jk})^2\right)}\left(\nu^U_{jk}\exp\{-\nu^E_{jk}|\widetilde{d}^0_{jk}|\} - \nu^E_{jk}\exp\{-\nu^U_{jk}|\widetilde{d}^0_{jk}|\}\right), \tag{12}$$

following the results of Proposition 1. Here $\widetilde{d}^0_{jk} = d^0_{jk} - \mathbf{e}_j^T\mathbf{G}^*_{jk} - (\mathbf{x}^0)^T\mathbf{B}^*_{jk}$.

Proposition 1. (Density of Sum of Two Independent DE Random Variables) *Assume that $X_1 \sim DE(0, b_1)$, $X_2 \sim DE(0, b_2)$, with densities $f(x_i) = 1/(2b_i)\exp\{-|x_i|/b_i\}, i = 1, 2$, and X_1 is independent of X_2. Let $Z = X_1 + X_2$. Then Z has zero mean, variance $2(b_1^2 + b_2^2)$, and characteristic function $\psi_z(t) = [(1 + b_1^2t^2)(1 + b_2^2t^2)]^{-1}$. The density of Z takes the form:*

$$f(z) = \begin{cases} \frac{1}{2(b_1^2 - b_2^2)}\left[b_1\exp\{-\frac{|z|}{b_1}\} - b_2\exp\{-\frac{|z|}{b_2}\}\right] & \text{when } b_1 \neq b_2, \\ \frac{1}{4b}\exp\{-\frac{|z|}{b}\}(1 + \frac{|z|}{b}) & \text{when } b_1 = b_2 = b. \end{cases}$$

The proof of proposition 1 can be done based on the results of Nadarajah and Kotz (2005) as well as the results of Griffin and Brown (2010). More details can be found in the supplementary materials. The formula of the density function can be verified using the moment generating function method.

If a block of correlated functions that are all sharing the same class is classified, when

the random effects are being integrated out, each $d_{jk}^{0,l}$ has a marginal density in the form of equation (12), with covariance matrix $2\mathbf{J}_L/(\nu_{jk}^U)^2) + 2\mathbf{I}_L/(\nu_{jk}^E)^2$ handling the within-block correlation. The analytical formula for the density of this multivariate distribution is not straightforward to obtain. Therefore, we recommend using numerical integration to approximate this density as follows:

$$\begin{aligned} & f(\mathbf{d}_{jk}^0 \mid c^0 = j, \mathbf{X}^0, \mathbf{\Theta}_{jk}) \\ = & \int \prod_{l=1}^{L} f(d_{jk}^{0,l} \mid c^0 = j, \mathbf{x}^{0,l}, u_{jk}^l, \nu_{jk}^E, G_{jk}^*, B_{jk}^*) f(u_{jk}^l | \nu_{jk}^U) du_{jk}^l \dots du_{jk}^L. \\ \approx & \; 1/N \sum_{s=1}^{N} \prod_{l=1}^{L} f(d_{jk}^{0,l} \mid c^0 = j, \mathbf{x}^{0,l}, u_{jk}^{l,s}, \nu_{jk}^\lambda, G_{jk}^*, B_{jk}^*), \end{aligned}$$

where $\{u_{jk}^{l,s}, s = 1, \dots, N\}$ are N samples generated from DE(0,$1/\nu_{jk}^\phi$) for $l = 1, \dots, L$. Since this approximation has to be done for each posterior sample $\mathbf{\Theta}_{jk}^{(t)}, t = 1, \dots, M$ to get an approximation for the final posterior predictive distribution, the computation can be intensive when both M and N are large. The computational burden can be ameliorated by making N, M small or taking subsamples of the M cases.

3.3 *Allowing the Covariance to Vary by Class*

In both the G-WFMM and R-WFMM discussed in Sections 3.1 and 3.2, we assumed that the distribution of $\mathbf{U}(t)$ and $\mathbf{E}(t)$ in model (1) were common for all classes. In some settings, one may wish the random effect and/or residual error covariances to vary across class, which would yield more flexible classification rules. This has been previously described for the G-WFMM (Morris and Carroll, 2006) and involves expanding the variance components $\{q_{j,k}^*\}, \{s_{j,k}^*\}$ to $\{q_{j,k}^{*,\,c}\}, \{s_{j,k}^{*,\,c}\}, c = 1, \dots, q$. For prediction, the posterior predictive probability needs to be adjusted so that the corresponding variance components of $c = j$ are used when the likelihood conditions on class label $c = j$. In the R-WFMM, we allow the population scale parameters $\{\nu_{jk}^E\}$ and $\{\nu_{jk}^U\}$ to be class specific, i.e., $\{\nu_{jk}^{E,c}\}, \{\nu_{jk}^{U,c}\}, c = 1, \dots, q$. Correspondingly, their Gamma hyper-prior parameters (α^E, β^E) and (α^U, β^U) would also be

indexed by c. This involves only minor changes of the previously described MCMC algorithm. Similarly, for prediction, we need to adjust the DE likelihood by plugging in the corresponding population scale parameters when conditioning on a particular $c = j$.

3.4 *Computing a Pointwise Discriminant Function*

While it is convenient to compute the posterior predictive probability for the test set in the wavelet domain directly, as we have described in Sections 3.1 and 3.2, at times one may wish to compute a pointwise (in the t domain) discriminant function $\zeta_j(t)$ that could be used as a descriptive summary of which regions of t were primary drivers of the classification of function j. Here, we define a *pointwise discriminate function* (denoted as $\zeta_j(t)$) as the pointwise log-odds of posterior predictive probability of belonging to class j versus 0,

$$\begin{aligned}
\zeta_j(t) &= \zeta_j(Y^0(t), \mathbf{x}^0, \mathbf{z}^0, \mathbf{Y}, \mathbf{V}, \mathbf{X}, \mathbf{Z}) \\
&= \log \frac{f(c^0 = j | Y^0(t), \mathbf{x}^0, \mathbf{z}^0, \mathbf{Y}, \mathbf{V}, \mathbf{X}, \mathbf{Z})}{f(c^0 = 0 | Y^0(t), \mathbf{x}^0, \mathbf{z}^0, \mathbf{Y}, \mathbf{V}, \mathbf{X}, \mathbf{Z})} \\
&= \log \frac{\int f(Y^0(t) | c^0 = j, \mathbf{x}^0, \mathbf{z}^0, \mathbf{\Theta}) f(\mathbf{\Theta} | \mathbf{Y}, \mathbf{V}, \mathbf{X}, \mathbf{Z}) d\mathbf{\Theta}}{\int f(Y^0(t) | c^0 = 0, \mathbf{x}^0, \mathbf{z}^0, \mathbf{\Theta}) f(\mathbf{\Theta} | \mathbf{Y}, \mathbf{V}, \mathbf{X}, \mathbf{Z}) d\mathbf{\Theta}} \\
&+ \log(f(c^0 = j)/f(c^0 = 0)). \qquad (13)
\end{aligned}$$

Note that although similar in notation, (13) is different from (2) in that for (13) we are computing the posterior predictive probability at time t while ignoring any correlation across t, while (2) is a general formula for computing the overall posterior predictive probabilities. To estimate $\zeta(\cdot)$, we need to first transform the the posterior samples of $\mathbf{\Theta}$ to the time domain by applying the IDWT to G^*, B^*, U^*, Q^*, and S^* for all of the M MCMC samples, after which the marginal likelihood of $Y^0(t)$ is computed for each t. It is because $\zeta_j(t)$ effectively ignores the correlations across t that it should be considered a descriptive summary measure and not itself used for classification.

In the G-WFMM, it is relatively easy to compute the predictive density in (13) because the IDWT of Gaussian distributions is again Gaussian. In the R-WFMM, it is not as straight-

forward. However, one can exploit the fact that the likelihood is Gaussian conditional on the individual scaling parameters λ_{jk}^0 to construct an approximate measure based on Monte Carlo numerical integration. Note that given $\mathbf{\Lambda}^0 = \text{Diag}\{\lambda_{jk}^0\}_{j,k}$, the residual covariance in the data space is given by $\Sigma_E^0 = W\mathbf{\Lambda}^0 W'$, where W is the DWT matrix. Thus, if we augment the numerator and denominator of (13) with λ_{jk}^0, we are left with

$$\int\int f\left(Y^0(t)|c^0 = j, \mathbf{x}^0, \mathbf{z}^0, \mathbf{\Theta}^-, \Sigma_E^0(t,t)\right) f\left(\Sigma_E^0(t,t)|\mathbf{\Theta}^-, \mathbf{Y}, \mathbf{V}, \mathbf{X}, \mathbf{Z}\right)\, d\,\Sigma_E^0(t,t) f\left(\mathbf{\Theta}^-|\mathbf{Y}, \mathbf{V}, \mathbf{X}, \mathbf{Z}\right) d\mathbf{\Theta}.$$

for the numerator, and an analogous expression for the denominator, where $\mathbf{\Theta}^-$ are the posterior samples for parameters excluding the $\{\lambda_{ijk}\}$. A Monte Carlo approximation of this density can be obtained by averaging over repeated sampling of the individual scaling parameters λ_{jk}^0 from their distribution $\text{DE}\{(\nu_{jk}^E)^2/2\}$ for each posterior sample of ν_{jk}^E.

4. Pancreatic Cancer Mass Spectrometry Application

We apply our FMM-based classification methods to predict cancer status using blood serum proteomics. Matrix assisted laser desorption and ionization, time-of-flight (MALDI-TOF) is a proteomic method that detects and measures the expression of hundreds of proteins. In a MALDI-TOF experiment, a biological sample of interest is first mixed with an energy-absorbing matrix substance, and the mixture is placed on a steel plate. The plate is then placed into a vacuum chamber, where a laser strikes the plate, desorbing ionized peptides from the sample. An electric field accelerates the particles into a potential free flight tube through which they travel at a constant velocity until striking a detector plate. The detector plate records the abundance of particles striking it over a series of short, fixed intervals of time indexed by $t = (t_1, \ldots, t_T)$, yielding a mass spectrum $Y(t)$. Using principles of basic physics, a quadratic transformation can be used to map the time axis t to a set of corresponding mass-to-charge ratios (m/z). Each spectrum is characterized by numerous peaks, which correspond to proteins or protein fragments present in the sample.

The data set we used for this paper was obtained from a pancreatic cancer experiment. In

this study, blood serum was taken from 139 patients with pancreatic cancer and 117 healthy controls. The blood serum was fractionated using 25% acetonitrile elutions optimized using myoglobin, then run on a MALDI-TOF instrument to obtain a proteomic spectrum for each sample. For this analysis, we consider the region of the spectra between $x = 4,000$ and $40,000$ Daltons, containing 6654 observations per spectrum. These 256 samples were run in four different time blocks over a period of several months. More specifics of the experiment can be found in Koomen et al. (2005). Our primary goal in this paper is to discriminate the cancer samples from the controls. It is well established that MALDI-TOF instruments are very sensitive, which leads to block effects that are manifest in systematic changes in both the intensities and locations of peaks (i.e., both the x and y axes), and are sometimes larger in magnitude than the biological effects of interest. Thus, it is important to adequately model these block effects to properly analyze the data.

The proposed model was applied to this data set for classification, and treats the time blocks as random effects associated with design matrix $\mathbf{Z}$ in model (1). Here $\mathbf{Z}$ has components $\mathbf{Z}_{i,j} = 1$ indicating the i^{th} spectra is measured in time block j, $j = 1, \ldots, 4$. For these data, we did not model any other fixed effects X on the functions. We used four-fold cross validation to assess the method, each time training the model using 3/4 of the data and testing it on the remaining 1/4. This was done in two different ways: (1) **in-block classification**, for which 3/4 of the samples from each of the 4 blocks were randomly selected for the training data, with the remaining 1/4 from each block serving as the test data, and (2) **out-of-block classification**, for which all samples from 3 out of the 4 time blocks were used for training, with validation done on all samples in the 4th time block. We considered classification based on all wavelet coefficients, and using compression, considered only the set of wavelet coefficients preserving at least 90% of the total energy for all of the functions, which in this case corresponded to a subset of 208 out of the 6655 total wavelet coefficients.

We sampled "virtual spectra" from the predictive distribution of the fitted model, and found that they looked just like real spectra (see plots in the supplementary materials), suggesting that the model is flexible enough to capture the salient features of the MALDI-TOF data.

We evaluated the classification results using several statistics: the area under the receiver operating characteristic (ROC) curve (AUC), the misclassification rate (MisR), the sensitivity (Sens) and the specificity (Spec). The ROC curve was generated by plotting Sens as a function of 1-Spec computed at all possible thresholds on the posterior probabilities. The AUC is the area under the ROC curve computed using numerical integration. The MisR, the Sens, and the Spec are those values computed when fixing the threshold at 0.5. We list these statistics for both in-block and out-of-block prediction in Table 1.

We compared the performance of our methods with three types of classification methods: the functional principal component (FPC) based methods introduced in Hall, Poskitt and Presnell (2001), the generalized functional linear models (GFLM) of Müller (2005) and Müller and Stadtmüller (2005), and peak-based methods (Peak) specifically adapted for mass spectrometry data (see, e.g. Koomen et al. (2005)). The first two types are functional data classification methods. The third type is not functional data based, but applies standard multivariate methods to peak intensities after performing peak detection. In the FPC methods, the dimension of the functions was reduced using truncated Karhunen-Loéve basis expansions. Classification was then performed using the resulting coefficients through either kernel density estimations (KDE) or quadratic discriminant analysis (QDA), which we denote as FPC-KDE and FPC-QDA, respectively. The GFLM method is based on a regression model with univariate response and functional predictors, where the functional predictors are approximated by the truncated Karhunen-Loéve expansion. The peak based methods were performed by first detecting the peaks of the mass spectra using the R package *msProcess* (Morris et al. 2005), and then using the detected peaks for classification with methods such as

penalized generalized linear model (GLM) with lasso (denoted as GLM-Lasso) and K-nearest neighbor (KNN). All model parameters, such as the penalty parameter in GLM-Lasso and the K parameter in KNN were determined using nested 4-fold cross validation based on the training set. Note that some classification methods intended for full rank multivariate data could not be directly applied to these data given the extremely high dimensionality of the functions ($T = 6654$) relative to the sample size ($N = 192$ for each 3/4 training split).

[Table 1 about here.]

Table 1 shows that the FMM-based classification methods compared favorably to the other methods considered, with higher AUCs and lower misclassification rates for both the in-block and out-of-block prediction. As expected, the R-WFMM outperformed the G-WFMM, and slightly improved results were obtained when applying wavelet compression. Among the comparison methods, the peak-based methods using GLM-Lasso outperformed others, and was competitive with the G-WFMM and not far behind the R-WFMM. The estimated ROC curves for in-block prediction are shown in Figure 1. The ROC plot for out-of-block prediction can be found in the supplementary materials.

[Figure 1 about here.]

As described in Section 3.4, pointwise discriminant functions provide summary measures that might help understand regions of the function that are influential to the classification of individual subjects. We estimated the discriminant function $\zeta(t)$ for each observation in the G-WFMM model. Figure 2 shows $\zeta(t)$ for four selected observations, two from the cancer class (top), two from control class (bottom). The regions above zero are those that suggested the curve be classified as class 1(cancer), and regions below zero are those that drove the curve to class 0 (control). Figure 2 shows that the regions of the spectra, and thus the proteins, driving the classification varied across subjects, as expected.

[Figure 2 about here.]

5. Discussion

We have proposed an FMM-based method for functional data classification. FMM-based modeling captures various types of important structures that might be present in the data, including the adjustment for covariates affecting the functions and the classes, the denoising of the function, and the modeling of design-induced between-function correlation that equips it for use to classify subjects based on multiple correlated functional predictors. Coupled with our Bayesian wavelet-space modeling approach, the method can handle spiky functions and quantitative image data, and scales up to large data sets using automated code, with the potential to be even more flexible by allowing different covariances between classes. When used with the R-WFMM, this method yields robust classification that can downweight outlying curves and regions of curves in building the discriminator. Another unique benefit of this FMM-based approach is that the same model fit can be used for both classification and unified inference on the fixed and random effect functions.

As opposed to functional regression models that seek to model $f\{c|Y(t)\}$, our method effectively uses a functional discriminant analysis involving modeling $f\{Y(t)|c\}$ and then using Bayes rule to invert the problem and compute $\Pr\{c^0 = j|Y(t)\}$ using posterior predictive probabilities. This approach is promising for classification in general, not just for functional predictors, and may be an underappreciated and under-recognized strategy for classification using Bayesian modeling. The inherent hierarchical modeling approach provides a natural way to extend this method to combine information across multiple functional, image, and scalar predictors in performing classification.

One potential downside of this approach is that the classification may be strongly dependent on the parametric assumptions made in the modeling of $f\{Y(t)|c\}$. Thus, it is important to ensure that the models used are flexible enough to capture the key functional

and distributional features of the data. The WFMM, especially paired with robust modeling, appears to be flexible enough for the data analyzed in this paper, as the data simulated from the model look just like real data.

Acknowledgements

Jeffrey S. Morris was supported by NCI grant CA-107304, and this work was done as part of the Analysis of Object Data program at the Statistical and Applied Mathematics Scientific Institute (SAMSI).

References

Aston, J. A. D., Chiou, J.-M., and Evans, J. P. (2010). Linguistic pitch analysis using functional principal component mixed effect models. *Journal of The Royal Statistical Society Series C* **59(2),** 297–317.

Bigelow, J. and Dunson, D. B. (2009). Bayesian semiparametric joint modeling of functional predictors. *Journal of the American Statistical Association*, **104**, 26–36.

Efron, B. (1975). The efficiency of Logistic regression compared to normal discriminant analysis. *Journal of the American Statistical Association* **70,** 892–898.

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

Griffin, J. E. and Brown, P. J. (2005). Alternative prior distributions for variable selection with very many more variables than observations, *CRiSM Working Paper No. 05-10*. University of Warwick.

Griffin, J. E. and Brown, P. J. (2010). Inference with normal-gamma prior distributions in regression problems. *Bayesian Analysis* **5,** 171–188.

Guo, W. (2002). Functional mixed effects models. *Biometrics* **58,** 121–128.

Hall, P., Poskitt, D. S., and Presnell, B. (2001). A functional data-analytic approach to signal discrimination. *Technometrics* **43,** 1–9.

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. (2002). Generalized linear models with functional predictors. *Journal of the Royal Statistical Society, Series B* **64,** 411–432.

Leng, X and Müller, H. (2005). Classification using functional data analysis for temporal gene expression data. *Bioinformatics* **22,** 68–76.

Li, B. and Yu, Q. (2008). Classification of functional data: A segmentation approach. *Computational Statistics & Data Analysis* **52,** 4790–4800.

Koomen, J. M., Shih, L. N., Coombes, K. R., Li, D., Xiao, L., Fidler, I. J., Abbruzzese, J. L., and Kobayashi, R. (2005). Plasma protein profiling for diagnosis of pancreatic cancer reveals the presence of host response proteins. *Clinical Cancer Research* **11,** 1110–1118.

Morris, J. S., Coombes, K. R., Kooman, J., Baggerly, K. A., and Kobayashi, R. (2005). Feature extraction and quantification for mass spectrometry data in biomedical applications using the mean spectrum. *Bioinformatics*, **21(9)**, 1764-1775.

Morris, J. S. and Carroll, R. J. (2006). Wavelet-based functional mixed models. *J. R. Statist. Soc. B* **68,** 179–199.

Morris, J. S., Baladandayuthapani, V., Herrick, R. C., Sanna, P., and Gutstein, H. (2011). Automated analysis of quantitative image data using isomorphic functional mixed models, with application to proteomics data (to appear). *Annals of Applied Statistics*, to appear.

Müller, H. and Stadtmüller, U. (2005). Generalized functional linear models. *The Annals of Statistics* **33,** 774–805.

Müller, H. (2005). Functional modeling and classification of longitudinal data. *Scandinavian Journal of Statistics* **32,** 223–240.

Nadarajah, S. and Kotz, S. (2005). On the linear combination of Laplace random variables. *Probability in the Engineering and Informational Sciences* **19,** 463–470.

Park, T. and Casella, G. (2008). The Bayesian Lasso. *Journal of the American Statistical Association* **103,** 681–686.

Ramsay, J. O. (2000). Functional components of variation in handwriting. *Journal of the American Statistical Association* **95,** 9–15.

Sorace, J. M. and Zhan, M. (2003). A data review and re-assessment of ovarian cancer serum proteomic profiling. *BMC Bioinformatics* **9,** 4–24.

Vidakovic, B. (1999). *Statistical Modeling by Wavelets*. New York: John Wiley & Sons, Inc.

Zhu, H., Vannucci, M., and Cox, D. D. (2010). A Bayesian hierarchical model for classification with selection of functional predictors. *Biometrics* **66,** 463–473.

Zhu, H., Brown, P. J., and Morris, J. S. (2011). Adaptive, robust functional regression in functional mixed model framework. *Journal of the American Statistical Association*, to appear.

Supplementary Materials

The Web Appendices referenced in Section 3 and 4 are available under the Paper Information link at the *Biometrics* website `http://www.biometrics.tibs.org`.

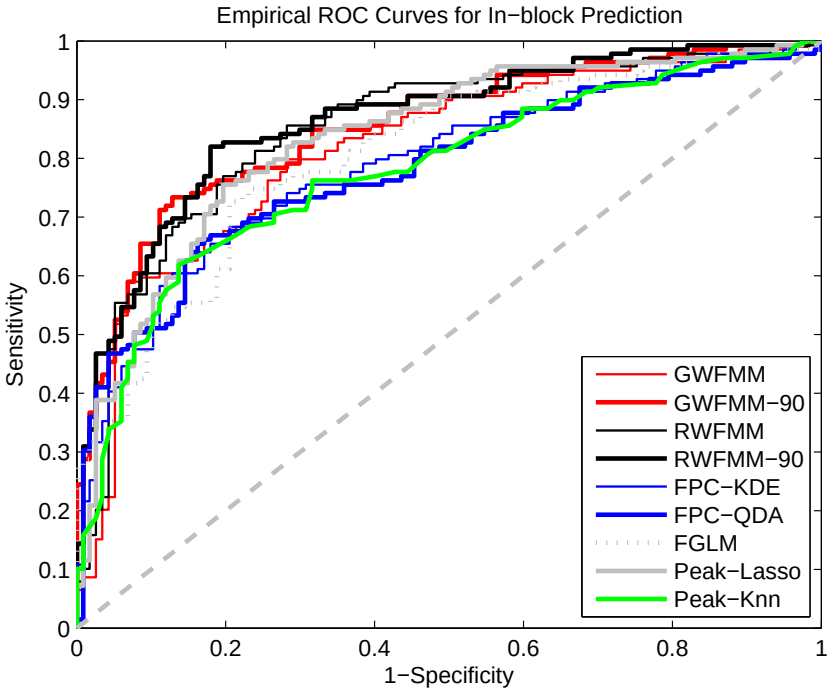

Figure 1. The empirical ROC curves for in-block prediction compared with other methods.

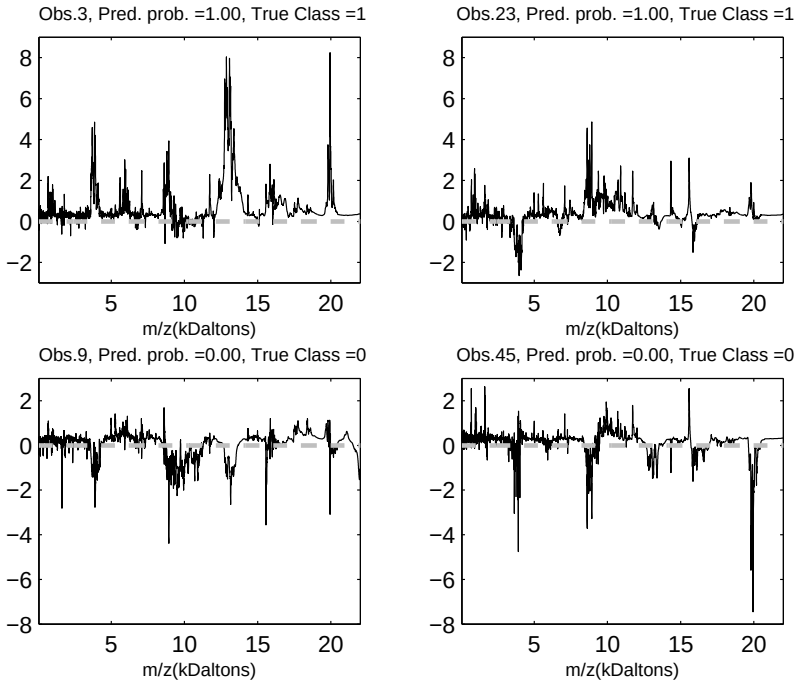

Figure 2. Plot of the estimated pointwise discriminant function $\zeta(t)$ of cancer (Class 1) vs. control (Class 0) for four selected observations, two from the cancer class (top), two from the control class (bottom).

Table 1
Comparison of Several Classification Approaches

	Methods	Model Name	AUC	MisR	Sens	Spec
In Block	FMM	GWFMM	0.816	0.270	0.669	0.812
		$GWFMM_{90}$	0.854	0.211	0.719	0.880
		RWFMM	0.850	0.231	0.705	0.846
		$RWFMM_{90}$	0.865	0.215	0.727	0.855
	FPC	FPC-KDE	0.790	0.379	0.331	0.966
		FPC-QDA	0.783	0.270	0.626	0.846
	FGLM	Logit Link	0.805	0.250	0.748	0.761
	Peak	GLM-Lasso	0.834	0.223	0.755	0.803
		KNN	0.774	0.273	0.633	0.838
Out Block	FMM	GWFMM	0.802	0.273	0.612	0.863
		$GWFMM_{90}$	0.815	0.254	0.655	0.855
		RWFMM	0.838	0.266	0.619	0.872
		$RWFMM_{90}$	0.830	0.242	0.705	0.829
	FPC	FPC-KDE	0.765	0.379	0.331	0.966
		FPC-QDA	0.804	0.285	0.619	0.821
	GFLM	Logit Link	0.766	0.281	0.741	0.692
	Peak	GLM-Lasso	0.813	0.273	0.719	0.735
		KNN	0.729	0.332	0.590	0.761