Functional Data Classification in Cervical Pre-cancer Diagnosis — A Bayesian Variable Selection Model

Hongxiao Zhu, Marina Vannucci, Dennis D. Cox
Department of Statistics, Rice University

Abstract

Fluorescence Spectroscopy provides a non-invasive tool for real time diagnosis of cervical pre-cancer. An important issue involved is to classify diseased tissue from normal using multiple functional data observations–the fluorescence excitation-emission matrices (EEMs). This paper proposes a Bayesian variable selection model to perform binary classification based on multiple functional covariates. The model contains two major steps. In the first step, functional principal component analysis or orthonormal basis expansion are used to approximate functional curves and reduce the high-dimensional functional covariates to a finite number of scores. In the second step, a Bayesian probit regression model is constructed to select the scores obtained from the first step and at the same time perform classification. The variable selection is performed through a mixture normal prior of the regression coefficients. And a latent variable is introduced to simplify computation. MCMC methods–a Gibbs sampler and an alternative Hybrid Gibbs/Metropolis-Hasting sampler, are used to obtain posterior samples of the parameters. Simulations show that this model can not only produce accurate variable selection and classification results, but also provide good estimate of the coefficient functions associated with the functional covariates. Application to spectroscopic data gives improved classification performance as compared with several other classification methods.

KEY WORDS: Fluorescence spectroscopy; Cervical cancer; Bayesian variable selection; Functional principal component analysis

1. Introduction

Cervical cancer is one of the leading causes of cancer deaths in women. The prevention of cervical cancer can be significantly improved by diagnosis at early stage of the disease using automatic, low cost screening devices. Among existing diagnosis tools, fluorescence spectroscopy has been shown promising as a non-invasive, real-time optical technology to quantitatively detect cervical pre-cancer. An important goal of fluorescence spectroscopy diagnosis is to classify the diseased observations from normal ones based on the fluorescence spectra measurement. However, since the underlying biochemical mechanisms associated with the fluorescence spectra differences between normal and dysplastic tissue are not fully understood, numerical algorithms need to be designed to find differentiating information from the spectra and perform diagnosis automatically.

One difficulty of designing efficient algorithms is that the spectra data are of functional form. They are smooth curves with high resolution. Most literatures to date on fluorescence spectroscopy diagnosis apply classification algorithms on "features" obtained from the spectra through dimension reduction methods. Commonly used dimension reduction methods are principal component analysis (Kamath et al. 2007, Palmer et al. 2003), or artificially selected intensity and shape information from the spectra (See Ramanujam et. al 1994). Classification algorithms such as K-nearest neighbor, neural network, support vector machine have been applied to the "features". All the above referred algorithms treat the fluorescence spectral curves and the reduced "features" equally when training the algorithm. This assumption is problematic since some information contained in the spectra is more disease-related hence plays more important role in classification (See Chang 2002 and Welch 1997). In this study, we look at the problem from a functional data point of view. A Bayesian variable selection model is proposed to perform binary classification based on multiple functional covariates–the fluorescence spectra obtained at each measurement. The proposed method contains two major steps. In the first step, functional principal component analysis or orthonormal basis expansion are used to approximate functional curves and reduce the high-dimensional functional covariates to a finite number of scores. In the second step, a Bayesian probit regression model is constructed to select the scores obtained from the first step and at the same time perform classification.

The data studied in this paper are drawn from a clinical study of using multiple fluorescence spectra to diagnose cervical abnormalities. To avoid possible confounding effects due to variabilities of device and tissue type, the data of consideration is obtained from a fixed instrument (called Fast EEM2) and all normal observations are from a fixed tissue type–squamous (ecto-cervix) tissue. Each observation consists of several spectral curves measured in the following way: an excitation light at certain fixed excitation wavelength is produced to illuminate the cervix tissue. The excitation light is absorbed by various endogenous fluorescent molecules in tissue, resulting in emission of fluorescent light. The emitted fluorescent light is measured by an optical detector and the spectrum is obtained as one smooth curve. The excitation light is varied at several different wavelengths and gives multiple spectral curves for each measurement. The left panel of figure 1 shows the plot of one measurement. It contains 16 spectral curves measured at excitation wavelengths ranging from 330 nm to 480 nm with increments of 10 nm. Each spectral curve contains fluorescence intensities recorded on a range of emission wavelengths between 385nm and 700nm. If we use a color

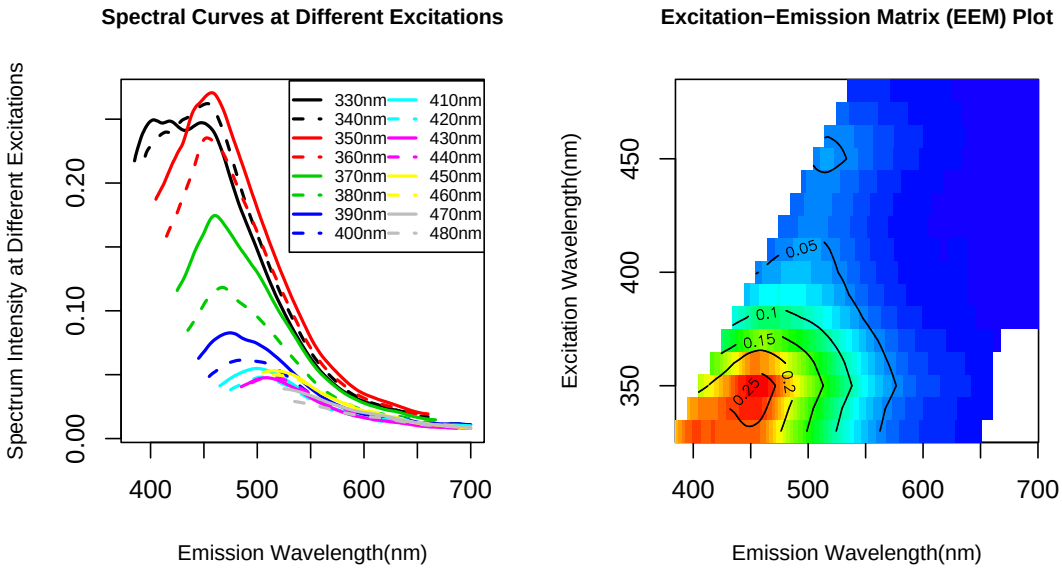

Figure 1: The left panel plots the fluorescence spectral curves at different excitation wavelength. The right panel is an image plot of the excitation-emission matrix(EEM).

plot to represent the intensities, we can stack all the 16 spectra and obtain an image as shown in the right panel of figure 1. We call such fluorescence spectroscopy data measurements excitation-emission matrices (EEMs).

Note that each EEM measurement contains 16 curves and each curve is of dimension around 220, which gives 3563 points in total (all curves are truncated at the edges). Point-wise variable selection used in Brown et al. (1998, 2002) are not practical due to the computation burden brought by multiple functional covariates. And another drawback of point-wise variable selection is that it will ignore the high correlation of contiguous points caused by the functional property of the covariates. In this paper, we apply the standard functional data analytical tools–functional principal component analysis or function approximation using orthonormal basis (Ramsay & Silverman 1997), to project functional data to the eigenspace or the space spanned by the orthonormal basis. A Bayesian variable selection model is constructed on the domain of projected scores for classification.

The structure of this paper is as follows: section 2 introduces the Bayesian probit model with variable selection for classification problems involving multiple functional covariates. Two simulation studies are conducted in section 3 to verify the effectiveness of the proposed model. Real data application results to the fluorescence spectroscopy data are presented in section 4. Further analysis and discussions of the model are shown in section 5.

2. A Bayesian probit model with variable selection for functional data classification

Suppose we observe n i.i.d. observations, each observation contains J functions. For $i = 1, \ldots, n$ and $j = 1, \ldots, J$, denote $x_{ij}(t)$ as the jth function observed from the ith observation, and without loss of generality, assume all functions have zero mean, i.e. $E[x_{ij}(t)] = \mu_j(t) \equiv 0$, $\forall i, \forall j$. Consider i.i.d. binary responses y_i which indicates the binary class that each observation belongs to. Similarly to James (2002) and Müller & Stadtmüller (2005), a generalized functional linear regression model for multiple functional predictors can be constructed by associating a univariate latent variable z_i with y_i through

$$y_i = \begin{cases} 1 & \text{if } z_i < 0, \\ 0 & \text{if } z_i \geq 0. \end{cases} \tag{1}$$

where

$$z_i = \beta_0 + \sum_{j=1}^{J} \int_{\mathcal{T}_j} x_{ij}(s)\beta_j(s)ds + \epsilon_i \tag{2}$$

and $\epsilon_i \sim N(0, 1)$ determines a probit link between y_i and z_i. Note that we assume $\mathcal{T}_j$s are compact domains of $x_{ij}(t)$. Based on above model setting, standard functional regression estimation paradigms, such as the EM algorithm in James (2002), or the estimating equation method in Müller & Stadtmüller (2005) can be performed to estimate the intercept β_0 and the coefficient functions $\beta_j(t)$s. However, when the $x_{ij}(t)$s contain large amount of redundant information which is unrelated to the responses, the efficiency of the model will be significantly reduced. Also, when J is large, the convergence of the estimation can hardly be guaranteed. This motivates us to consider variable selection method which selects a subset of the covariates as predictors. Due to the infinite dimensionality of functional data, point-wise selection from the predictors $x_{ij}(t)$ is not a practical choice. One can discretize $x_{ij}(t)$ on a finite grid and transform the problem to a multivariate model, but this will ignore the correlation between contiguous points on the grid. A natural choice is to apply standard functional dimension reduction methods to reduce the dimension first and conduct variable selection on the reduced space. If we assume $\forall j$, $x_{ij}(t) \in \mathcal{H}_j$ for some separable Hilbert space $\mathcal{H}_j$, we can expand $x_{ij}(t)$ on a set of com-

plete orthonormal basis $\{\phi_k^j\}_{k=1}^{\infty}$

$$x_{ij}(t) = \sum_{k=1}^{\infty} c_{ijk}\phi_k^j(t) \tag{3}$$

and the truncated version of (3) can be used to approximate $x_{ij}(t)$ since $\sum_{k=1}^{\infty} |c_{ijk}|^2 < \infty$. And similarly, we assume $\beta_j(t) \in \mathcal{H}_j$, thus

$$\beta_j(t) = \sum_{k=1}^{\infty} b_{jk}\phi_k^j(t) \tag{4}$$

Note that the orthonormal basis $\{\phi_k^j\}_{k=1}^{\infty}$ can be chosen to be a known basis such as a Fourier basis or a wavelet basis. If in addition, we assume $x_{ij}(t) \in L_2[\Omega]$ for the underlying sample space Ω, i.e. $E[x_{ij}(t)^2] < \infty, \forall t \in \mathcal{T}_j, \forall j$, Mercer's theorem and Karhunen-Loève theorem (Ash & Gardner 1975) suggests to take the orthonormal basis to be the eigenfunctions of the covariance operator K defined by

$$Kx(t) = \int x(s)k(s,t)ds, \quad k(s,t) = Cov(x(s), x(t)) \tag{5}$$

In this case, the coefficients $\{c_{ijk}, k = 1, \ldots, \infty\}$ are called functional principal component scores of $x_{ij}(t)$. Using functional principal component method is different from using known basis in that the eigenfunctions need to be estimated. Various estimating methods are proposed as in Ramsay & Silverman (1997), and in Hall, Müller & Wang (2006).

Once the orthonormal basis has been chosen or estimated, we can reduce equation (2) to

$$z_i = \beta_0 + \sum_{j=1}^{J}\sum_{k=1}^{p_j} c_{ijk}b_{jk} + \epsilon_i \tag{6}$$

where p_j is the truncation parameter for the jth functional predictor. We thus transfer the functional regression to multivariate regression. Variable selection can therefore be used to select among the reduced scores $\{c_{ijk}, j = 1, \ldots, J, k = 1, \ldots, p_j\}$. For convenience, we denote

$$C_i = (1, c_{i11}, \ldots, c_{i1p_1}, \ldots, c_{iJ1}, \ldots, c_{iJp_J})$$

$$\beta = (\beta_0, b_{11}, \ldots, b_{1p_1}, \ldots, b_{J1}, \ldots, b_{Jp_J})^T$$

Equation (6) can be simplified to

$$z_i = C_i\beta + \epsilon_i \tag{7}$$

Let $Z = (z_1, \ldots, z_n)^T$, $Y = (y_1, \ldots, y_n)^T$ and $X^T = (C_1^T, \ldots, C_n^T)$, then the conditional distribution of Z given β and Y are

$$\begin{aligned} f(Z|\beta, Y) &\propto \prod_{i=1}^{n} \phi(z_i - C_i\beta)\left(I_{\{z_i<0\}\cap\{y_i=1\}} + I_{\{z_i\geq 0\}\cap\{y_i=0\}}\right) \\ &\propto \exp\left\{-\frac{1}{2}(Z - X\beta)^T(Z - X\beta)\right\} \\ &\quad \times \prod_i \left(I_{\{z_i<0\}\cap\{y_i=1\}} + I_{\{z_i\geq 0\}\cap\{y_i=0\}}\right) \end{aligned} \tag{8}$$

where $\phi(\cdot)$ is the density function of $N(0,1)$ and $I_{\{\cdot\}}$ is the indicator function. And conditional on Z, we get a normal linear regression

$$Z = X\beta + \epsilon \tag{9}$$

The latent variables Z thus play the roles of simplifying computation by transferring the problem to a typical normal regression (Albert & Chib 1993). For the convenience of setting priors and MCMC sampling, we can standardize X in equation (9) by centering and scaling it to zero mean and unit variance. To perform variable selection, we introduce a hyper parameter τ to the priors of β by

$$\beta|\tau \sim N(0, \Sigma_\tau) \tag{10}$$

where $\Sigma_\tau = D_\tau R D_\tau$, R is the prior correlation matrix of β, and D_τ is the prior marginal standard deviation of β, which takes the form

$$D_\tau = Diag\{h, \tau_i\nu_{1i} + (1 - \tau_i)\nu_{0i}, i = 1, \ldots, K\} \tag{11}$$

where $K = \sum_j p_j$, h is a large number giving a large prior variance to the intercept term β_0, and $\nu_{1i} >> \nu_{0i} > 0$ for all i so that the corresponding component of β will cluster around 0 when $\tau_i = 0$ and have relatively large variances when $\tau_i = 1$. Priors of τ_i are set to be Bernoulli(ω_i), i.e. $f(\tau_i) = \omega_i^{\tau_i}\omega_i^{1-\tau_i}$. For simplicity, we assume τ_is are independent, and we can always set $\omega_i \equiv \omega$, $\nu_{1i} \equiv \nu_1$, and $\nu_{0i} \equiv \nu_0$ if no further information is known about the priority of selecting certain covariates. Generally, ω represents the prior belief for the proportion of covariates to be selected. The estimation of the vector $\tau = (\tau_1, \ldots, \tau_K)$ will indicate the selection of the variables. When R is taken to be the identity matrix, the priors of each component of β is independent. Correlated priors are suggested in George & McCulloch (1993, 1997). With the above setting of the priors, we get the joint posterior distribution of β, τ conditional on Z as

$$\begin{aligned} f(\beta, \tau|Z, Y) &\propto \pi(Z|\beta, \tau, Y)\pi(\beta|\tau)\pi(\tau) \\ &\propto \exp\left\{-\frac{1}{2}(Z - X\beta)^T(Z - X\beta)\right\} \\ &\quad \times |\Sigma_\tau|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\beta^T\Sigma_\tau^{-1}\beta\right\} \prod_{i=1}^{K} f(\tau_i) \end{aligned} \tag{12}$$

Based on equation (8) and (12), we propose the following MCMC algorithm using Gibbs sampler to obtain posterior samples:

Step 1: Set up the initial value $\beta^{(0)}$,$\tau^{(0)}$ and the priors ν_1,ν_0 and w.

For $j = 1, \ldots, nMc$, conduct step 2-4, where nMc is the total number of iterations.

Step 2: Conditional on $\beta^{(j-1)}$ and Y, sample $Z^{(j)}$ from truncated Normal distribution (8).

Step 3: Conditional on $Z^{(j)}$, update $\beta^{(j)}$ from a multivariate normal distribution.

$$\beta^{(j)}|Z^{(j)}, \tau, Y \sim N(T_\tau^{-1}X^T Z^{(j)}, T_\tau^{-1})$$

where $T_\tau = X^T X + \Sigma_\tau^{-1}$.

Step 4: Update $\tau^{(j)}$.

Case 1: In case of $R = I$, the posterior distribution for τ_k are independent for $k = 1, \ldots, K$, hence we can update $\tau^{(j)}$ marginally using posterior odds. i.e.

$$r_k = \frac{\pi(\tau_k = 1|\beta^{(j)}, Z^{(j)}, Y)}{\pi(\tau_k = 0|\beta^{(j)}, Z^{(j)}, Y)}$$

$k = 1, \ldots, K$. Update $\tau_k^{(j)} = 1$ with probability $r_k/(r_k + 1)$.

Case 2: In case of $R \neq I$, the posterior distribution for τ_i are not independent. Metropolis-Hasting can be used to propose the candidate $\tau^{(c)}$ similarly as Brown et al.(1998). i.e. Based on $\tau^{(j-1)}$, either change one 1 to 0, or change one 0 to 1, or swap one pair of 0 and 1 with certain pre-defined probability.

MCMC algorithm in case 2 is a hybrid Gibbs/Metropolis-Hasting sampling process since it combines Metropolis-Hasting updates with a larger Gibbs sampling iteration. Note that although $\tau_k^{(j)} = 0$ in iteration j indicates that the kth covariate is not selected, we do not remove this covariate in the $(j+1)$th iteration since the posterior sample for $b_{ik}^{(j)}$ will be close to 0 and thus the contribution of that covariate to the regression model will be negligible.

An alternative to the above proposed MCMC algorithm is to integrate β out from equation (12) so that τ can be updated independent of β. In this case, the posterior distribution of τ can not be marginally updated using posterior odds as in case 1, hence the stochastic search method using Metropolis-Hasting as in case 2 has to be used. However, since marginal updating of τ converges faster than the Metropolis-Hasting, the alternative method does not show too much advantage on mixing.

3. Simulation study

Two simulations are conducted to verify the performance of the proposed Bayesian variable selection model on functional data classification. Simulation 1 uses only one functional predictor, i.e. $J = 1$ in equation (2). Functional predictors are generated using 5 orthonormal cosine basis on the interval $[0, 1]$ so that the curves are simple enough and we can use a small number of orthonormal basis to approximate the functional predictor. Simulation 2 considers multiple functional predictors for each observation, i.e. $J = 20$ in equation (2). Thus the total number of variables to be selected is relatively large. The estimation results are shown and prediction results are compared with several other classifiers.

Simulation 1: Let the sample size to be $n = 1000$, we simulate a single functional predictor for each observation, i.e., $J = 1$ in equation (2) and the index j will be omitted in this simulation. Functional predictors $x_i(t)$ are generated using the first 5 cosine basis on the closed set $[0, 1]$, i.e. $\phi_0(t) = 1, \phi_k(t) = \sqrt{2}cos(k\pi t), k = 1, \ldots, 4$. The mean curve is predefined using cosine coefficients $c = (-1.12, -1.82, 7.77, 2.15, 3.25)$ and each functional predictor is generated by adding a random error $N(0, 1)$ to each component of c. For the true coefficient function $\beta(t)$, we set the first 5 cosine basis score to be $b_1 = b_3 = b_4 = 0$, and $b_2 = 5$, $b_5 = -4$, which corresponds to the true value of $\tau = (0, 1, 0, 0, 1)^T$. The latent variables z_i are generated using equation (2) by numerical integration. Note that $x_i(t)$ need to be centered first before integrating to satisfy the zero mean assumption, and β_0 is taken to be a scalar value 1.5. The binary responses y_i are generated from the sign of z_i. We randomly take 800 observations as the training set and the rest as the test set. The proposed model is applied to the above simulated data. For convenience of comparing the estimated coefficient β with the true values, we choose the same number of cosine basis to reduce the dimension of the functional predictor, which reduce the $x_i(t)$ to 5 scores. By choosing basis that coincides with the ones used to generate data, we minimized curve approximation error and it helps us to see the estimation performance in the variable selection step. Based on the reduced cosine basis scores, the model is trained using Gibbs sampler stated in section 2 with $\omega_i \equiv \omega = 0.4$, $R = I$, $\nu_1 = 20$, $\nu_0 = 0.02$, and 10000 MCMC iterations and 3000 burn-in period. By averaging the posterior samples of τ, we get the marginal posterior probability $P\{\tau_i = 1, i = 1, \ldots, 5\}$ to be $(0.007, 1.00, 0.018, 0.005, 1.000)^T$, which indicates that our algorithm has correctly picked out the second and the fifth cosine basis scores with enough accuracy. The estimation of reduced coefficient scores are obtained by the posterior sample mean, which are compared with the probit-link maximum likelihood estimation in table 1. Table 1 shows that the posterior estimation of the coefficient scores is as good as the maximum likelihood estimate. Posterior prediction of coefficient curve $\beta(t)$ can be easily computed by inverse cosine transform of the posterior samples of coefficient scores. Figure 2 shows the posterior estimation of $\beta(t)$ and the corresponding 95% credible interval computed point-wisely using 2.5% and 97.5% quantile of the inverse transformed posterior samples.

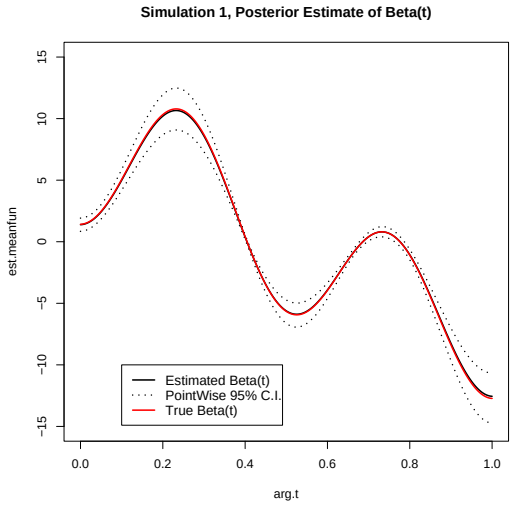

Figure 2: Simulation 1– the posterior estimation of $\beta(t)$

Table 1: Simulation 1– the estimation of vector β compared with maximum likelihood estimation(MLE). Note that $\omega_i$indicates $P\{\tau_i = 1\}$. BVS: The Bayesian variable selection model proposed in section 2.

True		*MLE*		*BVS*				
τ	β	$\hat{\beta}$	*S.E.*	$\hat{\beta}$	*S.E.*	95% *C.I.* 2.5%	97.5%	$\hat{\omega_i}$
-	-0.02	-0.09	0.11	0.05	0.10	-0.15	0.27	-
0	0	-0.09	0.11	0.00	0.02	-0.04	0.04	0.007
1	5	4.94	0.48	4.93	0.41	4.17	5.77	1.000
0	0	0.21	0.11	0.01	0.04	-0.03	0.05	0.018
0	0	-0.08	0.11	0.00	0.02	-0.04	0.04	0.005
1	-4	-3.91	0.39	-3.96	0.33	-4.67	-3.34	1.000

Posterior prediction for the test data can be easily computed by applying the posterior samples of β vector to the cosine basis scores of test set. Note that since the training set has been centered and scaled to zero mean and unit variance, the test data need to be centered and scaled using the mean and standard deviation obtained from the training set. If treating $y_i = 1$ as diseased and $y_i = 0$ as normal class, the out-of-sample prediction of test set provides sensitivity 90% and specificity 98% with a total misclassification rate 6% and area under the ROC curve 0.986 (See Zweig & Campbell (1993) for ROC Curves).

Instead of using cosine basis to reduce the dimension, we also tried to use functional principal component to approximate the functional predictor, the proposed model are trained and applied to the test data, the resulting prediction sensitivity is 91% and specificity is 96%, total misclassification error is 6.5% and the area under the ROC curve to be 0.987. These results indicates that using functional principal component to reduce the dimension will produce as accurate prediction as using cosine basis, although the data are not generated using eigenfunctions.

Simulation 2: In this simulation, we evaluate the performance of the model with multiple functional predictors. Let $J = 20$ in equation (2). Functional predictors are generated similarly as simulation 1 using the first 5 cosine basis. Thus the total number of scores in equation (6) is $K = J \times p = 100$. For the coefficient scores β, we randomly choose 15 out of 100 components to be nonzero, which take values from a uniform distribution with support $[-7, 7]$, and let the intercept $\beta_0 = 5$. We use the same way as in simulation 1 to generate latent variables and binary responses, and to split the training and test sets.

To apply the proposed model, we also choose 5 cosine basis to approximated the functional predictors as in simulation 1. Under 10000 MCMC iterations and 2000 burn-in period, and with the prior setting to be $w = 0.15$,$R = I$, $\nu_1 = 20$ and $\nu_0 = 0.02$, we get prediction result of 95% sensitivity and 99% specificity with total misclassification rate 2.5%. 14 out of 15 nonzero β scores are correctly picked out with posterior probability of τ_i equals 1. The other 1 have posterior probability 0.974, and all those βs with true value 0s have estimated posterior probability of $\tau_i = 1$ less than 0.03. These results show that, even with fairly large number of predictors $J = 20$,

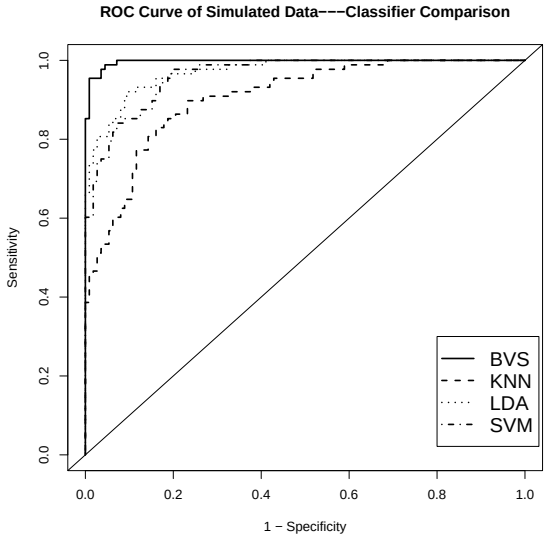

Figure 3: Simulation 2– the ROC curves of different classification models. BVS: the proposed Bayesian variable selection model. KNN: K-nearest neighbor. LDA: Linear Discriminant Analysis. SVM: support vector machine. Note that all classifiers are based on first 5 cosine basis scores

the proposed model can obtain accurate estimation of both τ. The prediction performance of the proposed Bayesian variable selection model is compared with three other classifiers by comparing the empirical ROC curves in figure 3. All the 4 methods are based on the same dimension reduction method, i.e. the first 5 cosine basis scores. Note that the number of neighbors k used in KNN is the optimal k determined by 10 block cross validation using the training data. The detailed prediction results are compared and reported in table 2. Note that the sensitivities and specificities in table 2 are the points picked from the ROC curves by using thresholds so that the sums of sensitivity and specificity are maximized. Figure 3 and table 2 show that,if there exists redundant information in the functional predictor, the variable selection model offers better prediction results than models using all the covariates equally.

Table 2: Simulation 2– the comparison of classificat methods on the test set. AUC: Area under the ROC cu Sens: sensitivity; Spec: specificity; MisR: misclassificat rate. The BVS, KNN, LDA, SVM are defined same as in ure 3.

Method	*AUC*	*Sens*	*Spec*	*MisR*
BVS	0.997	95%	99%	2.5%
KNN	0.905	83%	84%	16.5%
LDA	0.971	92%	90%	9.0%
SVM	0.963	98%	80%	12.0%

4. Fluorescence spectroscopy data classification

Totally 1013 EEM measurements were made from 521 tients. Measurements were taken from different sites of the cervix and may include repeated measurements at the same site. All the measurements were made using the same instrument called FastEEM2. And all normal measurements were those measured on squamous tissue, which reduced the confounding effects due to the tissue type. The curves were preprocessed by background correction, smoothing and registration procedures. Data are split into a training set and a test set randomly with 607 measurements in the training set and 406 in the test set. The proportions of diseased cases within each set are 0.096, 0.080, respectively. Both cosine basis approximation and functional principal components are chosen as the way to reduce the dimension of functional predictors. The reduced scores are centered and scaled to zero mean and unit variance in the training set. To reduce possible bias, the reduced scores of the test set is computed and normalized based on only information obtained from the training set. For example, the eigenfunctions used for computing functional principal component scores of the test set are estimated from training set. And the mean and standard deviation used for normalizing the test set are those estimated from the training set.

The proposed Bayesian variable selection model is applied to the scores obtained from cosine basis expansion and functional principal component analysis. 5 scores per curve are used in both cases. For both types of scores, we set the priors as $w = 0.3, \nu_1 = 20, \nu_0 = 0.02$, $R = I$ with 10000 MCMC iterations and 2000 burn-in period. The posterior probability of $\tau_i = 1$ in the functional principal component case are plotted as an image plot in figure 4. The x-axis in figure 4 indicates the five functional principal component scores from a single excitation curve. The y-axis indicates the spectroscopy curves. Figure 4 shows that, out of 80 principal component scores, only 5 have posterior probability greater than 0.5. And 4 of these 5 scores are the third or fourth principal components. The posterior prediction results to the test set are compared with three other classifiers similarly as in simulation 2. Table 3 shows a comparison between the 4 different classifiers used in simulation 2. Figure 5 shows the corresponding ROC curves for the test data prediction. Both table 3 and figure 5 shows that the proposed Bayesian variable selection model performs better than the other three classifiers on both types of curve approximation methods. Using 5 principal components

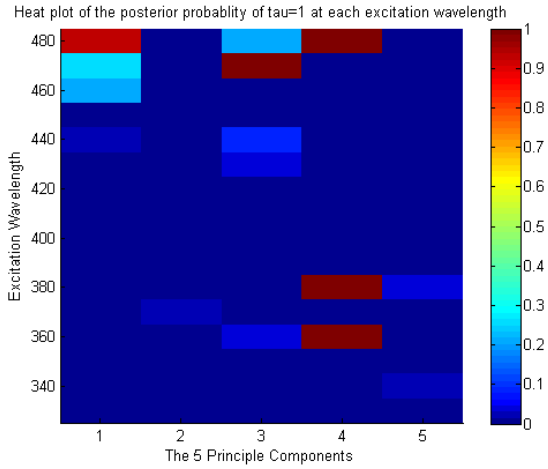

Figure 4: The posterior probability of $\tau_i = 1$ for all the scores obtained using 5 functional principal components.

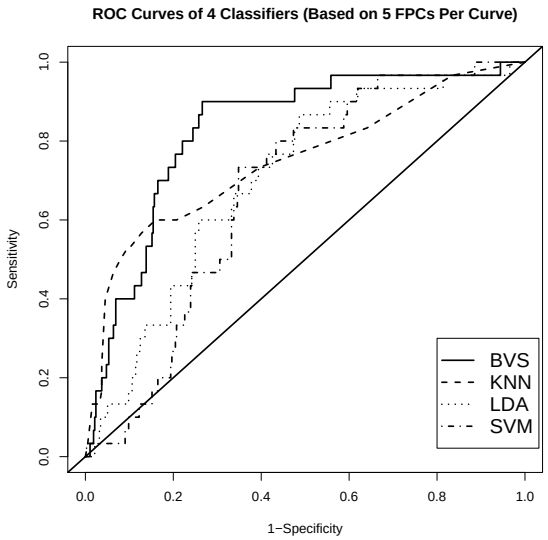

Figure 5: Empirical ROC curve for the test set.

reached sensitivity 90% and specificity 70% with area under the curve 0.83. The results for cosine basis expansion give higher specificity but lower sensitivity.

In summary, the proposed Bayesian variable selection provides us an efficient classification algorithm on fluorescence spectroscopy data as compared with several other classifiers. Rather than treating all scores equally, it reduces the dimension of the data to some scores by function approximation or functional principal component analysis, and selects useful scores for classification. For this particular data set, although the total misclassification rate is at 20% level (which is no better than the k-nearest neighbor method), it reduces the risk of false negative diagnosis by providing sensitivity as high as 90%, and at the same time remains a reasonable specificity at 70% level.

Table 3: A comparison four classification methods. FPCA(5): Using the first 5 functional principal components. Cosine(5): Using first 5 cosine basis. Sens,Spec,MisR and BVS, KNN, LDA, SVM are defined same as in table 2.

	FPCA(5)				*Cosine(5)*			
Method	*AUC*	*Sens*	*Spec*	*MisR*	*AUC*	*Sens*	*Spec*	*MisR*
BVS	0.83	90%	73%	26%	0.83	73%	79%	21%
KNN	0.71	60%	84%	22%	0.71	60%	85%	18%
LDA	0.70	87%	51%	46%	0.73	90%	54%	44%
SVM	0.68	73%	65%	34%	0.68	83%	60%	39%

5. Discussion

The number of basis functions used in (2) for curve approximation is one concern of our study. Since the main purpose of the study is prediction, we compute the prediction results for different number functional principal components and plotted them as in figure 6. Different criterion may result in different choice. But in all, we can see that using around 6 functional principal components gets sensitivity 87% and specificity 74% with area under the ROC curve 0.83, which indicates a fairly good prediction. Small misclassification rate may not be the best criterion to choose the number of basis since classifiers with higher specificity but lower sensitivity will tend to have lower misclassification rate because the number of diseased observation is fairly small as compared with the number of normal observations. And in real application, we wish to obtain enough sensitivity to reduce the risk of false negative diagnosis.

On setting up priors to the proposed model, since the scores obtained by curve approximation will be normalized before applying variable selection, the priors of β vector will not be influenced by the absolute levels of the data. And it turns out that as long as the priors satisfy $\nu_1 >> \nu_0 > 0$, the estimation results of β and τ won't be influenced much. Another issue of concern is the convergence of MCMC. The convergence of MCMC in this study is confirmed by running multiple chains starting from different initial values.

In summary, we have proposed a model on functional data classification using Bayesian variable selection. This model uses probit link to connect the binary or ordinal responses to the covariates and automatically selects informative covariates through a mixture normal prior of the regression coefficients. The model shows advantages over classical classification methods such as K nearest neighbor, linear discriminant analysis and support vector machine.

Future work to improve the precision of classification includes trying more dimension reduction methods such as Bayesian nonparametric techniques. Another possible improvement is to build random effects caused by instruments, tissue type and menopausal status into the variable selection model to account for effects coming from these factors. Reducing the dimension of EEM measurement by each excitations curve is just a convenient way of data pre-processing. EEM plot in figure 1 shows obvious continuity across the excitation curves. Hence the EEM data can be treated as 2-d functions. But since the data on the y-axis is very sparse, methods involving variable selection on both directions need

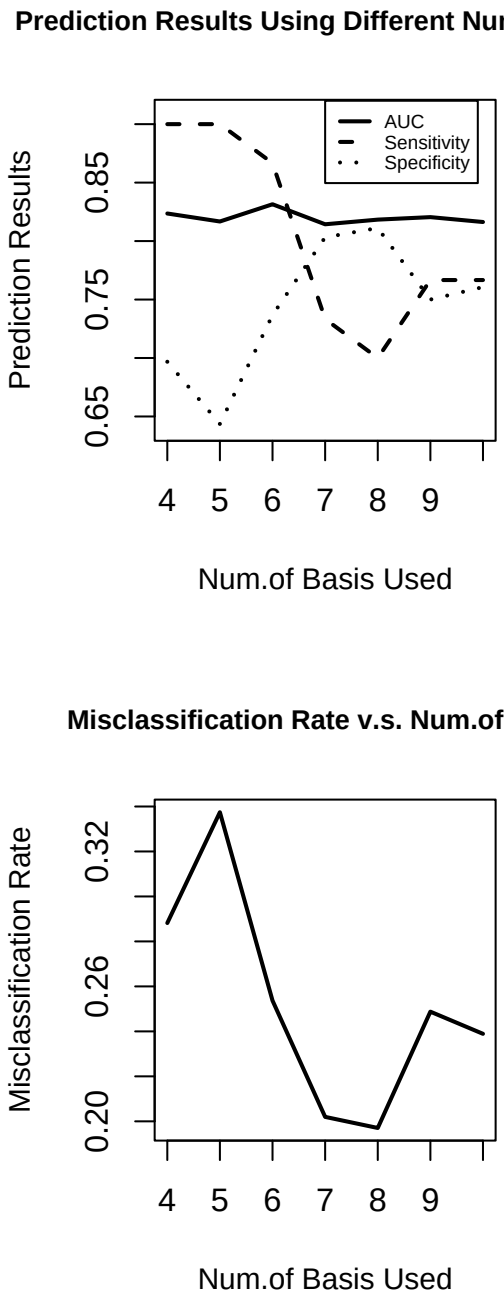

Figure 6: Prediction results under different number of functional principal components

to be explored.

6. Acknowledgement

This research was supported by National Cancer Institute grant PO1-CA82710.

REFERENCES

Albert, J. H. and Chib, S. (1993), "Bayesian analysis of binary and polychotomous response data," *Journal of the American Statistical Society*, **88**, 669–679.

Ash, R. B. (1975) ,*Topics in Stochastic Processes*, Academic Press, New York.

Brown, P. J. , Vannucci, M. and Fearn, T. (1998) , "Bayes model averaging with selection of regressors," *Journal of the Royal Statistical Society, Series B*, **60**, 627–641.

Brown, P. J. , Vannucci, M. and Fearn, T. (2002) , "Bayes model averaging with selection of regressors," *Journal of the Royal Statistical Society, Series B*, **64**, 519–536.

Chang, S. K. ,Follen, M. , Malpica, A. , Utzinger, U. , Staerkel, G. , Cox, D., Atkinson, E.N., MacAulay, C. and Richards-Kortum, R.(2002), "Optimal excitation wavelengths for discriminantion of cervical neoplasia," *IEEE Transactions on Biomedical Engineering*, **49**,1102–1110.

George, E. I. and McCulloch, R. E. (1993), "Variable selection via gibbs sampling," *Journal of the American Statistical Association*, **88**, 881–889.

George, E. I. and McCulloch, R. E. (2002)," Approaches for bayesian variable selection," *Statistics Sinica*, **7**, 339–373.

Hall, P. , Müller, H. and Wang, J. (2006), "Properties of principle component methods for functional and longitudinal data analysis," *The Annals of Statistics*, **34**, 1493–1517.

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

Kamath, S. D. and Mahato, K.K. (2007), "Optical pathology using oral tissue fluorescence spectra: classification by principle component analysis and k-means nearest neighbor analysis," *Journal of Biomedical Optics*, **12**.

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

Palmer, G. M., Zhu, C. , Breslin, T. M. , Xu, F. , Gilchrist, K. W. and Ramanujam, N. (2003), "Comparison of multiexcitation fluorescence and diffuse reflectance spectroscopy for the diagnosis of breast cancer," *IEEE Transactions on Biomedical Engineering*, **50**.

Ramanujam, N., Mitchell, M.F., Mahadevan, A., Thomsen, S., Malpica, A., Wright, T., Atkinson, N. and Richards-Kortum, R. (1996), "Spectorscopic diagnosis of cervical intraepithelial neoplasia (cin) in vivo using laser induced fluorescence spectra at multiple excitation wavelengths," *Lasers Surg. Med.*, **19** 63–67

Ramanujam, N., Mitchell, M.F., Mahadevan, A., Warren, S., Thomsen, S., Silva, E. and Richards-Kortum,R. (1994) "*In vivo* diagnosis of cervical intraepithelial neoplasia using 337-nm-excited laser-induced fluorescence," *Proc. Natl. Acad. Sci.*, **91**, 10193–10197.

Ramsay, J. and Silverman, B. (1997), *Functional Data Analysis*, Springer-Verlag, New York.

Welch, A. J., Gardner, C. , Richards-Kortum, R. , Chan, E. , Criswell, G. , Pfefer, J. and Warren, S.(1997), "Propagation of fluorescent light," *Lasers in Surgery and Medicine*, **21**, 166–178.

Zweig, M.H. and Campbell G. (1993), "Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine," *Clinical Chemistry*, **39**, 561-577.