Available online at www.sciencedirect.com

SCIENCE DIRECT

Chemometrics and Intelligent Laboratory Systems 77 (2005) 139–148

Chemometrics and intelligent laboratory systems

www.elsevier.com/locate/chemolab

NIR and mass spectra classification: Bayesian methods for wavelet-based feature selection

Marina Vannucci[a,*], Naijun Sha[b], Philip J. Brown[c]

[a]*Department of Statistics, Texas A&M University, United States*
[b]*Department of Mathematical Sciences, University of Texas at El Paso, United States*
[c]*Institute of Mathematics and Statistics, University of Kent, Canterbury, UK*

Received in revised form 10 September 2004; accepted 10 October 2004
Available online 4 March 2005

Abstract

Here we focus on classification problems that involve functional predictors, specifically spectral data. One of our practical contexts involves the classification of three wheat varieties based on 100 near infra-red absorbances. The dataset consists of a total 117 samples of wheat collected during a study that aimed at exploring the possibility of using NIR spectra to assign unknown samples to the correct variety. In another example we look at serum spectra from 162 ovarian cancer and 91 control subjects generated through surface enhanced laser desorption ionization time-to-flight mass spectrometry (SELDI-TOF). We employ wavelet transforms as a tool for dimension reduction and noise removal, reducing spectra to wavelet components. We then use probit models and Bayesian methods that allow the simultaneous classification of the samples as well as the selection of the discriminating features of the spectra. In both examples our method is able to find very small sets of features that lead to good classification results.

Keywords: Bayesian variable selection; Discrimination; Multinomial probit models; NIR spectra; Proteomic data; Wavelet transforms

1. Introduction

Classification problems with functional predictor data has become prominent in many fields in recent years. Near infrared (NIR) spectroscopy is used in many areas, such as in the analysis of food and drink and of pharmaceutical products. A NIR spectrum of a sample is typically measured by modern scanning instruments at hundreds of equally spaced wavelengths. The information in the curve is used to predict the chemical composition of the sample by extracting the relevant information from many overlapping peaks. Osborne et al. [1] describe standard approaches, such as linear discriminant analysis (LDA). These methods can fail with many variables and different approaches need to be taken, see for example Krzanowski et al. [2]. A common solution is to reduce the dimension of the predictor matrix by using principal components and then apply LDA.

Recent advances in functional genomics have made it possible to measure the expression levels of thousands of genes or proteins. Proteomic technology is non-invasive and requires small amounts of biological material (tissue or blood samples). Proteomics is often used for biomarker discovery, to identify proteins linked to disease status, response to therapy, or clinical prognosis. A mass spectrum can be represented as a curve where the x-axis indicates the ratio of the weight of a specific molecule to its electrical charge (m/z) and the y-axis is the signal intensity for the same molecule as a measure of the abundance of that molecule in the sample. Proteomic spectra are characterized by many peaks, most of which correspond to proteins or protein fragments (peptides). The identification of peaks related to a specific outcome, for example peaks that discriminate samples or that predict a clinical response, is

* Corresponding author.
E-mail addresses: mvannucci@stat.tamu.edu (M. Vannucci), nsha@utep.edu (N. Sha), Philip.J.Brown@kent.ac.uk (P.J. Brown).

0169-7439/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.chemolab.2004.10.009

often the goal of the analysis. Some studies have focused on developing methods for diagnosing cancer using the proteomic profile of a serum sample [3]. Very few contributions have appeared on statistical analyses of proteomic data. Wu et al. [4] compare several methods for classification based on mass spectra, including linear and quadratic discriminant analysis and classification trees methods. In their conclusions the authors emphasize the need for methods to remove noise from the data and select relevant features. Proteomic spectra may have, in fact, as many as 60,000 observations per spectrum. Qu et al. [5] look at wavelet transforms and apply standard discriminant analysis to the denoised wavelet coefficients.

Here we look at probit models for classification combined with Bayesian variable selection methods to simultaneously classify the samples and identify the features of the spectra that characterize the different classes. We employ orthogonal wavelet transformations as an effective tool for dimension reduction and, in the analysis of proteomic data, for noise removal. Wavelets have been successfully employed for thresholding, i.e. the removal of noise from the data, following the seminal work of Donoho and Johnstone [6,7]. Chang et al. [8] have used wavelets in classification to approximate Bayesian classifiers. Recently, wavelet-based methods for modelling functional data based on feature selection have been put forward by Brown et al. [9] and Morris et al. [10]. Here we extend some of their ideas to classification settings. The probit modelling we use for the selection strategy has been previously introduced in Sha et al. [11] in the functional genomic context to classify samples based on gene expression profiles. There the methodology is applied to the simpler case of binary responses only, and to the selection of the original variables. Here we point out the relevance of the modelling for the analysis of functional data and extend the methodologies to the selection of derived wavelet components. We provide applications to multinomial data arising in NIR spectroscopy to the diagnosis of ovarian cancer based on mass spectra.

2. Wavelet regression

Wavelets are families of orthonormal basis functions that can be used to parsimoniously represent other functions. For example, in $L^2(\mathbb{R})$, an orthogonal wavelet basis is obtained by dilating and translating a *mother wavelet* ψ as $\psi_{jk}(x)=2^{j/2}\psi(2^j x-k)$ with j, k integers. A function f can then be represented by the wavelet series $f(x) = \sum_{j,k\in\mathbb{Z}} d_{jk}\psi_{jk}(x)$, with wavelet coefficients $d_{jk} = \int f(x)\psi_{jk}(x)dx$ describing features of the function f at the spatial locations indexed by k and scales indexed by j.

Wavelets have been extremely successful as a tool for the analysis and synthesis of discrete data. Let $\mathbf{X}=(x_1, \ldots, x_n)'$ be samples of a function taken at n equally spaced points. A fast algorithm, the discrete wavelet transform (DWT), exists for decomposing $\mathbf{X}$ into a set of n wavelet coefficients [12] in only $O(n)$ operations. Although it operates in practice by means of linear recursive filters, the DWT can be also represented in matrix form as $\mathbf{D}=\mathbf{WX}$ with $\mathbf{W}$ an orthogonal matrix corresponding to the discrete wavelet transform and $\mathbf{D}$ the vector of wavelet coefficients. An algorithm for the inverse reconstruction, the IDWT, also exists. Wavelets possess many useful properties. Daubechies [13] proposed a class of wavelet families which have compact support and maximum number of vanishing moments for any given smoothness. These properties allow an effective and parsimonious representation of functions with local behavior. Daubechies wavelets are extensively used in applications.

Wavelet shrinkage, i.e. the estimation of a function from noisy observations, is probably the most successful application of wavelets. There a wavelet transform is applied to the data and the noise is removed by thresholding or shrinking the smallest wavelet coefficients, Donoho and Johnstone [6,7]. Bayesian approaches have also been proposed that use mixture priors on the wavelet coefficients. The recent review paper of Antoniadis et al. [14] provides an exhaustive review of the different approaches, classical and Bayesian, and related extensions. All these contributions are limited to the single function setting. Wavelet-based methods for the analysis of multiple curves are in Brown et al. [9] who considered regression models that relate a multivariate response to functional predictors, applied wavelet transforms to the curves, and used Bayesian selection methods to identify features that best predict the responses. Vannucci et al. [15] used decision theoretical methods in the same multivariate regression setting. Also, Morris et al. [10] extended ideas of wavelet regression to the setting of nested functional modelling.

3. Bayesian variable selection for discrimination

Here we use probit models for classification purposes. Albert and Chib [16] proposed a Bayesian approach to inference that uses data augmentation and introduces latent responses into the model. Let ($\mathbf{Z}$, $\mathbf{X}$) indicate the observed data, with $\mathbf{X}_{n\times p}$ the predictor matrix and $\mathbf{Z}_{n\times 1}$ a (categorical) response vector coded as 0,. . ., $J-1$, for J classes. Let $\mathbf{Y}_{n\times q}$ with $q=J-1$, be a latent matrix for the $\mathbf{Z}_{n\times 1}$ observed categorical vector. The element $y_{i,j}$ is the unobserved propensity of the ith subject to belong to the jth class. Let us assume a multivariate normal distribution for $\mathbf{Y}$ with common covariance across the different groups

$$\mathbf{Y}_i = \alpha' + \mathbf{X}_i'\mathbf{B} + \varepsilon_i, \varepsilon_i \sim N(0, \Sigma), i = 1, \ldots, n \qquad (1)$$

with $\mathbf{Y}_i=(y_{i,1}, \ldots, y_{i,q})$ the row vector of $\mathbf{Y}$ corresponding to the ith subject, $\mathbf{X}_i$ the vector of p predictor values and $\mathbf{B}$ the $p\times q$ matrix of regression coefficients. Model (1) consists therefore of q regressions on p variables, with p the number

of data points in the curves. The relationship between z_i and the unobserved $\mathbf{Y}_i$ becomes

$$z_i = \begin{cases} 0 \text{ if } \max\limits_{1\le k\le J-1}\{y_{i,k}\}\le 0 \\ j \text{ if } \max\limits_{1\le k\le J-1}\{y_{i,k}\}>0 \text{ and } y_{i,j} = \max\limits_{1\le k\le J-1}\{y_{i,k}\} \end{cases} \quad (2)$$

In Sha et al. [11] Bayesian variable selection is done by elaborating the prior on B through the introduction of a binary p-vector γ with the jth element γ_j either 1 or 0 according to whether the jth variable is included or not in the model. For this selection prior each column of B is modeled as having a mixture of a point mass distribution at zero with a conjugate normal distribution on those coefficients corresponding to the elements γ_j=1. Conjugate normal and inverse Wishart distributions are imposed on α and $\sum$, respectively. The simplest form of the prior distribution on γ results in a binomial distribution for the number of nonzero elements with expectation pw where w is the probability of inclusion of a single variable. A further Beta prior distribution can be imposed on w allowing it to be either concentrated or widely dispersed according to the choice of the hyperparameters.

3.1. Wavelet component selection

When a wavelet transform is applied to each row of $\mathbf{X}$ the model becomes

$$\mathbf{Y}_i = \alpha' + \mathbf{D}_i'\tilde{\mathbf{B}} + \varepsilon_i, \varepsilon_i \sim N(0, \sum), i = 1, \ldots, n \quad (3)$$

with $\mathbf{D}=\mathbf{XW}$ the matrix of wavelet coefficients and $\tilde{\mathbf{B}}=\mathbf{W}'\mathbf{B}$ the matrix of transformed regression coefficients. The prior structure described in the previous section nicely transforms to priors in the wavelet domain. Shrinkage mixture priors are now imposed on the transformed regression coefficients. Suitable prior covariance structures $\mathbf{H}$ can be specified in the domain of the data and transformed to modified priors on the wavelet coefficients using results from Vannucci and Corradi [17] for computations of quantities such as $\mathbf{W}'\mathbf{H}\mathbf{W}$. The priors on α and $\sum$ are unchanged.

3.2. Posterior inference

Conditioned on the latent responses, the model is equivalent to that of Brown et al. [18,19] for regression models with multivariate responses, although inference in this setting is complicated by the presence of the unknown latent responses. Sha et al. [11] proposed a fast scheme for posterior inference, essentially integrating out α, $\mathbf{B}$ and $\sum$ from the joint posterior. The latent variable matrix $\mathbf{Y}(n\times q)$ is treated as missing and imputed from its marginal truncated distribution. The vector γ is sampled by using a Metropolis algorithm as done in Brown et al. [19,20]. The method visits a sequence of models that differ successively in one or two variables. At a generic step, given the previous visited vector, the algorithm randomly chooses among a set of transition moves. Typical moves add or delete one variable or swap two variables. We adopt the same inferential scheme in our wavelet regression model (3). As an additional feature, we also consider moves that specifically take into account possible correlation among adjacent variables. We therefore add moves that swap one variable with the neighbour by choosing independently at random a 1 and swapping its value with one of the two adjacent variables. This type of moves seems especially appropriate when considering the correlation structure of wavelet coefficients.

4. NIR spectra classification

4.1. Wheat variety data

We seek to classify three wheat varieties based on 100 near infra-red absorbances. The spectra are measured on samples of ungrounded wheat using a Tecator Infratec Grain Analyzer which measures transmission through the wheat sample of radiation at 100 wavelengths from 850 to 1048 nm in steps of 2 nm. Each wheat sample is classified into one of 3 named varieties, on the basis of known provenance. The data set consists of 117 samples of wheat. Fig. 1 shows the NIR spectra for the three varieties. The study aims at exploring the possibility of using NIR spectra to assign unknown samples to the correct variety. Dimension reduction induces robustness into the problem by virtue of concentrating on regions where clear predictive advantages occur and are less prone to contamination. Our methods find small sets of spectral features that lead to good classification results.

4.2. Classical analyses

The dataset is analyzed in Fearn et al. [21]. There the authors assume multivariate normal distributions for the predictors with a common variance within groups and achieve variable selection via a Bayesian decision theory approach that balances costs for variables against an error classification loss. The dataset was deliberately chosen because it is hard to predict grain type given the spectral data. Fearn et al. [21] model spectral data given grain type whereas we model grain type given spectral data. The authors randomly split the whole data set into 3 sets, a tuning set with (10, 7, 6) observations in the three classes, respectively, used to estimate some of the parameters of the model, a training set with (32, 22, 17) and a validation set with (10, 7, 6) observations.

Using the tuning set to estimate a loss function, Fearn et al. [21] found a best (in terms of their "cost" function) subset of six variables that leads to a misclassification rate of 5/23 in the validation set, that is 5 cases out of the 23 for validation are incorrectly classified. A slightly better error rate of 3/23 was found with a larger set that includes 12

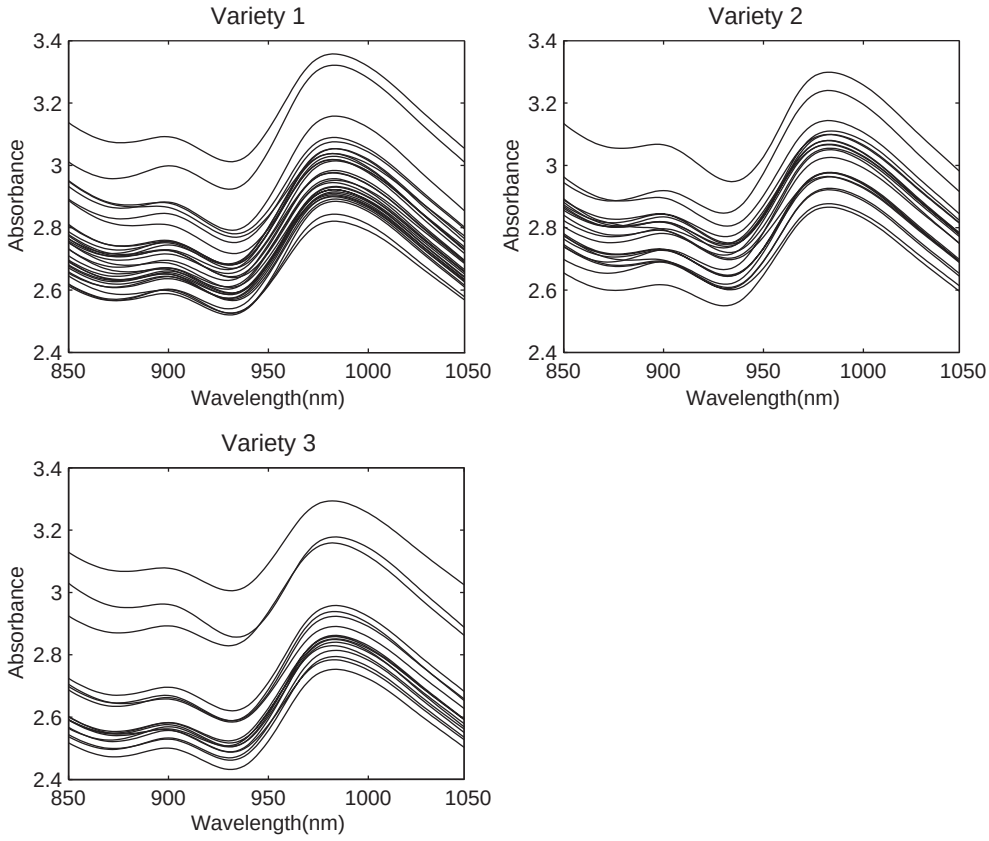

Fig. 1. Wheat data. NIR spectra for three wheat varieties.

variables. The authors also report on results they obtained with a more standard analysis where principal components are computed first using the covariance matrix of the training data and then a linear discriminant analysis retains the first s scores. For all choices of s in s=14, . . . ,18 LDA achieved an error rate of 4/23 in the validation set. In addition, we tried a quadratic discriminant analysis (QDA) and got a best misclassification rate of 7/23 with the first 8–10 components.

4.3. Wavelet component selection

For our Bayesian analysis we used the same training and validation sets as in Fearn et al. [21]. We applied a wavelet transform to each spectrum, using Daubechies wavelets with 4 vanishing moments. We set suitable vague priors on the intercept parameter vector α and on the error covariance $\sum$. Conjugate proper priors avoid identifiability problems. Also we specified the diagonal elements of the expected value of $\sum$ as equal to one, therefore imposing an identifiability constraint at a second stage of the model.

We used Bernoulli priors on γ with an additional Beta prior as previously described. Based on our experience with similar applications we chose to have an expectation of 10 variables and obtained a relatively vague specification by imposing the sum of the two parameters of the Beta to be equal to two. The prior distribution for the regression coefficients, given γ, depends on a covariance matrix $\mathbf{H}_\gamma$. Brown et al. [20] discuss relative merits and drawbacks of different specifications. Here we use $\mathbf{H}=c\mathrm{Diag}((\mathbf{D}'\mathbf{D})^{-})=c\mathrm{Diag}(\mathbf{W}'(\mathbf{X}'\mathbf{X})^{-}\mathbf{W})$, a diagonalized version of a full g-prior $\mathbf{H}=c(\mathbf{D}'\mathbf{D})^{-}$. The parameter c regulates the amount of shrinkage in the model (as does the prior on the number of non-zero regression coefficients). In practice sensitivity analyses should be performed. Generally speaking, we want to choose a value that leads to moderate shrinkage, therefore avoiding very small values, that would lead to too much regularization, as well as large values, that could induce nonlinear shrinkage as a result of the Lindley's paradox, Lindley [22]. In Sha et al. [11] some guidelines are provided on how to compute a suitable range of c values. Results here presented were obtained with c=.1.

We used four MCMC runs with 20,000 iterations after a burn-in of 5000 each. Starting γ vectors were with (i) 1, (ii) 20, (iii) 50 and (iv) 20 randomly selected coefficients included, respectively.

4.4. Results

From the MCMC output the missing value $\mathbf{Y}$ can be imputed using the mean of all sampled $\mathbf{Y}$'s and, conditional on this estimate, the normalized conditional posterior probabilities for all distinct γ's visited by the MCMC can

be computed. Marginal probabilities of inclusion of single variables can be obtained by averaging over the components of γ. Fig. 2 shows plots of these marginal probabilities for the four different MCMC runs. Spikes correspond to coefficients with high posterior probability that should therefore be important in distinguishing the samples. Notice how the plots are fairly similar, despite the different starting values. In fact, there appears to be a group of four wavelet coefficients, with indices around index 20 that have high posterior probability in all plots. We will refer to this group of four coefficients as "group A" in the sequel. The spike nearby index 40 (group B) that appears in one of the four plots corresponds to wavelet coefficients that belong to the second level of the transform and have locations corresponding to those of the coefficients in group A. Same applies to the spike nearby index 70 (group C) that shows up in one of the four plots. The wavelet coefficients of these three groups (A, B and C) are therefore describing the same features of the original spectral data, but at different resolution levels. An additional feature shows up in two of the plots, captured by the spike nearby index 50 in plot (*c*) (group D) and nearby index 30 in plot (*b*) (group E), again describing the same feature of the spectra but at two different resolution levels.

To locate regions of the spectra described by the selected wavelet coefficients we can exploit the linearity of the wavelet transforms and of the least squares prediction equation, in a similar manner to Brown et al. [9] for the regression model case. This allows us to compute the vectors of regression coefficients that will be applied to the original spectral data for future prediction. These vectors are obtained by applying the inverse wavelet transform to the least squares estimates of the regression coefficients obtained for a given model in the wavelet domain, i.e. to the columns of $\hat{B}^{(\gamma)}$ whose nonzero coefficients are calculated as $(\mathbf{D}_\gamma'\mathbf{D}_\gamma)^{-1}\mathbf{D}_\gamma'\mathbf{Y}$. Fig. 3 shows the regression coefficient vectors resulting by retaining only the four wavelet coefficients with high marginal probability in group A, plots (a) and (c), and the nine wavelet coefficients in groups A, B, C, D and E, plots (b) and (d). The range 900–970 nm is identified also by Fearn et al. [21] as the most well represented in all subsets they selected.

For inference we pooled together the distinct models visited by the four chains. We then looked at the predictive performances of the sets of discriminating variables obtained simply by considering the coefficients with marginal posterior probability greater than a certain value. Interesting sets can be also found by exploring the "best" models as those models visited by the different searches that have the highest values of the joint posterior probability. We found good agreement between variables with high marginal posterior probabilities and those selected by the best models. Performances of the selected models were assessed via prediction on the validation set.

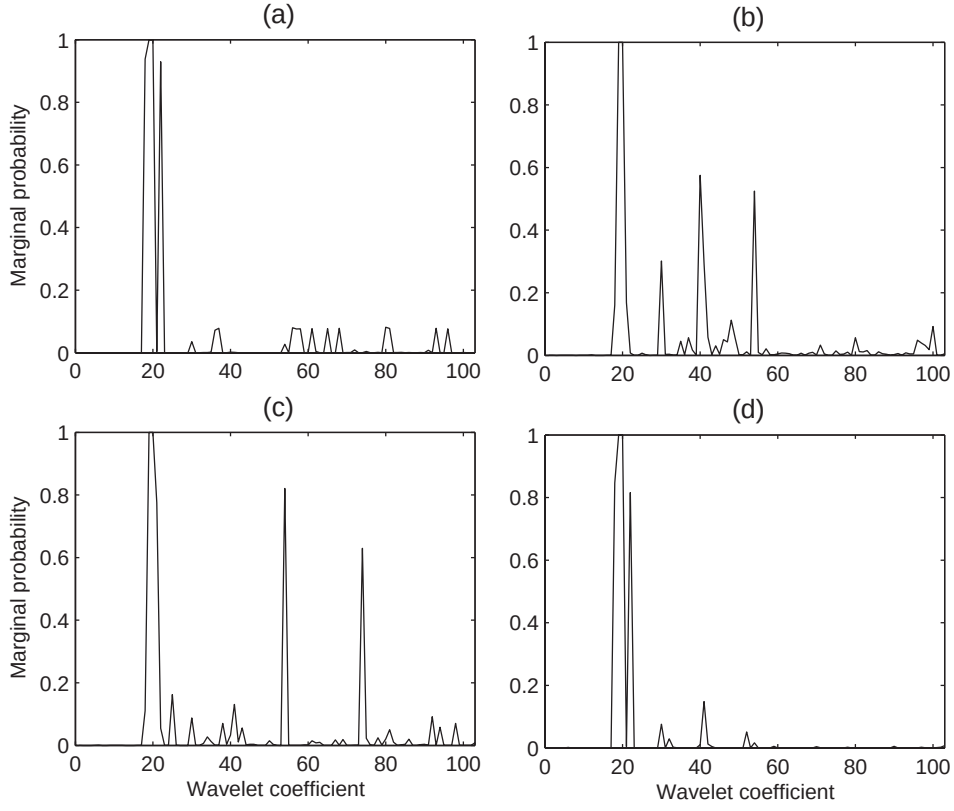

Fig. 2. Wheat data. Marginal posterior probabilities of single wavelet coefficients for 4 different MCMC runs.

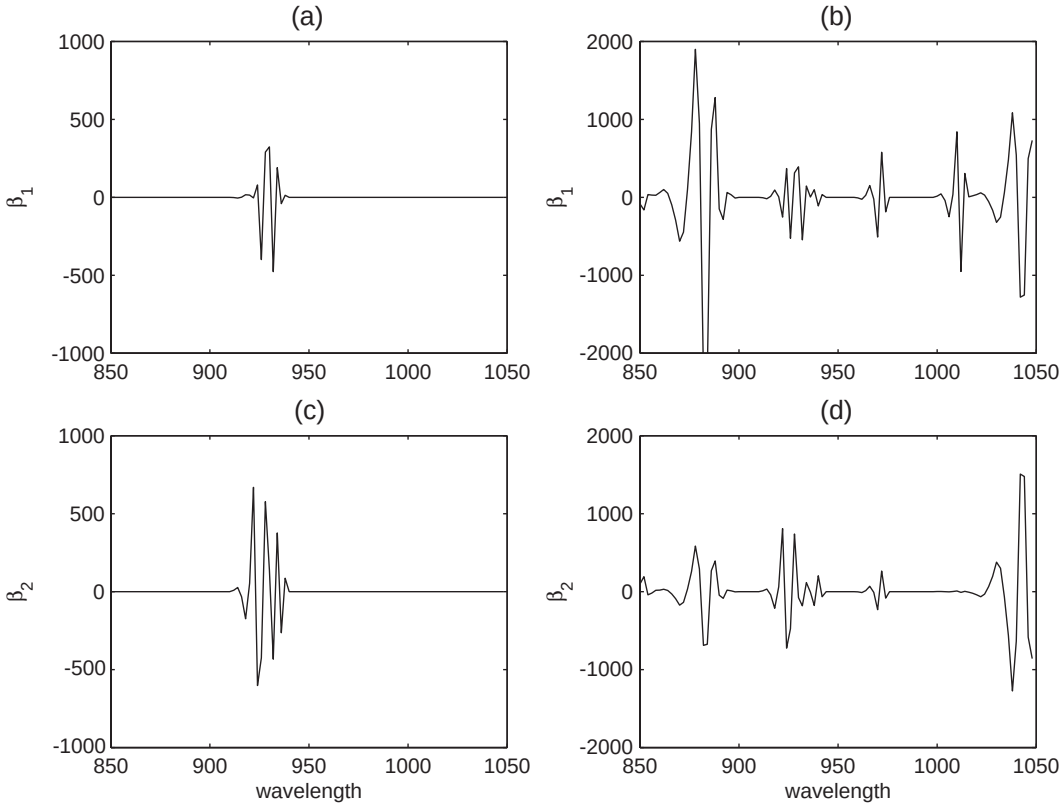

Fig. 3. Wheat data. Estimated regression coefficients vectors with 2 different wavelet models. Plots (a) and (c) are obtained using 4 wavelet coefficients and plots (b) and (d) using 9 coefficients.

We found single models LS predictions comparable with the best results of Fearn et al. [21], though obtained with smaller sets of variables. For example, we achieved a misclassification error of 4/23 when using the model with the four coefficients of group A and 3/23 when using the nine coefficients of groups A, B, C, D and E.

5. Mass spectra classification

5.1. Proteomic data

Recent advances in functional genomics have made it possible to measure the expression levels of thousands of genes or proteins. Proteomic methods include 2-d gel electrophoresis (2DE) and mass spectrometry methods like electrospray ionization (ESI-MS) and matrix-assisted laser desorption and ionization (MALDI-MS). Proteomic technology requires small amounts of biological material (tissue or blood samples) and it is often used for biomarker discovery, to identify proteins linked to disease status, response to therapy, or clinical prognosis. A mass spectrum is a curve where the x-axis indicates the ratio of the weight of a specific molecule to its electrical charge (m/z, in Daltons per unit charge) and the y-axis is the signal intensity for the same molecule as a measure of the abundance of that molecule in the sample. Proteomic spectra are characterized by many peaks, most of which correspond to proteins or protein fragments (peptides).

Here we analyse proteomic data from a recent study, see Petricoin et al. [3], where mass-spectra are used to detect ovarian cancer using blood serum samples. Data are available from http://home.ccr.cancer.gov/ncifdaproteomics. There are three different datasets and we use the most recent one dated 08-07-02, see Alexe et al. [23] for a description of all datasets and of various analyses. The dataset comprises 162 cancer samples and 91 control cases. Each mass-spectrum curve represents the expression profile of 15,154 peptides defined by their m/z ratios. Data are plotted in Fig. 4. Criticisms have been raised on the design of the Petricoin data, see Baggerly et al. [24]. Here we use these data with the only purpose of demonstrating the methodologies we propose.

5.2. Data pre-processing

Certain preprocessing steps must be performed before analyzing the spectra, including removal of baseline, noise elimination and normalization to calibrate the spectra from different samples. We performed baseline correction on all spectra by using a *loess* procedure, Cleveland et al. [25]. Noise removal and normalization was done instead on wavelet coefficients, by first interpolating the spectral data on a grid of equi-spaced

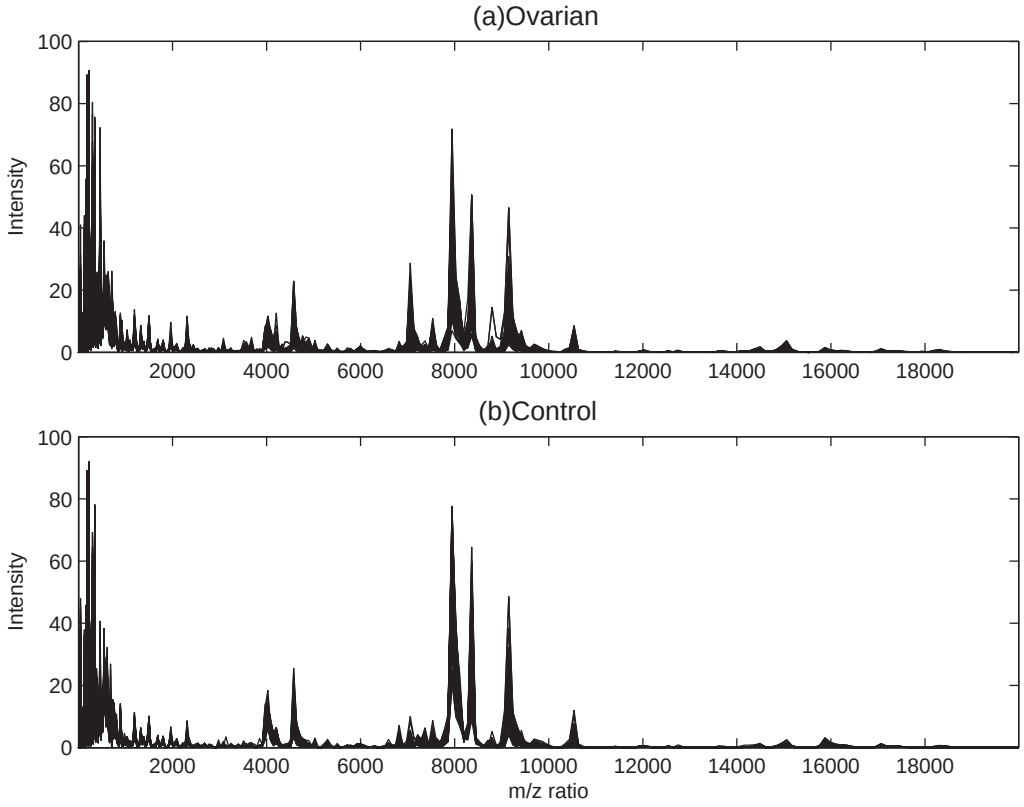

Fig. 4. Petricoin data. 253 mass-spectra with baseline subtracted.

m/z values, then transforming the spectra to wavelet coefficients and finally applying the wavelet shrinkage procedure suggested by Donoho and Johnstone [6,7]. Wavelet thresholding proves to be particularly advantageous with proteomic data, in that, as a result of the thresholding procedure, many wavelet coefficients will be zero, and can therefore be eliminated from further analyses. Here we applied to each spectrum the Donoho and Johnstone sureshrink wavelet shrinkage with adaptive threshold and Daubechies wavelets with 4 vanishing moments. Only around 900–1100 wavelet coefficients for each spectrum survived the thresholding step. When experimenting with different thresholding procedures we noticed that the universal threshold, a fixed values for all coefficients, would remove a lot more coefficients, but also attenuate some of the distinctive features (peaks) of the spectra. The coefficients that survived the thresholding were normalized by dividing each non-zero coefficient by the sum of the squares of all coefficients for any given spectrum.

5.3. Wavelet component selection

We modelled the coefficients that survived the thresholding with our probit model and used Bayesian variable selection methods to select those coefficients that discriminate the samples. We split the data into training and validation, leaving half of the data to assess the prediction performances of the selected models. Since the non-zeros wavelet coefficients that survived the thresholding procedure were slightly different from spectrum to spectrum, we considered as common set those that were non-zero in at least half of the samples for either classes (normal and/or ovarian) in the training set only. This resulted in 1001 coefficients to which we fit the probit model with variable selection.

We used Bernoulli priors on γ with an expectation of 10 coefficients ($w=10/p$). In order to widely explore the posterior space of possible models, we ran four parallel searches with 50,000 iterations and very different starting vectors. We discarded the first 10,000 iterations to eliminate dependence from the starting points. We used $c=1$. The training data were centered, the validation data were centered on the training means. See Section 4.3 for further remarks on such hyperparameter choices.

5.4. Results

Each search visited around 15,000–20,000 different models, after burn-in, with models with high posterior probabilities with about 8–10 coefficients. All best single models identified by the four chains achieved a sensitivity of 100% and a specificity of up 97% with an overall best misclassification rate of 1%. Fig. 5 shows the marginal posterior probabilities of inclusion of the single coefficients for the four chains.

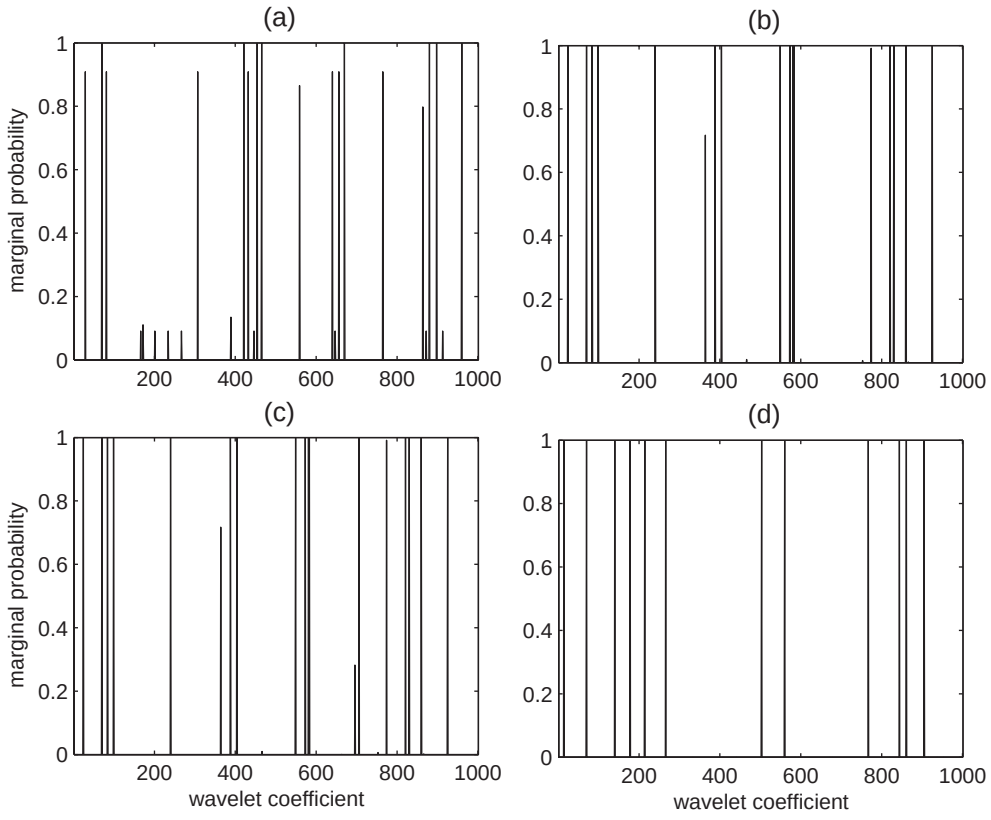

Fig. 5. Proteomic data. Marginal plots from 4 chains.

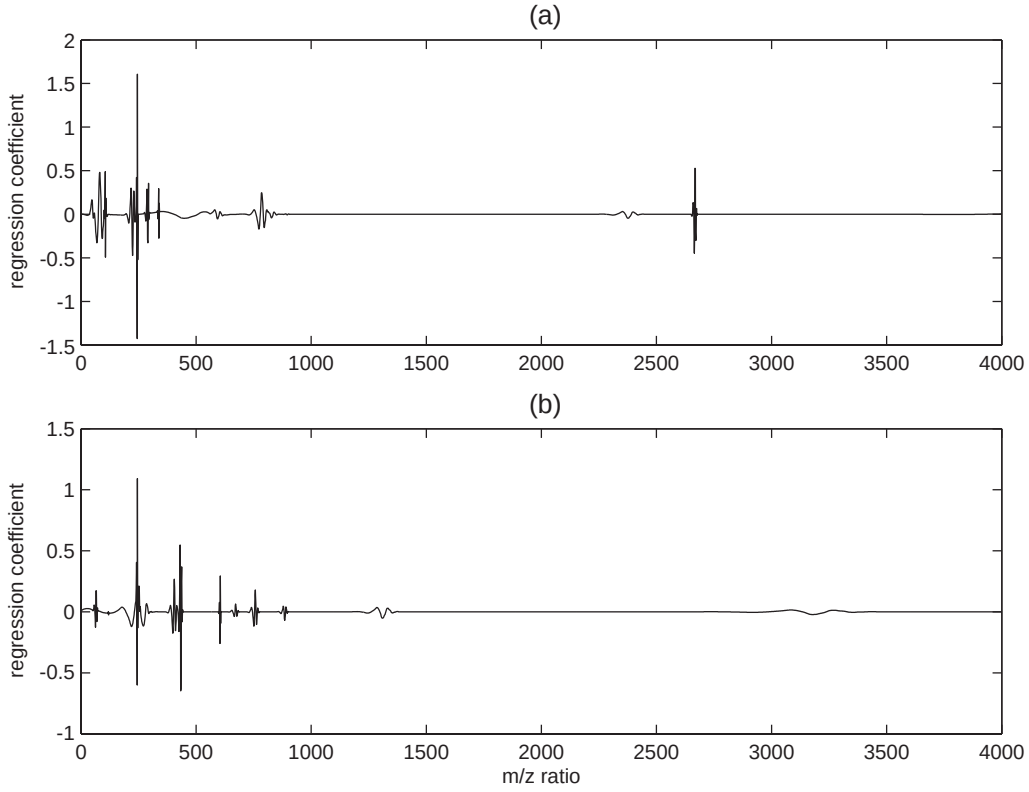

Fig. 6. Proteomic data. Regression coefficient vector estimates based on the "best" wavelet model, plot (a), and on "mean subset" over the best 500 wavelet models, plot (b).

For prediction we can again exploit the linearity of the wavelet transforms and of the least squares prediction equation to compute the vector of regression coefficients that will be applied to the original spectral data for future prediction. This provides useful information on the predictive features of the mass spectra. The regression coefficient vector, versus the m/z values, computed based on the 18 wavelet coefficients of the overall best model, i.e. the wavelet model with the highest marginal probability among all models visited by the four chains, is displayed in Fig. 6(a). The plot shows only the first 3000 regression coefficients, to allow a better view of the m/z range of interest. The other coefficients were essentially zero. In addition, we looked into model averaging estimates by computing a "mean subset" coefficient vector estimate, see Ojelund et al. [26], as $\hat{\beta}_{\mathrm{MS}}=\sum_\gamma w_\gamma \hat{\beta}^{(\gamma)}$ where the normalized weights are computed as $w_\gamma=\mathrm{SS}_\gamma^{-n/2}$ with $\mathrm{SS}_\gamma=\mathbf{Y}'\mathbf{Y}-\mathbf{Y}'\mathbf{D}_\gamma(\mathbf{D}_\gamma'\mathbf{D}_\gamma)^{-1}\mathbf{D}_\gamma'\mathbf{Y}$ the residual sum of squares, and where the nonzero coefficients of $\hat{\beta}^{(\gamma)}$ are computed as $(\mathbf{D}_\gamma'\mathbf{D}_\gamma)^{-1}\mathbf{D}_\gamma'\mathbf{Y}$. The vector estimate obtained using the best 500 wavelet models identified by the MCMC chains is displayed in Fig. 6(b).

In Alexe et al. [23] the authors identify a set of 9 peptides using an optimization-based procedure of logical analysis of data (LDA) that provided up to 100% sensitivities and specificities using cross-validation predictions. 8 out of the 9 peptides selected have m/z values in the range 200–700, the 9th peptide has m/z value 4004.8. Petricoin et al. [3] point out that restricting the range of the m/z values to the interval 235–500 does not affect accuracy of the predictions. Our results appear to be fairly coherent with these findings. For example, the regression coefficient with largest magnitude in the plots of Fig. 6 is the one with index 184 and corresponds to the m/z value 245.3, which is one of the 9 peptides identified by Alexe et al. [23]. The peptides that correspond to the largest nine coefficients of Fig. 6(b) are (245.3, 433.2, 434.6, 243.9, 430.6, 241.3, 437.2, 605.2, 431.9).

6. Summary

We have presented a wavelet-based method for classification based on functional data that uses probit models with latent variables and Bayesian mixture priors for variable selection. We have applied the method to the classification of three wheat varieties based on 100 near infra-red absorbances and to ovarian cancer discrimination based on mass-spectra. In the applications we have employed wavelet transforms as a tool for dimension reduction and noise removal, reducing spectra to wavelet components. In the examples our method has been able to identify small sets of coefficients that capture the discriminatory information of the spectral data.

In future work we plan to employ alternative transformations of the data, such as translation invariant wavelet transforms, that better preserve the alignment of the spectral features in the wavelet domain. Also, different types of thresholding techniques, such as block shrinkage methods, Cai [27], will be investigated. For the proteomics application we notice that, although Alexe et al. [23] perform their analyses on the entire range of m/z values, it has been suggested that low m/z values may be unreliable with the current technology, because they can be influenced by the chemicals used to ionize the proteins. We plan on looking at the effect on prediction results when different m/z ranges are considered.

Acknowledgements

We thank Tom Fearn, University College, London, for providing the wheat variety dataset. Vannucci's research is supported by National Science Foundation, CAREER award number DMS-0093208. Sha's research is supported by BBRC/RCMI NIH grant 2G12RR08124 and a University Research Institute (UTEP) grant.

References

[1] B.G. Osborne, T. Fearn, P.H. Hindle, Practical NIR Spectroscopy, Longman, Harlow, U.K, 1993.
[2] W.J. Krzanowski, P. Jonathan, W.V. McCarthy, M.R. Thomas, Discriminant analysis with singular covariance matrices: methods and applications to spectroscopic data, Applied Statistics 44 (1995) 105–115.
[3] E.F. Petricoin, A.M. Ardekani, B.A. Hitt, P.J. Levine, V.A. Fusaro, S.M. Steinberg, G.B. Mills, C. Simone, D.A. Fishman, E.C. Kohn, L.A. Liotta, Use of proteomic patterns in serum to identify ovarian cancer, Lancet 359 (2002) 572–577.
[4] B. Wu, T. Abbott, D. Fishman, W. McCurray, G. Mor, K. Stone, D. Ward, K. Williams, H. Zhao, Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data, Bioinformatics 19 (2003) 1636–1643.
[5] Y. Qu, B.L. Adam, M. Thornquist, J.D. Potter, M.L. Thompson, Y. Yasui, J. Davis, P.F. Schellhammer, L. Cazares, M.A. Clements, G.L. Wright Jr., Z. Feng, Data reduction using a discrete wavelet transform in discriminant analysis of very high dimensionality data, Biometrics 59 (2003) 143–151.
[6] D. Donoho, I. Johnstone, Ideal spatial adaption via wavelet shrinkage, Biometrika 81 (1994) 425–455.
[7] D. Donoho, I. Johnstone, Minimax estimation via wavelet shrinkage, Annals of Statistics 26 (3) (1998) 879–921.
[8] W. Chang, S. Kim, B. Vidakovic, Wavelet-based estimation of a discriminant function, Applied Stochastic Models in Business and Industry 19 (2003) 185–198.
[9] P.J. Brown, T. Fearn, M. Vannucci, Multivariate Bayesian variable selection and prediction, Journal of the American Statistical Association 96 (2001) 398–408.
[10] J.S. Morris, M. Vannucci, P.J. Brown, R.J. Carroll, Wavelet-based nonparametric modeling of hierarchical functions in colon carcinogenesis (with discussion), Journal of the American Statistical Association 98 (2003) 573–597.
[11] N. Sha, M. Vannucci, M.G. Tadesse, P.J. Brown, I. Dragoni, N. Davies, T.C. Roberts, A. Contestabile, N. Salmon, C. Buckley, F. Falciani, Bayesian variable selection in multinomial probit models to identify molecular signatures of disease stage, Biometrics 60 (3) (2004) 812–819.

[12] S.G. Mallat, Multiresolution approximations and wavelet orthonormal bases of $L_2(R)$, Transactions of the American Mathematical Society 315 (1) (1989) 69–87.
[13] I. Daubechies, Ten Lectures on Wavelets, in: SIAM, Conference Series, vol. 61, 1992.
[14] A. Antoniadis, J. Bigot, T. Sapatinas, Wavelet estimators in non-parametric regression: a comparative simulation study, Journal of Statistical Software 6 (2001) 1–83.
[15] M. Vannucci, P.J. Brown, T. Fearn, A decision theoretical approach to wavelet regression on curves with a high number of regressors, Journal of Statistical Planning and Inference 112 (1–2) (2003) 195–212.
[16] J.H. Albert, S. Chib, Bayesian analysis of binary and polychotomous response data, Journal of the American Statistical Association 88 (1993) 669–679.
[17] M. Vannucci, F. Corradi, Covariance structure of wavelet coefficients: theory and models in a Bayesian perspective, Journal of the Royal Statistical Society. Series B 61 (4) (1999) 971–986.
[18] P.J. Brown, M. Vannucci, T. Fearn, Multivariate Bayesian variable selection and prediction, Journal of the Royal Statistical Society. Series B 60 (3) (1998) 627–641.
[19] P.J. Brown, M. Vannucci, T. Fearn, Bayesian wavelength selection in multicomponent analysis, Journal of Chemometrics 12 (1998) 173–182.
[20] P.J. Brown, M. Vannucci, T. Fearn, Bayes model averaging with selection of regressors, Journal of the Royal Statistical Society. Series B 64 (3) (2002) 519–536.
[21] T. Fearn, P.J. Brown, P. Besbeas, A Bayesian decision theory approach to variable selection for discrimination, Statistics and Computing 12 (3) (2002) 253–260.
[22] D.V. Lindley, A statistical paradox, Biometrika 44 (1957) 187–192.
[23] G. Alexe, S. Alexe, L.A. Liotta, E. Petricoin, M. Reiss, P.L. Hammer, Ovarian cancer detection by logical analysis of proteomic data, Proteomics 4 (2004) 766–783.
[24] K.A. Baggerly, J.S. Morris, K.R. Coombes, Reproducibility of SELDI-TOF protein patterns in serum: comparing data sets from different experiments, Bioinformatics 20 (2004) 777–785.
[25] R.B. Cleveland, W.S. Cleveland, J.E. McRae, I. Terpenning, STL: a seasonal trend decomposition procedure based on LOESS (with discussion), Journal of Official Statistics 6 (1990) 3–73.
[26] H. Ojelund, P.J. Brown, H. Madsen, P. Thyregod, Prediction based on mean subset, Technometrics 44 (2002) 369–378.
[27] T.T. Cai, Adaptove wavelet estimation: a block thresholding and oracle inequality approach, Annals of Statistics 27 (1999) 898–924.