# Bayesian Wavelet Regression on Curves with Application to a Spectroscopic Calibration Problem

P. J. Brown[*], T. Fearn[†] and M. Vannucci[‡]

November 16, 2000

[*] University of Kent at Canterbury, Institute of Mathematics and Statistics, Cornwallis Building, Canterbury, CT2 7NF, UK

[†] Department of Statistical Science, University College London, Gower Street, London WC1E 6BT, UK

[‡] Department of Statistics, Texas A&M University, College Station, Texas 77843-3143, USA

## Author Footnote

P.J. Brown is Pfizer Professor of Medical Statistics, Institute of Mathematics and Statistics, University of Kent, Canterbury, Kent, CT2 7NF, U.K. (E-mail:philip.j.brown@ukc.ac.uk, URL www.ukc.ac.uk/ IMS/statistics/people/P.J.Brown). T. Fearn is Professor of Applied Statistics, Department of Statistical Science, University College London, Gower Street, London WC1E 6BT, U.K. (E-mail tom@stats.ucl.ac.uk). M. Vannucci is Assistant Professor of Statistics, Department of Statistics, Texas A&M University, College Station, Texas 77843-3143, USA (E-mail:mvannucci@stat.tamu.edu, URL http://stat.tamu.edu/people/faculty/mvannucci.html/).
This work was supported by the UK Engineering and Physical Sciences Research Council with a grant under the Stochastic Modelling in Science and Technology Initiative, Grant number GK/K73343. Vannucci also acknowledges support from the Texas Higher Advanced Research Program, grant number 010366-0075, from Texas A&M International Research Travel Assistant Program and from NSF, CAREER award number DMS-0093208. The spectroscopic calibration problem was provided by the Flour Milling and Baking Research Association. The authors wish to thank the associate editor, and a referee, as well as Mike West, Duke University, and Adrian Raftery, University of Washington, for suggestions that helped to improve the paper.

## Abstract

Motivated by calibration problems in near infrared spectroscopy, we consider the linear regression setting where the many predictor variables arise from sampling an essentially continuous curve at equally spaced points, and where there may be multiple predictands. We tackle this regression problem by calculating the wavelet transforms of the discretized curves, and then applying a Bayesian variable selection method using mixture priors to the multivariate regression of predictands on wavelet coefficients. For prediction purposes, we average over a set of likely models. Applied to a particular problem in near infrared spectroscopy this approach was able to find subsets of the wavelet coefficients with overall better predictive performance than the more usual approaches.

In the application the predictors available are measurements of the near infrared reflectance spectrum of biscuit dough pieces at 256 equally spaced wavelengths. The aim is to predict the composition, ie the fat, flour, sugar and water content, of the dough pieces using the spectral variables. Thus we have a multivariate regression of 4 predictands on 256 predictors with quite high intercorrelation amongst the predictors. A training set of 39 samples is available to fit this regression. Applying a wavelet transform replaces the 256 measurements on each spectrum with 256 wavelet coefficients that carry the same information. The variable selection method was able to use subsets of these coefficients that gave good predictions for all four compositional variables on a separate test set of samples. Selecting in the wavelet domain rather than from the original spectral variables is appealing in this application because a single wavelet coefficient can carry information from a band of wavelengths in the original spectrum; a band that can be narrow or wide depending on the scale of the wavelet selected.



**KEY WORDS:** Markov chain Monte Carlo; Mixture prior; Model averaging; Multivariate regression; NIR spectroscopy; Variable selection.

# 1 INTRODUCTION

This paper presents a new way of tackling linear regression problems where the predictor variables arise from sampling an essentially continuous curve at equally spaced points. The work was motivated by calibration problems in near infrared spectroscopy, of which the following example is typical.

## 1.1 NIR Spectroscopy of Biscuit Doughs

Quantitative near infrared (NIR) spectroscopy is used in areas as diverse as the analysis of food and drink, pharmaceutical products and petrochemicals. The NIR spectrum of a sample of, eg, wheat flour is a continuous curve that is measured by modern scanning instruments at hundreds of equally spaced wavelengths. The information in this curve can be used to predict the chemical composition of the sample. The problem lies in extracting the relevant information from possibly thousands of overlapping peaks. Osborne et al. (1993) describe applications in food analysis and review some of the standard approaches to the calibration problem.

The example studied in detail here arises from an experiment to test the feasibility of using NIR spectroscopy for measuring the composition of biscuit dough pieces, for example formed but unbaked biscuits, for possible on-line implementation. For a full description of the experiment see Osborne et al. (1984). Briefly, two similar sample sets were made up, with the standard recipe varied to provide a large range for each of the four constituents under investigation: fat, sucrose, dry flour and water. The calculated percentages of these four ingredients represent the $q = 4$ responses. There were $n = 39$ samples in the calibration or training set, sample number 23 having been excluded from the original 40 as an outlier, and a further $m = 39$ in the separate prediction or validation set, again after the exclusion of one outlier. Thus $Y$ and $Y_f$, the matrices of compositional data for the training and validation sets respectively, are both of dimension $39 \times 4$.

An NIR reflectance spectrum is available for each dough piece. The original spectral data comprise 700 points measured from 1100 to 2498 nanometers (nm) in steps of 2nm. For our analyses using wavelets we have chosen to reduce the number of spectral points to save on computational time. The first 140 and last 49 wavelengths, which were thought to contain little useful information, were removed leaving a wavelength range from 1380nm to 2400nm over which we took every other point, so increasing the gap to 4nm and reducing the number of points to $p = 256$. The matrices $X$ and $X_f$ of spectral data are then $39 \times 256$. A sample of 3 centered spectra is presented in the left column of Figure 1.

The aim is to derive an equation that will predict the response values $Y$ from the spectral data $X$ for future samples where $Y$ is unknown but $X$ can be measured cheaply and rapidly.

## 1.2 Standard Analyses

The most commonly used approaches to this calibration problem regress $Y$ on $X$, with the linear form

$$Y = XB + E$$

being justified either by appeals to the Beer-Lambert law (Osborne et al. 1993) or on the grounds that it works in practice. In Section 6 we investigate also logistic transformations of the responses, showing that overall their impact on the prediction performance of the model is not beneficial.

The problem is not straightforward because there are many more predictor variables than training samples, 256 and 39 respectively in our example. The methods most commonly used for overcoming this difficulty fall into two broad classes: variable selection and factor-based methods. When scanning NIR instruments first appeared the standard approach was to select, typically using a stepwise procedure, predictors at a small number of wavelengths and use multiple linear regression with this subset (Hrushka 1987). Later this approach was largely superseded by methods that reduce the $p$ spectral variables to scores on a much smaller number of factors and regress on these scores. Two variants, principal components regression (PCR) (Cowe and McNicol 1985) and partial least squares regression (PLS) (Wold et al. 1983) are both now widely used, with equal effectiveness, as the standard approaches. As computers have become more powerful there has been a renewal of research interest in wavelength selection, now using computer-intensive search methods.

## 1.3 Selecting wavelet coefficients

The approach we investigate here involves selection of variables, but we select from derived variables. The idea is to transform each spectrum into a set of wavelet coefficients, the whole of which would suffice to reconstruct the spectrum, and select good predictors from amongst these. There are good reasons for thinking that this approach might have advantages over selecting from the original variables.

In previous work (Brown et al. 1998a,b, 1999) we have explored Bayesian approaches to the problem of selecting predictor variables in this multivariate regression context, and we shall apply this methodology to wavelet selection here.

We know that wavelets can be used successfully for compression of curves like the spectra in our example, in the sense that the curves can be accurately reconstructed from a fraction of the full set of wavelet coefficients (Trygg and Wold 1998, Walczak and Massart 1997). Further, the wavelet decomposition of the curve is a local one, so that if the information relevant to our prediction problem is contained in a particular part or parts of the curve, as typically it is, this information will be carried by a very small number of wavelet coefficients. Thus we may expect selection to work. The ability of wavelets to model the curve at different levels of resolution means that we have the option of selecting from our curve at a range of bandwidths. In some situations it may be advantageous to select a sharp band, as we do when we select one of the original variables; in others a broad band, averaging over many adjacent points, may be preferable. Selecting from the wavelet coefficients gives us both of these options automatically and in a very computationally efficient framework.

Not surprisingly, Fourier analysis has also been successfully used for data compression and denoising of NIR spectral data (McClure et al. 1984). For this purpose, there is probably very little to choose between Fourier and wavelet approaches. When it comes to selecting small numbers of coefficients for prediction however, the local nature of wavelets makes them the obvious choice.

It is worth emphasising that what we are doing here is not what is commonly described as wavelet regression. We are not fitting a smooth function to a single noisy curve, either by using thresholding or shrinkage of wavelet coefficients, see Donoho et al. (1995), and Clyde and George (2000) and the references therein for Bayesian approaches that use mixture modelling. Unlike these authors we have several curves, the spectra of 39 dough pieces, each of which is transformed to wavelet coefficients. We then select some of these wavelet coefficients (the same ones for each spectrum), not because they give a good reconstruction of the curves, which they do not, or to remove noise, of which there is very little to start with, from the curves, but because the selected coefficients are useful for predicting some other quantity measured on the dough pieces. One consequence of this is that it is not necessarily the large wavelet coefficients that will be useful, small coefficients in critical regions of the spectrum may carry important predictive information. Thus the standard thresholding or shrinkage approaches are just not relevant to this problem.

# 2 PRELIMINARIES

## 2.1 Wavelet Bases and Wavelet Transforms

Wavelets are families of functions that can accurately describe other functions in a parsimonious way. In $L^2(\mathbb{R})$, for example, an orthonormal wavelet basis is obtained as translations and dilations of a *mother wavelet* $\psi$ as $\psi_{j,k}(x) = 2^{j/2}\psi(2^j x - k)$ with $j, k$ integers. A function $f$ is then represented by a wavelet series as

$$f(x) = \sum_{j,k \in \mathbb{Z}} d_{j,k}\psi_{j,k}(x), \tag{1}$$

with wavelet coefficients $d_{j,k} = \int f(x)\psi_{j,k}(x)dx$ describing features of the function $f$ at the spatial location $2^{-j}k$ and frequency proportional to $2^j$ (or scale $j$).

Daubechies (1992) proposed a class of wavelet families which have compact support and maximum number of vanishing moments for any given smoothness. They are extensively used in statistical applications.

Wavelets have been extremely successful as a tool for the analysis and synthesis of discrete data. Let $Y = (y_1, \ldots, y_n)$, $n = 2^J$, be a sample of a function at equally spaced points. This vector of observations can be viewed as an approximation to the function at the fine scale $J$. A fast algorithm exists for decomposing $Y$ into a set of wavelet coefficients (Mallat 1989) in only $O(n)$ operations. This algorithm is known as the *Discrete Wavelet Transform* (DWT). It operates in practice by means of linear recursive filters. For illustration purposes we can write it in matrix form as $Z = WY$ with $W$ an orthogonal matrix corresponding to the discrete wavelet transform and $Z$ a vector of wavelet coefficients describing features of the function at scales from the fine $J-1$ to a coarser one, say $J-r$. An algorithm for the inverse construction also exists.

Wavelet transforms are extremely fast to compute and have good compression properties. Because they are localised in both time and frequency wavelets have the ability to represent many classes of functions in a sparse form by describing important features with few coefficients.

For a general exposition of wavelet theory see Daubechies (1992).

## 2.2 Matrix-Variate Distributions

In what follows we shall use the notation for matrix-variate distributions due to Dawid (1981). We write

$$V - M \sim \mathcal{N}(\Gamma, \Sigma)$$

when the random matrix $V$ has a matrix-variate normal distribution with mean $M$ and covariance matrices $\gamma_{ii}\Sigma$ and $\sigma_{jj}\Gamma$ for its $i$-th row and $j$-th column, respectively. Such a $V$ could be generated as $V = M + A'UB$ where $M, A, B$ are fixed matrices such that $A'A = \Gamma$ and $B'B = \Sigma$ and $U$ is a random matrix having independent standard normal entries. This notation has the advantage of preserving the matrix structure instead of reshaping $V$ as a vector. It also makes for much easier formal Bayesian manipulation.

The other notation we shall use is

$$W \sim \mathcal{IW}(\delta; \Sigma)$$

for a random matrix $W$ having an inverse Wishart distribution with scale matrix

$\Sigma$ and shape parameter $\delta$. The shape parameter differs from the more conventional degrees of freedom, again making for very easy Bayesian manipulations. With $U$ and $B$ defined as above, and if $U$ is $n \times p$ with $n > p$, then $W = B'(U'U)^{-1}B$ has an inverse Wishart distribution with shape parameter $\delta = n - p + 1$ and scale matrix $\Sigma$. The expectation of $W$ exists for $\delta > 2$ and is then $\Sigma/(\delta - 2)$.

More details of these notations, and a corresponding form for the matrix-variate $T$, can be found in Dawid (1981) or Brown (1993, Appendix A).

# 3 MODELLING

## 3.1 Multivariate Regression Model

The basic setup we consider is a multivariate linear regression model, with $n$ observations on a $q$-variate response and $p$ explanatory variables. Let $Y$ denote the $n \times q$ matrix of observed values of the responses and let $X$ be the $n \times p$ matrix of predictor variables. Our special concern is with functional predictor data, ie the situation where each row of $X$ is a vector of observations of a curve $x(t)$ at $p$ equally spaced points.

The standard multivariate normal regression model has, conditional on $\alpha, B, \Sigma$ and $X$

$$Y - 1_n\alpha' - XB \sim \mathcal{N}(I_n, \Sigma), \tag{2}$$

with $1_n$ an $n \times 1$ vector of ones, $\alpha$ a $q \times 1$ vector of intercepts and $B = (\beta_1, \ldots, \beta_q)$ a $p \times q$ matrix of regression coefficients. Without loss of generality we assume that the columns of $X$ have been centered by subtracting their means.

The unknown parameters are $\alpha$, $B$ and the $q \times q$ error covariance matrix $\Sigma$. A conjugate prior for this model is as follows: first, given $\Sigma$,

$$\alpha' - \alpha_0' \sim \mathcal{N}(h, \Sigma), \tag{3}$$

and, independently,

$$B - B_0 \sim \mathcal{N}(H, \Sigma). \tag{4}$$

The marginal distribution of $\Sigma$ is then

$$\Sigma \sim \mathcal{IW}(\delta; Q). \tag{5}$$

Notice that the priors on both $\alpha$ and $B$ have covariances dependent on $\Sigma$ in a way that directly extends the univariate regression natural conjugate prior distributions.

In practice we will let $h \to \infty$ to represent vague prior knowledge about $\alpha$, and we will take $B_0 = 0$, leaving the specification of $H$, $\delta$ and $Q$ to incorporate prior knowledge about our particular application.

## 3.2 Transformation to Wavelets

We now transform the predictor variables by applying to each row of $X$ a wavelet transform, as described in Section 2.1. In matrix form, multiplying each row of $X$ by the same matrix $W$ is equivalent to multiplying $X$ on the right hand side by $W'$.

The wavelet transform is orthogonal, ie $W'W = I$, and therefore model (2) can be written as

$$Y - 1_n\alpha' - XW'WB \sim \mathcal{N}(I_n, \Sigma). \tag{6}$$

We can now express the model in terms of wavelet transformations of the predictors as

$$Y - 1_n\alpha' - Z\tilde{B} \sim \mathcal{N}(I_n, \Sigma) \tag{7}$$

where $Z = XW'$ is now a matrix of wavelet coefficients and $\tilde{B} = WB$ a matrix of regression coefficients. The transformed prior on $\tilde{B}$, in the case of inclusion of all predictors, is

$$\tilde{B} \sim \mathcal{N}(\tilde{H}, \Sigma), \tag{8}$$

where $\tilde{H} = WHW'$ and the parameters $\alpha$ and $\Sigma$ are unchanged by the orthogonal transformations, as are the priors (3) and (5).

In practice, wavelets exploit the recursive application of filters, and the $W$-matrix notation is more useful for explanation than for computation. Vannucci and Corradi

(1999) have proposed a fast recursive algorithm for computations of quantities such as $WHW'$. Their algorithm has a useful link to the two-dimensional discrete wavelet transform (DWT2), making computations simple. The matrix $WHW'$ can be computed from $H$ with an $O(n^2)$ algorithm. For more details see Sections 3.1 and 3.2 of Vannucci and Corradi (1999).

### 3.3 A Framework for Variable Selection

In order to do selection in the wavelet coefficient domain, we further elaborate the prior on $\tilde{B}$ by introducing a latent binary $p$-vector $\gamma$. The $j$th element of $\gamma$, $\gamma_j$, may be either 1 or 0 according to whether the $j$th column of $Z$ is included or not in the model. When $\gamma_j$ is unity then the covariance matrix of the corresponding row of $\tilde{B}$ is 'large', when $\gamma_j$ is zero then the covariance matrix is a zero matrix. We have assumed that the prior expectation of $\tilde{B}$ is zero, and so $\gamma_j = 0$ effectively deletes the $j$th explanatory variable (or wavelet coefficient) from the model. This gives, conditional on $\gamma$,

$$\tilde{B}_\gamma \sim \mathcal{N}(\tilde{H}_\gamma, \Sigma), \tag{9}$$

where $\tilde{B}_\gamma$ and $\tilde{H}_\gamma$ are just $\tilde{B}$ and $\tilde{H}$ with the rows and, in the case of $\tilde{H}$, columns for which $\gamma = 0$ deleted. Under this prior, each row of $\tilde{B}$ is modeled as having a scale mixture of the type

$$[\tilde{B}]_{[j:]} \sim (1-\gamma_j)I_0 + \gamma_j N(0, \tilde{h}_{jj}\Sigma) \tag{10}$$

with $\tilde{h}_{jj}$ equal to the $j$th diagonal element of the matrix $\tilde{H} = WHW'$ and $I_0$ a distribution placing unit mass on the $1 \times q$ zero vector. Notice that rows of $\tilde{B}$ are not independent.

A simple prior distribution $\pi(\gamma)$ for $\gamma$ takes the $\gamma_j$ to be independent with $\text{Prob}(\gamma_j = 1) = w_j$, $\text{Prob}(\gamma_j = 0) = 1 - w_j$, with hyperparameters $w_j$ to be specified, for $j = 1, \ldots, p$. In our example we shall take all the $w_j$ equal to a common $w$, so that the number of nonzero elements of $\gamma$ will have a binomial distribution with expectation $pw$.

Mixture priors have been widely used for variable selection in the original model space, originally by Leamer (1978) and more recently by Mitchell and Beauchamp (1988) and George and McCulloch (1997) for the linear multiple regression case. Geweke (1996), Carlin and Chib (1995) and Chipman (1996), among others, concentrated on special features of these priors. Clyde et al. (1996) used model mixing in prediction problems with correlated predictors when expressing the space of models in terms of an orthogonalization of the design matrix. Their methods are not directly applicable to our situation because the wavelet transforms do not leave us with an orthogonal design. The use of mixture priors for selection in the multivariate regression setup has been investigated by Brown et al. (1998a,b).

# 4 SELECTING WAVELET COEFFICIENTS

## 4.1 Posterior Distribution of $\gamma$

The posterior distribution of $\gamma$ given the data, $\pi(\gamma|Y,Z)$, gives a posterior probability to each $\gamma$ vector and thus to each possible subset of predictors (wavelet coefficients). This posterior arises from the combination of a likelihood that gives high weight to subsets explaining a high proportion of the variation in the responses $Y$ and a prior for $\gamma$ that penalises large subsets. It can be computed by integrating out $\alpha, B, \Sigma$ from the joint posterior distribution of these parameters and $\gamma$ given the data. With the vague ($h \to \infty$) prior for $\alpha$, this parameter is essentially estimated by the mean $Y$ in the calibration data, see Smith (1973), and to simplify the formulae that follow we now assume that the columns of $Y$ have been centered. Full details of the prior to posterior analysis are given in Brown et al. (1998b), which also considers other prior structures. After some manipulation we have

$$\pi(\gamma|Y,Z) \propto g(\gamma) = \{|\tilde{H}_\gamma||Z'_\gamma Z_\gamma + \tilde{H}_\gamma^{-1}|\}^{-q/2}|Q_\gamma|^{-(n+\delta+q-1)/2}\pi(\gamma) \qquad (11)$$

with $Q_\gamma = Q + Y'Y - Y'Z_\gamma(Z'_\gamma Z_\gamma + \tilde{H}_\gamma^{-1})^{-1}Z'_\gamma Y$, and $Z_\gamma$ is $Z$ with the columns for which $\gamma = 0$ deleted. Care is needed in computing (11), and the alternative forms below are useful.

A simplifying feature of this setup is that all the computations can be formulated as least squares problems with modified $Y$ and $Z$ matrices. By writing

$$\tilde{Z}_\gamma = \begin{pmatrix} Z_\gamma \tilde{H}_\gamma^{\frac{1}{2}} \\ I_{p_\gamma} \end{pmatrix}, \quad \tilde{Y} = \begin{pmatrix} Y \\ 0 \end{pmatrix},$$

where $\tilde{H}_\gamma^{\frac{1}{2}}$ is a matrix square root of $\tilde{H}_\gamma$ and $p_\gamma$ is the number of ones in $\gamma$, the relevant quantities entering into (11) can be computed as

$$|\tilde{H}_\gamma||Z'_\gamma Z_\gamma + \tilde{H}_\gamma^{-1}| = |\tilde{Z}'_\gamma \tilde{Z}_\gamma| \tag{12}$$

and

$$Q_\gamma = Q + \tilde{Y}'\tilde{Y} - \tilde{Y}'\tilde{Z}_\gamma(\tilde{Z}'_\gamma \tilde{Z}_\gamma)^{-1}\tilde{Z}'_\gamma \tilde{Y} \tag{13}$$

that is, $Q_\gamma$ is given by $Q$ plus the residual sum of products matrix from the least squares regression of $\tilde{Y}$ on $\tilde{Z}_\gamma$. The $QR$ decomposition can then be used, see for example Seber (1984, chapter 10, section 1.1b), which avoids 'squaring' as in (12), (13).

## 4.2 Metropolis Search

Equation (11) gives the posterior probability of each of the $2^p$ different $\gamma$ vectors, and thus of each choice of wavelet coefficient subsets. What remains to be done is to look for 'good' wavelet components by computing these posterior probabilities.

When $p$ is much greater than about 25 there are too many subsets for this to be feasible. Fortunately we can use simulation methods that will find the $\gamma$ vectors with relatively high posterior probabilities. One can then quickly identify useful coefficients which have high marginal probabilities of $\gamma_j = 1$.

Here we use a Metropolis search, as suggested for model selection by Madigan and York (1995) and applied to variable selection for regression by George and McCulloch (1997), Raftery et al. (1997) and Brown et al. (1998a).

The search starts from a randomly chosen $\gamma^0$ and then moves through a sequence of further values of $\gamma$. At each step the algorithm generates a new candidate $\gamma$ by randomly modifying the current one. Two types of move are used:

Add or delete a component by choosing at random one component in the current $\gamma$ and changing its value. This move is chosen with probability $\phi$.

Swap two components by choosing independently at random a 0 and a 1 in the current $\gamma$ and changing both of them. This move is chosen with probability $1 - \phi$.

The new candidate model, $\gamma^*$, is accepted with probability

$$\min\{\frac{g(\gamma^*)}{g(\gamma)}, 1\}. \tag{14}$$

Thus a more probable $\gamma^*$ is always accepted, a less probable one may be. There is scope for further ingenuity in designing the sequence of random moves. For example, moves that add or subtract or swap two or three or more at a time, or a combination of these, may be useful.

The sequence of $\gamma$'s generated by the search is a realisation of a Markov chain, and the choice of acceptance probabilities ensures that the equilibrium distribution of this chain is the distribution given by (11). In typical uses of such schemes the realizations are monitored after a suitable burn in period to see that they appear stationary. Here we have a closed form for the posterior distribution, and we are using the chain simply to explore this distribution. Thus we have not been so concerned about strict convergence of the Markov chain. As in Brown et al. (1998a) we adopt a strategy of running the chain from a number of different starting points, four here, and looking at the four marginal distributions provided by the computed $g(\cdot)$ values of the visited $\gamma$'s. We also look for good indication of mixing and explorations with returns. Because we know the relative probabilities we do not need to worry about using a burn-in period.

Notice that, given the form of the acceptance probability (14), $\gamma$ vectors with high posterior probability have more chance of appearing in the sequence, and we might expect that a long run of such a chain will therefore visit many of the best subsets.

# 5 PREDICTION

Suppose now that we wish to predict $Y_f$, an $m \times q$ matrix of further $Y$-vectors given the corresponding $X$- vectors, $X_f$, $(m \times p)$. First we treat $X_f$ exactly as the training data have been treated, by subtracting the training data means and transforming to wavelet coefficients $Z_f$, $(m \times p)$. The model for $Y_f$, following the model for the training data (6), is

$$Y_f - 1_m \alpha' - Z_f \tilde{B} \sim \mathcal{N}(I_m, \Sigma). \tag{15}$$

If we believe our Bayesian mixture model, then logically we should apply the same latent structure model to prediction as well as training. This has the practical appeal of providing averaging over a range of likely models, Madigan and Raftery (1994).

Results in Brown et al. (1998b) show that, with the columns of $Y_f$ centered using the mean $Y$ from the training set, the expectation of the predictive distribution $p(Y_f|\gamma, Z, Y)$ is given by $Z_f \hat{B}_\gamma$ with

$$\hat{B}_\gamma = (Z'_\gamma Z_\gamma + \tilde{H}_\gamma^{-1})^{-1} Z'_\gamma Y = \tilde{H}_\gamma^{\frac{1}{2}} (\tilde{Z}'_\gamma \tilde{Z}_\gamma)^{-1} \tilde{Z}'_\gamma \tilde{Y}. \tag{16}$$

Averaging over the posterior distribution of $\gamma$ gives

$$\hat{Y}_f = \sum_\gamma Z_f \hat{B}_\gamma \pi(\gamma|Z, Y), \tag{17}$$

and we might choose to approximate this by some restricted set of $\gamma$ values, perhaps the $r$ most likely values from the Metropolis search.

# 6 APPLICATION TO NIR SPECTROSCOPY OF BISCUIT DOUGHS

We now apply the methodology developed above to the spectroscopic calibration problem described in 1.1. First, however, we report the results of some other analyses of these data.

## 6.1 Analysis by Standard Methods

For all the analyses carried out here, both compositional and spectral data were centered by subtraction of the training set means from the training and the validation data. The responses, but not the spectral data, were also scaled, to give each of the four variables unit variance in the training set.

Mean squares prediction errors were converted back to the original scale by multiplying by the training sample variances. The pre-processing of the data makes no difference to the standard analyses, which treat the response variables separately, but simplifies the prior specification for our multivariate wavelet analysis.

Osborne et al. (1984) derived calibrations by multiple regression, using various stepwise procedures to select wavelengths for each constituent separately. The mean square error of predictions on the 39 validation samples for their calibrations are reported in the first row of Table 1.

Brown et al. (1999) also selected small numbers of wavelengths to find calibration equations for this example. Their Bayesian decision theory approach differed from that of Osborne in being multivariate, ie in trying to find one small subset of wavelengths good for predicting all four constituents, and in employing a more extensive search, using simulated annealing. The results for this alternative wavelength selection approach are given in the second row of Table 1.

Table 1 near here

For purposes of comparison we have carried out two further analyses, using partial least squares regression (PLS) and principal components regression (PCR). These approaches, both of which construct factors from the full spectral data and then regress constituents on the factors, are very much the standard tools in NIR spectroscopy, see for example Geladi and Martens (1996) and Geladi et al. (1996). For the computations we used the PLS Toolbox 2.0 of Wise and Gallagher (Eigenvector Research, Manson, WA). Although there are multivariate versions of PLS, we took the usual approach of calibrating for each constituent separately. The number of factors used, selected in each case by cross-validation on the training set, was 5 for each of the

PLS equations and 6 for each of the PCR equations. The results in rows three and four of Table 1 show that, as usual, there is little to choose between the two methods. These results are for PLS and PCR using the reduced 256 point spectrum that we used for our wavelet analysis. When we repeated the analyses using the original 700 point spectrum, the results were very similar for PLS and somewhat worse than those reported for PCR.

Because we shall shortly have to specify a prior distribution on regression coefficients, it is interesting to examine those from one of the factor-type approaches. Combining the coefficients for the regression of constituent on factor scores with the loadings that produce scores from the original spectral variables gives the coefficient vector that would be applied to a measured spectrum to give a prediction. These vectors are plotted for the PLS equations for the four constituents in Figure 2, showing the smoothness in the 256 coefficients that we shall try to reflect in our prior distribution for $B$.

## 6.2 Wavelet Transforms of Spectra

To each spectrum we apply a wavelet transform, converting it to a set of 256 wavelet coefficients. The MATLAB toolbox *Wavbox 4.3* (Copyright Carl Taswell 1995) was used for this. Using spectra with $2^m$, $m$ integer, points is not a real restriction here. Methods exist to overcome the limitation allowing the discrete wavelet transform to be applied on any length of data. We used MP(4), Daubechies (1992, p194), wavelets with 4 vanishing moments. The Daubechies wavelets have compact support, important for good localization, and maximum number of vanishing moments for a given smoothness. A large number of vanishing moments leads to high compressibility, since the fine scale wavelet coefficients will be essentially zero where the functions are smooth. On the other hand, the support of the wavelets increases with the number of vanishing moments so there is a trade off with the localization properties. Some rather limited exploration suggested that the chosen wavelet family is a good compromise for these data.

The right hand column of Figure 1 presents the wavelet transforms corresponding to the 3 NIR spectra in the left column. Coefficients are ordered from coarsest to finest.

## 6.3 Prior Settings

We need to specify the values of $H$, $\delta$ and $Q$ in (3), (4) and (5) and the prior probability $w$ that an element of $\gamma$ is 1.

We wish to put in weak but proper prior information about $\Sigma$. We choose $\delta = 3$ because this is the smallest integer value such that the expectation of $\Sigma$, $E(\Sigma) = Q/(\delta-2)$, exists. The scale matrix $Q$ is chosen as $Q = kI_q$ with $k = 0.05$, comparable in size to the expected error variances of the standardised $Y$ given $X$. With $\delta$ small the choice of $Q$ is unlikely to be critical.

Much more likely to be influential are the choices of $H$ and $w$ in the priors for $B$ and $\gamma$. To reflect the smoothness in the coefficients $B$, as exemplified in Figure 2, whilst keeping the form of $H$ simple, we have taken $H$ to be the variance matrix of a first-order autoregressive process, with $h_{ij} = \sigma^2\rho^{|i-j|}$. The values $\sigma^2 = 254$ and $\rho = 0.32$ were derived by maximising a type II likelihood, Good (1965). Integrating $\alpha$, $B$ and $\Sigma$ from the joint distribution given by (2), (3), (4) and (5) for the regression on the full untransformed spectra, with $h \to \infty$ and $B_0 = 0$ we get

$$f \propto |K|^{-q/2}|Q|^{(\delta+q-1)/2}|Q + Y'K^{-1}Y|^{-(\delta+n+q-1)/2}, \tag{18}$$

where

$$K = I_n + XHX',$$

and the columns of $Y$ are centered, as in Section 4.1. With $k = 0.05$ and $\delta = 3$ already fixed, (18) is a function, via $H$, of $\sigma^2$ and $\rho$. We used for our prior the values of these hyperparameters that maximise (18). Possible underestimation due to the use of the full spectra was taken into account by multiplying the estimate of $\sigma^2$ by the inflation factor 256/20 reflecting our prior belief on the expected number of included coefficients.

Figure 3 shows the diagonal elements of the matrix $\tilde{H} = WHW'$ implied by our choice of $H$. The variance matrix of the $i$th column of $\tilde{B}$ in (8) is $\sigma_{ii}\tilde{H}$, so this plot shows the pattern in the prior variances of the regression coefficients when the predictors are wavelet coefficients. The wavelet coefficients are ordered from coarsest to finest, so the decreasing prior variance means that there will be more shrinkage at the finer levels. This is a logical consequence of the smoothness that we have tried to express in the prior distribution. The spikes in the plot at the level transitions are due to the boundary condition problems of the discrete wavelet transform.

We know from experience that good predictions can usually be obtained with 10 or so selected spectral points in examples of this type. Having no previous experience of selecting wavelet coefficients, and wanting to induce a similarly 'small' model without constraining the possibilities too severely, we chose $w$ in the prior for $\gamma$ so that the expected model size was $pw = 20$. We have given equal prior probability here to coefficients at the different levels. Although we considered the possibility of varying $w$ in blocks, we had no strong prior opinions about which levels were likely to provide the most useful coefficients, apart from a suspicion that neither the coarsest nor the finest levels would feature strongly.

## 6.4 Implementation of the Metropolis Search

We chose widely different starting points for the 4 Metropolis chains, by setting to 1: (i) the first 1, (ii) the first 20, (iii) the first 128 (ie half) and (iv) all elements of $\gamma$. There were 100,000 iterations in each run, where an iteration consisted of either adding/deleting or swapping as described in Section 4.2. The two types of move were chosen with equal probability $\phi = 1/2$. Acceptance of the possible move by (14) was by generation of a Bernoulli random variable. Computation of $g(\gamma)$ and $g(\gamma^*)$ was done using the QR decomposition of MATLAB (copyright by MathWorks, Inc, Natick, Mass., USA).

For each chain we recorded the visited $\gamma$'s and their corresponding relative probability $g(\gamma)$. No 'burn-in' was necessary, since relatively unlikely $\gamma$'s would automatically be downweighted in our analysis. There were around 38,000-40,000 successful moves for each of the 4 runs. Of these moves around 95% were swaps. The relative probabilities of the set of distinct visited $\gamma$ were then normalized to 1 over this set.

The marginal probabilities for components of $\gamma$, $P(\gamma_j = 1)$, $j = 1, \ldots, 256$ are plotted in (i)-(iv) of Figure 4. The spikes show where regressor variables have been included in subsets with high probability. For one of the runs, (iii), two further plots are given, the number of 1's, Figure 5($a$), and the log- relative probabilities, $\log(g(\gamma))$, of the visited $\gamma$, Figure 5($b$), plotted over the 100,000 iterations. The other runs produced very similar plots, quickly moving towards models of similar dimensions and posterior probability values.

Despite the very different starting points, there are clear similarities between the regions explored by the 4 chains in that plots of marginals are overall broadly similar. However there are also clear differences, with some chains making frequent use of variables not picked up by others. Although we would not claim convergence, all 4 chains arrive at some similarly 'good' but different subsets. With mutually correlated predictor variables, as we have here, there will always tend to be many solutions to the problem of finding the best predictors. We adopt the pragmatic stance that all we are seeking to do is to identify some of the good solutions. If we happen to miss some other good ones then it is unfortunate but not disastrous.

## 6.5 Results

We pooled the distinct $\gamma$'s visited by the 4 chains, normalised the relative probabilities and ordered them according to probability. Then we predicted the further 39 unseen samples using Bayes model averaging. Mean square prediction errors converted to the original scale were $0.063, 0.449, 0.348$ and $0.050$. These results use the best 500 models, accounting for almost 99% of the total visited probability and using 219 wavelet coefficients. They considerably improve on any of the standard methods reported in Table 1. The single 'best' subset among the visited ones had 10 coefficients, accounted for 9% of the total visited probability and gave least squares predictions of 0.059, 0.466, 0.351 and 0.047.

It is interesting to look at the scales of the selected coefficients. The model making the best predictions used coefficients (10, 11, 14, 17, 33, 34, 82, 132, 166, 255), which include (0%, 0%, 0%, 37%, 6%, 6%, 1%, 2%) of all the coefficients at the 8 levels from coarsest to finest scales. The most useful coefficients seem to be in the middle of the scale, some more towards the finer end. Notice that the very coarsest ones, which would be essential to any reconstruction of the spectra, are not used, despite the smoothness implicit in $H$ and the consequent increased shrinkage associated with finer scales as seen in Figure 3.

Some idea of the locations of the wavelet coefficients selected by the modal model can be obtained from Figure 6. Here we have used the linearity of the wavelet transform and the linearity of the prediction equation to express the prediction equation as a vector of coefficients to be applied to the original spectral data. This vector is obtained by applying the inverse wavelet transform to the columns of the matrix of the least squares estimates of the regression coefficients. Since selection has discarded unnecessary details, most of the coefficients are very close to zero. We therefore display only the 1600-1800nm range, allowing a better view of the features of coefficients in the range of interest. These coefficient vectors can be compared directly with those in Figure 2. They are applied to the same spectral data to produce predictions and, despite the different scales, some comparisons are possible. The two coefficient vectors for fat can be interpreted easily. Fats and oils have absorption bands at around 1730 and 1765nm (Osborne et al. 1993), and strong positive coefficients in this region are the major features of both plots. The wavelet-based equation (Figure 6) is simpler, the selection having discarded much unnecessary detail. The other coefficient vectors show less resemblance, and are also harder to interpret. One thing to bear in mind in interpreting these plots is that the four constituents add up to 100%. Thus it should not be surprising that we find the fat measurement peak in all eight plots nor that the coefficient vectors for sugar and flour are so strongly inversely related.

Finally, we briefly comment on results we obtained by investigating logistic transformations of the data. Our response variables are in fact percentages and are constrained to sum up to 100, therefore lie on a simplex, rather than the full $q$-dimensional space. Sample ranges are $18 \pm 3$ for fat, $17 \pm 7$ for sugar, $51 \pm 6$ for flour, $14 \pm 3$ for water.

Following Aitchison (1986), we transformed the original $Y'$s into log ratios of the form

$$Z_1 = \ln(Y_1/Y_3), \quad Z_2 = \ln(Y_2/Y_3), \quad Z_3 = \ln(Y_4/Y_3). \tag{19}$$

The choice of the third ingredient at the denominator was the most natural, in that flour is the major constituent and also because often ingredients in recipes are expressed as a ratio to flour. We centered and scaled the $Z$ variables and re-computed empirical Bayes estimates for $\sigma^2$ and $\rho$. The other hyperparameters were not affected by the logistic transformation. We then ran 4 Metropolis chains using the starting points previously used with the data in the original scale. Diagnostic plots and plots of the marginals appeared very similar to those of Figures 4 and 5. The four chains visited 151,183 distinct models. We finally computed Bayes model averaging and least squares predictions with the best model, unscaling the predicted values and transforming them back to the original scale as

$$Y_i = \frac{100\exp(Z_i)}{\sum_{j=1}^{3}\exp(Z_j)+1}, \quad i=1,2 \quad Y_3 = \frac{100}{\sum_{j=1}^{3}\exp(Z_j)+1}, \quad Y_4 = \frac{100\exp(Z_3)}{\sum_{j=1}^{3}\exp(Z_j)+1}.$$

The best 500 visited models accounted for 99.4% of the total visited probability, used 214 wavelet coefficients and gave Bayes mean square prediction errors of 0.058, 0.819, 0.457 and 0.080. The single 'best' subset among the visited ones had 10 coefficients, accounted for 15.8% of the total visited probability and gave least squares prediction errors of 0.091, 0.793, 0.496 and 0.119. The logistic transformation does not seem to have a positive impact on the predictive performance and, overall, the simpler linear model seems to be adequate. Our approach gives Bayes predictions on the original scale satisfying the constraint of summing exactly to 100. This stems from the conjugate prior, linearity in $Y$, zero mean for the prior distribution of the regression coefficients, and vague prior for the intercept. It is easily seen that with $X$ centered $\hat{\alpha}'1 = 100$ and $\hat{B}'1 = 0$ for either least squares or Bayes estimates and hence that all predictions sum to 100. Additionally had we imposed a singular distribution to deal with the four responses summing to 100, then a proper analysis would suggest eliminating one component but then the predictions for the remaining three components would be as we have derived. Thus our analysis is Bayes for the singular problem even though superficially it ignores this aspect. This does not address the positivity constraint, but the composition variables are all so far from the boundaries compared with residual error that this is not an issue. Our desire to stick with the original scale is supported by the Beer-Lambert law which relates absorbance linearly to composition. The logistic transform distorts this linear relationship.

# 7 DISCUSSION

In specifying the prior parameters for our example we made a number of arbitrary choices. In particular the choice of $H$ as the variance matrix of an autoregressive process is a rather crude representation of the smoothness in the coefficients. It might be interesting to try to model this in more detail. However, while there may be scope for improvement, the simple structure used here does appear to work well.

Another area for future investigation is the use of more sophisticated wavelet systems in this context. The additional flexibility of wavelet packets (Coifman, Meyer and Wickerhauser 1992) or m-band wavelets (Mallet, Coomans, Kautsky and De Vel 1997) might lead to improved predictions, or might just be an unnecessary complication.

# REFERENCES

Aitchison, J. (1986), *The Statistical Analysis of Compositional Data*, London: Chapman and Hall.

Brown, P. J. (1993), *Measurement, Regression, and Calibration*, Oxford, U.K.: Clarendon Press.

Brown, P. J., Fearn, T., and Vannucci, M. (1999), "The Choice of Variables in Multivariate Regression: a Bayesian Non-Conjugate Decision Theory Approach," *Biometrika*, 86, 635–648.

Brown, P.J., Vannucci, M., and Fearn, T. (1998a), "Bayesian Wavelength Selection in Multicomponent Analysis," *Journal of Chemometrics*, 12, 173–182.

Brown, P.J., Vannucci, M., and Fearn, T. (1998b), "Multivariate Bayesian Variable Selection and Prediction," *Journal of the Royal Statistical Society,* Ser. B, 60, 627–641.

Carlin, B.P., and Chib, S. (1995), "Bayesian Model Choice Via Markov Chain Monte Carlo," *Journal of the Royal Statistical Society,* Ser. B, 57, 473–484.

Chipman, H. (1996), "Bayesian Variable Selection with Related Predictors," *Canadian Journal of Statistics*, 24, 17–36.

Clyde, M., DeSimone, H., and Parmigiani, G. (1996), "Prediction Via Orthogonalized Model Mixing," *Journal of the American Statistical Association*, 91, 1197–1208.

Clyde, M., and George, E.I. (2000), "Flexible Empirical Bayes Estimation for Wavelets," *Journal of the Royal Statistical Society*, Ser. B, 62, 681–698.

Coifman, R.R., Meyer, Y., and Wickerhauser, M.V. (1992), "Wavelet Analysis and Signal Processing," in *Wavelets and Their Applications*, eds. M.B. Ruskai, G. Beylkin, R. Coifman, I. Daubechies, S. Mallat, Y. Meyer and L. Raphael, Boston: Jones and Bartlett, pp. 153–178.

Cowe, I.A., and McNicol, J.W. (1985), "The Use of Principal Components in the Analysis of Near Infrared Spectra," *Applied Spectroscopy*, 39, 257–266.

Daubechies, I. (1992), *Ten Lectures on Wavelets*, Volume 61, CBMS-NSF Regional Conference Series in Applied Mathematics, Philadephia, Pennsylvania: Society for Industrial and Applied Mathematics.

Dawid, A.P. (1981), "Some Matrix-Variate Distribution Theory: Notational Considerations and a Bayesian Application," *Biometrika*, 68, 265–274.

Donoho, D., Johnstone, I., Kerkyacharian, G., and Picard, D. (1995), "Wavelet Shrinkage: Asymptopia?" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 57, 301–369.

Geladi, P., and Martens, H. (1996), "A Calibration Tutorial for Spectral Data. Part 1. Data Pretreatment and Principal Component Regression Using Matlab," *Journal of Near Infrared Spectroscopy*, 4, 225–242.

Geladi, P., Martens, H., Hadjiiski, L., and Hopke, P. (1996), "A Calibration Tutorial for Spectral Data. Part 2. Partial Least Squares Regression Using Matlab and Some Neural Network Results," *Journal of Near Infrared Spectroscopy*, 4, 243–255.

George, E.I., and McCulloch, R.E. (1997), "Approaches for Bayesian Variable Selection," *Statistica Sinica*, 7, 339–373.

Geweke, J. (1996), "Variable Selection and Model Comparison in Regression," in *Bayesian Statistics 5*, eds. J.M. Bernardo, J.O. Berger, A.P. Dawid, and A.F.M. Smith, Oxford, U.K.: Clarendon Press, pp. 609–620.

Good, I.J. (1965), *The Estimation of Probabilities. An Essay on Modern Bayesian Methods*, Cambridge, Mass: MIT Press.

Hrushka, W.R. (1987), Data Analysis: Wavelength Selection Methods. In *Near-Infrared Technology in the Agricultural and Food Industries*, eds. P. Williams and K. Norris, St Paul, Minnesota: American Association of Cereal Chemists, pp. 35–55.

Leamer, E.E. (1978), "Regression Selection Strategies and Revealed Priors," *Journal of the American Statistical Association*, 73, 580–587.

Madigan, D., and Raftery, A.E. (1994), "Model Selection and Accounting for Model Uncertainty in Graphical Models Using Occam's Window," *Journal of the American Statistical Association*, 89, 1535–1546.

Madigan, D., and York, J. (1995), "Bayesian Graphical Models for Discrete Data," *International Statistical Review*, 63, 215–232.

Mallat, S.G. (1989), "Multiresolution Approximations and Wavelet Orthonormal Bases of $L_2(r)$," *Transactions of the American Mathematical Society*, 315, 69–87.

Mallet, Y., Coomans, D., Kautsky, J., and De Vel, O. (1997), "Classification Using Adaptive Wavelets for Feature Extraction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19, 1058–1066.

McClure, W.F., Hamid, A., Giesbrecht, F.G., and Weeks, W.W. (1984), "Fourier-Analysis Enhances NIR Diffuse Reflectance Spectroscopy," *Applied Spectroscopy*, 38, 322–329.

Mitchell, T.J., and Beauchamp, J.J. (1988), "Bayesian Variable Selection in Linear Regression," *Journal of the American Statistical Association*, 83, 1023–1036.

Osborne, B.G., Fearn, T., and Hindle, P.H. (1993), *Practical NIR Spectroscopy*, Harlow, U.K.: Longman.

Osborne, B.G., Fearn, T., Miller, A.R., and Douglas, S. (1984), "Application of Near Infrared Reflectance Spectroscopy to Compositional Analysis of Biscuits and Biscuit Doughs," *Journal of the Science of Food and Agriculture*, 35, 99–105.

Raftery, A.E., Madigan, D., and Hoeting, J.A. (1997), "Bayesian Model Averaging for Linear Regression Models," *Journal of the American Statistical Association*, 92, 179–191.

Seber, G.A.F. (1984), *Multivariate Observations*, New York: Wiley.

Smith, A.F.M. (1973), "A General Bayesian Linear Model," *Journal of the Royal Statistical Society*, Ser. B, 35, 67–75.

Taswell, C. (1995), "Wavbox 4: A Software Toolbox for Wavelet Transforms and Adaptive Wavelet Packet Decompositions," in *Wavelets and Statistics*, eds. A. Antoniadis and G. Oppenheim, New York: Springer Verlag, pp. 361–375.

Trygg, J., and Wold, S. (1998), "PLS Regression on Wavelet Compressed NIR Spectra," *Chemometrics and Intelligent Laboratory Systems*, 42, 209–220.

Vannucci, M., and Corradi, F. (1999), "Covariance Structure of Wavelet Coefficients: Theory and Models in a Bayesian Perspective," *Journal of the Royal Statistical Society*, Ser. B, 61, 971–986.

Walczak, B., and Massart, D.L. (1997), "Noise Suppression and Signal Compression using the Wavelet Packet Transform," *Chemometrics and Intelligent Laboratory Systems*, 36, 81–94.

Wold, S., Martens, H., and Wold, H. (1983), "The Multivariate Calibration Problem in Chemistry Solved by PLS," in *Matrix Pencils*, eds. A. Ruhe and B. Kagstrom, Lecture Notes in Mathematics, Heidelberg: Springer, pp. 286–293.

| Method | Fat | Sugar | Flour | Water |
|---|---|---|---|---|
| Stepwise MLR | 0.044 | 1.188 | 0.722 | 0.221 |
| Decision theory | 0.076 | 0.566 | 0.265 | 0.176 |
| PLS | 0.151 | 0.583 | 0.375 | 0.105 |
| PCR | 0.160 | 0.614 | 0.388 | 0.106 |

Table 1: Mean squared errors of prediction on the 39 biscuit dough pieces in the validation set using four calibration methods

## Figure Captions

Figure 1. *Left column:* Original spectra. *Right column:* Wavelet transforms.

Figure 2. Coefficient vectors from 5-factor PLS equations.

Figure 3. Diagonal elements of $\tilde{H} = WHW'$, $\tilde{h}_{ii}$ is plotted against $i$.

Figure 4. Marginal probabilities of components of $\gamma$ for 4 runs.

Figure 5. Plots in sequence order for run (iii): (a) the number of ones, (b) log relative probabilities.

Figure 6. Coefficient vectors from the 'best' wavelet equations.

Figure 1

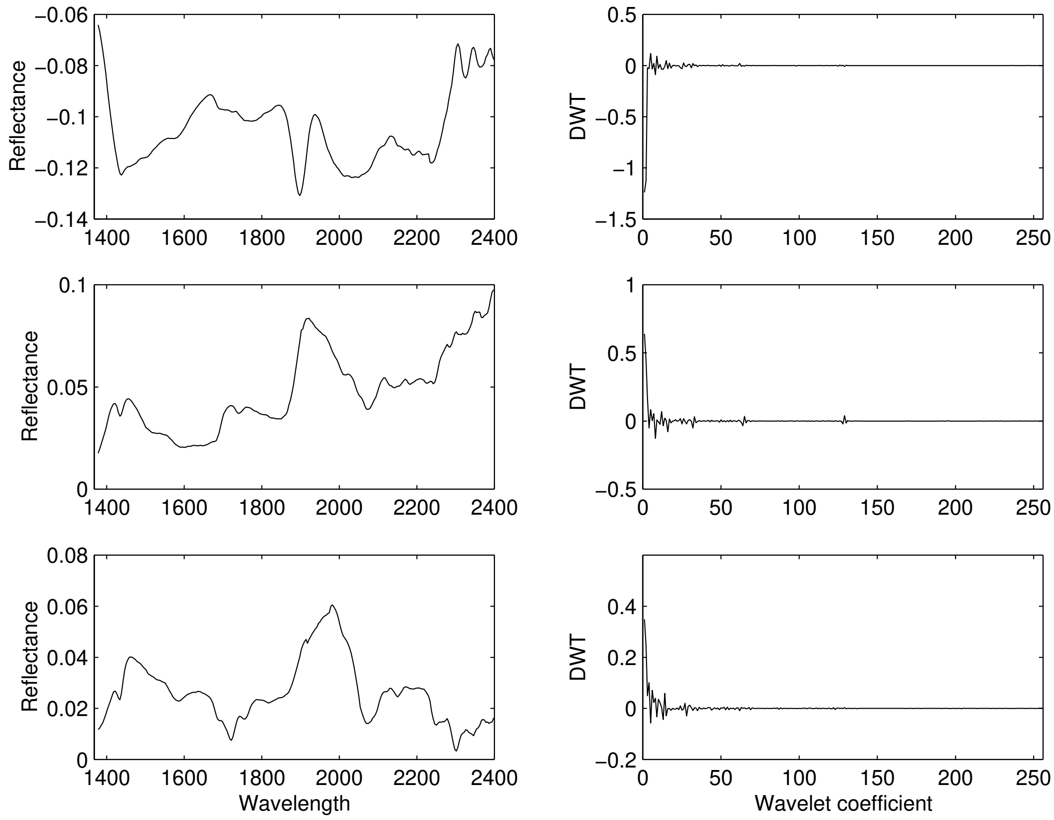


Figure 2

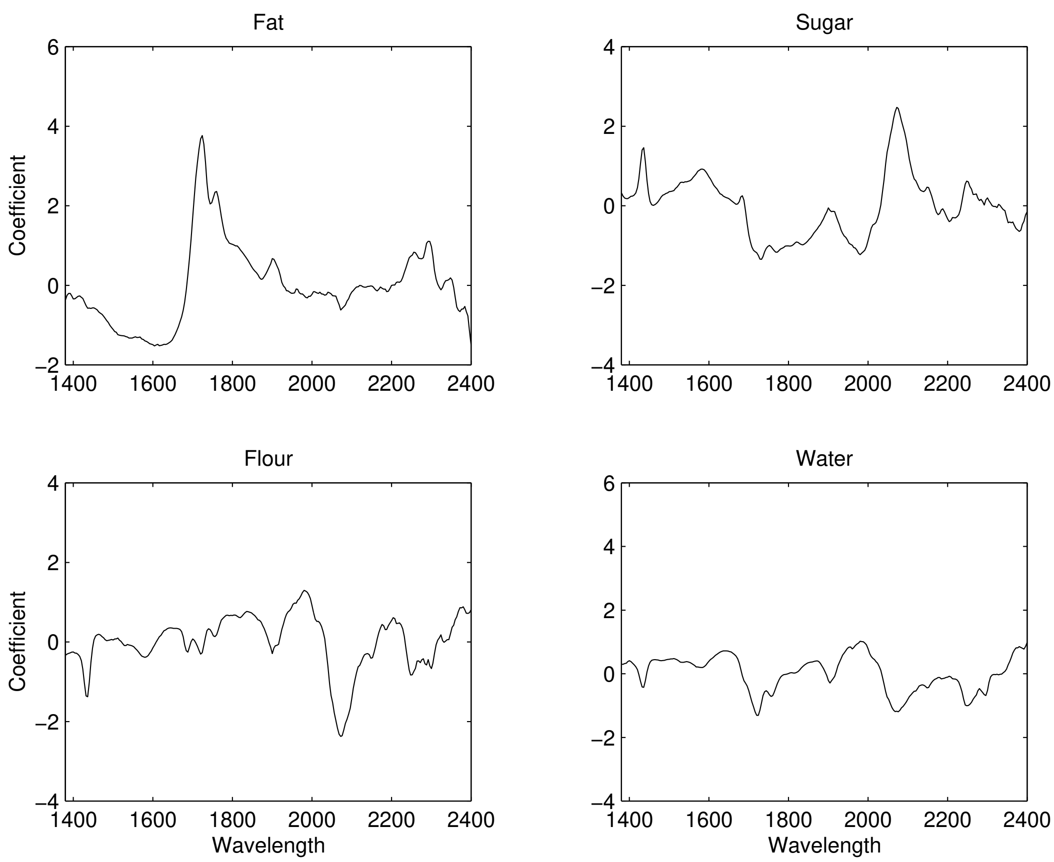

Fat
Sugar
Flour
Water
Coefficient
Coefficient
Wavelength
Wavelength
1400 1600 1800 2000 2200 2400
6
4
2
0
−2
−4

Figure 3

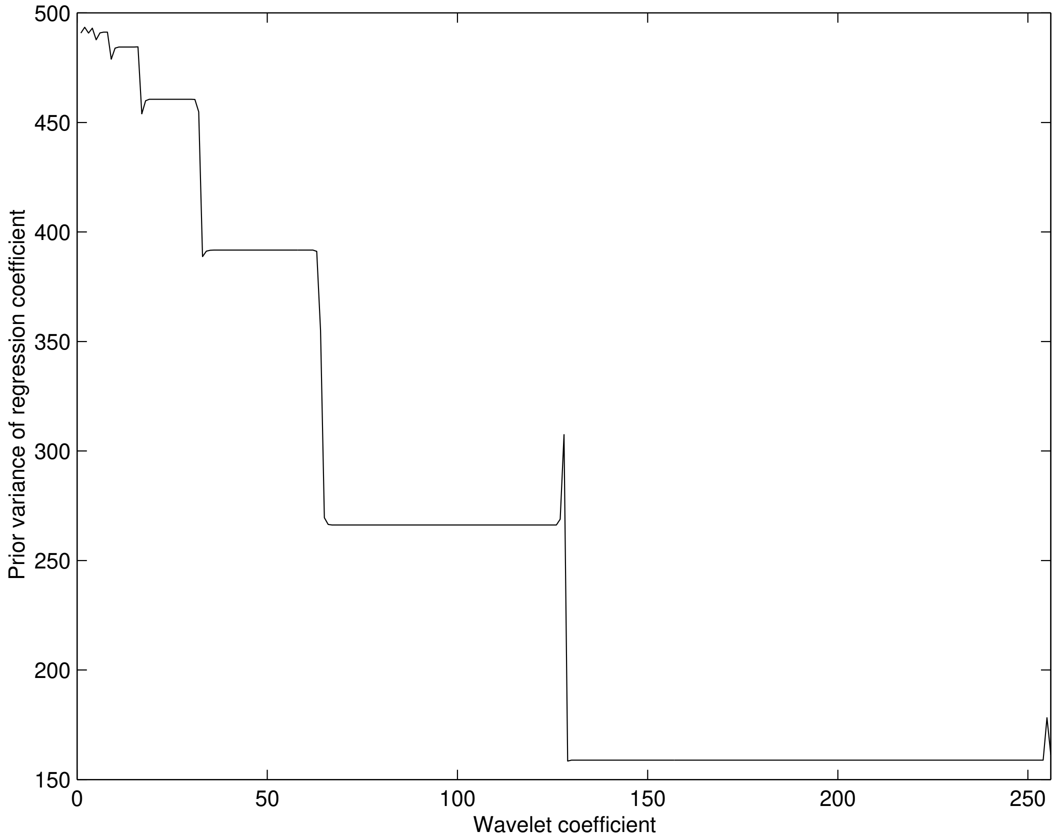

Prior variance of regression coefficient
Wavelet coefficient
500
450
400
350
300
250
200
150
0
50
100
150
200
250

Figure 4

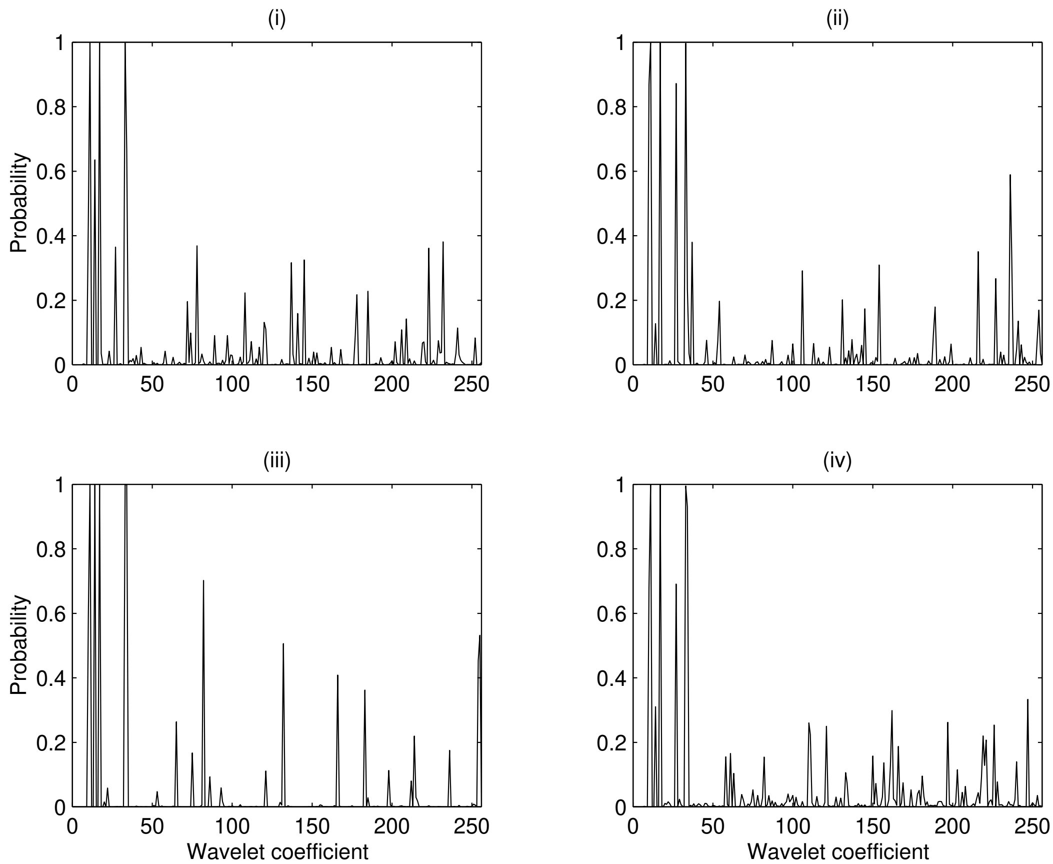

(i)
(ii)
(iii)
(iv)
Probability
Probability
Wavelet coefficient
Wavelet coefficient
0
0.2
0.4
0.6
0.8
1
0
50
100
150
200
250

Figure 5

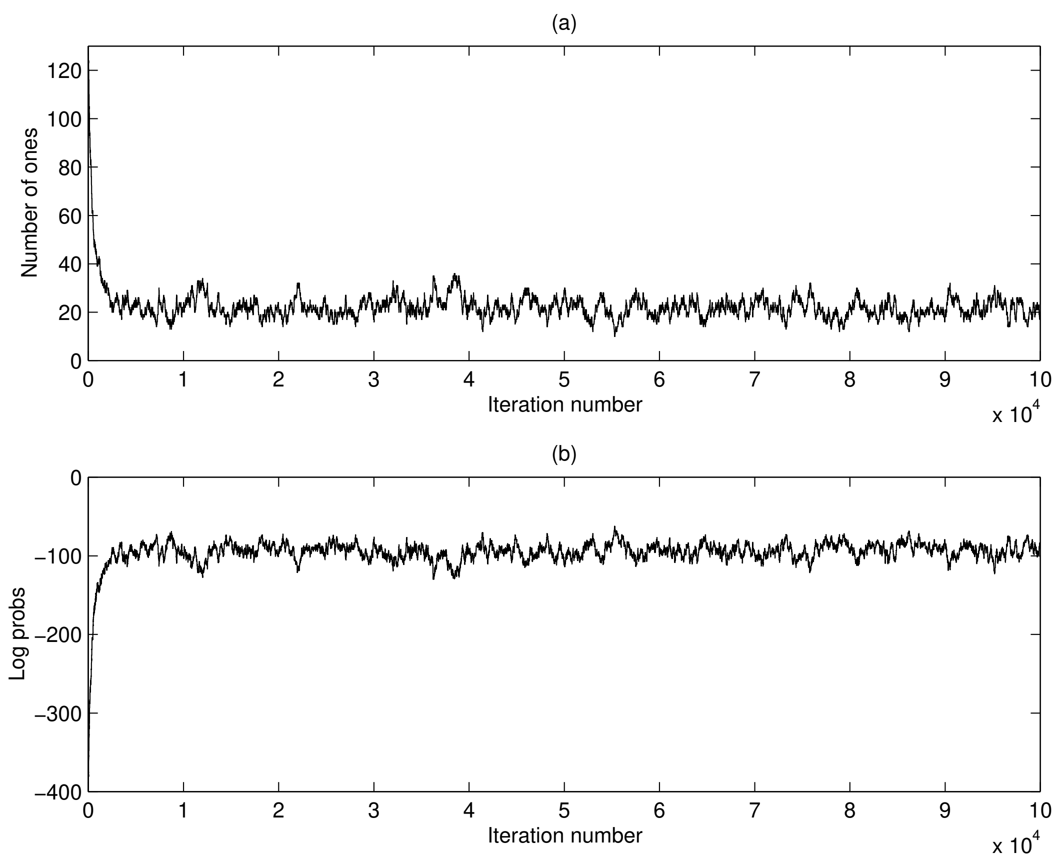
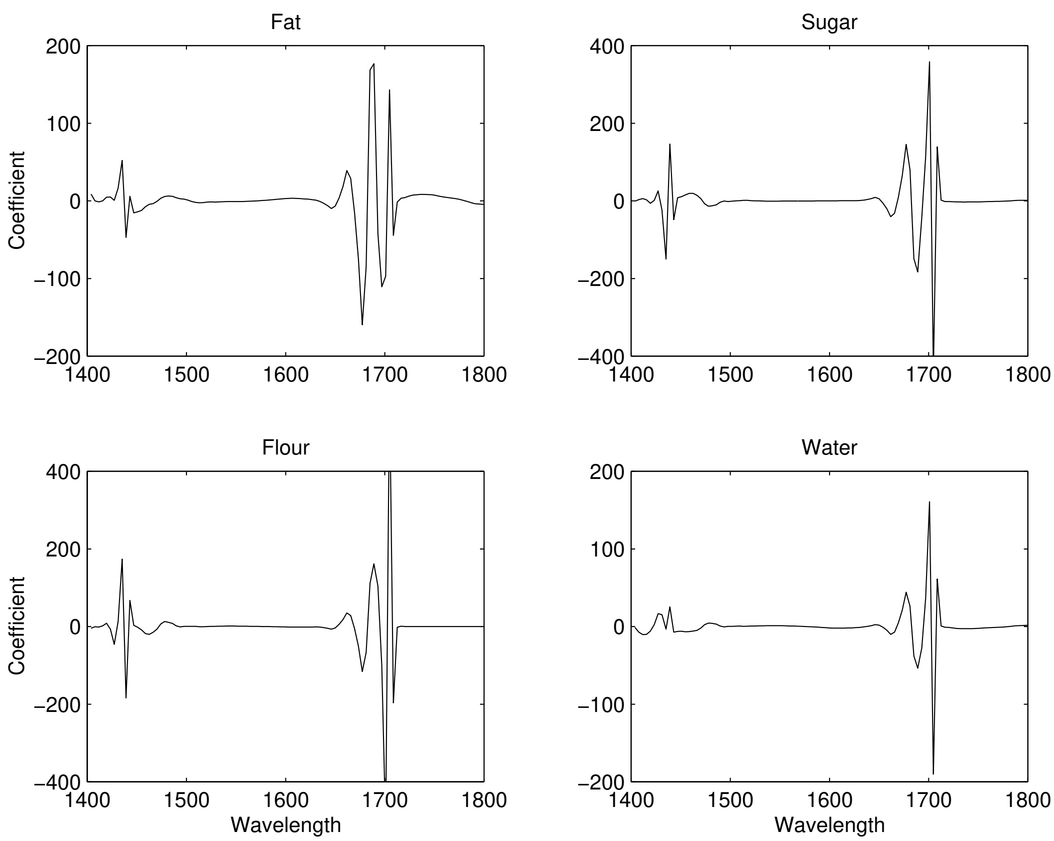

(a)
Number of ones
Iteration number
x 10^4
(b)
Log probs
Iteration number
x 10^4

Figure 6



Fat
Sugar
Flour
Water
Coefficient
Coefficient
Wavelength
Wavelength
200
100
0
−100
−200
400
200
0
−200
−400
1400
1500
1600
1700
1800