Statistica Sinica **22** (2012), 465-488
doi:http://dx.doi.org/10.5705/ss.2010.141

BAYESIAN WAVELET-BASED CURVE CLASSIFICATION VIA DISCRIMINANT ANALYSIS WITH MARKOV RANDOM TREE PRIORS

Francesco C. Stingo, Marina Vannucci and Gerard Downey

University of Texas MD Anderson Cancer Center, Rice University, and Teagasc Food Research Centre

Abstract: Discriminant analysis is an effective tool for the classification of experimental units into groups. When the number of variables is much larger than the number of observations it is necessary to include a dimension reduction procedure in the inferential process. Here we present a typical example from chemometrics that deals with the classification of different types of food into species via near infrared spectroscopy. We take a nonparametric approach by modeling the functional predictors via wavelet transforms and then apply discriminant analysis in the wavelet domain. We consider a Bayesian conjugate normal discriminant model, either linear or quadratic, that avoids independence assumptions among the wavelet coefficients. We introduce latent binary indicators for the selection of the discriminatory wavelet coefficients and propose prior formulations that use Markov random tree (MRT) priors to map scale-location connections among wavelets coefficients. We conduct posterior inference via MCMC methods, we show performances on our case study on food authenticity, and compare results to several other procedures.

Key words and phrases: Bayesian variable selection, classification and pattern recognition, Markov chain Monte Carlo, Markov random tree prior, wavelet-based modeling.

1. Introduction

Discriminant analysis, sometimes called supervised pattern recognition, is a statistical technique used to classify observations into groups. For each case in a given training set a $p\times 1$ vector of observations, $\mathbf{x}_i$, and a known assignment to one of G possible groups are available. Let the group indicators be stored in a $n \times 1$ vector $\mathbf{y}$. On the basis of the $\mathbf{X}$ and $\mathbf{y}$ data we wish to derive a classification rule that assigns future cases to their correct groups. When the distribution of $\mathbf{X}$, conditional on the group membership, is assumed to be a multivariate normal then this statistical methodology is know as discriminant analysis. Here we focus in particular on situations in which the number of observed variables is considerably large, often larger than the number of samples. We present a

typical example from chemometrics that deals with the classification of different types of food into species via near infrared spectral data, i.e., curve predictors.

Different approaches can cope with the high dimensionality of a data matrix. One approach is to use a dimension reduction technique, for example principal component analysis, and then use only the first k components to classify units into groups via linear or quadratic discriminant analysis, see for example Jolliffe (1986). Another approach is to first select a subset of the variables, essentially removing noisy ones, and then perform discriminant analysis using only the selected variables. This, for example, is the approach taken by Fearn, Brown, and Besbeas (2002) who developed a Bayesian decision theory approach to linear discriminant analysis that balances costs of variables against a loss due to classification errors. Also, more recently, Murphy, Dean, and Raftery (2010) proposed a frequentist model-based approach to discriminant analysis in which variable selection is achieved by imposing constraints on the form of the covariance matrices. These authors use a specifically designed search algorithm that compares models using BIC approximations of log Bayes factors.

In this paper we take a dimension reduction approach and model the functional predictors in a nonparametric way by means of wavelet series representations. Wavelets can accurately describe local features of curves in a parsimonious way via a small number of coefficients. Because of this compression ability, selecting the important wavelet coefficients, rather than original variables, may be expected to lead to improved classification performances. We therefore apply wavelet transforms, reducing curves to wavelet coefficients, and then perform discriminant analysis in the wavelet domain while simultaneously employing a selection scheme of the relevant wavelet coefficients. We consider a Bayesian conjugate normal discriminant model, either linear or quadratic, and introduce latent binary indicators for the selection of the discriminatory coefficients. Unlike current literature on wavelet-based modeling, where models are fit one wavelet coefficient at a time, our model formulation avoids independence assumptions among the wavelet coefficients. We go one step further and, additionally, propose a prior model formulation that includes Markov random tree (MRT) priors to map scale-location connections among wavelet coefficients. We achieve dimension reduction by building a stochastic search variable selection procedure for posterior inference. We investigate performances of the proposed method on our case study on food authenticity, and compare results to several other procedures.

Our approach to selection builds on the extensive literature on Bayesian methods for variable selection. For example, we introduce a latent binary vector γ for the identification of the discriminating variables (wavelet coefficients) and use stochastic search MCMC techniques to explore the space of variable subsets.

This is also the approach taken by George and McCulloch (1993, 1997), among many others, for linear regression models, and by Brown, Fearn, and Vannucci (2001) in a wavelet approach to curve regression. However unlike linear settings, where γ is used to induce mixture priors on the regression coefficients of the model, in mixture models, like the one we work with, the elements of the matrix $\mathbf{X}$ are viewed as random variables and γ is used to index the contribution of the different variables to the likelihood term of the model, Tadesse, Sha and Vannucci (2005). The method for variable selection we adopt can be used either for linear discriminant analysis, where all groups share the same covariance matrix, or for quadratic discriminant analysis, where different groups are allowed to have different covariance matrices. We illustrate the model for the quadratic discriminant analysis case and report the modification needed to perform linear discriminant analysis in the Appendix.

The rest of the paper is organized as follows: we complete this discussion with a brief review of wavelet series representations and wavelet transforms. We also introduce the concept of zero-tree wavelet structures that map wavelet coefficients with the same spatial locations but at different resolution scales. In Section 2 we describe how to perform discriminant analysis under the Bayesian paradigm and how to re-parameterize the model in the wavelet domain. We also discuss likelihood and prior distributions that allow us to implement a variable selection mechanism in the wavelet domain. We then describe how to incorporate into the prior model information about the connections among wavelet coefficients at the same spatial locations. We present the MCMC algorithm for posterior inference in Section 3, where we also address the case of having samples with missing labels. Finally, in Section 4, we apply our method to the NIR spectral data for food classification and compare results to several other competing procedures.

1.1. Wavelet representations of curves

The basic idea behind wavelets is to represent a general function in terms of simpler functions (building blocks), defined as scaled and translated versions of an oscillatory function, describing local features in a parsimonious way. The existence of fast and efficient transformations to calculate coefficients of wavelet expansions have made wavelets a simple tool that can be used for a great variety of applications. Indeed, wavelets have been extremely successful in, for example, the compression or denoising of signals and images, see for example Antoniadis, Bigot, and Sapatinas (2001) and Gonzalez and Woods (2002), and references therein.

In $L^2(\mathbb{R})$, for example, an orthonormal wavelet basis is obtained as translations and dilations of a "mother" wavelet ψ 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

$$f(x) = \sum_{j,k \in \mathbb{Z}} d_{j,k} \psi_{j,k}(x), \tag{1.1}$$

with wavelet coefficients $d_{j,k} = \int f(x)\psi_{j,k}(x)dx$ describing features of the function f at spatial locations indexed by k and scales indexed by j. Because of their localization properties, i.e., fast decay in both the time and the frequency domains, wavelets have the ability to represent many classes of functions in a sparse form by describing important features with a relatively small number of coefficients.

Wavelets have been extremely successful as a tool for the analysis and synthesis of discrete data. Let $\mathbf{x} = (x_1, \ldots, x_p)^T$ be a sample of the function $f(\cdot)$ at equally spaced points and let p be a power of 2, $p = 2^J$. This vector of observations can be viewed as an approximation of f at scale J. A fast algorithm exists, the *Discrete Wavelet Transform* (DWT), that permits decomposition of $\mathbf{x}$ into a set of wavelet coefficients, Mallat (1989). This algorithm operates in practice by means of linear recursive filters; for illustrative purposes, it is useful to write the DWT in matrix form as

$$\mathbf{z} = \mathbf{W}\mathbf{x}, \tag{1.2}$$

with $\mathbf{W}$ an orthogonal matrix corresponding to the discrete wavelet transform, and $\mathbf{z}$ a vector of wavelet coefficients describing features of the function at scales from the fine $J-1$ to a coarser one, say $J-r$. Figure 1 illustrates the wavelet decomposition of one of the near infrared curves analyzed in Section 4: the NIR curve is shown at the top of the figure and the wavelet coefficients at the individual scales are depicted below, from coarsest to finest. An algorithm for the inverse construction, the *Inverse Wavelet Transform* (IWT), also exists.

In this paper we use Daubechies wavelets; they have compact support and a maximum number of vanishing moments for any given smoothness. They are used extensively in statistical applications. A detailed description of the construction of these wavelets, together with a general exposition of the wavelet theory, can be found in Daubechies (1992). Some of the early applications of wavelets in statistics are described in Vidakovic (1999).

1.2 The zero-tree wavelet structure

Because of the recursive nature of the DWT, the resulting wavelet coefficients tend to share certain properties. This is true, in particular, for coefficients that map to the same spatial locations but at different scales. When a decimation is performed at every iterative step of the DWT, the transform naturally leads

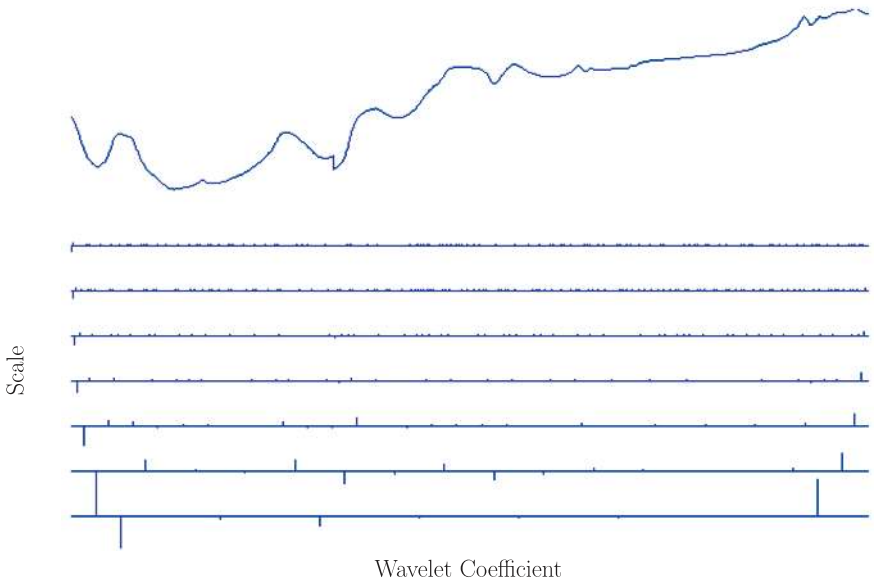


Figure 1. Discrete wavelet transform of one of the NIR curves analyzed in Section 4. The NIR curve is shown at the top of the figure and the wavelet coefficients at the individual scales are depicted below, from finest to coarsest.

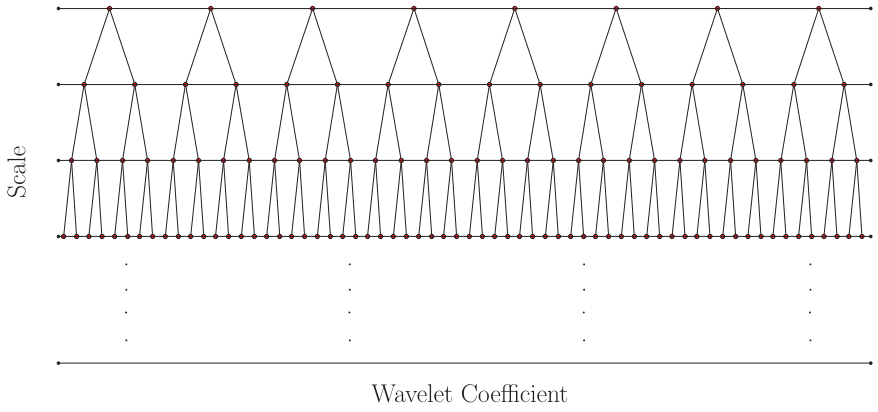


Figure 2. Schematic representation of the wavelet tree structure, where each coefficient serves as "parent" for up to two "children" nodes.

to a tree structure in which each coefficient at a given scale serves as "parent" for up to two "children" nodes at the finer scale. The wavelet coefficients of the coarsest scale compose the "root node" and those at the finest scale the "leaf nodes". Figure 2 provides a schematic representation of the wavelet tree structure.

In general, for most signals and images, if a coefficient at a particular scale is negligible (or large) then its children will likely tend to be negligible (or large) as well. This simple intuition has led to the construction of "zero-tree" structures for signal and image compression, where coefficients belonging to sub-trees are all considered negligible, see Shapiro (1993). Many commercial software packages for image coding and compression routinely incorporate zero-tree structures in their algorithms, like for example JPEG2000. In statistical modeling, He and Carin (2009) have used zero-tree structures in a Bayesian approach to compressive sensing for signals and images that are sparse in the wavelet domain.

2. Wavelet-based Bayesian Discriminant Analysis

We start by describing the model for discriminant analysis. We assume that each observation comes from one of the G possible groups, each with distribution $N(\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)$. We represent the data from each group by the $n_g \times p$ matrix

$$\mathbf{X}_g - \mathbf{1}_{n_g}\boldsymbol{\mu}_g^T \sim \mathcal{N}(\mathbf{I}, \boldsymbol{\Sigma}_g) \tag{2.1}$$

with $g = 1, \ldots, G$, and where $\boldsymbol{\mu}_g$ and $\boldsymbol{\Sigma}_g$ are the mean and the covariance matrix of the g-*th* group, respectively. Here the notation $\mathbf{V} - \mathbf{M} \sim \mathcal{N}(\mathbf{A}, \mathbf{B})$ indicates a matrix normal variate $\mathbf{V}$ with matrix mean $\boldsymbol{M}$ and with variance matrices $b_{ii}\mathbf{A}$ for its generic i-*th* column and $a_{jj}\mathbf{B}$ for its generic j-*th* row. This notation was proposed by Dawid (1981) and has the advantage of preserving the matrix structure instead of reshaping $\mathbf{V}$ as a vector. It also makes for much easier formal Bayesian manipulation and has become quite standard in the Bayesian literature.

Model (2.1) is completed by imposing a conjugate multivariate normal distribution on $\boldsymbol{\mu}_g$ and an Inverse-Wishart prior on the covariance matrix $\boldsymbol{\Sigma}_g$,

$$\begin{aligned} \boldsymbol{\mu}_g &\sim N(\mathbf{m}_g, h_g\boldsymbol{\Sigma}_g), \\ \boldsymbol{\Sigma}_g &\sim IW(\delta_g, \boldsymbol{\Omega}_g), \end{aligned} \tag{2.2}$$

where $\boldsymbol{\Omega}_g$ is a scale matrix and δ_g a shape parameter.

In discriminant analysis the predictive distribution of a new observation $\mathbf{x}^f$ $(1 \times p)$ is used to classify the new sample into one of the G possible groups. This distribution is a multivariate T-student, see Brown (1993) among others,

$$\mathbf{x}^f - \widetilde{\boldsymbol{\mu}}_g \sim \mathcal{T}(\delta_g^*, a_g, \boldsymbol{\Omega}_g^*), \tag{2.3}$$

where $\widetilde{\boldsymbol{\mu}}_g = \pi_g\mathbf{m}_g + (1-\pi_g)\bar{\mathbf{x}}_g$, $\delta_g^* = \delta_g + n_g$, $a_g = 1 + (1/h_g + n_g)^{-1}$, and $\boldsymbol{\Omega}_g^* = \boldsymbol{\Omega}_g + \mathbf{S}_g + (h_g + 1/n_g)^{-1}(\bar{\mathbf{x}}_g - \mathbf{m}_g)^T(\bar{\mathbf{x}}_g - \mathbf{m}_g)$ with $\pi_g = (1 + h_g n_g)^{-1}$ and $\mathbf{S}_g = (\mathbf{X}_g - 1_{n_g}\bar{\mathbf{x}}_g^T)^T(\mathbf{X}_g - 1_{n_g}\bar{\mathbf{x}}_g^T)$. The probability that a future observation, given the observed data, belongs to the group g is then

$$\pi_g(y^f|\mathbf{X}) = p(y^f = g|\mathbf{x}^f, \mathbf{X}),$$

where y^f is the group indicator of the new observation. By estimating the probability that one observation comes from group g as $\hat{\pi}_g = n_g/n$ the previous distribution can be written in closed form as

$$\pi_g(y^f|\mathbf{X}) = \frac{p_g(\mathbf{x}^f)\hat{\pi}_g}{\sum_{i=1}^{G} p_i(\mathbf{x}^f)\hat{\pi}_i}, \tag{2.4}$$

where $p_g(\mathbf{x}^f)$ indicates the predictive distribution defined at (2.3). The underling assumption of exchangeability of the training data with the future data is required, we assume that observations from training and validation sets arise in the same proportions from the groups, see also Fearn, Brown, and Besbeas (2002) A new observation is then assigned to the group with the highest posterior probability.

2.1. Model in the wavelet domain

We approach dimension reduction by transforming curves into wavelet coefficients. Because of the compression ability of wavelets, selecting the important wavelet coefficients, rather than original variables, is expected to lead to improved classification performances. Our wavelet-based approach incorporates both selection and synthesis of the data.

We show how to apply wavelet transforms to the data. We have n curves and want to apply the same wavelet transform to them all. In matrix form, a DWT can be applied to multiple curves as $\mathbf{Z}_g = \mathbf{X}_g\mathbf{W}^T$, with $\mathbf{W}$ the orthogonal matrix representing the discrete wavelet transform, see Brown, Fearn, and Vannucci (2001). We therefore rewrite model (2.1) as

$$\mathbf{Z}_g - \mathbf{1}_{n_g}\tilde{\boldsymbol{\mu}}_g^T \sim \mathcal{N}(\mathbf{I}, \tilde{\boldsymbol{\Sigma}}_g),$$

where $\tilde{\boldsymbol{\Sigma}}_g = \mathbf{W}\boldsymbol{\Sigma}_g\mathbf{W}^T$ and $\tilde{\boldsymbol{\mu}}_g^T = \boldsymbol{\mu}_g^T\mathbf{W}^T$. The prior model on $\tilde{\boldsymbol{\Sigma}}_g$ transforms into

$$\tilde{\boldsymbol{\Sigma}}_g \sim IW(\delta_g, \tilde{\boldsymbol{\Omega}}_g)$$

with $\tilde{\boldsymbol{\Omega}}_g = \mathbf{W}\boldsymbol{\Omega}_g\mathbf{W}^T$. Vannucci and Corradi (1999) have derived an algorithm to compute variance and covariance matrices like $\tilde{\boldsymbol{\Sigma}}_g$ and $\tilde{\boldsymbol{\Omega}}_g$ that makes use of the recursive filters of the DWT and avoids multiplications by the matrix W. In the sequel we drop the tilde notation and simply refer to $\boldsymbol{\Sigma}_g$, $\boldsymbol{\Omega}_g$, and $\boldsymbol{\mu}_g$.

It is worth noticing that our model formulation avoids any independence assumption among the wavelet coefficients of a given curve. Such assumptions are often made in the current literature on wavelet-based Bayesian modeling as a convenient working model that allows to fit models one wavelet coefficient at a time, and are heuristically justified by the whitening properties of the wavelet

transforms, see for example Morris and Carroll (2006) and Ray and Mallick (2006).

2.2. A prior model for variable selection

In the wavelet domain we achieve dimension reduction by selecting the discriminating wavelet coefficients. We do this by extending to the discriminant analysis framework an approach to variable selection proposed by Tadesse, Sha and Vannucci (2005) for model-based clustering. As they do, we introduce a $(p \times 1)$ latent binary vector $\boldsymbol{\gamma}$ to index the selected variables, with $\gamma_j = 1$ if the jth wavelet coefficient contributes to the classification of the n units into the corresponding groups, and $\gamma_j = 0$ otherwise. We then use the latent vector $\boldsymbol{\gamma}$ to index the contribution of the different wavelet coefficients to the likelihood term of the model. Unlike Tadesse, Sha and Vannucci (2005) we avoid independence assumptions among the variables by defining a likelihood that allows separation of the discriminating coefficients from the noisy ones as follows:

$$L(\mathbf{Z}, \mathbf{y}; \cdot) = \prod_{i=1}^{n} p(\mathbf{z}_{i(\boldsymbol{\gamma}^c)} | \mathbf{z}_{i(\boldsymbol{\gamma})}) \prod_{g=1}^{G} w_g^{n_g} \prod_{i=1}^{n_g} p_g(\mathbf{z}_{i(\boldsymbol{\gamma})}), \tag{2.5}$$

where w_g is the prior probability that unit i belongs to group g, $\mathbf{z}_{i(\boldsymbol{\gamma}^c)}$ is the $|\boldsymbol{\gamma}^c| \times 1$ vector of the non-selected wavelet coefficients, and $\mathbf{z}_{i(\boldsymbol{\gamma}^c)}$ is the $|\boldsymbol{\gamma}| \times 1$ vector of the selected ones, for the i-*th* subject. The first factor of the likelihood refers to the non-important variables, while the second is formed by variables able to classify observations into the correct groups. Under the assumption of normality of the data the likelihood becomes

$$\prod_{i=1}^{n} N_{|\gamma^c|}(\mathbf{z}_{i(\boldsymbol{\gamma}^c)} - \mathbf{B}\mathbf{z}_{i(\boldsymbol{\gamma})}; \boldsymbol{\mu}_{0(\boldsymbol{\gamma}^c)}, \boldsymbol{\Sigma}_{0(\boldsymbol{\gamma}^c)}) \prod_{g=1}^{G} w_g^{n_g} \prod_{i=1}^{n_g} N_{|\boldsymbol{\gamma}|}(\mathbf{z}_{i(\boldsymbol{\gamma})}; \boldsymbol{\mu}_{g(\boldsymbol{\gamma})}, \boldsymbol{\Sigma}_{g(\boldsymbol{\gamma})}), \tag{2.6}$$

where $\mathbf{B}$ is a matrix of regression coefficients resulting from the linearity assumption on the expected value of the conditional distribution $p(\mathbf{z}_{i(\boldsymbol{\gamma}^c)} | \mathbf{z}_{i(\boldsymbol{\gamma})})$, and where $\boldsymbol{\mu}_{0(\boldsymbol{\gamma}^c)}$ and $\boldsymbol{\Sigma}_{0(\boldsymbol{\gamma}^c)}$ are the mean and covariance matrix, respectively, of $\mathbf{z}_{i(\boldsymbol{\gamma}^c)} - \mathbf{B}\mathbf{z}_{i(\boldsymbol{\gamma})}$. Note that all parameters of the distribution of the non-discriminatory variables are assumed independent of the cluster indicators and that the (conditional) distribution of the non-discriminatory variables $z_{i(\gamma^c)}$ is therefore not a mixture-type. This assumption also implies that the covariance between $z_{i(\gamma)}$ and $z_{i(\gamma^c)}$ is not group dependent. These choices reflect our intention of having a likelihood that factorizes into two parts, the first with group specific parameters and the second with parameters that are common to all the non-discriminatory variables. Murphy, Dean, and Raftery (2010) use a similar

likelihood formulation in a frequentist approach to variable selection in discriminant analysis.

For the parameters corresponding to the non-selected wavelet coefficients we again choose conjugate priors:

$$\begin{aligned} \boldsymbol{\mu}_{0(\boldsymbol{\gamma}^c)}|\boldsymbol{\Sigma}_{0(\boldsymbol{\gamma}^c)} &\sim N(\mathbf{m}_{0(\boldsymbol{\gamma}^c)}, h_0\boldsymbol{\Sigma}_{0(\boldsymbol{\gamma}^c)}), \\ \mathbf{B}-\mathbf{B}_0|\boldsymbol{\Sigma}_{0(\boldsymbol{\gamma}^c)} &\sim \mathcal{N}(\mathbf{H}_{\boldsymbol{\gamma}}, \boldsymbol{\Sigma}_{0(\boldsymbol{\gamma}^c)}), \\ \boldsymbol{\Sigma}_{0(\boldsymbol{\gamma}^c)} &\sim IW(\delta_c, \boldsymbol{\Omega}_{0(\boldsymbol{\gamma}^c)}). \end{aligned} \tag{2.7}$$

This parametrization allows us to create a computationally efficient variable selection algorithm, see Section 3. We also assume $\boldsymbol{\Omega}_{0(\boldsymbol{\gamma}^c)} = k_0\mathbf{I}_{|\boldsymbol{\gamma}^c|}$, a specification implying shrinkage toward the case of no correlation between non-selected variables, also adopted by Tadesse, Sha and Vannucci (2005) in model-based clustering, and Dobra et al. (2004) for graphical model settings.

We complete the prior model by specifying an improper non-informative prior on the vector $\mathbf{w} = (w_1, \ldots, w_G)$ using a Dirichlet distribution, $\mathbf{w} \sim Dirichlet$ $(0, \ldots, 0)$. With this prior, marginalizing over $\mathbf{w}$ in the predictive distribution $\pi_g(y^f|Z)$, with the integration done over the posterior distribution $\mathbf{w} \sim Dirichlet(n_1, \ldots, n_G)$, is equivalent to estimating $\hat{\pi}_g = n_g/n$, as we have done in (2.4). Note that, with the inclusion of the variable selection mechanism, the predictive distribution does not change because it depends only on the selected variables. We discuss prior models for $\boldsymbol{\gamma}$ in the next Section.

2.3. Markov random tree priors for zero-tree structures

Although our model allows for dependencies among variables through the choice of the priors (2.2) and (2.7), it is not straightforward to specify dependence structures known a priori on the prior covariance matrices. We take a different approach and show how available information can be incorporated into the model via the prior distribution on $\boldsymbol{\gamma}$.

In defining our prior construction we follow the original idea of Shapiro (1993) and use a zero-tree structure to specify a dependence network among wavelet coefficients. In this network, given the wavelet decomposition of a signal, an individual wavelet coefficient at a given scale is directly "linked" to the two coefficients at the next finer scale that correspond to the same spatial location. We encode this network structure into our model via a Markov random tree (MRT) prior on $\boldsymbol{\gamma}$. This is a type of Markov random field (MRF), in which the distribution of a set of random variables follows Markov properties that can be described by an undirected graph. In a MRF, variables are represented by nodes and relations between them by edges; in our context, the nodes are the wavelet coefficients and the edges represent the relations encoded in the zero-tree structure.

We adopt the parametrization suggested by Li and Zhang (2010) for the global MRF distribution for $\boldsymbol{\gamma}$,

$$p(\boldsymbol{\gamma}|\mathbf{d},\mathbf{E}) \propto \exp(\mathbf{d}^T\boldsymbol{\gamma} + \boldsymbol{\gamma}^T\mathbf{E}\boldsymbol{\gamma}), \tag{2.8}$$

where $\mathbf{d} = d\mathbf{1}_p$, with $\mathbf{1}_p$ the unit vector of dimension p, and where $\mathbf{E}$ is a matrix with elements $\{e_{ij}\}$ usually set to a constant e for the connected nodes and to 0 for the non connected ones. Note that the connections among wavelet coefficients on the MRT prior are a characteristic of the data by definition of the wavelet transform. This feature implies that there is no uncertainty on the links of the MRT.

The parameter d in (2.8) represents the expected prior number of significant wavelet coefficients and controls the sparsity of the model, while e affects the probability of selecting a variable according to its neighbor values. This is more evident by noting that the conditional probability

$$P(\gamma_j|d,e,\gamma_k,k\in N_j) = \frac{\exp(\gamma_j(d+e\sum_{k\in N_j}\gamma_k))}{1+\exp(d+e\sum_{k\in N_j}\gamma_k)}, \tag{2.9}$$

with N_j the set of direct neighbors of variable j in the MRF, increases as a function of the number of selected neighbors. Note that if a variable does not have any neighbor, then its prior distribution reduces to an independent Bernoulli with parameter $p = \exp(d)/[1+\exp(d)]$, which is a logistic transformation of d.

Although the parametrization above is somewhat arbitrary, some care is needed in deciding whether to put a prior distribution on e. First, allowing e to vary can lead to a *phase transition* in which the expected number of variables equal to 1 increases massively for small increments of e. This can happen because (2.9) can only increase as a function of the number of z_j's equal to 1. A clear description of the phase transition is given by Li and Zhang (2010). In brief, Ising models undergo transitions between an ordered and a disordered underlying state, from a model with most of the variables equal to 1 to a model with most of the variables equal to 0, at or near the phase transition boundary, that depends on the parameter specifications. Phase transition has consequences such as the loss of model sparsity, and consequently a critical slow down of the MCMC. In Bayesian variable selection with large p, phase transition leads to a drastic change in the proportion of included variables, for example, from $< 5\%$ to $> 90\%$, near the phase transition boundary.

The most effective way to obtain an empirical estimate of the phase transition value is to sample from (2.8), using the algorithm proposed by Propp and Wilson (1996) to obtain an estimate of the expected model size for different values of d over a range of values for e. The value of e for which the expected

model size shows a dramatic increase can be considered a good estimate of the phase transition point. However, even if an estimate is obtained, inference on the parameters d and e requires special techniques in order to handle non-tractable priors of type (2.8), which are known only up to normalizing constant, and substantially increases the computational complexity of our model. In this paper we have therefore opted for fixing the parameters d and e. Similar choices have been made by Li and Zhang (2010). We provide some guidelines for choosing these parameters, when we perform a sensitivity analysis.

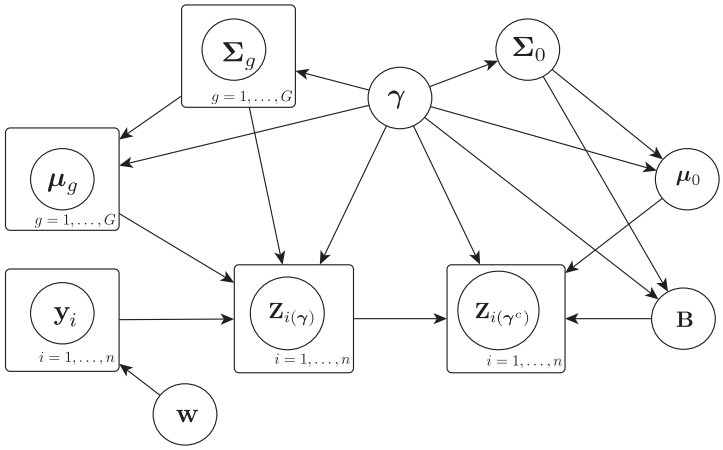
Figure 3 provides a graphical representation of our proposed model, illustrating the probabilistic dependencies among the observed variables and the model parameters. The different layers of the proposed hierarchical model are also summarized.

3. MCMC for Posterior Inference

We concentrate on the posterior distribution on $\boldsymbol{\gamma}$, which allows us to achieve variable selection. This distribution cannot be obtained in closed form and an MCMC is required. We illustrate the procedure for the quadratic discriminant analysis case, and report the modification needed to perform linear discriminant analysis in the Appendix.

The inferential procedure can be greatly simplified by integrating out the parameters $w_g, \mathbf{B}, \Sigma_0, \boldsymbol{\mu}_0, \boldsymbol{\mu}_g$, and $\boldsymbol{\Sigma}_g$. In the MCMC procedure we describe, a single variable is added and/or removed at every iteration. Therefore, without loss of generality, one can simplify the prior parametrization by assuming that the set of non-selected variables is formed by only one variable, so that the matrix $\boldsymbol{\Sigma}_{0(\boldsymbol{\gamma}^c)}$ reduces to a scalar σ^2 and the $|\boldsymbol{\gamma}^c| \times |\boldsymbol{\gamma}|$ matrix $\mathbf{B}$ to a $|\boldsymbol{\gamma}| \times 1$ vector $\boldsymbol{\beta}$, that is to the row of $\mathbf{B}$ corresponding to the regression coefficients between the selected variables and the variable prosed to be added or removed. Our priors therefore reduce to $\sigma^2 \sim \text{Inv-Gamma}(\delta_c/2, k_0/2)$ and $\boldsymbol{\beta} \sim N(\boldsymbol{\beta}_0, \sigma^2 \mathbf{H}_{\boldsymbol{\gamma}})$. Integrating out the parameters $w_g, \boldsymbol{\beta}, \sigma^2, \boldsymbol{\mu}_0, \boldsymbol{\mu}_g$ and $\boldsymbol{\Sigma}_g$ leads to the marginal likelihood

$$
\begin{aligned}
p(\mathbf{Z}|\mathbf{y}, \boldsymbol{\gamma}) \propto\ & (k_0 + (\mathbf{z}_{(\boldsymbol{\gamma}^c)} - \mathbf{1}_{p_{\boldsymbol{\gamma}}} m_{0(\boldsymbol{\gamma}^c)} - \mathbf{Z}_{(\boldsymbol{\gamma})} \boldsymbol{\beta}_0)^T \\
& \times (\mathbf{I}_n + h_0 \mathbf{1}_n \mathbf{1}_n^T + \mathbf{Z}_{(\boldsymbol{\gamma})} \mathbf{H}_{\boldsymbol{\gamma}} \mathbf{Z}_{(\boldsymbol{\gamma})}^T)^{-1} \\
& \times (\mathbf{z}_{(\boldsymbol{\gamma}^c)} - \mathbf{1}_{p_{\boldsymbol{\gamma}}} m_{0(\boldsymbol{\gamma}^c)} - \mathbf{Z}_{(\boldsymbol{\gamma})} \boldsymbol{\beta}_0))^{-(n+\delta)/2} \\
& \times \prod_{g=1}^{G} \mathbf{K}_{g(\boldsymbol{\gamma})} |\boldsymbol{\Omega}_{g(\boldsymbol{\gamma})}|^{(\delta + p_{\boldsymbol{\gamma}} - 1)/2} |\boldsymbol{\Omega}_{g(\boldsymbol{\gamma})} + \mathbf{S}_{g(\boldsymbol{\gamma})}|^{-(n_g + \delta + p_{\boldsymbol{\gamma}} - 1)/2}, \qquad (3.1)
\end{aligned}
$$



Likelihood:

$$L(\mathbf{Z}, \mathbf{y}; \cdot) = \prod_{i=1}^{n} N_{|\boldsymbol{\gamma}^c|}(\mathbf{Z}_{i(\boldsymbol{\gamma}^c)} - \mathbf{B}\mathbf{Z}_{i(\boldsymbol{\gamma})}; \boldsymbol{\mu}_{0(\boldsymbol{\gamma}^c)}, \boldsymbol{\Sigma}_{0(\boldsymbol{\gamma}^c)}) \prod_{g=1}^{G} \prod_{i=1}^{n_g} w_g^{n_g} N_{|\boldsymbol{\gamma}|}(\mathbf{Z}_{i(\boldsymbol{\gamma})}; \boldsymbol{\mu}_{g(\boldsymbol{\gamma})}, \boldsymbol{\Sigma}_{g(\boldsymbol{\gamma})})$$

Model parameters:

Non-selected variables		*Selected variables*	
$\boldsymbol{\Sigma}_{0(\boldsymbol{\gamma}^c)}$	$\sim IW(\delta_c, k_0 I_{\lvert\boldsymbol{\gamma}^c\rvert})$	$\boldsymbol{\Sigma}_{g(\boldsymbol{\gamma})}$	$\sim IW(\delta_g, \boldsymbol{\Omega}_{g(\boldsymbol{\gamma})})$
$\boldsymbol{\mu}_{0(\boldsymbol{\gamma}^c)} \lvert \boldsymbol{\Sigma}_{0(\boldsymbol{\gamma}^c)}$	$\sim N(\mathbf{m}_{0(\boldsymbol{\gamma}^c)}, h_0 \boldsymbol{\Sigma}_{0(\boldsymbol{\gamma}^c)})$	$\boldsymbol{\mu}_{g(\boldsymbol{\gamma})} \lvert \boldsymbol{\Sigma}_{g(\boldsymbol{\gamma})}$	$\sim N(\mathbf{m}_g, h_g \boldsymbol{\Sigma}_{g(\boldsymbol{\gamma})})$
$\mathbf{B} - \mathbf{B}_0 \lvert \boldsymbol{\Sigma}_{0(\boldsymbol{\gamma}^c)}$	$\sim \mathcal{N}(\mathbf{H}_{\boldsymbol{\gamma}}, \boldsymbol{\Sigma}_{0(\boldsymbol{\gamma}^c)})$	$\mathbf{w}$	$\sim Dirichlet(0, \ldots, 0)$

Variable selection parameters:

$$p(\boldsymbol{\gamma} | \mathbf{d}, \mathbf{E}) \propto \exp(\mathbf{d}^T \boldsymbol{\gamma} + \boldsymbol{\gamma}^T \mathbf{E} \boldsymbol{\gamma})$$

Figure 3. Graphical model representation and hierarchical formulation of the proposed probabilistic model.

where

$$\mathbf{K}_{g(\boldsymbol{\gamma})} = (h_1 n_g + 1)^{-p_{\boldsymbol{\gamma}}/2} \prod_{j=1}^{p_{\boldsymbol{\gamma}}} \frac{\Gamma((n_g + \delta + p_{\boldsymbol{\gamma}} - j)/2)}{\Gamma((\delta + p_{\boldsymbol{\gamma}} - j)/2)},$$

$$\mathbf{S}_{g(\boldsymbol{\gamma})} = \sum_{i | \boldsymbol{\gamma}_i = 1} (\mathbf{z}_{i(\boldsymbol{\gamma})} - \bar{\mathbf{z}}_{g(\boldsymbol{\gamma})})(\mathbf{z}_{i(\boldsymbol{\gamma})} - \bar{\mathbf{z}}_{g(\boldsymbol{\gamma})})^T + \frac{n_g}{h_0 n_g + 1} (\mathbf{m}_{0(\boldsymbol{\gamma})} - \bar{\mathbf{z}}_{g(\boldsymbol{\gamma})})(\mathbf{m}_{0(\boldsymbol{\gamma})} - \bar{\mathbf{z}}_{g(\boldsymbol{\gamma})})^T,$$

and $p_{\boldsymbol{\gamma}}$ is the number of selected variables. We implement a Stochastic Search

Variable Selection (SSVS) algorithm that has been used successfully in variable selection, see Madigan and York (1995) for graphical models, Sha et al. (2004) for classification settings, and Tadesse, Sha and Vannucci (2005) for clustering, among others. This is a Metropolis-type algorithm:

- with probability ϕ, add or delete one variable by choosing at random one component in the current $\boldsymbol{\gamma}$ and changing its value;
- with probability $1-\phi$, swap two variables by choosing independently at random a 0 and a 1 in the current $\boldsymbol{\gamma}$ and changing their values.

The proposed $\boldsymbol{\gamma}^{new}$ is then accepted with a probability that is the ratio of the relative posterior probabilities of the new versus the current model:

$$\min\left[\frac{p(\mathbf{Z}|\mathbf{y},\boldsymbol{\gamma}^{new})\pi(\boldsymbol{\gamma}^{new})}{p(\mathbf{Z}|\mathbf{y},\boldsymbol{\gamma}^{old})\pi(\boldsymbol{\gamma}^{old})},1\right]. \tag{3.2}$$

Because these moves are symmetric, the proposal distribution does not appear in the previous ratio. Here we also simplify the computation of the acceptance probability using a factorization of the marginal likelihood adopted by Murphy, Dean, and Raftery (2010) in their calculation of the ratio between the BIC statistics of two nested models,

$$p(\mathbf{Z}_{(\boldsymbol{\gamma}^c)}|\mathbf{y},\mathbf{Z}_{(\boldsymbol{\gamma})},\mathbf{Z}_{(prop)})p(\mathbf{Z}_{(\boldsymbol{\gamma})},\mathbf{Z}_{(prop)}|\mathbf{y}),$$

where $\mathbf{Z}_{(prop)}$ represents the variable in an add/delete move, or the two variables in a swap move, whose indicator element(s) have been proposed to change. The first factor of this marginal likelihood simplifies when calculating the ratio in (3.2), while the second one can assume either the form of the joint marginal distribution of the selected variable(s) or it can be written as (3.1) where the set $\boldsymbol{\gamma}^c$ is formed by only the proposed variable(s). In details, the acceptance probability of the Metropolis step is

- $\min\left[\frac{p_m(\mathbf{Z}|\boldsymbol{\gamma}^{p-})p_m(\mathbf{Z}|\mathbf{y},\boldsymbol{\gamma}^{new})\pi(\boldsymbol{\gamma}^{new})}{p_m(\mathbf{Z}|\mathbf{y},\boldsymbol{\gamma}^{old})\pi(\boldsymbol{\gamma}^{old})},1\right]$, for a remove move,
- $\min\left[\frac{p_m(\mathbf{Z}|\mathbf{y},\boldsymbol{\gamma}^{new})\pi(\boldsymbol{\gamma}^{new})}{p_m(\mathbf{Z}|\boldsymbol{\gamma}^{p+})p_m(\mathbf{Z}|\mathbf{y},\boldsymbol{\gamma}^{old})\pi(\boldsymbol{\gamma}^{old})},1\right]$, for an add move,
- $\min\left[\frac{p_m(\mathbf{Z}|\boldsymbol{\gamma}^{p-})p_m(\mathbf{Z}|\mathbf{y},\boldsymbol{\gamma}^{new})\pi(\boldsymbol{\gamma}^{new})}{p_m(\mathbf{Z}|\boldsymbol{\gamma}^{p+})p_m(\mathbf{Z}|\mathbf{y},\gamma^{old})\pi(\gamma^{new})},1\right]$, for a swap move.

Here $p_m(\mathbf{Z}|\boldsymbol{\gamma}^{p-})$ is the marginal likelihood of the variable selected to be removed from the set of significant variables, while $p_m(\mathbf{Z}|\boldsymbol{\gamma}^{p+})$ is the marginal likelihood of the variable selected to be added to the set of significant variables. Note that, in all of the possible moves, the part of the marginal likelihood that involves the non-significant variables is one-dimensional.

The MCMC procedure results in a list of visited models, $\boldsymbol{\gamma}^{(0)}, \ldots, \boldsymbol{\gamma}^{(T)}$, and their corresponding posterior probabilities. Variable selection can then be achieved either by looking at the $\boldsymbol{\gamma}$ vectors with largest joint posterior probabilities among the visited models or, marginally, by calculating frequencies of inclusion for each γ_j and then choosing those γ_j's with frequencies exceeding a given cut-off value. Finally, new observations are assigned to one of the G groups according to (2.4).

3.1. Handling missing labels

Murphy, Dean, and Raftery (2010) show how to handle unlabeled data, situations where the group information for some of the samples is missing, using an EM algorithm within their frequentist approach to discriminant analysis. Although we do not have unlabeled samples in our case study data, for completeness we show how to adapt our Bayesian method to handle such situations.

In Section 2 we implicity defined the distribution of $\mathbf{y}$ as $P(y_i = g) = w_g$, and then performed all inference conditioning upon the observed group indicators. Within this framework, missing labels can be handled by considering latent variables that we can sample via an additional MCMC step. Let $\{y_k,\ k \in \mathbf{S}\}$ indicate the set of unlabeled observations. To simplify the sampling of the y_k's, we do not integrate the mixture weights w_g's but sample them from their full conditional distribution that can be derived in closed form as $\mathbf{w}|\mathbf{y} \sim Dirichlet(n_1, \ldots, n_G)$. Note that the n_g's now depend on the sampled values of the missing y_i's. As a consequence of not having integrated out w, (3.1) also changes. In particular, $P(\mathbf{Z}|\mathbf{y}, \boldsymbol{\gamma}, \mathbf{w})$ is obtained as in (3.1) by replacing $\mathbf{K}_{g(\boldsymbol{\gamma})}$ with

$$\mathbf{K}'_{g(\boldsymbol{\gamma})} = w_g^{n_g}(h_1 n_g + 1)^{-p_{\boldsymbol{\gamma}}/2} \prod_{j=1}^{p_{\boldsymbol{\gamma}}} \frac{\Gamma((n_g + \delta + p_{\boldsymbol{\gamma}} - j)/2)}{\Gamma((\delta + p_{\boldsymbol{\gamma}} - j)/2)}.$$

Given the sampled w_g's, the full conditional distribution of $y_i|y_{-i}, \mathbf{Z}, \boldsymbol{\gamma}$ is the predictive distribution (2.4) with the only difference that we replace $\hat{\pi_g}$ with the sampled value for w_g. This step is equivalent to sampling from a multinomial distribution with probabilities that depend on the selected variables and the group indicators, both observed and sampled.

4. An Application to NIR Spectral Data

Discriminant analysis is frequently used to classify units into groups based on near infrared (NIR) spectra, see Fearn, Brown, and Besbeas (2002) and Dean, Murphy, and Downey (2006) for approaches that incorporate variable selection. Food authenticity studies are concerned with establishing whether foods are authentic or not. NIR spectroscopy provides a quick and efficient method of collecting the data, see Downey (1996). Correct identification of food via analysis of

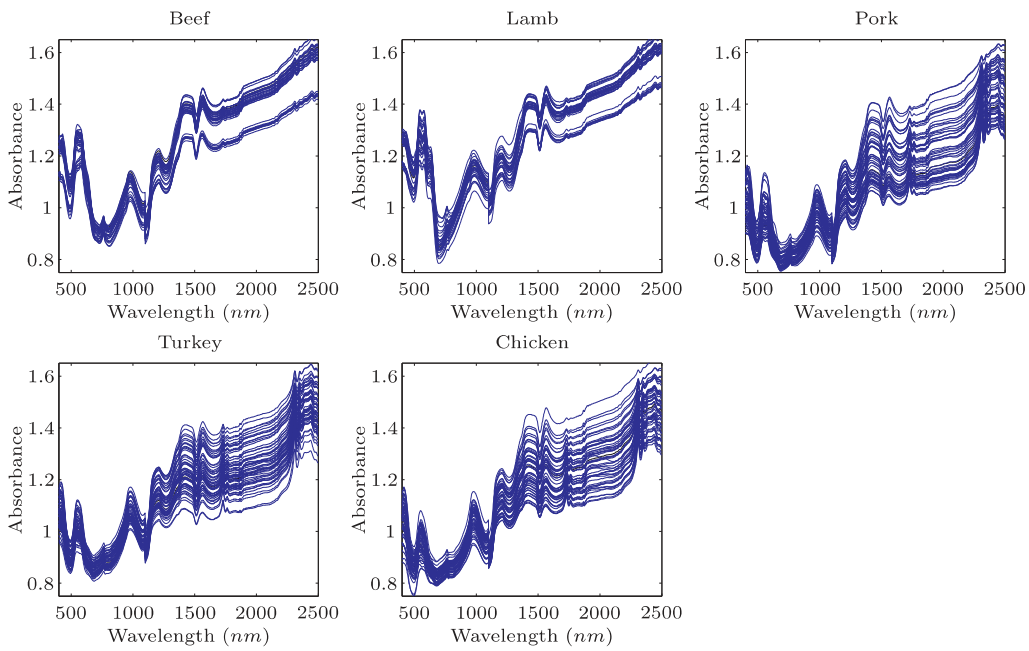


Figure 4. NIR spectral data from 231 food samples of five different species of meat.

NIR spectroscopy data is important in order to avoid potential fraud. Food producers, regulators, retailers, and consumers need to be assured of the authenticity of food products.

We analyze a data set that consists of combined visible and near-infrared spectroscopic measurements from 231 homogenized samples of five different species of meat (Beef, Chicken, Lamb, Pork and Turkey). The NIR data were collected in reflectance mode using a NIRSystems 6,500 instrument over the range 400-2,498nm at intervals of 2nm. These are shown in Figure 4. These data have been analyzed by Dean, Murphy, and Downey (2006) and recently by Murphy, Dean, and Raftery (2010). A two-step approach was adopted by Dean, Murphy, and Downey (2006), who first applied the standard wavelet thresholding of Donoho and Johnstone (1994), and then performed discriminant analysis on selected subsets of wavelet coefficients. Murphy, Dean, and Raftery (2010) also reported results on such other techniques as Transductive Support Vector Machine (SVM), Random Forest, AdaBoost, Bayesian Multinomial Regression, Factorial Discriminant Analysis (FDA), k-nearest neighbors, discriminant partial least squares (PLS) regression, and soft independent modeling of class analogy (SIMCA).

We removed the first 27 wavelengths to obtain curves observed at 1,024 equispaced points (in nm). We transformed the curves into wavelet coefficients

using DWT and Daubechies wavelets with three vanishing moments. This gave us 8 scaling coefficients and 1,016 wavelet coefficients for each curve. Because the scaling coefficients carry information on the global features of the data, we decided to discard them and performed our selection only on the standardized wavelet coefficients. Previous experience with wavelet-based modeling of NIR spectra has confirmed the intuition that the important predictive information of the data is represented by local features and therefore captured by the wavelet coefficients, see Brown, Fearn, and Vannucci (2001).

We randomly split the data into a training set of 117 observations (16 samples of Beef, 28 of Chicken, 17 of Lamb, 28 of Pork, and 28 of Turkey) and a validation set of 114 observations (16 samples of Beef, 27 of Chicken, 17 of Lamb, 27 of Pork, and 27 of Turkey). We assumed $\delta_1 = \ldots = \delta_G = \delta_c = \delta$ and set $\delta = 3$, the minimum value such that the expectation of $\mathbf{\Sigma}$ exists. We set each element of $\mathbf{m}_g$ and $\mathbf{m}_0$ to the corresponding interval midpoint of the observed wavelet coefficients. We let $\beta_0 = 0$, a standard choice when no additional information is available. As suggested by Tadesse, Sha and Vannucci (2005), we specified $h_1 = 100$, $h_0 = 1{,}000$, and $\mathbf{H}_\gamma = 100 \cdot \mathbf{I}_{|\gamma|}$. A good rule of thumb for these values is to set them in the range 10 to $1,000$ to obtain fairly flat priors over the region where the data are defined. A diagonal specification for $\mathbf{H}_\gamma$ still allows posterior dependence among regressions coefficients, mostly depending on the covariance structure of the selected wavelet coefficients. Alternative specifications are also possible. For example, Brown, Fearn, and Vannucci (2001) adopted a first-order autoregressive structure in the data domain and then transformed the variance-covariance matrix via the wavelet transform, similarly to what done in Section 2.1 to define $\tilde{\Sigma}_g$ and $\tilde{\Omega}_g$. Because of the decorrelation properties of the wavelet transform, the transformed matrix has a nearly block-diagonal form.

Some care is needed in the choice of $\mathbf{\Omega}_g$ and k_0, since the posterior inference is sensitive to the setting of these parameters. This was originally noted in Kim, Tadesse, and Vannucci (2006) where guidelines for the specification of these parameters are provided. In general, these parameters need to be specified in the range of variability of the data. A data-based specification, in particular, ensures that the prior distributions overlap with the likelihood, resulting in well-behaved posterior densities. Other authors have reported that noninformative or diffuse priors produce undesirable posterior behavior in finite mixture models, see Richardson and Green (1997) and Wasserman (2000), among others, and Kass and Wasserman (1996) for a nice discussion on prior specifications and their effects on posterior inference. In our application we set $\mathbf{\Omega}_g = k\mathbf{I}_{|\gamma|}$ with $k = 3^{-1}$, a value close to the standard deviation of the means of the columns of $\mathbf{Z}$. We also specified $k_0 = 10^{-1}$, a value in the same scale of magnitude of $\mathbf{\Omega}_g$.

The hyperparameters of the MRT were set to $d = -2.5$ and $e = 0.3$. The choice of d reflects our prior expectation about the number of significant variables, in this case equal to 7.5%, while a moderate value was chosen for e to avoid the phase transition problem. In general, any value of e below the phase transition point can be considered a reasonable choice. However, a value close to the phase transition point would result in a high prior probability of selection for those nodes whose neighbors are already selected, particulary in a sparse network. Consequently the data would play a much less important role in the selection of the wavelet coefficients. The approach we adopt considers also that the prior probability of a wavelet coefficient should not vary enormously according to the selection of its neighbors. Specifically we set e so that the prior probability of inclusion of a wavelet coefficient with all its three neighbors already selected is roughly twice the prior probability of a coefficient that does not have any of its neighbors selected.

We assumed unequal covariance matrices across the groups. We ran MCMCs by setting $\phi = 0.5$, therefore giving equal probability to the add/delete and the swap moves. We used two chains, one that started from a model with two randomly-selected variables, the other one from a model with ten included variables. We ran the chains for 200,000 iterations, using the first 1,000 as burn-in. We observed fast convergence, therefore selecting a relative short burn-in. The stochastic searches mostly explored models with 16-18 wavelet coefficients and then quickly settled down to models with similar numbers of variables. In our Matlab implementation, the MCMC algorithm needs only a few minutes to run.

The results we report here were obtained by pooling together the output of the two chains. Figure 5 shows the marginal probability of inclusion of the individual wavelet coefficients. A threshold of 0.4 on these probabilities selected a subset of 14 wavelet coefficients. This threshold corresponds to an expected false discovery rate (Bayesian FDR) of 36.7%, which we calculated according to the formulation suggested by Newton et al. (2004). As we expected, the selected wavelet coefficients belonged to the intermediate scales, with 12 out of 14 belonging to scales 5-6-7 (in our decompositions wavelet coefficients ranged from scale 3, the coarsest, to scale 9, the finest). Additionally, we found that 4 of these 12 coefficients were directly connected in the prior tree structure. Increasing the threshold to 0.5 selected 7 wavelet coefficients, corresponding to a Bayesian FDR of 17.9%.

Figure 6 shows the posterior probabilities of class memberships for the 114 observations of the validation set, calculated based on the selected 14 wavelet coefficients, and Table 1 summarizes the classification results according to these probabilities and a threshold of 0.4. Overall, the model is able to classify 110 of the 114 food samples of the validation set. Predictions worsen to 102 out of 114

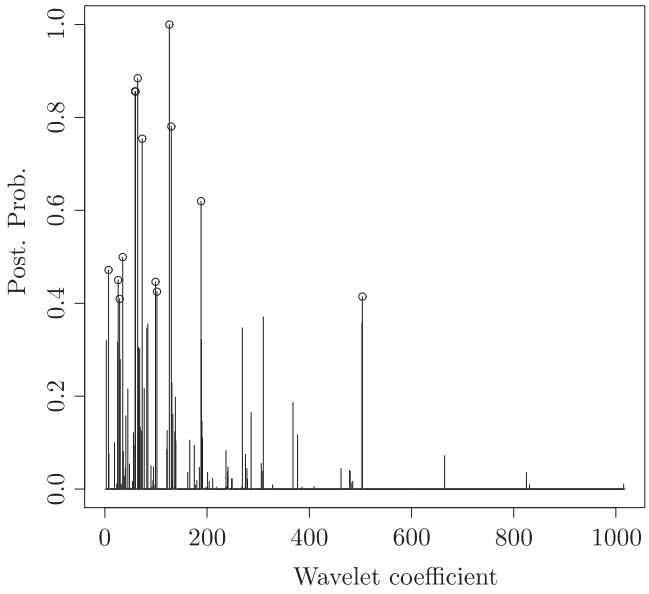


Figure 5. Marginal posterior probabilities of inclusion for single wavelet coefficients.

Table 1. Validation set: Classification results for the five different species of meat using a threshold of 0.4 for the posterior of inclusion. The number of misclassified units is reported in parentheses.

	Predicted				
Truth	Beef	Lamb	Pork	Turkey	Chicken
Beef	100.0	0.0	0.0	0.0	0.0
Lamb	5.9 (1)	94.1	0.0	0.0	0.0
Pork	0.0	0.0	100.0	0.0	0.0
Turkey	0.0	0.0	0.0	92.6	7.4 (2)
Chicken	0.0	0.0	0.0	3.7 (1)	96.3

by using the 7 wavelet coefficients selected with the 0.5 threshold on the marginal probabilities of inclusion.

We also looked into comparisons of our results with alternative procedures. A standard approach in Chemometrics is to apply linear or quadratic discriminant analysis on selected principal components. We therefore calculated the principal components of our NIR curves, in the original domain of the data, and computed classical LDA and QDA by selecting different numbers of principal components. We found best results in terms of misclassification rate by using 13-14 principal components with LDA, achieving a total misclassification rate of 5.3%, and 11-13 components with QDA, achieving a total misclassification rate of 6.1%. With a total misclassification rate of 3.4% our method compares

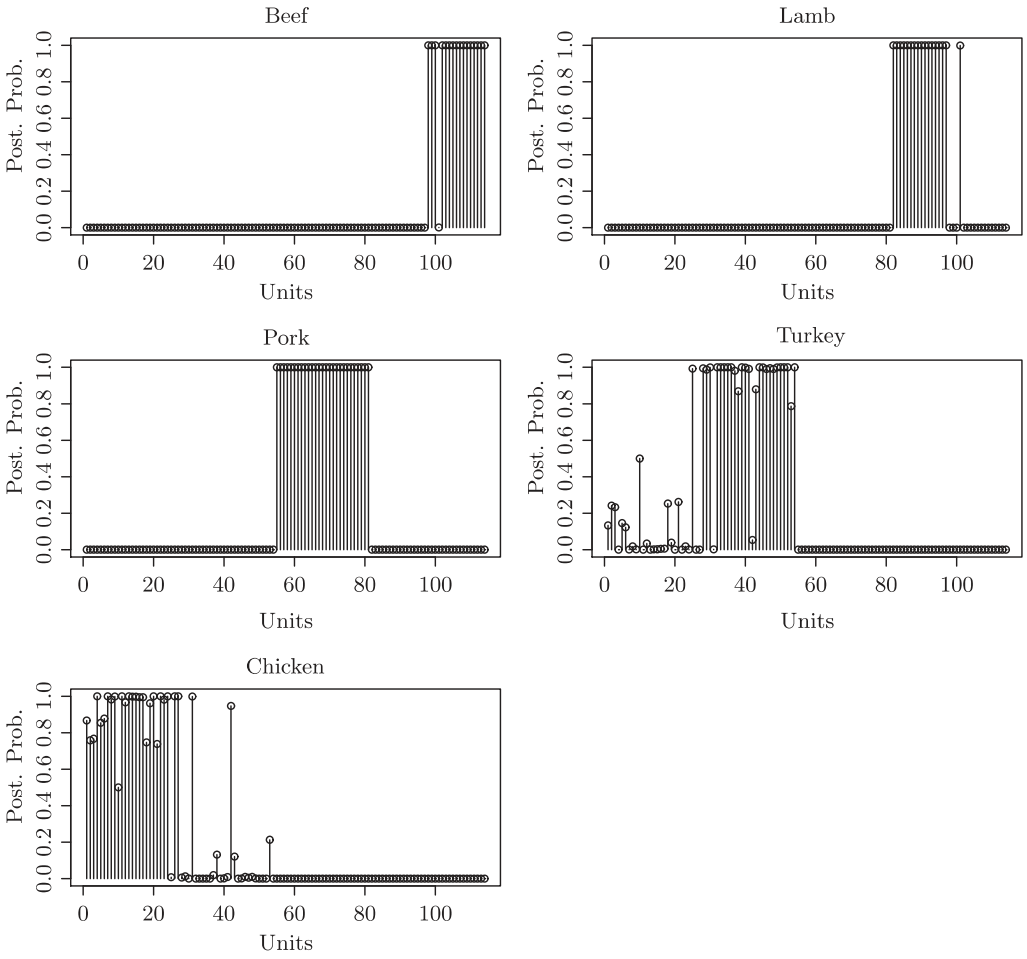


Figure 6. Posterior probabilities of group memberships for the 114 observations in the validation set.

favorably with these more standard techniques. Our method also outperforms the results reported by Murphy, Dean, and Raftery (2010), obtained with their proposed method, and those obtained with competing alternatives as reported by these authors. It needs to be pointed out, however, that these authors performed variable selection by fitting their model to all 231 samples and eliminating at random half of the labels, which they predicted; this procedure was repeated 50 times and an average misclassification rate was then reported. With respect to their results, our method, in particular, obtains a consistently better separation between Turkey and Chicken samples.

We conclude this section by briefly commenting on additional analyses we performed to understand the sensitivity of the model to the specification of the

parameters d and e. In particular, we re-analyzed the NIR data using values of d in the range -3 to -2 (implying a proportion of expected significant wavelet coefficients from 5% to 12%) and values of e in the range 0.6 to 1. A value $e = 0.6$ implies that the prior probability of selecting a wavelet coefficient with all its three neighbors already selected is roughly 4 times the prior probability of a wavelet coefficient that does not have any of its neighbors selected, while with $e = 1$ this ratio is almost 10 when $d = -3$, and between 6 and 7 when $d = -2$. Our method performed well in all four settings. With a threshold of 0.4 for the posterior probabilities our method correctly classified on average 110 of the 114 samples in the validation set. The best predictive power was achieved with $e = 1$ and $d = -2$, with 112 correctly classified samples. The number of selected wavelet coefficients was stable among the different scenarios, with 14 to 17 selected coefficients that mostly overlapped, except for some coefficients that represented the same feature of the data at close locations. As a general behavior, we noticed that, while different settings of the parameters slightly affect the magnitude of the posterior probabilities of inclusion of the wavelet coefficients, their ordering tends to remain largely unaffected.

5. Conclusion

We have put forward a wavelet-based approach to discriminant analysis for curve classification. We have employed wavelet transforms as an effective tool for dimension reduction that reduces curves into wavelet coefficients. We have illustrated how to perform variable selection in the wavelet domain within a Bayesian paradigm for discriminant analysis. We have considered linear and quadratic discriminant analysis and have constructed Markov random field priors that map scale-location connections among wavelet coefficients. Unlike current literature on Bayesian wavelet-based modeling, our model formulation avoids any independence assumption among wavelet coefficients. For posterior inference we have achieved dimension reduction using a stochastic search variable selection procedure that selects the discriminatory wavelet coefficients. We have presented a typical example from chemometrics that deals with the classification of different types of food into species via near infrared spectroscopy. Our method has performed well in comparison with several alternative procedures. As already noticed by Tadesse, Sha and Vannucci (2005), in applications a careful setting of $\mathbf{\Omega}_g$ and k_0 is needed, due to the sensitivity of the model to these hyperparameters.

A possible extension of our model is to allow for a third group of variables in the likelihood factorization (2.5) formed by variables that are marginally independent of the significant ones. The factor of the marginal likelihood corresponding to this third group of variables simplifies in the Metropolis-Hasting steps proposed in Section 3, while new moves that allow variables to be assigned to or

removed from this third set are needed. We have implemented this approach but did not see any significant difference in the selection of the wavelet coefficients and in the corresponding predictions.

The Gaussian assumption is commonly made in applications with the type of spectral datasets we have considered in this paper, see Dean, Murphy, and Downey (2006) and Oliveri et al. (2011), among many others. However, readers may wonder whether the performances of our method depend on this. To investigate this aspect we designed a small simulation study with non-Gaussian data. First we selected one curve $\mathbf{x}^B$ at random among the observed data analyzed in Section 4 and generated a set of n observations belonging to two groups using the following steps. (1) For each data point x_{ij} we added some uniform noise as $x_{ij} = x_j^B + u$, with $u \sim Uniform(-c, c)$, $i = 1, \ldots, n$, $j = 1, \ldots, p$. This implies that the distribution at each data point x_{ij} is uniform in the original data domain. (2) We applied the DTW, using Daubechies wavelets with three vanishing moments, to the generated curves. (3) We randomly selected three wavelet coefficients belonging to intermediate scales. For samples belonging to group 1 we added a constant c_w to these coefficients, while we subtracted the same constant c_w for samples belonging to group 2. (4) For samples belonging to group 1 we added a constant c_n to all neighbors of the three wavelet coefficients selected at (3). We subtracted the same constant c_n for samples belonging to group 2. Since each of the wavelet coefficients selected at (3) had three neighbors in the MRT, there were 12 wavelet coefficients that had discriminatory power among the two groups. We considered three different scenarios. In the first scenario we set $(c, c_w, c_n) = (0.02, 0.06, 0.03)$, in the second $(c, c_w, c_n) = (0.015, 0.02, 0.04)$, and in the third $(c, c_w, c_n) = (0.05, 0.08, 0.04)$. The small values for (c, c_w, c_n) we considered did not alter the original shape of the observed curve $\mathbf{x}^B$. Note that the first two scenarios result in $c_n < 2c < c_w$, while the third one is more challenging, with $c_n < c_w < 2c$. With the same hyperparameter settings adopted in Section 4 for the analysis of the NIR spectral data, in the first scenario a threshold of 0.5 on the marginal posterior probabilities resulted in the selection of all the 12 significant wavelet coefficients without any false positives. In the second scenario the same threshold led to the selection of 11 significant wavelet coefficients without any false positives. In the third scenario 8 coefficient were correctly selected without any false positives.

Acknowledgement

We thank the Editor, an Associate editor and one referee for suggestions that led to a significant improvement of the paper.

Appendix. Linear Discriminant Analysis

If all G groups share the same covariance matrix $\mathbf{\Sigma}$, then the most appropriate technique is linear discriminant analysis (LDA). Using an Inverse-Wishart prior on this matrix, $\mathbf{\Sigma} \sim IW(\delta, \mathbf{\Omega})$, and leaving all the other settings unchanged, the marginal likelihood used in the MCMC algorithm, corresponding to (3.1) for the quadratic case, is

$$\begin{aligned} p(\mathbf{Z}|\mathbf{y}, \boldsymbol{\gamma}) \propto & (k_0 + (\mathbf{Z}_{(\boldsymbol{\gamma}^c)} - \mathbf{1}_{p_{\boldsymbol{\gamma}}} m_{0(\boldsymbol{\gamma}^c)} - \mathbf{Z}_{(\boldsymbol{\gamma})}\boldsymbol{\beta})^T (\mathbf{I}_n + h_0 \mathbf{1}_n \mathbf{1}_n^T + \mathbf{Z}_{(\boldsymbol{\gamma})} \mathbf{H}_{\boldsymbol{\gamma}} \mathbf{Z}_{(\boldsymbol{\gamma})}^T)^{-1} \\ & (\mathbf{Z}_{(\boldsymbol{\gamma}^c)} - \mathbf{1}_{p_{\boldsymbol{\gamma}}} m_{0(\boldsymbol{\gamma}^c)} - \mathbf{Z}_{(\boldsymbol{\gamma})}\boldsymbol{\beta}))^{-(n+\delta)/2} \\ & \mathbf{K}_{(\boldsymbol{\gamma})} |\mathbf{\Omega}_{(\boldsymbol{\gamma})}|^{(\delta + p_{\boldsymbol{\gamma}} - 1)/2} |\mathbf{\Omega}_{(\boldsymbol{\gamma})} + \sum_{g=1}^{G} \mathbf{S}_{g(\boldsymbol{\gamma})}|^{-(n+\delta+p_{\boldsymbol{\gamma}}-1)/2}, \end{aligned}$$

with

$$\mathbf{K}_{(\gamma)} = \big[\prod_{g=1}^{G} (h_1 n_g + 1)^{-p_\gamma/2} \big] \prod_{j=1}^{p_\gamma} \frac{\Gamma((n_g + \delta + p_\gamma - j)/2)}{\Gamma((\delta + p_\gamma - j)/2)}$$

and $S_{g(\gamma)}$ defined as in Section 3. The predictive distribution for the LDA case is a multivariate T-student, see Brown (1993),

$$\mathbf{z}^{\mathbf{f}} - \widetilde{\boldsymbol{\mu}}_g \sim \mathcal{T}(\delta^*, a_g, \mathbf{\Omega}^*),$$

where $\widetilde{\boldsymbol{\mu}}_g = \pi_g \mathbf{m}_g + (1 - \pi_g)\bar{\mathbf{z}}_g$, $\delta^* = \delta + n$, $a_g = 1 + (1/h_g + n_g)^{-1}$, and $\mathbf{\Omega}^* = \mathbf{\Omega} + \mathbf{S} + (h_g + 1/n_g)^{-1}(\bar{\mathbf{Z}} - \mathbf{M})^T(\bar{\mathbf{Z}} - \mathbf{M})$ with $\pi_g = (1 + h_g n_g)^{-1}$ and $\mathbf{S} = (\mathbf{Z} - \mathbf{J}\bar{\mathbf{Z}})^T(\mathbf{Z} - \mathbf{J}\bar{\mathbf{Z}})$; $\mathbf{J}$ consists of G dummy vectors identifying the group of origin of the observation, $\bar{\mathbf{Z}}$ is the $G \times p$ matrix of the sample group means and $\mathbf{M} = (\mathbf{m}_1, \ldots, \mathbf{m}_G)^T$. Note that the part relative to the non-selected variables does not change compared to (3.1). When handling units with missing label, $\mathbf{K}_{(\gamma)}$ can be written as:

$$\mathbf{K}_{(\gamma)} = \big[\prod_{g=1}^{G} w_g^{n_g} (h_1 n_g + 1)^{-p_\gamma/2} \big] \prod_{j=1}^{p_\gamma} \frac{\Gamma((n_g + \delta + p_\gamma - j)/2)}{\Gamma((\delta + p_\gamma - j)/2)},$$

while the predictive distribution remains unchanged after having assigned unlabeled units to groups.

References

Antoniadis, A., Bigot, J. and Sapatinas, T. (2001). Wavelet estimators in nonparametric regression: A comparative simulation study. *J. Statist. Software* **6**, 1-83.

Brown, P. (1993). *Measurement, Regression, and Calibration*. Oxford University Press.

Brown, P., Fearn, T. and Vannucci, M. (2001). Bayesian wavelet regression on curves with application to a spectroscopic calibration problem. *J. Amer. Statist. Assoc.* **96**, 398-408.

Daubechies, I. (1992). *Ten Lectures on Wavelets*. SIAM, Conference Series.

Dawid, A. P. (1981). Some matrix-variate distribution theory: notational considerations and a bayesian application. *Biometrika* **68**, 265-274.

Dean, N., Murphy, T. and Downey, G. (2006). Using unlabelled data to update classification rules with applications in food authenticity studies. *J. Roy. Statist. Soc. Ser. C* **55**, 1-14.

Dobra, A., Jones, B., Hans, C., Nevins, J. and West, M. (2004). Sparse graphical models for exploring gene expression data. *J. Multivariate Anal.* **90**, 196-212.

Donoho, D. and Johnstone, I. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **81**, 425-455.

Downey, G. (1996). Authentication of food and food ingredients by near infrared spectroscopy. *J. Near Infrared Spectroscopy* **4**, 47-61.

Fearn, T., Brown, P. and Besbeas, P. (2002). A Bayesian decision theory approach to variable selection for discrimination. *Statist. Comput.* **12**, 253-260.

George, E. I. and McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *J. Amer. Statist. Assoc.* **88**, 881-889.

George, E. I. and McCulloch, R. E. (1997). Approaches for Bayesian variable selection. *Statist. Sinica* **7**, 339-373.

Gonzalez, R. and Woods, R. (2002). *Digital Image Processing*. Prentice Hall.

He, L. and Carin, L. (2009). Exploiting structure in wavelet-based Bayesian compressive sensing. *IEEE Trans. Signal Process.* **57**, 3488-3497.

Jolliffe, I. (1986). *Principal Component Analysis*. Springer-Verlag.

Kass, R. and Wasserman, L. (1996). The selection of prior distributions by formal rules. *J. Amer. Statist. Assoc.* **91**, 1343-70.

Kim, S., Tadesse, M. and Vannucci, M. (2006). Variable selection in clustering via Dirichlet process mixture models. *Biometika* **93**, 877-893.

Li, F. and Zhang, N. (2010). Bayesian variable selection in structured high-dimensional covariate space with application in Genomics. *J. Amer. Statist. Assoc.* **105**, 1202-14.

Madigan, D. and York, J. (1995). Bayesian graphical models for discrete data. *Internat. Statist. Rev.* **63**, 215-232.

Mallat, S. (1989). Multiresolution approximations and wavelet orthonormal bases of $l^2(\mathbb{R})$'. *Trans. Amer. Math. Soc.* **315**, 69-87.

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

Murphy, T., Dean, N. and Raftery, A. (2010). Variable selection and updating in model-based discriminant analysis for high dimensional data with food authenticity applications. *Ann. Appl. Statist.* **4**, 396-421.

Newton, M., Noueiry, A., Sarkar, D. and Ahlquist, P. (2004). Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics* **5**, 155-176.

Oliveri, P., Di Egidio, V., Woodcock, T. and Downey, G. (2011). Application of class-modelling techniques to near infrared data for food authentication purposes. *Food Chemistry* **125**, 1450-1456.

Propp, J. and Wilson, D. (1996). Exact sampling with coupled markov chains and applications to statistical mechanics. *Random Structures Algorithms* **9**, 223-252.

Ray, S. and Mallick, B. (2006). Functional clustering by Bayesian wavelet methods. *J. Roy. Statist. Soc. Ser. B* **68**, 305-332.

Richardson, S. and Green, P. (1997). On Bayesian analysis of mixtures with an unknown number of components (with discussion). *J. Roy. Statist. Soc. Ser. B* **59**, 731-92.

Sha, N., Vannucci, M., Tadesse, M. G., Brown, P. J., Dragoni, I., Davies, N., Roberts, T. C., Contestabile, A., Salmon, N., Buckley, C. and Falciani, F. (2004). Bayesian variable selection in multinomial probit models to identify molecular signatures of disease stage. *Biometrics* **60**, 812-19.

Shapiro, J. (1993). Embedded image coding using Zeotrees of Wavelet coefficients. *IEEE Trans. Signal Process.* **41**, 3445-3462.

Tadesse, M., Sha, N. and Vannucci, M. (2005). Bayesian variable selection in clustering high-dimensional data. *J. Amer. Statist. Assoc.* **100**, 602-617.

Vannucci, M. and Corradi, F. (1999). Covariance structure of wavelet coefficients: Theory and models in a Bayesian perspective. *J. Roy. Statist. Soc. Ser. B* **61**, 971-986.

Vidakovic, B. (1999). *Statistical Modeling by Wavelets*. Wiley.

Wasserman, L. (2000). Asymptotic inference for mixture models using data-dependent priors. *J. Roy. Statist. Soc. Ser. B* **62**, 159-80.

Department of Biostatistics, University of Texas MD Anderson Cancer Center, Houston, TX 77230-1402, U.S.A.

E-mail: fra.stingo@gmail.com

Department of Statistics, Rice University, Houston, TX 77251, U.S.A.

E-mail: marina@rice.edu

Food Chemistry and Technology Department, Teagasc Food Research Centre, Ashtown, Dublin 15, Ireland.

E-mail: gerard.downey@teagasc.ie

(Received June 2010; accepted May 2011)