Bayesian nonparametric classification for spectroscopy data

Luis Gutiérrez[*], Eduardo Gutiérrez-Peña[†], Ramsés H. Mena[†]

April 12, 2014

Abstract

High-dimensional spectroscopy data are increasingly common in many fields of science. Building classification models in this context is challenging, due not only to high dimensionality but also to high autocorrelations. A two-stage classification strategy is proposed. First, in a data pre-processing step, the dimensionality of the data is reduced using one of two distinct methods. The output of either of these methods is then used to feed a classification procedure that uses a multivariate density estimate from a Bayesian nonparametric mixture model for discrimination purposes. The model employed is based on a random probability measure with decreasing weights. This nonparametric prior is chosen so as to ease the identifiability and label switching problems inherent to these models. This simple and flexible classification strategy is applied to the well-known 'meat' data set. The results are similar or better than previously reported in the literature for the same data.

[*] Escuela de Salud Pública, Fac. de Medicina, U. de Chile, e-mail:luisgutierrez@med.uchile.cl
[†] IIMAS-UNAM, Mexico

Key Words: Discriminant Analysis, Food Authentication, Gaussian Process, Geometric Weights Prior.

1 Introduction

Spectroscopy data are increasingly common in many fields of science. Spectroscopy is the study of the interaction between radiation and matter as a function of wavelength. It measures the reflectance or absorbance values mainly in the visible and near-infrared region of the electromagnetic spectrum. The values of reflectance are produced by vibrations in the chemical bonds in the substance analyzed. Of particular interest in this work are the spectroscopies corresponding to different types of meats. Figure 1 shows $p = 1050$ wavelength points, reporting reflectance measurements for 110 samples of homogenized meat: $n_1 = 55$ samples of chicken (gray lines) and $n_2 = 55$ samples of turkey (green lines). Reflectance measurements were taken in the range 400-2498nm at 2nm intervals. These data were first reported and analyzed by McElhinney et al. (1999) with the purpose of classification between different types of meats in the context of food authentication. Authentication is the process by which food or beverages are verified to match their label description (Winterhalter; 2007).

Indeed, statistical methods for spectroscopy data are appealing in a variety of fields. See, for example, the recent statistical and computational methods for spectroscopy data in Lee and Cox (2010) and Chakraborty (2012) in the context of smoothing spectra and multiple response kernel regression with spectroscopy predictors, respectively. Spectroscopy is the source of information in many biomedical and pharmaceutical research such as cardiovascular radiology, brain imaging, quality/process control and clinical trials, among others. In particular, in the context of food authentication, discriminant analysis methods constitute a key tool (Brown et al.; 1999; Dean et al.; 2006;

Toher et al.; 2007; Gutiérrez et al.; 2011). However, as is evident from Figure 1, these data feature high autocorrelation and non-linearity, hence requiring a robust model able to capture such behaviour (Murphy et al.; 2010).

In a nutshell, high dimensional spectroscopic data are characterized by: 1) The number of samples n is typically much smaller than the number of variables or measurements p ($n << p$); 2) The curve trajectories are nonlinear; 3) The data are evenly spaced as a function of wavelength, usually every 2 nanometers; 4) The data exhibit high positive autocorrelation; 5) Due to this autocorrelation, the curve trajectories tend to be smooth.

There are two common approaches to deal with the high dimensionality of data in classification contexts. The first one is to use a dimension reduction technique such as principal component analysis (Jollife; 1986), and then use only the first p^* components to assign units to groups by means of a classification method for $p^* < n$. The second approach is to first select a subset of p^* variables, useful for discrimination, and then perform the classification with a method for $p^* < n$. In both approaches, the classification method for $p^* < n$ commonly used is linear or quadratic discriminant analysis; see, for example, Fearn et al. (2002); Datta (2008); Murphy et al. (2010); Stingo et al. (2012). However, linear and quadratic discriminant methods are based on the assumption of normality and hence are not always robust.

Motivated by the above data set, generated for food authentication purposes, a Bayesian nonparametric classification approach for spectroscopy data is proposed. The proposal is based on a two-stage procedure: First, a dimension reduction of the spectroscopy curves, from $p >> n$ to $p^* < n$ dimensions, is conducted. Secondly, with the selected p^* variables, a robust version of quadratic discriminant analysis is used. In the first stage, two approaches for dimension reduction are explored: principal component

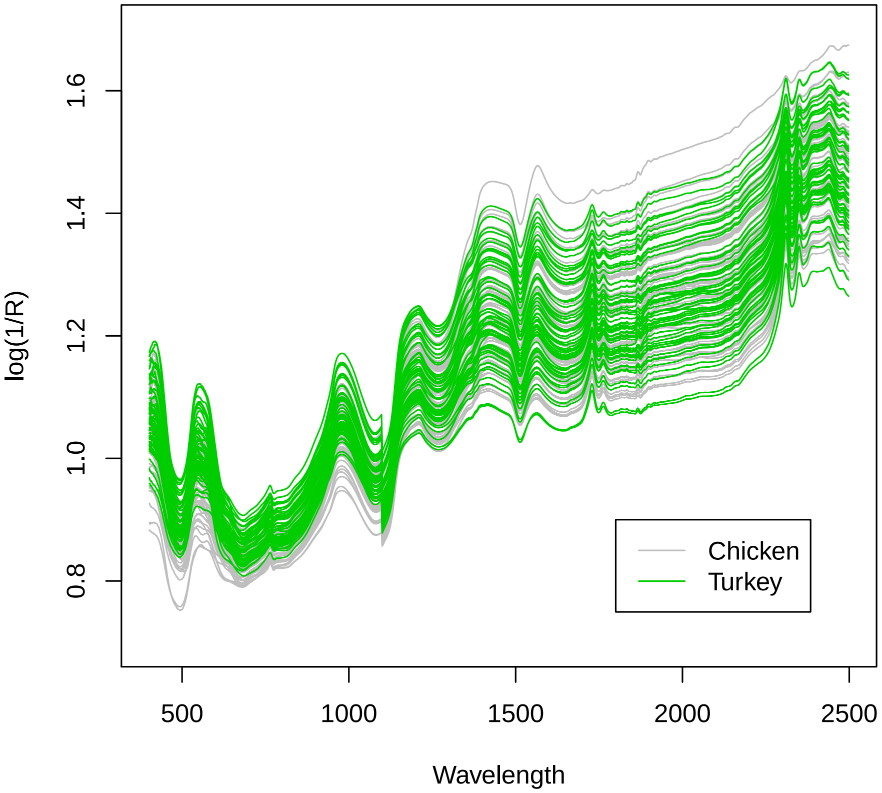


Figure 1: Spectra curves

analysis and a simple variable selection algorithm. The latter is based on a Gaussian process with random effects which jointly models the spectra in order to identify wavelength regions that are informative about group membership.

In stage two, a flexible discrimination procedure based on a multivariate Bayesian nonparametric mixture model with geometric weights is used (Fuentes-García et al.; 2010; Mena et al.; 2011; Mena; 2013). An appealing feature of this nonparametric prior is its ordered weights construction, which leads to a more adequate distribution of the latent allocation variables reducing also some identifiability issues (Mena and Walker;

2013). Furthermore, as discussed in Fuentes-García et al. (2010), such a feature results in better posterior density estimates and therefore in better discrimination results.

Indeed, for one dimensional density estimation problems, this model has proved to be more efficient than stick-breaking based mixture models such as the Dirichlet process mixture (DPM) model (Fuentes-García et al.; 2010). Hence, as a by-product of the present analysis, the efficiency of this model in multivariate settings is validated, which to the best of our knowledge has not been done elsewhere.

When this mixture is based on a Gaussian kernel, the proposed classification model can be seen as a robust generalization of Quadratic Discriminant Analysis (QDA) since it allows for asymmetries and multimodality in the distribution of the responses. Supervised classification based on mixtures of Dirichlet process has been discussed by De la Cruz-Mesía et al. (2007) in a biomedical context, and by Gutiérrez and Quintana (2011) in food authentication. The present proposal has the advantage that it is considerably simpler than other approaches based on finite mixtures or DPMs, and at the same time does not compromise any of the appealing nonparametric features.

This work is organized as follows. In Section 2, a general Bayesian classification approach is presented. This section also develops the classification equations and decision rules employed in the sections that follow. In Section 3, the dimension reduction strategy is described. Section 4 discusses the nonparametric prior distribution employed and also contextualizes it within the flexible multivariate classification model. Section 5 presents an example with a real spectroscopy data set in the context of food authentication. Finally, Section 6 contains some concluding remarks.

2 Bayesian Classifier

In a classification context, it is common to have a training data set comprising n units $\{(\mathbf{y}_i, \mathbf{x}_i, g_i), i = 1, \ldots, n\}$. Here $\mathbf{y}_i = (y_{i1}, \ldots, y_{ip})' \in R^p$ is the observed response vector for the i-th unit (reflectance for spectroscopy data), $\mathbf{x}_i = (x_{i1}, \ldots, x_{iq})' \in R^q$ is a vector of covariates for the i-th unit (wavelengths for spectroscopy data) and g_i denotes the known group label or class for the i-th unit, $g_i \in \{1, \ldots, u\}$. For the meat spectroscopy data set, g_i is the type of meat (Chicken, Turkey, Pork, Beef, Lamb). Let $\mathbf{y}^{(tra)} = (\mathbf{y}_1, \ldots, \mathbf{y}_n, \mathbf{x}_1, \ldots, \mathbf{x}_n, g_1, \ldots, g_n)$ denote the training data set. Let $\mathbf{y}^{(new)} = (\mathbf{y}_{n+1}, \mathbf{x}_{n+1})$ be the observed data vector of a new future unit for which the corresponding label g_{n+1} is unknown. Assuming a predictive approach to classification, the focus is on inference about g_{n+1} i.e. the interest is in estimating $P(g_{n+1} = k \mid \mathbf{y}^{(tra)}, \mathbf{y}^{(new)})$, $k = 1, \ldots, u$. Following De la Cruz-Mesía and Quintana (2007), it is useful to consider an augmented model with marginal prior $P(g_i = k) = \pi_k$ for $k = 1, \ldots, u$. For instance, the π_k's could be taken as the empirical group proportions if these were representative of the relevant population, or they could be taken as $\pi_k = 1/u$ when there is no prior information about these quantities. Let $\boldsymbol{\theta}$ denote the vector of all possible parameters and hyperparameters. The classification probabilities are obtained by weighting the posterior group probabilities, conditional on $\boldsymbol{\theta}$, with respect to the posterior distribution $p(\boldsymbol{\theta} \mid \mathbf{y}^{(tra)})$. Specifically, the classification probability that a new unit $\mathbf{y}^{(new)}$ belongs to the k-th group is

$$P(g_{n+1} = k \mid \mathbf{y}^{(new)}, \mathbf{y}^{(tra)}) = \int \frac{\pi_k p(\mathbf{y}_{n+1} \mid \boldsymbol{\theta}_k, \mathbf{x}_{n+1})}{\sum_{l=1}^{u} \pi_l p(\mathbf{y}_{n+1} \mid \boldsymbol{\theta}_l, \mathbf{x}_{n+1})} p(\boldsymbol{\theta} \mid \mathbf{y}^{(tra)}) d\boldsymbol{\theta}, \tag{1}$$

for details see Gutiérrez et al. (2011). In practice, direct evaluation of (1) is impossible so it is necessary to resort to posterior simulation methods. Given a sample

$\{\boldsymbol{\theta}^{(c)}, c = 1, \ldots, C\}$ from the posterior distribution $p(\boldsymbol{\theta} \mid \mathbf{y}^{(tra)})$, Equation (1) is then approximated by

$$P(g_{n+1} = k|\mathbf{y}^{(new)}, \mathbf{y}^{(tra)}) \approx \frac{1}{C} \sum_{c=1}^{C} \frac{\pi_k p(\mathbf{y}_{n+1}|\boldsymbol{\theta}_k^{(c)}, \mathbf{x}_{n+1})}{\sum_l \pi_l p(\mathbf{y}_{n+1}|\boldsymbol{\theta}_l^{(c)}, \mathbf{x}_{n+1})}. \tag{2}$$

Thus, it is common to classify a future unit, $\mathbf{y}^{(new)}$, using

$$\hat{g}_{n+1} = \arg\max_k P(g_{n+1} = k|\mathbf{y}^{(new)}, \mathbf{y}^{(tra)}), \tag{3}$$

i.e. assigning the label to the category that maximizes the classification probability (Hastie et al.; 2001).

The classification equation (1) is based on a probability model for the response $\mathbf{y}$ given the covariates $\mathbf{x}$ and the unknown parameters $\boldsymbol{\theta}$. Thus, in the above general Bayesian classification approach, the effort needs to focus on an adequate model for the responses, namely $p(\mathbf{y}|\boldsymbol{\theta}_k)$. As pointed out before, for spectroscopy data we need a flexible model able to capture different distributional features. Such will be the case of the nonparametric model described in Section 4.

3 Data pre-processing

The high-dimensional aspects of spectroscopy data demand strategies for dimension reduction. Here two approaches are proposed, namely principal component analysis (PCA) (Jollife; 1986) and a simple variable selection algorithm. PCA is widely used in high-dimensional classification problems. Illustrations of the use of PCA for mass spectrometry data are given in Fearn (2008) and Hoefsloot et al. (2008). Examples of dimension reduction using variable selection as a pre-processing step prior to classification can be found in Datta (2008) and Heidema and Nagelkerke (2008).

3.1 Dimension reduction with variable selection

The variable selection algorithm uses three criteria for identifying and selecting informative wavelengths for classification purposes: (a) A candidate wavelength to be selected should be informative about group membership; (b) The correlation between the selected wavelengths should be as low as possible; (c) The resulting number of variables p^* (wavelengths) should be less than the number of samples ($p^* < n$). Criterion (a) is obvious and necessary. Criterion (b) is desirable, because if the algorithm selects highly correlated wavelengths, the resulting information is redundant and the dimensionality of the multivariate classification model is unnecessarily increased. Criterion (c) is required to get stable estimations of the covariances in the multivariate classification model.

To meet criterion (a), the classification for each wavelength $(1, \ldots, p)$ is obtained separately using a univariate approach (at each wavelength) with a relatively simple model. In principle, such a procedure would demand p univariate classification models (e.g. p univariate LDAs). To avoid this, we propose to build a joint model with common parameters among wavelengths, and then use this model in a univariate way as requested by criterion (a). The simple joint model is given as follows:

Suppose that each data unit (reflectance, wavelengths) $\{\mathbf{y}_i, \mathbf{x}_i : i = 1, \ldots, n\}$ is generated from the model $y_{ij} = f(x_{ij}) + \epsilon_{ij}, \quad j = 1, \ldots, p$, where $f(\cdot)$ is some real-valued function of the wavelengths x_{ij}. With a slight abuse of notation, the model is written as

$$\mathbf{y}_i = f(\mathbf{x}_i) + \boldsymbol{\epsilon}_i,$$

where the $\{\boldsymbol{\epsilon}_i\}$ are independent, normally distributed random vectors. Considering the special characteristics of spectroscopy data (smoothness, high autocorrelation, non-linear trajectories, evenly spaced data points) the proposed joint model is:

$$\begin{aligned}
\mathbf{Y}_{i_{p\times 1}} &= P_q(\mathbf{x}_i) + \delta_{(\mathbf{x}_i)} + \mathbf{Z}_{i_{p\times m}}\mathbf{b}_{i_{m\times 1}} + \boldsymbol{\epsilon}_{i_{p\times 1}}, && (4)\\
\delta_{(\mathbf{x}_i)} &\sim GP(0, \Sigma_{\delta_{(\mathbf{x}_i)}}), && (5)\\
\mathbf{b}_i &\sim N_m(\mathbf{0}, \boldsymbol{\Sigma}_b) \text{ and } \boldsymbol{\Sigma}_b = \sigma_b^2\mathbf{I}_m, &&\\
\boldsymbol{\epsilon}_i &\sim N_p(\mathbf{0}, \sigma^2\mathbf{I}_p),\;\; i = 1, \ldots, n. &&
\end{aligned}$$

Here, $\mathbf{Y}_i \in R^p$ is the response vector, $\mathbf{x}_i \in R^p$ is the vector of evenly spaced wavelengths, $P_q(\mathbf{x}_i)$ is a penalized spline function of wavelength $\mathbf{x}_i$, i.e. $P_q(\mathbf{x}_i) = \mathbf{X}_i\boldsymbol{\beta}$, with $\mathbf{X}_i = \tilde{\mathbf{X}}_i\Phi_q^{-1/2}$ and $\boldsymbol{\beta} = \Phi_q^{1/2}\boldsymbol{u}$. Also, Φ_q is a penalty matrix with (r, s)th entry given by $|\ell_r - \ell_s|^3$, where $\ell_1, \ldots, \ell_q$ are fixed knots, and $\tilde{\mathbf{X}}_i$ is a matrix with $j-$th row $(j = 1, \ldots, p)$ given by $\{|x_{ij} - \ell_1|^3, \ldots, |x_{ij} - \ell_q|^3\}$. Finally, $\boldsymbol{u}$ is a parameter with $E(\boldsymbol{u}) = 0$ and $cov(\boldsymbol{u}) = \sigma_u^2\Phi_q^{-1}$. Thus σ_u^2 controls the degree of penalization of the fixed matrix Φ_q. With this specification, $E(\boldsymbol{\beta}) = 0$ and $cov(\boldsymbol{\beta}) = \boldsymbol{\Sigma}_\beta = \sigma_u^2\mathbf{I}_q$. For further details, see the Bayesian analysis for penalized spline regression described in Crainiceanu et al. (2005). An important concern that arises with splines functions is how to determine the optimal number of knots and their corresponding locations. To simplify these choices, a formulation with two nonparametric functions is proposed. $P_q(\mathbf{x}_i)$ is intended to capture gross features, and thus a relatively small number of knots can be used, choosing their locations to correspond to evenly-spaced percentiles of the wavelengths. In order to capture more refined aspects of $f(\cdot)$, the term $\delta_{(\mathbf{x}_i)}$ which follows a Gaussian process is added (Blight and Ott; 1975). Finally, simple linear adjustments to $P_q(\mathbf{x}_i) + \delta_{(\mathbf{x}_i)}$ are used. These linear adjustments are given by the parameters $\mathbf{b}_i$ with $\mathbf{Z}_i$ its corresponding $p \times m$ design matrix. These parameters are introduced into the model as random effects. Random effects allow for specific features for each curve, and induce a heteroscedastic covariance structure when integrated out of the model. In order to reduce the number of parameters, a simple structure for the covariance matrix of the random effects in (5) is proposed. Other structures could be used, e.g. $\mathbf{\Sigma}_b^{-1} \sim Wish_m(df, \mathbf{S})$, which would be useful when the intercept and slope are expected to be correlated *a priori*.

A key aspect of Gaussian process described in (5) is the variance-covariance matrix $\mathbf{\Sigma}_{\delta(\mathbf{x}_i)}$, whose entries are selected to have the form

$$\sigma_{rs} = \tau^2 \rho^{\|x_{ir} - x_{is}\|^2}, \qquad r, s = 1, 2, \ldots, p,$$

where $\tau^2 > 0$ is a scale parameter and $\rho \in (0, 1)$. The above choice is infinitely differentiable everywhere and thus yields smooth estimates of $f(\cdot)$. The Bayesian formulation of model (4) is completed assuming that

$$\begin{aligned} \boldsymbol{\beta} &\sim N_k(0, \mathbf{\Sigma}_\beta), \\ \sigma^{-2} &\sim Ga(\alpha_0, \gamma_0), \\ \sigma_b^{-2} &\sim Ga(\lambda_0, \delta_0). \end{aligned} \tag{6}$$

Here, Ga denotes the gamma distribution with expected value given by α_0/γ_0. To avoid identifiability problems due to the multiple variance components $(\sigma^2, \mathbf{\Sigma}_b, \mathbf{\Sigma}_{\delta(\mathbf{x}_i)})$, the parametric space of the scale parameter τ^2 can be restricted. One simple strategy consists in fixing τ^2 at a suitable value. Another strategy is to use an informative prior distribution for τ^2; see Section 5 for further details.

Posterior inference for model (4) is based on Gibbs sampling. The analysis is conjugate except for ρ. In order to reduce the serial correlation in the sampling algorithm the parameters are updated in blocks (Chib and Bradley; 1999). Details of the Gibbs sampling algorithm are given in the Appendix.

To obtain the classification for each wavelength using the joint model, equation (4) is used after integrating out the random effects, that is

$$[\,\mathbf{Y}_i \mid \boldsymbol{\beta}, \mathbf{X}_i, \delta_{(\mathbf{x}_i)}, \boldsymbol{\Sigma}_b, \sigma^2\,] \;\sim\; N_p(\boldsymbol{\mu}, \boldsymbol{\Omega}), \tag{7}$$

where $\boldsymbol{\mu} = \mathbf{X}_i\boldsymbol{\beta} + \delta_{(\mathbf{x}_i)}$ and $\boldsymbol{\Omega} = \mathbf{Z}_i\boldsymbol{\Sigma}_b\mathbf{Z}_i^t + \sigma^2\mathbf{I}_p$. Note that $\boldsymbol{\mu}$ and $\boldsymbol{\Omega}$ are invariant across $i = 1, \ldots, n$ because, for spectroscopy data, the wavelengths are the same for all curves. Model (7) provides a reasonable estimation of the mean curve and also it captures possible heteroscedastic covariance structures. The univariate distributions derived from (7) are given by

$$[\,Y_{ij} \mid \boldsymbol{\beta}, \mathbf{X}_i, \delta_{(x_j)}, \Omega_{jj}\,] \sim N_1(\mu_j, \Omega_{jj}), \quad j = 1, \ldots, p, \tag{8}$$

where Ω_{jj}, $j = 1, \ldots, p$ are the elements of the main diagonal of $\boldsymbol{\Omega}$. With (8) and (1), the classification is performed at each wavelength using (3). The identification of informative wavelengths for classification purposes is then straightforward, since for each $j = 1, \ldots, p$ the classification performance is available (that is, the proportion of correctly classified samples) for each category $k = 1, \ldots, u$. To meet criteria (b) and (c), the variable selection algorithm has the following steps:

- Select the wavelengths that attain the best classification for each category or group under consideration.

- Compute the sample correlation ϕ between the wavelengths selected in the previous step and discard those variables that have higher correlation ($\phi > 0.99$, say).

The last step of the algorithm considers the computation of correlations. The Pearson correlation coefficient is proposed for this step. The decision about which wavelength needs to be discarded depends on the specific classification problem. The idea is to keep the wavelengths that give more information for those populations that are more difficult to differentiate. For a specific example, see the discussion in Section 5.1.1.

The final result of this stage is a subset of p^* variables useful for discrimination, which in turn will be the input for the multivariate flexible classification model of stage two.

3.2 Dimension reduction with PCA

A standard approach in Chemometrics is to apply principal component analysis to reduce dimensionality. Given a data matrix $\mathbf{Y}_{n\times p}$, the principal component analysis is carried out by calculating the eigenvalues and the corresponding eigenvectors of the sample covariance matrix $\mathbf{S}_{p\times p}$, which is estimated from the data. Further details concerning PCA can be found in Jollife (1986) and in classical textbooks of multivariate analysis such as Mardia et al. (1979). The key idea behind the PCA is to summarize the data, losing as little information as possible in the process. The first p^* components (linear combinations of the original variables) are chosen to explain a large proportion of the original variability of the data and, in most cases, are subsequently used as input for other statistical methods. In our case, this procedure will produce $p^* < n$ components that we will use in the next section to feed our robust classification model.

4 A multivariate nonparametric mixture model

In this section, a flexible multivariate classification model is proposed. The model uses as input the p^* variables or p^* components selected by means of one of the methods discussed in the previous section.

Classification methods for $p^* < n$ such as LDA and QDA do not allow for asymmetries or multimodality in the distribution of the responses Y. Indeed, the possibilities are varied and unlikely to be captured by a parametric model. One way to overcome this issue is by considering Bayesian nonparametric mixture models (Lo; 1984), i.e. through random densities of the type

$$f(\mathbf{y}) = \int f(\mathbf{y} \mid \boldsymbol{\theta}) P(d\boldsymbol{\theta}), \tag{9}$$

where P denotes a suitably chosen random probability measure. In particular, if P is an almost-surely discrete random probability measure having full support and $f(\cdot \mid \boldsymbol{\theta})$ is a Gaussian density, the random mixture (9) will be able to replicate any possible density with positive probability.

The benchmark model for P is the Dirichlet process (Ferguson; 1973); however, currently there is a vast literature concerning alternative priors, see Lijoi and Prünster (2010) for an up-to-date account. A particularly appealing model, due to its relative simplicity and good performance in density estimation, is the geometric-weights prior, $GWP(a, b)$, introduced by Fuentes-García et al. (2010). This model can be represented as

$$P(\cdot) = \sum_{i=1}^{\infty} \lambda (1-\lambda)^{i-1} \delta_{\boldsymbol{\theta}_i}(\cdot), \tag{10}$$

with $\boldsymbol{\theta}_i \overset{\text{iid}}{\sim} P_0$ independent of $\lambda \sim Be(a,b)$, $a,b > 0$. Here $P_0 := E[P]$ is chosen to be a non-atomic probability measure and $Be(a,b)$ denotes the Beta distribution with mean $a/(a+b)$. P_0 is an important component in the specification of (10) and is sometimes referred to as the prior guess at the shape of P or as the baseline measure.

Model (10) is considerably simpler than the Dirichlet process, since the weights have only one source of randomness, i.e. through λ, instead of an infinite number of Beta random variables. Surprisingly, such a simplification does not compromise the full-support property. Furthermore, another appealing feature of the random probability measure (10) is the ordered-weights property, a feature that allows the underlying MCMC algorithms to converge much faster than those available for other models such as the Dirichlet process. This is due to the fact that having ordered weights reduces the impact of the typical identifiability and label switching issues as seen in Mena and Walker (2013) and deduced from Yao and Lindsay (2009)

A more detailed account of geometric-weights priors can be found in Mena (2013). In what follows, the algorithm for posterior inference is briefly reviewed.

First, notice that the random probability measure (10) can be embedded in a wider class of models given by

$$P(\cdot) = E_\pi \left[\frac{1}{\eta} \sum_{i=1}^{\eta} \delta_{\boldsymbol{\theta}_i} \right],$$

where the expectation is taken with respect to $\eta \sim \pi$ and, as before, $\boldsymbol{\theta}_i \overset{\text{iid}}{\sim} P_0$. Choosing $\pi := \text{Neg-Bin}(2,\lambda)$, $0 < \lambda < 1$, leads back to (10). Therefore, using this representation of the random probability measure (10), the random density (9) can be written as

$$f(\mathbf{y}) = \sum_{\eta=1}^{\infty} \frac{1}{\eta} \sum_{i=1}^{\eta} f(\mathbf{y} \mid \boldsymbol{\theta}_i) \pi(\eta \mid \lambda),$$

where $\pi(\eta \mid \lambda)$ denotes the density of $\pi := \text{Neg-Bin}(2, \lambda)$. As shown in Fuentes-García et al. (2010), posterior inference based on the above representation can be carried out via a Gibbs sampler algorithm based on a slice sampler. Specifically, to avoid the internal summation, a latent variable d (which, given η, indicates the component $f(\cdot \mid \boldsymbol{\theta}_d)$ that better represents the mass at y) can be introduced. With this in mind, a hierarchical representation for the random variable $\mathbf{Y}_i$ modelled through the nonparametric mixture model is given by

$$\begin{aligned}
\mathbf{Y}_i \mid \boldsymbol{\theta}^n, d_i, \eta_i &\overset{\text{ind}}{\sim} f(\mathbf{y}_i \mid \boldsymbol{\theta}_{d_i}),\\
d_i \mid \eta_i &\overset{\text{ind}}{\sim} U\{1, \ldots, \eta_i\},\\
\eta_i &\overset{\text{iid}}{\sim} \text{Neg-Bin}(\cdot \mid 2, \lambda),\\
\lambda &\sim Be(a, b).
\end{aligned}$$

Therefore, for a sample $\mathbf{Y}^{(n)} := (\mathbf{Y}_1, \ldots, \mathbf{Y}_n)$ inference can be performed via a Gibbs sampler algorithm based on the following full conditional distributions

$$\begin{aligned}
f(\boldsymbol{\theta}_j \mid \ldots) &\propto P_0(\boldsymbol{\theta}_j) \prod_{d_i = j} f(\mathbf{y}_i \mid \boldsymbol{\theta}_j), \text{ for } j = 1, \ldots, M \qquad (11)\\
\mathbb{P}(d_i = l \mid \ldots) &\propto f(\mathbf{y}_i \mid \boldsymbol{\theta}_l) \mathbb{I}(l \in \{1, \ldots, \eta_i\}),\\
\mathbb{P}(\eta_i = j \mid \ldots) &= \lambda(1-\lambda)^{j-1} \mathbb{I}(j \geq d_i),\\
f(\lambda \mid \ldots) &= Be(\lambda \mid a + 2n, b + \sum_{i=1}^{n} \eta_i - n),
\end{aligned}$$

for $i = 1, \ldots, n$, where $M = \max\{\eta_1, \ldots, \eta_n\}$. Clearly, when f and P_0 are a conjugate pair the full conditional (11) can be simplified. In particular, within the context of the application in Section 5, f is defined as $f := N_{p^*}(\cdot | \boldsymbol{\theta}_j, \boldsymbol{\Sigma}_j)$ and $P_0 := N_{p^*}(\cdot | \boldsymbol{\theta}_0, \boldsymbol{\Sigma}_{\boldsymbol{\theta}}) IW_{p^*}(\cdot | \nu, \mathbf{A})$, where, IW_{p^*} denotes the inverse-Wishart distribution with expected value given by $E(\mathbf{\Sigma}_j) = (\nu - p^* - 1)^{-1}\mathbf{A}$. Hence, under these considerations the full conditional posterior distributions simplifies as

$$N_{p^*}(\tilde{\boldsymbol{\theta}}_j, \tilde{\mathbf{\Sigma}_{\boldsymbol{\theta}}})\, IW_{p^*}(\kappa, \boldsymbol{\xi}),$$

where

$$\begin{aligned}
\tilde{\mathbf{\Sigma}_{\boldsymbol{\theta}}} &= [n_j \mathbf{\Sigma}_j^{-1} + \mathbf{\Sigma}_{\boldsymbol{\theta}}^{-1}]^{-1}, \qquad \tilde{\boldsymbol{\theta}}_j = \tilde{\mathbf{\Sigma}_{\boldsymbol{\theta}}}[\mathbf{\Sigma}_j^{-1} \sum_{i:d_i=j} \mathbf{y}_i + \mathbf{\Sigma}_{\boldsymbol{\theta}}^{-1}\boldsymbol{\theta}_0], \qquad \kappa = \nu + n_j \text{ and} \\
\boldsymbol{\xi} &= \sum_{i:d_i=j} (\mathbf{y}_i - \boldsymbol{\theta}_j)(\mathbf{y}_i - \boldsymbol{\theta}_j)^t + \mathbf{A}, \quad \text{with } n_j = \sum_{i=1}^{n} \mathbf{1}_{\{d_i=j\}}.
\end{aligned}$$

As pointed out before, this model together with Equation (1) can be regarded as a robust Bayesian nonparametric version of Quadratic Discriminant Analysis.

5 Illustration

In this section the proposed two-stage classification strategy is illustrated in the context of food authentication mentioned in the introduction. A data set previously discussed by McElhinney et al. (1999) is considered. The data consists of reflectance values for $p = 1050$ wavelengths (400-2498nm at 2nm intervals) for five types of homogenized meat (Chicken, $n_1 = 55$; Turkey, $n_2 = 55$; Pork, $n_3 = 55$; Beef, $n_4 = 32$; and Lamb, $n_5 = 34$). The purpose of the analysis is to classify among different types of meat using the spectroscopy information as input. For details about the data collection process see McElhinney et al. (1999).

5.1 Data pre-processing

5.1.1 Dimension reduction with variable selection

Here, the variable selection algorithm of Section 3 is applied to the meat data set. First, model (4) is fitted independently for each type of meat. When a joint model for all meat types is considered, the model must share some parameters between meats, implying that some aspects of the distribution for each meat are the same. Although this assumption reduces the number of parameters, it is not a good strategy in classification problems because, in discriminant analysis, the aim is to discover those aspects of the distribution that highlight the differences between the populations to be classified. A spline term with $q = 11$ knots was employed, choosing the location of the knots to correspond to evenly spaced percentiles of the wavelengths. The random effects $\mathbf{b}_i$ were defined as $\mathbf{b}_i = (b_{0i}, b_{1i})^t$, where b_{0i} is the intercept and b_{1i} is the slope for the i-th curve. The values of the hyperparameters were set to $\mathbf{\Sigma}_\beta = 10^6\mathbf{I}_q$, and $\alpha_0 = \gamma_0 = \lambda_0 = \delta_0 = 10^{-6}$, implying proper but vague prior distributions and representing lack of genuine prior information about the parameters. The (discretized) prior distribution for ρ (see Appendix) is given by a uniform discrete distribution with support on the set $\{0.1, \ldots, 0.9, 0.99, 0.999, 0.9991, 0.9992, \ldots, 0.9999\}$. Different values for τ^2 (0.001, 0.1, 0.25, 1, 5, 10) were tested and it was found that values between 0.1 to 1 worked well. Informative Gamma priors for τ^2 around each of these values were also tried out. Small values of τ^2 generated excessively smooth estimates, while very large values produced rather rough estimates. In summary, and mostly for simplicity, the value of τ^2 was fixed at 0.25. Finally, the values of π_k were fixed at 1/5 for all $k \in \{1, \ldots, 5\}$, that is, the same prior probability for all types of meat.

Figure 2 shows the spectra curves for Chicken together with the posterior mean and standard deviations. The joint model produces a reasonable estimation of the mean curve; also, the heteroscedastic effect is captured by the variance measure.

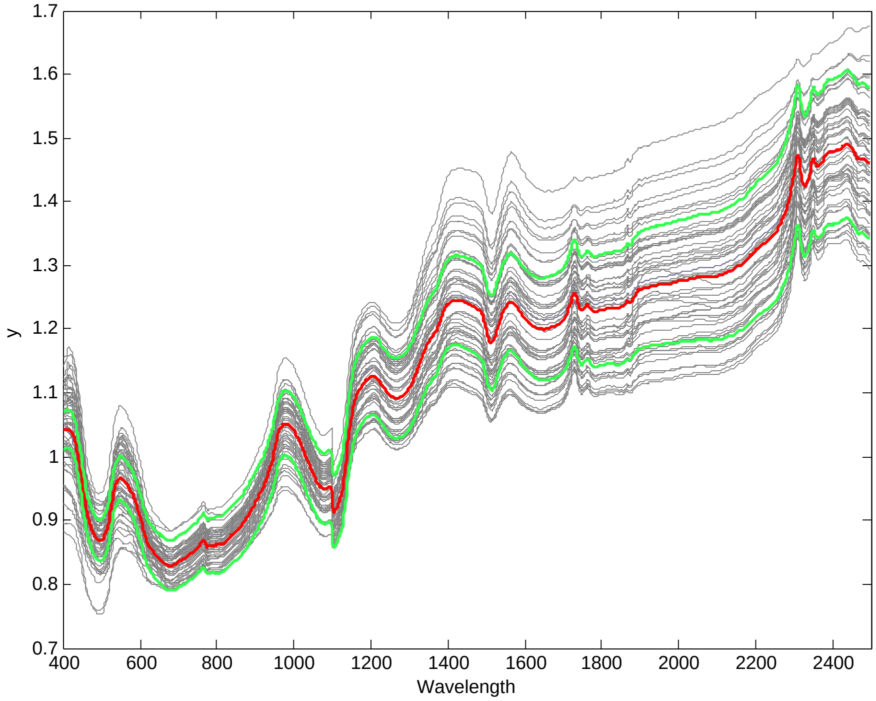


Figure 2: Posterior mean ± standard deviation (Chicken)

The classification for each wavelength was obtained using (8) as well as the classification rule (3), as described in Section 3. Figure 3 shows the classification performance (that is, the proportion of correctly classified samples for each type of meat) for each wavelength. The red arrows indicate the most informative regions for each type of meat. For instance, for Chicken the most informative region for classification is between wavelengths 574 and 584. Similarly, Turkey has an informative region between wavelengths 636 and 642, while the informative region for Pork is between wavelengths 794 and 800. Although there is a high classification rate between wavelengths 1098 and 1100 for Pork, this region corresponds to a sudden jump in the spectra, which may be due to instrumental noise. Finally, informative regions for Beef and Lamb are wavelengths 1056 to 1080, and 636 to 640, respectively.

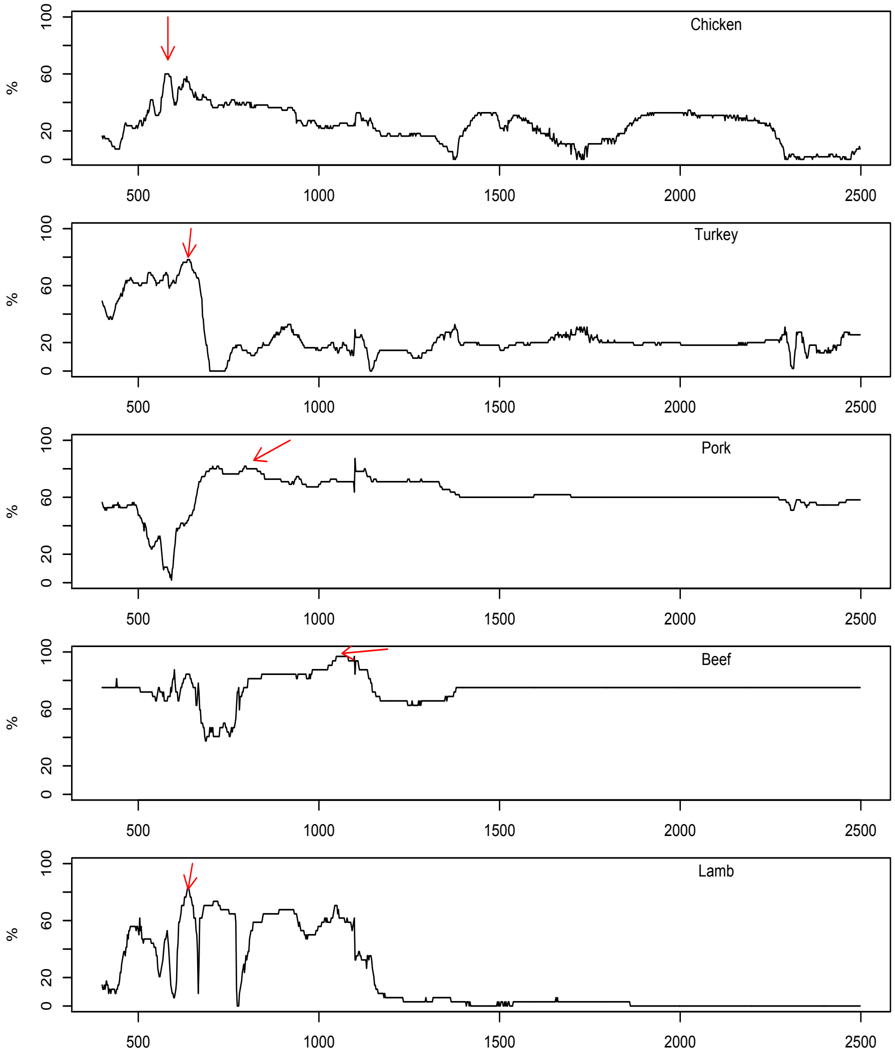


Figure 3: Percentage of meat samples correctly classified at each wavelength. Arrows indicates the most informative regions for each type of meat

Considering the identified regions for each case, the most informative wavelength for each type of meat was selected, that is the wavelength that reached the highest classification rate. The selected wavelengths by type of meat are shown in Table 1.

Category	Wavelength	% Classification
Chicken	w582	60.0%
Turkey	w638	78.2%
Pork	w796	81.8%
Beef	w1056	96.9%
Lamb	w632	82.4%

Table 1: Maximum classification performance by wavelength for each category

For illustrative purposes, the classification using the joint model of equation (7) together with the classification rule (3) was explored. The results are shown in Table 2. The joint model shows good performance for red meats (Beef and Lamb). Note also that white meats are not confused with red meats, but the classification between white meats (Chicken, Turkey and Pork) leaves a lot to be desired. The highest miss-classification rate appears between Chicken and Turkey.

	Chicken	Turkey	Pork	Beef	Lamb	% Class.
Chicken	33	15	7	0	0	60.0%
Turkey	15	40	0	0	0	72.7%
Pork	3	10	42	0	0	76.4%
Beef	0	0	0	32	0	100.0%
Lamb	0	0	0	1	33	97.1 %
% Total Class.						**77.9%**

Table 2: Classification matrix for the joint model

Continuing with the variable selection algorithm, Table 3 shows the correlations between the selected wavelengths of Table 1. Following the algorithm of Section 3.2, wavelength 632 was discarded because its correlation with wavelength 638 is larger than 0.99. Furthermore, from Table 2 it follows that Turkey was more difficult to differentiate from other types of meat compared with Lamb. For this reason, the wavelength that provides classification information for Turkey instead of Lamb was retained. In summary, the selected wavelengths for the next step were: w582, w638, w796 and w1056.

	w582	**w632**	**w638**	**w796**	**w1056**
w582	1.00				
w632	0.96	1.00			
w638	0.96	> 0.99	1.00		
w796	0.76	0.79	0.79	1.00	
w1056	0.76	0.80	0.80	0.94	1.00

Table 3: Correlations

5.1.2 Dimension reduction with PCA

A PCA analysis of the meat data set was performed. The first $p^* = 4$ components explain 99.2% of the total variability of the data. Recall that, when the variable selection algorithm was applied, 4 variables were selected. For the sake of comparison, the first 4 components were retained as input for the classification model. Thus, in both cases (PCA and variable selection) the classification model has four dimensions.

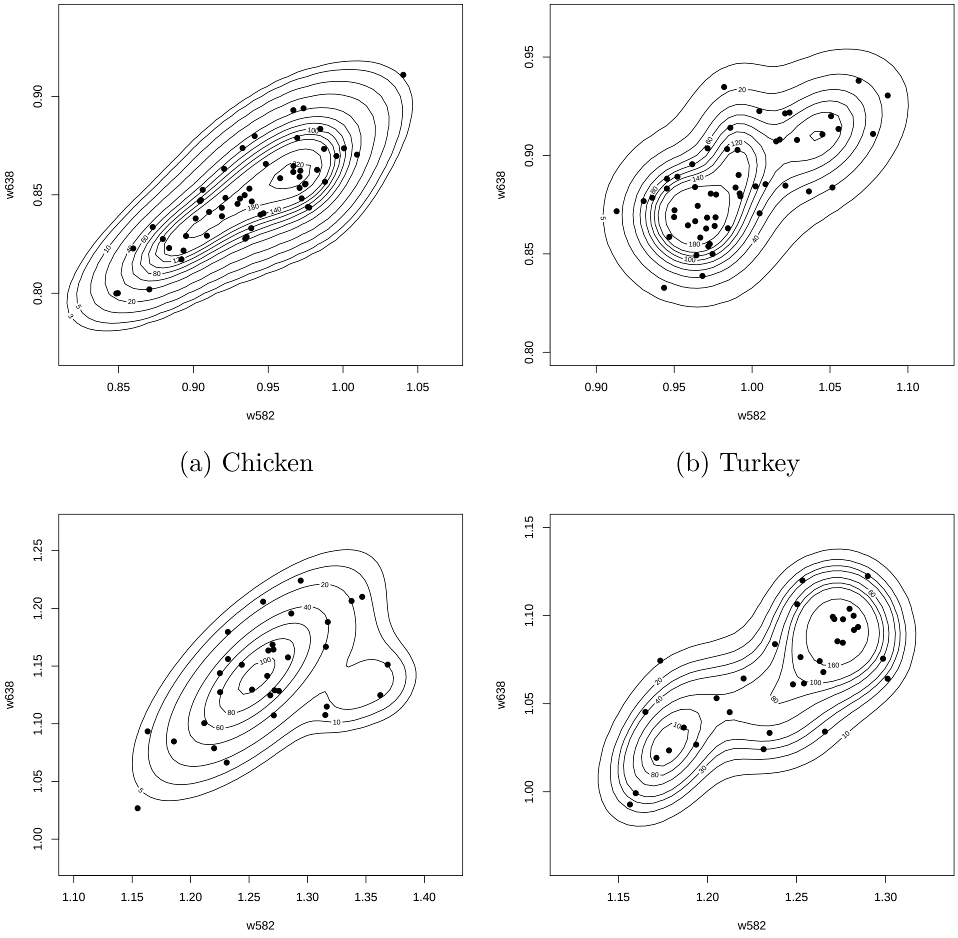
5.2 Model-based classification

5.2.1 Classification based on variable selection

The flexible classification model described in Section 4 was applied to the $p^* = 4$ wavelengths selected in 5.1.1. The values of the hyperparameters were fixed at $\boldsymbol{\theta}_0 = \mathbf{0}$, $\boldsymbol{\Sigma}_{\boldsymbol{\theta}} = 10^5\mathbf{I}_4$, $\mathbf{A} = 0.001\mathbf{I}_4$, $\nu = 8$, $a = b = 1$ and $\pi_k = 1/5$ for all $k \in \{1, \ldots, 5\}$. These hyperparameter values imply proper but vague prior distributions, specially for $\boldsymbol{\theta}$ and λ. The prior expected value for the variance matrix Σ was taken to be diagonal. The

Gibbs sampling algorithm was implemented in R (R Development Core Team; 2012). 10,000 iterations were generated. After a burn-in of 2,000 iterations, samples were collected every 8 iterations so as to obtain uncorrelated samples. Finally, $C = 1,000$ samples were considered for calculating the posterior summaries of interest.

Figure 4 shows the joint posterior predictive densities (for wavelengths w582 and w638) for various types of meat. Note that the nonparametric model with geometric weights is able to capture the multimodality as well as the asymmetries of the data.



(a) Chicken

(b) Turkey

(c) Beef

(d) Lamb

Figure 4: Posterior predictive densities for wavelengths w582 and w638

Similarly, Figure 5 shows the joint posterior predictive densities for wavelengths w582 and w1100. In order to facilitate the comparison, the posterior densities for all five types of meat were included in the same graph. The red meats (Beef and Lamb) are well apart from the white meats. Also, Beef and Lamb are well separated from each other. On the other hand, Chicken and Turkey overlap, while Pork is closer to the Chicken-Turkey group than the red meats. The above configurations are reasonable because the composition of Chicken is similar to that of Turkey (both are poultry) and Pork is more similar to poultry than to red meats. Finally, Lamb and Beef are similar to each other and different from white meats.

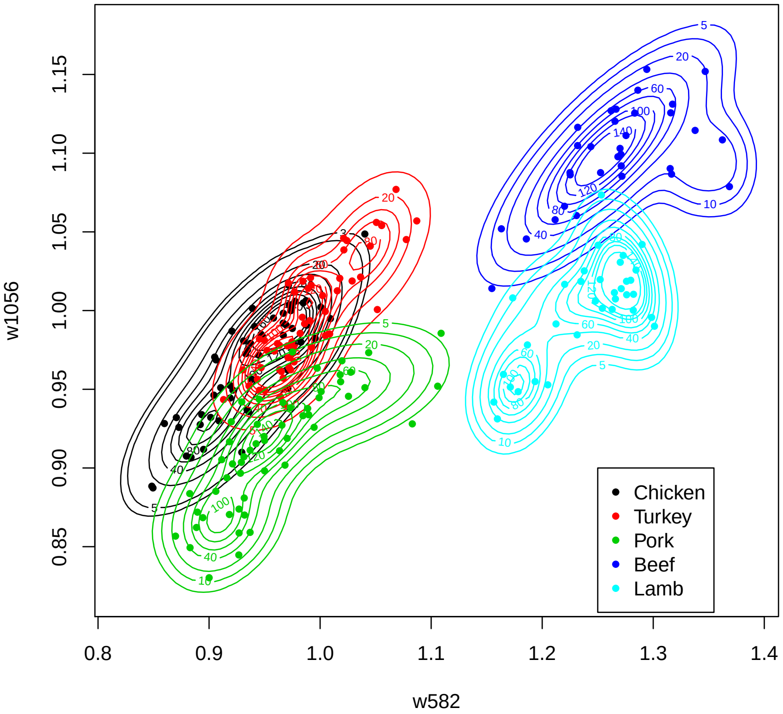


Figure 5: Posterior predictive densities

Table 4 shows the classification obtained by applying the classifier to the same data from which it was computed as well as a leave-one-out cross-validation (LOOCV) approach. The latter values are within brackets. The LOOCV was chosen because the data set is relatively small. For moderately large data sets, a series of random partitions of the data into two components could be considered: one set reserved for deriving the classification rule (the training sample) and the other to assess the rule's performance (the test sample). The estimated classification rate was 93.1% based on the training sample and 90.9% with LOOCV. For the sake of comparison, LDA and QDA were also applied to the same wavelengths. The classification rates were 85.7% (84.4%) for LDA and 88.7% (85.7%) for QDA, based on the training sample and with LOOCV respectively. The flexible classification model performs better with the same data. The improvement is due to its ability to capture the multimodality and the asymmetries which are typical of this kind of data. When the classification results of Tables 4 and 2 are compared, it is possible to conclude that the inclusion of too many variables that are not informative about group membership increases the complexity of the analysis and may degrade the overall classification performance.

	Chicken	**Turkey**	**Pork**	**Beef**	**Lamb**	**% Class rate.**
Chicken	47 (45)	7 (9)	1 (1)	0 (0)	0 (0)	85.5% (81.8%)
Turkey	4 (5)	50 (49)	1 (1)	0 (0)	0 (0)	90.9% (89.1%)
Pork	0 (0)	2 (3)	53 (52)	0 (0)	0 (0)	96.4% (94.5%)
Beef	0 (0)	0 (0)	0 (0)	32 (31)	0 (1)	100% (96.9%)
Lamb	0 (0)	0 (0)	0 (0)	1 (1)	33 (33)	97.1% (97.1%)
% Class rate.						**93.1% (90.9%)**

Table 4: Classification matrix for the flexible model based on variable selection. The values in brackets were obtained using the leave-one-out cross-validation approach

5.2.2 Classification based on dimension reduction with PCA

The flexible classification model described in Section 4 was applied to the four selected principal components. The values of the hyperparameter used here were the same as in Section 5.2.1 except for $\mathbf{A}$ (here $\mathbf{A}$ was fixed to $\mathbf{A} = 0.1\mathbf{I}_4$). Figure 6 shows the joint posterior predictive densities for the first and second principal components. From Figure 6 it can be seen how the model also captures asymmetries and bimodality of information.

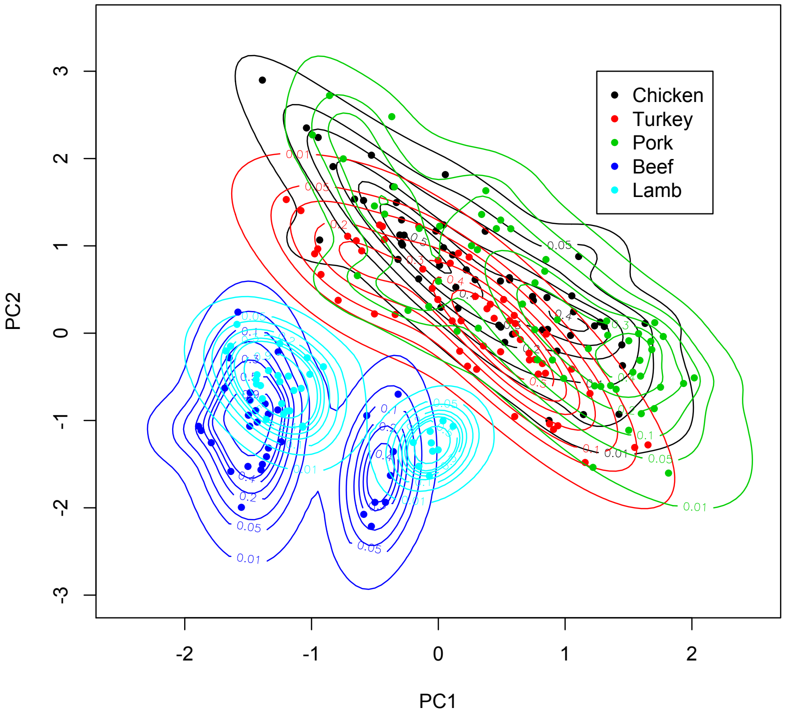


Figure 6: Posterior predictive densities

Table 5 shows the classification of the flexible model based on the four principal components. The results for Pork, Beef and Lamb are very similar to those reported in Table 4 with variable selection. Turkey got better classification here and the Chicken classification rate is lower than that reported in Table 4. For the sake of comparison, LDA and QDA were also applied to the four principal components as is the standard approach in Chemometrics. The overall classification rates were 85.3% (83.9%) for LDA and 87.0% (84.8%) for QDA, based on the training sample and with LOOCV respectively. Again, the proposed flexible model performs better with the same information, the classification rate being 93.9% (87.4%).

	Chicken	**Turkey**	**Pork**	**Beef**	**Lamb**	**% Class rate.**
Chicken	46 (41)	7 (10)	2 (4)	0 (0)	0 (0)	83.6% (74.5%)
Turkey	3 (11)	52 (44)	0 (0)	0 (0)	0 (0)	94.5% (80.0%)
Pork	1 (1)	0 (2)	54 (52)	0 (0)	0 (0)	98.2% (94.5%)
Beef	0 (0)	0 (0)	0 (0)	32 (31)	0 (1)	100% (96.9%)
Lamb	0 (0)	0 (0)	0 (0)	1 (1)	33 (33)	97.1% (97.1%)
% Class rate.						**93.9% (87.4%)**

Table 5: Classification matrix for the flexible model based on PCA. The values in brackets were obtained using the leave-one-out cross-validation approach

6 Discussion

A general classification strategy for high-dimensional spectroscopy data was proposed. The strategy is based on dimension reduction together with a flexible classification model. The results show that the inclusion of variables that are not informative about group membership increases the complexity of the analysis and degrades the classification performance.

Two approaches for dimension reduction were compared: PCA (which is a standard method in Chemometrics) and a simple variable selection algorithm. Both approaches reached similar overall rates of classification. If the user is interested only in classification, then PCA is a reasonable method for dimensionality reduction. On the other hand, if the user is also interested in identifying regions of the electromagnetic spectrum that provide useful information for classification, then the proposed variable selection algorithm is a simple alternative. In either case, a flexible classification model must then be used.

The nonparametric model with geometric weights employed here constitutes a novel alternative to other mixture models widely used in the literature, such as those based on Dirichlet processes or generalizations thereof. In particular, since the model is based on a random probability measure with geometric (i.e. ordered) weights, it results in good and efficient density estimations (Mena; 2013) which come at a relatively small computational cost and without compromising the appealing full-support property of nonparametric priors. Indeed, having good density estimations is of prime importance for Bayesian classifiers. Within the context at issue, such a classification approach can be regarded as a robust Bayesian version of Quadratic Discriminant Analysis. The flexible classification model performs better than LDA and QDA with the same data, and could be useful in other multivariate classification problems.

When these results are compared with other proposals for the same data set, it can be seen that the proposed strategy performs better than or similarly to the proposals of McElhinney et al. (1999) (86.1-92.7%) and Murphy et al. (2010) (90.7%) (Variable selection only). The results of Dean et al. (2006) (94.4%), Murphy et al. (2010) (Variable selection and updating) (93.9%) (variable selection (greedy) and updating) (94.9%) and Stingo et al. (2012) (96.5%) are slightly better than ours, but their proposal considers more complex algorithms for variable selection. In fact, the proposal of Murphy et al. (2010) builds a discriminant rule in a stepwise manner by considering the inclusion of extra variables into the model and also considering removing existing variables as in Raftery and Dean (2006) based on the Bayesian Information Criterion (BIC); the algorithm is iterated until no further variables are added or removed. The proposal of Stingo et al. (2012) is based on wavelet transforms, variable selection and discriminant analysis in the wavelet domain. Our dimension reduction methods are straightforward; PCA is a standard method, and the variable selection algorithm needs to be applied only once because it uses as input the classification of the joint model at each wavelength.

Thus, all the information in the spectra is explored.

Acknowledgements

The authors are grateful to the Associate Editor and to two anonymous referees for their valuable comments which greatly improved the presentation of the paper. The first author was partially funded by Program U-INICIA VID 2011, grant U-INICIA 02/12A, University of Chile. The work of the second author was partially supported by the Sistema Nacional de Investigadores (Mexico) and PAPIIT project IN106114–3. The third author gratefully acknowledges the support by CONACYT project 131179.

7 Appendix

Gibbs sampling algorithm for model (4)

Due to the conditional Gaussian structure of model (4) we have

$$\mathbf{Y}_i \mid \boldsymbol{\beta}, \boldsymbol{\Sigma}_{\delta(\mathbf{x}_i)}, \boldsymbol{\Sigma}_b, \sigma^2 \sim N_p(\mathbf{X}_i\boldsymbol{\beta}, \boldsymbol{\Omega}_1), \tag{12}$$

$$\mathbf{Y}_i \mid \boldsymbol{\beta}, \delta_{(\mathbf{x}_i)}, \boldsymbol{\Sigma}_b, \sigma^2 \sim N_p(\mathbf{X}_i\boldsymbol{\beta} + \delta_{(\mathbf{x}_i)}, \boldsymbol{\Omega}_2), \tag{13}$$

$$\mathbf{Y}_i \mid \boldsymbol{\beta}, \delta_{(\mathbf{x}_i)}, \mathbf{b}_i, \sigma^2 \sim N_p(\mathbf{X}_i\boldsymbol{\beta} + \delta_{(\mathbf{x}_i)} + \mathbf{Z}_i\mathbf{b}_i, \boldsymbol{\Omega}_3), \tag{14}$$

where $\boldsymbol{\Omega}_1 = \boldsymbol{\Sigma}_{\delta(\mathbf{x}_i)} + \mathbf{Z}_i\boldsymbol{\Sigma}_b\mathbf{Z}_i^t + \sigma^2\mathbf{I}_p$, $\boldsymbol{\Omega}_2 = \mathbf{Z}_i\boldsymbol{\Sigma}_b\mathbf{Z}_i^t + \sigma^2\mathbf{I}_p$ and $\boldsymbol{\Omega}_3 = \sigma^2\mathbf{I}_p$. Following Chib and Bradley (1999) we update $\boldsymbol{\beta}$, $\boldsymbol{\delta}$ and $\{\mathbf{b}_i\}$ in one block as follows:

Algorithm 1

1. Sample $\boldsymbol{\beta}$, $\boldsymbol{\delta}$ and $\{\mathbf{b}_i\}$ from $[\boldsymbol{\beta}, \boldsymbol{\delta}, \{\mathbf{b}_i\} \mid y, \sigma^2, \boldsymbol{\Sigma}_b, \boldsymbol{\Sigma}_{\delta(\mathbf{x}_i)}]$ by sampling

   (a) $\boldsymbol{\beta}$ from $[\boldsymbol{\beta} \mid \mathbf{y}, \sigma^2, \boldsymbol{\Sigma}_b, \boldsymbol{\Sigma}_{\delta(\mathbf{x}_i)}]$

(b) $\boldsymbol{\delta}$ from $[\boldsymbol{\delta} \mid \mathbf{y}, \boldsymbol{\beta}, \sigma^2, \mathbf{\Sigma}_b, \mathbf{\Sigma}_{\delta(\mathbf{x}_i)}]$

(c) $\mathbf{b}$ from $[\mathbf{b} \mid \mathbf{y}, \boldsymbol{\beta}, \boldsymbol{\delta}, \sigma^2, \mathbf{\Sigma}_b, \mathbf{\Sigma}_{\delta(\mathbf{x}_i)}]$

2. Sample σ_b^2 from $[\sigma_b^2 \mid \mathbf{y}, \boldsymbol{\beta}, \boldsymbol{\delta}, \mathbf{b}, \sigma^2]$

3. Sample σ^2 from $[\sigma^2 \mid \mathbf{y}, \boldsymbol{\beta}, \boldsymbol{\delta}, \mathbf{b}, \sigma_b^2]$

4. Sample ρ from $[\rho \mid \mathbf{y}, \boldsymbol{\beta}, \boldsymbol{\delta}, \mathbf{b}, \sigma^2, \sigma_b^2]$

5. Repeat Steps 1-4 using the most recent values of the conditioning variables.

Using (12), (13) and (14), together with the prior distributions (6), we find that the full conditional posterior distributions for step (1) of Algorithm 1 are given by

$$\boldsymbol{\beta} \mid \mathbf{y}, \sigma^2, \mathbf{\Sigma}_b, \mathbf{\Sigma}_{\delta(\mathbf{x}_i)} \sim N_q(\tilde{\boldsymbol{\beta}}, \tilde{\mathbf{\Sigma}}_\beta),$$

where $\tilde{\mathbf{\Sigma}}_\beta = [\sum_{i=1}^n \mathbf{X}_i^t \mathbf{\Omega}_1^{-1} \mathbf{X}_i + \mathbf{\Sigma}_\beta^{-1}]^{-1}$ and $\tilde{\boldsymbol{\beta}} = \tilde{\mathbf{\Sigma}}_\beta[\sum_{i=1}^n X_i^t \mathbf{\Omega}_1^{-1} \mathbf{Y}_i]$;

$$\boldsymbol{\delta} \mid \mathbf{y}, \boldsymbol{\beta}, \sigma^2, \mathbf{\Sigma}_b, \mathbf{\Sigma}_{\delta(\mathbf{x}_i)} \sim N_p(\tilde{\boldsymbol{\delta}}, \tilde{\mathbf{\Sigma}}_\delta),$$

where $\tilde{\mathbf{\Sigma}}_\delta = [n\mathbf{\Omega}_2^{-1} + R^{-1}]^{-1}$, $R = \{\tau^2 \rho^{\|\mathbf{x}_i - x_j\|}\}$ and $\tilde{\boldsymbol{\delta}} = \tilde{\mathbf{\Sigma}}_\delta[\sum_{i=1}^n \mathbf{\Omega}_2^{-1}(\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta})]$;

$$\mathbf{b}_i \mid \mathbf{y}, \boldsymbol{\beta}, \boldsymbol{\delta}, \sigma^2, \mathbf{\Sigma}_b, \mathbf{\Sigma}_{\delta(\mathbf{x}_i)} \sim N(\tilde{\mathbf{b}}_i, \tilde{\mathbf{\Sigma}}_{\mathbf{b}_i}),$$

where $\tilde{\mathbf{\Sigma}}_{\mathbf{b}_i} = [\mathbf{Z}_i^t \mathbf{\Omega}_3^{-1} \mathbf{Z}_i + \mathbf{\Sigma}_b^{-1}]^{-1}$ and $\tilde{\mathbf{b}}_i = \tilde{\mathbf{\Sigma}}_{\mathbf{b}_i}[\mathbf{\Omega}_3^{-1}(\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta} - \boldsymbol{\delta})]$. The full conditionals for steps 2 and 3 are given by

$$\sigma^{-2} \mid \mathbf{y}, \boldsymbol{\beta}, \boldsymbol{\delta}, \mathbf{b}, \sigma_b^2 \;\sim\; Ga\left(\frac{1}{2}np + \alpha_0, \frac{1}{2}\sum_{i=1}^n (\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta} - \boldsymbol{\delta} - \mathbf{Z}_i\mathbf{b}_i) + \gamma_0\right),$$

and

$$\sigma_b^{-2} \mid \mathbf{y}, \boldsymbol{\beta}, \boldsymbol{\delta}, \mathbf{b}, \sigma^2 \quad \sim \quad Ga\left(\frac{1}{2}nq + \lambda_0, \frac{1}{2}\sum_{i=1}^{n} \mathbf{b}_i^t \mathbf{b}_i + \delta_0\right).$$

Generating ρ from $[\rho \mid \mathbf{y}, \boldsymbol{\beta}, \boldsymbol{\delta}, \mathbf{b}, \sigma^2, \sigma_b^2] = [\rho \mid \boldsymbol{\delta}, \tau^2]$ (step 4) is quite difficult because the form of $[\rho \mid \boldsymbol{\delta}, \tau^2]$ can change dramatically from one iteration to the next, and each iteration involves inverting and calculating the determinant of a $p \times p$ matrix. A simple, yet useful approximation is obtained by discretizing $[\rho \mid \boldsymbol{\delta}, \tau^2]$ over the $(0, 1)$ interval (Gutiérrez-Peña and Smith; 1998). This approximation is appropriate for bounded parameter spaces. With the above strategy the computational effort is significantly reduced. A recent efficient approach for Gaussian process estimation can be found in Banerjee et al. (2013).

References

Banerjee, A., Dunson, D. and Tokdar, S. (2013). Efficient Gaussian process regression for large datasets, *Biometrika* . DOI:10.1093/biomet/ass068.

Blight, B. and Ott, L. (1975). A Bayesian approach to model inadequacy for polynomial regression, *Biometrika* **62**: 79–88.

Brown, P., Fearn, T. and Haque, M. (1999). Discrimination with many variables, *Journal of the American Statistical Association* **94**: 1320–1329.

Chakraborty, S. (2012). Bayesian multiple response kernel regression model for high dimensional data and its practical applications in near infrared spectroscopy, *Computational Statistics and Data Analysis* **56**: 2742–2755.

Chib, S. and Bradley, C. (1999). On MCMC sampling in hierarchical longitudinal models, *Statistics and Computing* **9**: 17–26.

Crainiceanu, C., Ruppert, D. and Wand, M. (2005). Bayesian analysis for penalized splines regression using WinBUGS, *Journal of Statistical Software* **14**(14): 1–24.

Datta, S. (2008). Classification of breast cancer versus normal samples from mass spectrometry profiles using linear discriminant analysis of important feautres selected by random forest, *Statistical Applications in Genetics and Molecular Biology* **7**(2): Article 7.

De la Cruz-Mesía, R. and Quintana, F. (2007). A model-based approach to Bayesian classification with applications to predicting pregnancy outcomes from longitudinal, *Biostatistics* **8**: 228–238.

De la Cruz-Mesía, R., Quintana, F. and Müller, P. (2007). Semiparametric Bayesian classification with longitudinal markers, *Journal of the Royal Statistical Society, Series C* **56**(2): 119–137.

Dean, N., Murphy, T. and Downey, G. (2006). Using unlabelled data to update classification rules with applications in food authenticity studies, *Journal of the Royal Statistics Society. Series C. Applied Statistics* **55**(1): 1–14.

Fearn, T. (2008). Principal component discriminant analysis, *Statistical Applications in Genetics and Molecular Biology* **7**(2): Article 6.

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

Ferguson, T. (1973). A Bayesian analysis of some nonparametric problems, *The Annals of Statistics* **1**(2): 209–230.

Fuentes-García, R., Mena, R. and Walker, S. (2010). A new Bayesian nonparametric mixture model, *Communications in Statistics- Simulation and Computation* **39**: 669–682.

Gutiérrez, L. and Quintana, F. (2011). Multivariate Bayesian semiparametric models for authentication of food and beverages, *Annals of Applied Statistics* **5**(4): 2385–2402.

Gutiérrez, L., Quintana, F., von Baer, D. and Mardones, C. (2011). Multivariate Bayesian discrimination for varietal authentication of Chilean red wine, *Journal of Applied Statistics* **38**(10): 2099–2109.

Gutiérrez-Peña, E. and Smith, A. (1998). Aspects of smoothing and model inadequacy in generalized regression, *Journal of statistical planing and inference* **67**: 273–286.

Hastie, T., Tibshirani, R. and Friedman, J. (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer, New York.

Heidema, A. and Nagelkerke, N. (2008). Developing a discrimination rule between breast cancer patients and controls using proteomic mass spectrometry data: a three-step approach, *Statistical Applications in Genetics and Molecular Biology* **7**(2): Article 5.

Hoefsloot, H., Smit, S. and Smilde, A. (2008). A classification model for the Lieden proteomics competition, *Statistical Applications in Genetics and Molecular Biology* **7**(2): Article 8.

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

Lee, J. and Cox, D. D. (2010). Robust smoothing: Smoothing parameter selection and applications to fluorescence spectroscopy, *Computational Statistics and Data Analysis* **54**: 3131–3143.

Lijoi, A. and Prünster, I. (2010). Models beyond the dirichlet process., *in* N. Hjort, C. Holmes, P. Müller and S. G. Walker (eds), *Bayesian Nonparametrics*, Cambridge Univ. Press., pp. 80–136.

Lo, A. (1984). On a class of Bayesian nonparametric estimates I. Density estimates., *Ann. Statist.* **12**: 351–357.

Mardia, K., Kent, J. and Bibby, J. (1979). *Multivariate analysis*, Academic Press.

McElhinney, J., Downey, G. and Fearn, T. (1999). Chemometric processing of visible and near infrared reflectance spectra for species identification in selected raw homogenized meats, *Journal of Near Infrared Spectroscopy* **7**: 145–154.

Mena, R. H. (2013). Geometric weight priors and their applications, *in* P. Damien, P. Dellaportas, N. G. Polson and D. A. Stephens (eds), *Bayesian Theory and Applications*, Oxford Univ. Press., pp. 271–296.

Mena, R., Ruggiero, M. and Walker, S. (2011). Geometric stick-breaking processes for continuos-time Bayesian nonparametric modeling, *Journal of Statistical Planning and Inference* **141**: 3217–3230.

Mena, R. and Walker, S. (2013). On the Bayesian mixture model and identifiability, *Submitted manuscript* .

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, *The Annals of Applied Statistics* **4**(1): 396–421.

R Development Core Team (2012). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
URL: *http://www.R-project.org/*

Raftery, A. and Dean, N. (2006). Variable selection for model-based clustering, *Journal of the American Statistical Association* **101**(473): 168–178.

Stingo, F., Vannucci, M. and Downey, G. (2012). Bayesian wavelet-based curve classification via discriminant analysis with markov random tree priors, *Statistica Sinica* **22**: 465–488.

Toher, D., Downey, G. and Brendan, T. (2007). A comparison of model-based and regression classification techniques applied to near infrared spectroscopic data in food authentication studies, *Chemometrics and Intelligent Laboratory Systems* **89**: 102–115.

Winterhalter, P. (2007). *Authentification of food and wine*, ACS SYMPOSIUM SERIES 952. American Chemical Society Washington DC.

Yao, W. and Lindsay, B. (2009). Bayesian mixture labeling by highest posterior density, *J. Am. Stat. Assoc.* **104**: 758–767.