MULTIVARIATE BAYESIAN METHODS FOR AUTHENTICATION OF FOOD AND BEVERAGES

By

Luis Alberto Gutiérrez Inostroza

SUBMITTED IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
AT
PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE
SANTIAGO, CHILE
JANUARY 2011

PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE
DEPARTMENT OF
STATISTICS

The undersigned hereby certify that they have read and recommend to the Faculty of Mathematics for acceptance a thesis entitled "**Multivariate Bayesian Methods For Authentication of Food and Beverages**" by **Luis Alberto Gutiérrez Inostroza** in partial fulfillment of the requirements for the degree of **Doctor of Philosophy**.

Dated: January 2011

External Examiner: _____
Dietrich von Baer

Research Supervisor: _____
Fernando Quintana

Examing Committee: _____
Alejandro Jara

_____
Rolando De La Cruz

PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE

Date: **January 2011**

Author: **Luis Alberto Gutiérrez Inostroza**

Title: **Multivariate Bayesian Methods For Authentication of Food and Beverages**

Department: **Statistics**

Degree: **Ph.D.** Convocation: **January** Year: **2011**

Permission is herewith granted to Pontificia Universidad Católica de Chile to circulate and to have copied for non-commercial purposes, at its discretion, the above title upon the request of individuals or institutions.

Signature of Author

To Paola and Emilio.

Table of Contents

Abstract

Food and beverage authentication is the process where food or beverages are verified as complying with their label description. From the viewpoint of consumers' acquisition, the mislabeling of foods represents a commercial fraud. Authentication is important for foods and beverages with high commercial value, like honey, wines or olive oils, since their prices depend on their quality, variety or origin. Then, it could be possible that these products will be mixed with similar of lower quality substances to get a better price. Misleading labeling might also have negative health implications, especially when food have not declared allergenic compounds.

The common way to deal with an authentication process is to measure a number of attributes on samples of food and then use these as input for a classification problem. In this context, the present thesis proposes multivariate hierarchical models, parametric and semiparametric; these models are based on fixed and random effects in order to model the mean response and different covariance matrices for each category to be classified. The semiparametric model has the advantage of not having to assume any parametric form, which may be particularly difficult to check in multivariate cases. Furthermore, the model is formulated under the formalism of dependent random probability measures for increasing its flexibility.

In many authentication applications there may be several types of measurable attributes. Then, an important problem consists of determining which of these would provide the best information, in the sense of achieving the highest possible classification accuracy at the lowest cost. We approach the problem under a decision theoretic

strategy. We adapted and applied two approaches for taking optimal decisions proposed in a biomedical context, in order to solve the problem of selecting optimal information.

The proposed models and methodology were applied to a dataset consisting of concentration measurements of a number of chemical markers in samples of Chilean red wines. The dataset includes determinations of nine Anthocyanins on 399 wine samples, of which 228 were declared by the producers as Cabernet Sauvignon, 76 as Merlot and 95 samples as Carménère. The data set also includes determinations of six Flavonols and four Organic acids, on 149 samples for which the anthocyanin were also determined. The grape varieties in this subset were Cabernet Sauvignon (101 samples), Merlot (19 samples) and Carménère (29 samples). All wine samples has registered its valley and vintage. In the case of the semiparametric proposal, the model was applied to a simulated dataset too.

Acknowledgements

I would like to thank Fernando Quintana, my supervisor, for his many suggestions and constant support during this research. I am also thankful to Dietrich von Baer for his guidance through my early years in food authenticity studies.

Professor Glenn Hofmann expressed his interest in my work and he gave me the first lessons of Statistics and modeling. He also encouraged me to study a Ph.D in Statistics.

The *CONICYT Scholarship*, which was awarded to me for the period 2006–2009, was crucial to the successful completion of this project.

Of course, I am grateful to my wife Paola and my son Emilio for their patience and *love*. Without them this work would never have come into existence (literally).

Finally, I wish to thank the following friends: Guillermo, Felipe, Manuel, Ricardo, Juan and Mauricio.

Santiago, Chile
January 1, 2011

Luis Gutiérrez Inostroza

Chapter 1

Introduction

Consumers increasingly demand reassurance of the origin and content of their food and beverages. The process through which food or beverages are verified as complying with its label description is called food authentication (Winterhalter; 2007). From the viewpoint of consumers' acquisition, the mislabeling of foods represents commercial fraud (Mafra et al.; 2008). Food authentication is important for foods and beverages of high commercial value, like honey, wines or olive oil, because their prices depend of their quality, variety or origin. It is then important to uncover unscrupulous sellers who decide to increase their profit by adulterating these products with similar but lower quality substances. Misleading labeling might also have negative health implications, especially when the food has undeclared allergenic compounds.

Because of the growing demand from consumers of clarity and certainty in food origins and contents, the importance of food authentication has substantially increased in recent years. The wine industry has been using the authentication procedure for a long time. Substantial research efforts have been put into this particular topic. Chilean wine represents an important part of Chile's worldwide exports, which have increased from 52 to 1,256 million U.S. dollars over the period 1997-2007. von Baer

et al. (2005) report that some containers of Chilean red wine have been rejected in Germany because they did not satisfy the parameters applied there to verify wine varieties. These problems have a direct impact on producers and their income. The main red wine varieties produced in Chile are Merlot, Carménère and Cabernet Sauvignon. Therefore, it is important for sustainable long-term growth to develop a reliable system to verify product authenticity. In this sense, various authors have proposed to differentiate among red wine varieties using their anthocyanin profiles (Eder et al.; 1994; Holbach et al.; 1997; Berente et al.; 2000; Holbach et al.; 2001; Otteneder et al.; 2002, 2004; von Baer et al.; 2005; Revilla et al.; 2001; von Baer et al.; 2007). Anthocyanins are a group of chemical compounds present in red wine, which confer to this beverage its characteristic red color and are transferred from the grape skins to wine during the winemaking process.

Holbach et al. (2001) and von Baer et al. (2007) additionally proposed combining anthocyanin profiles with shikimic acid concentrations to differentiate between red wine varieties. Fischerleitner et al. (2005) concluded that among Austrian wines, Cabernet Sauvignon is the only variety that can be completely identified by its shikimic acid content. The reason for this is that Cabernet Sauvignon concentrations are far above those for other Austrian varieties. However, most authors consider only simple relations between these compounds. The method approved by the International Organization of Vine and Wine OIV in 2003 is also based on this principle (OIV; 2003). More sophisticated exploratory statistical methods for classification purposes, based on anthocyanin profiles, have been proposed by Berente et al. (2000), Otteneder et al. (2002), von Baer et al. (2005), de Villiers et al. (2005), and von Baer

et al. (2007). Linear discriminant analysis and some variations of this methods (forward or backward selection) have been used by de Villiers et al. (2005) and Aleixandre et al. (2002). Other approaches include neural networks (Beltrán et al.; 2005; Kruzlicova et al.; 2009) and similarity index based on mid-infrared spectroscopy data (Bevin et al.; 2006).

Probabilistic modeling for discrimination and authentication purposes was proposed by Brown et al. (1999), who used Bayesian methods to discriminate 39 microbiological taxa using their reflectance spectra. In the special case of longitudinal data analysis, Bayesian discrimination has been discussed and used by Brown et al. (2001), De la Cruz-Mesía and Quintana (2007), De la Cruz et al. (2007b), De la Cruz (2008) and De la Cruz et al. (2008b). Binder (1978) describes a general class of normal-mixture models, discussing some aspects of the use of such models for Bayesian classification, clustering and discrimination. Mixture models are extensively reviewed in McClachlan and Peel (2000). Lavine and West (1992) describe Bayesian methods for classification and discrimination using Gibbs sampling. Mallick et al. (2005) discussed Bayesian classification using gene expression data, concluding from their comparison with other methods, that the Bayesian classification approach performed better than other popular alternatives. Rigby (1997) carries out a thorough comparison between Bayesian and classical estimates of P, the probability that a new observation belongs to one of two multivariate normal populations with equal covariance matrices. The conclusion was that Bayesian methods generally provide less extreme and more reliable estimates of P. Similar conclusions were found by Brown et al. (1999) when comparing Bayesian classification methods with classical alternatives such as linear or

quadratic discriminant analysis. More recently, Agrawal et al. (2009) consider an incremental framework for feature selection and Bayesian classification for multivariate normal groups. In the present thesis we propose a model-based classification approach in order to verify that a food matches with its label description. The problem of optimal information selection in an authentication process is also addressed.

1.1 The Motivating Dataset

We consider a dataset consisting of concentration measurements of a number of chemical markers in samples of Chilean red wines. The dataset includes determinations of Anthocyanins, Organic acids and Flavonols. All wine samples came directly from wineries and include the grape variety as declared by the producer, the year of harvest and the geographic origin or valley. Anthocyanins are a group of chemical compounds present on the grape skins. They are transferred to the wine during the winemaking process and confer to this beverages its characteristic red color. The dataset includes measurements of nine anthocyanins (listed in Table 1.1) on 399 wine samples, of which 228 were declared by the producers as Cabernet Sauvignon, 76 as Merlot and 95 samples as Carménère. The vintages included in the anthocyanin determinations were 2001-2004. The valleys included in the anthocyanins determination sorted from north to south of Chile are: Aconcagua, Maipo, Rapel, Curicó, Maule, Itata and Bío-Bío. The Valleys range from 33 to 38 degrees latitude south, and provide a wide range of soil types and weather conditions. Vinification was made at production scale and samples were taken after malolactic fermentation, but before blending. Anthocyanin determination was made by reverse phase HPLC (High Performance Liquid Chromatography) based on method described by Holbach et al. (1997), Otteneder et al.

(2002) and OIV (2003) with minor modifications. More details about anthocyanin determination can be found in von Baer et al. (2005) and von Baer et al. (2007).

Anthocyanin	Abbreviation
delphinidin-3-glucoside	DP
cyanidin-3-glucoside	CY
petunidin-3-glucoside	PT
peonidin-3-glucoside	PE
malvidin-3-glucoside	MV
peonidin-3-acetylglucoside	PEAC
malvidin-3-acetylglucoside	MVAC
peonidin-3-coumaroylglucoside	PECU
malvidin-3-coumaroylglucoside	MVCU

Table 1.1: Description of measured anthocyanins.

Flavonol and Organic acid are antioxidant compounds. The dataset include determinations of six flavonol and four organic acids (listed in Table 1.2), on 149 samples for which the anthocyanin were also determined. The grape varieties in this subset were Cabernet Sauvignon (101 samples), Carménère (29 samples) and Merlot (19 samples) and the included valleys were Aconcagua, Maipo, Rapel, Curicó and Maule. Most of the samples come from 2004 harvest and some of them come from 2002 harvest. Flavonols were determined by HPLC based on the methodology of McDonald et al. (1998) with minor modifications. Organic acids were determined by a combination of reverse phase and ion exclusion chromatography in series, as described by Holbach et al. (2001) and OIV (2004). More details about Flavonols and Organic acid determination can be found in von Baer et al. (2007).

Organic Acids	Flavonol
Tartaric	Myricetin
Shikimic	Quercetin
Lactic	Total myricetin
Acetic	Total quercetin
	Conjugate myricetin
	Conjugate quercetin

Table 1.2: Measured compounds

1.2 A Bayesian Classification Approach for Solving Authentication Problems

We assume that an authentication problem can be solved by a classification approach. In that context, we assume a training dataset comprising n units $\{(y_i, x_i, g_i), i = 1, \ldots, n\}$. Here $y_i = (y_{i1}, \ldots, y_{ip})' \in R^p$ is the observed response vector for the *ith* unit, $x_i = (x_{i1}, \ldots, x_{iq})' \in R^q$ is the vector of covariates for the *ith* unit and g_i denotes the known group label or class for the *ith* unit, $g_i \in \{1, \ldots, g\}$. Let $y^n = (y_1, \ldots, y_n, x_1, \ldots, x_n, g_1, \ldots, g_n)$ denote the complete data. Let $y^{n+1} = (y_{n+1}, x_{n+1})$ be the observed data vector for a future unit, for which the corresponding label g_{n+1} is unknown. We adopt a predictive approach for classification, so the focuss is on the inference about g_{n+1} i.e. we are interested in estimating $P(g_{n+1} = k \mid y^n, y^{n+1})$, $k = 1, \ldots, g$. The above probability can be approximated by

$$P(g_{n+1} = k | y_{n+1}, y^n) \approx \frac{1}{C} \sum_{c=1}^{C} \frac{\pi_k p(y_{n+1} | \theta_k^{(c)})}{\sum_l \pi_l p(y_{n+1} | \theta_l^{(c)})}. \tag{1.2.1}$$

for details see Chapter 2. We propose classifying an existing unit, i, and a future one, $n+1$, using

$$\hat{g}_i = \arg\max_k P(g_i = k|y^n) \qquad \text{and} \qquad \hat{g}_{n+1} = \arg\max_k P(g_{n+1} = k|y^n, y_{n+1}). \quad (1.2.2)$$

i.e. assigning the label as the category that maximizes the classification probability. In practice, the authentication problem can be solved by computing the probability that the product complies with its label description. To do so, we need a probability model that adequately accounts for all the problem-specific features. We consider for group k a generic hierarchical model of the form

$$y_{ik} \mid \theta_{ik}, x_{ik} \sim p(y_{ik} \mid \theta_{ik}, x_{ik}), \qquad \theta_{ik} \sim G(\theta_{ik} \mid \phi_k). \quad (1.2.3)$$

In simple words, the data vector y_{ik} for the *ith* sampling unit in group k are sampled from a probability model parameterized by a vector θ_{ik}. Here x_{ik} is vector of covariates. The parameter vector θ_{ik} can be partitioned into a common fixed effect θ_k^F and unit-specific random effects θ_{ik}^R. When the θ_{ik}^R are assumed to be generated from a distribution parameterized by ϕ_k that belongs to a finite dimensional space, the resulting model is of parametric type, which is the focus of **Chapter 2**. When ϕ_k belongs to an infinite dimensional space, a nonparametric model for random effects is implied, and this is the focus of **Chapter 3**.

When more than one group of chemical compounds are available for food authentication, the dimension p of vector y_{ik} can be changed based on the available information. For example, in the wine dataset, $p = 9$ when we use the anthocyanin compounds, $p = 4$ when we use the Organic acid, $p = 6$ for flavonols, $p = 19$ when we use a combinations of the three groups of compounds, but in all cases the dimension of x_{ik} remains constant, so the covariates are the same for all models. Let $\mathcal{M}_{p_j}$ be a

model of the form (1.2.3) with the response vector $y_{ik} \in R^{p_j}$, $j = 1, 2, \ldots$. There are costs c_j associated with model $\mathcal{M}_{p_j}$, and losses in making wrong decisions. Selecting a particular model $\mathcal{M}_{p_j}$ implies selecting the compounds or combinations of them that reached the best performance. For that, we mean that the cost or expenses c_j of determining the compounds should be low and the accuracy of the results should be good. We propose a solution that implies the definition of a loss function that combines the penalty associated to a wrong decision with the cost c_j of each model $\mathcal{M}_{p_j}$, and this is the focus of **Chapter 4**. In **Chapter 5** we consider possible future research directions.

This thesis addresses an issue that has been developed outside the statistics field. In this context, the solutions proposed so far are mainly related to exploratory and descriptive tools for data analysis. Early approaches to authentication problems from a probabilistic point of view were made by Brown et al. (1999). Over the years there are new studies such as Dean et al. (2006) and Toher et al. (2007). We propose a Bayesian classification approach which is general enough for different authentication problems. Our approach allows to incorporate covariate information in the modeling. In addition, parametric assumptions are avoided supposing a flexible nonparametric distribution.

When we address the problem of optimal information search, we propose a decision theory approach. This approach is a standard tool, objective and is applied in many decision problems in different fields. In a context of food authentication there are no references about the use of decision theory in similar problems. Therefore, we believe that our proposal is novel in the context of research in food authentication.

Finally, from a statistical point of view, this thesis is a good example of the

application of the Bayesian methods and concepts to solve real problems.

The Chapters in this thesis can be read independently, because they have an abstract, introduction, development and they finish with the conclusions. In the next sections we give some background material, with basic concepts that will be used in the next chapters.

1.3 Prior Distributions on Probability Distributions

Semi-parametric models have both a parametric and a nonparametric part. The parametric part of the model has parameters that belong to a finite dimensional space, and the parameters of the nonparametric part belong to an infinite dimensional space. Nonparametric Bayesian models are used mainly to avoid critical dependence on parametric assumptions, and one their main applications arise when modeling random effects distributions in hierarchical models, where often little is known about the specific form of the random effects distributions (Müller and Quintana; 2004). To handle the nonparametric part of the model we need to define a random measure on the space of distribution functions. The most popular random measure on the space of distributions functions is the Dirichlet process (DP) (Ferguson; 1973). This process is defined by Ferguson (1973) as follows. Let Ω be a space and $\mathcal{A}$ a σ-field of subsets of Ω, and G_0 a probability measure on $(\Omega, \mathcal{A})$, where M is a scalar such that $M > 0$. The stochastic process G indexed by elements A of $\mathcal{A}$ is a DP on $(\Omega, \mathcal{A})$ with parameter $MG_0(\cdot)$ if for any partition $(A_1, \ldots, A_k)$ of Ω the random vector $(G(A_1), \ldots, G(A_k))$ follows a Dirichlet distribution with parameters $(MG_0(A_1), \ldots, MG_0(A_k))$. We denote this by

$G \sim DP(M, G_0)$. A key property of the DP is that if we have a sample $x_1, \ldots, x_n$ i.i.d. from G and $G \sim DP(M, G_0)$, then the posterior distribution $G \mid x_1, \ldots, x_n$ is of the same type, namely $DP(M + n, \tilde{G})$, where $\tilde{G} \propto G_0 + \sum_{i=1}^{n} \delta_{x_i}$ and δ_x denotes the measure giving mass one at the point x.

An important property of a DP, specially for computational purposes, is the Polya urn representation by Blackwell and MacQueen (1973). This representation was used by Escobar (1994) for estimating the mean of a normal distribution using a semi-parametric model. Many of the posterior developments are based on the same representation.

1.4 Dependent Dirichlet Processes

In a context of food authentication, it is common to collect food samples from different regions of origin or some which were put through different processing technologies, then, if a vector of responses is measured on these samples and also a covariates vector given by the origin or technology is recorded, it is reasonable to assume that the distribution that generates the responses may depend on the level of covariates. In the above context, we introduce below the Dependent Dirichlet Processes (DDP).

Suppose we have a response vector Y_i, a vector of covariates X_i, we are interesting in modeling the distribution of Y_i to include dependence on X_i. Then, if we think about a model $(Y_i \mid \theta_i, X_i) \sim F_{Y|X,\theta}(\cdot \mid \theta_i, X_i)$ and $(\theta_i \mid G) \sim G_i$ it will be necessary that G_i were dependent on the X_i level. Dirichlet Processes that include dependence on covariates were proposed by MacEachern (1999). The main idea, following a discrete covariates reasoning, was as follows. If a single distribution is assumed for all G_i and a nonparametric prior placed on this distribution, then $G_{x_1} = \ldots = G_{x_d}$; the

other extreme approach to account for differences in G_i is to placed d nonparametric prior distributions, the results is that $G_{x_1}, \ldots, G_{x_d}$ are mutually independent.

MacEachern (1999) stated that in the first approach, the d distributions may be allowed to differ by a small number of parameters, perhaps locations and scales, but the distributions are identical in many ways; in the second approach, the d distributions may be linked together through hyperparameters, but conditional on these hyperparameters, the realized distributions are independent, so what is needed is a modelling strategy that allows the set of random effects, distributions to be similar, but not identical.

To introduce the definition of Dependent Dirichlet Processes (DDP), it is necessary to present Sethuraman's representation of DPs (Sethuraman; 1994). Assume $G \sim DP(MG_0)$. Then G admits a stick-breaking representation as

$$G(\cdot) = \sum_{h=1}^{\infty} p_h \delta_{\theta_h}(\cdot), \tag{1.4.1}$$

where δ_θ is a probability measure concentrated at θ, $p_h = V_h \prod_{l=1}^{h-1}(1 - V_l)$ for $h \geq 1$ and all $V_h's$ and $\theta_h's$ are independent, with V_h i.i.d $Be(1, M)$ and θ_h i.i.d. G_0.

Definition 1. Dependent Dirichlet processes are defined by the relation

$$G_\chi \sim DDir(M_\chi, G_{0,\chi}, Z_\chi, U_\chi, T_{Z,\theta;\chi}, T_{U,V;\chi}), \tag{1.4.2}$$

where M_χ is the mass parameter, $0 < M_\chi < \infty$ for all $x \in \chi$, $G_{0,\chi}$ is the base measure, Z_χ and U_χ are stochastic processes providing draws that are turned into locations and probabilities, respectively, $T_{Z,\theta;\chi}$ and $T_{U,V;\chi}$ are transformations specifying a mapping of Z_x into θ_x and from U_x into V_x for each $x \in \chi$, respectively, and at each $x \in \chi$, the

distribution G_x is defined by

$$G_x = \sum_{h=1}^{\infty} p_{hx}\delta_{\theta_{hx}}. \tag{1.4.3}$$

MacEachern's proposal allows the weights p_{hx}, $(h = 1, \ldots, \infty)$ and atoms θ_{hx} ,$(h = 1, \ldots, \infty)$ to vary with x according to a stochastic process. DDPs where p_h is assumed to be fixed with respect to x have been successfully applied to the analysis of variance (De Iorio et al.; 2004), spatial modeling with a Gaussian process for the atoms (Gelfand et al.; 2005), times series (Caron et al.; 2006), classification (De la Cruz et al.; 2007b), dynamic density estimation (Rodriguez and ter Horst; 2008), inferences on stochastic ordering (Dunson and Peddada; 2008), quantile regression (Kottas and Krnjajić; 2009), survival analysis (De Iorio et al.; 2009) and recently, by Jara et al. (2010) who proposed a Poisson-Dirichlet process for the analysis of a data set coming from a dental longitudinal study. Griffin and Steel (2006) argue that allowing only the values of θ_h to depend on the covariates will guide to certain problems with points far from the observed data in the domain. In particular, MacEachern noted that the distribution of G can then be expressed as a mixture of Dirichlet processes. The posterior process will have an updated mass parameter $M + n$, where n is the sample size, at all values of the index. Griffin and Steel (2006) think that the above property is counterintuitive, because it would be desirable that the process reduces to the prior distribution (with mass parameter M) at points in the domain far from the observed data. Therefore, they proposed an approach that avoids this property by resorting to local updates of the process. Their proposal basically consist of inducing dependence in the weights through similarities in the ordering of the atoms, by viewing the atoms as marks in a point process and implementing such orderings through distance measure. Other works where covariate dependence

is introduced in the weights are Dunson et al. (2007), and Dunson and Park (2008). Müller et al. (1996) considered a completely different approach for inducing dependence in G. They used a DP mixture of normals for the joint distribution of y and z, and then focused on the implied conditional density of y given z for estimating the mean regression function. Finally, a recent reference about DDPs is Chung and Dunson (2011), who proposed the Local Dirichlet process to allow predictor dependence. The almost sure discreteness of the Dirichlet process makes it inappropriate as a model for a continuous quantity y. A standard procedure for overcoming this difficulty is to introduce an additional convolution, with a continuous kernel, so that

$$H(y) = \int f(y \mid \theta) dG(\theta) \qquad \text{with} \qquad G \sim DP(M, G_0). \tag{1.4.4}$$

Such models are known as DP mixtures (DPM) (Antoniak; 1974). The mixture model Hjort et al. (2010) (1.4.4) can be equivalently written as a hierarchical model by introducing latent variables θ_i and breaking the mixture as

$$y_i \mid \theta_i \sim f(y_i \mid \theta_i), \qquad \theta_i \sim G, \qquad \text{and} \qquad G \sim DP(M, G_0). \tag{1.4.5}$$

For the majority of food authentication problems the responses are continuous multivariate and covariates are discrete. This is the case for the wine dataset, so the ANOVA-DDP approach of De Iorio et al. (2004) is a natural way to build the desired dependence. Thus we will adopt the popular semiparametric modeling strategy that consists of introducing dependence in the random effects distribution and then adding a convolution with a continuous kernel.

1.5 MCMC Methods in Conjugate Dirichlet Process Mixtures Models

In this section we provide a brief discussion on the computational aspects for posterior sampling of Conjugate Dirichlet Process Mixtures models, because this is the class of models that we will employ in the next Chapters. Basically, we focuss the attention on Markov Chain Monte Carlo (MCMC) algorithms (Escobar (1994); Escobar and West (1995); Dey et al. (1998); Neal (2000)), because they have been used successfully in the posterior sampling under Dirichlet Process priors, and they provide a mechanism for fitting a wide class of hierarchical models. Consider a hierarchical generic model

$$\begin{aligned} Y_i \mid \theta_i &\overset{ind}{\sim} F(\cdot \mid \theta_i), \quad i = 1, \ldots, n \\ \theta_1, \ldots, \theta_n \mid G &\overset{iid}{\sim} G, \\ G \mid M, \lambda &\sim DP(MG_\lambda), \\ (M, \lambda) &\sim p(M)p(\lambda). \end{aligned} \tag{1.5.1}$$

Here, $Y_1 \ldots, Y_n$ are part of an infinite exchangeable sequence, or equivalently, as being independently drawn from some unknown distribution. The Y_i may be multivariate, as our applications in Chapter 3. The model from which the Y_i's are drawn, is a mixture of distributions of the form $F(\cdot \mid \theta)$, with the mixing distributions over θ being G. We let the prior for this mixing distribution be a Dirichlet process with concentration parameter M and base distribution G parameterized by λ. Now, we show the first MCMC approach for DP priors proposed by Escobar (1994). The Escobar's algorithm simplify the use of the Dirichlet Process integrating G over its prior distribution, the sequence of θ_i's follows a general Polya urn scheme (Blackwell

and MacQueen; 1973); that is

$$\begin{aligned}
\theta_1 \quad &| \lambda \sim G_\lambda, \\
\theta_n \quad &| \theta_1, \ldots, \theta_{n-1}, \lambda, M \begin{cases} = \theta_j, & \text{with probability} \frac{1}{M+n-1}, \quad \text{for } j = 1, \ldots, n-1 \\ \sim G_\lambda, & \text{with probability} \frac{M}{M+n-1}. \end{cases}
\end{aligned}$$

With the above scheme, it is easy to sample a sequence $\theta_1, \ldots, \theta_n$ given G_λ and M. The conditional distribution for θ_j given $\theta^{(j)} = (\theta_1, \ldots, \theta_{j-1}, \theta_{j+1}, \ldots, \theta_n)^T$, M and λ is given by

$$dP(\theta_j \mid \theta^{(j)}, M, \lambda) \propto MG_\lambda(d\theta_j) + \sum_{i \neq j} \delta(\theta_i, d\theta_j),$$

where $\delta(\theta, \cdot)$ is a measure defined by

$$\delta(\theta, B) = \begin{cases} 1, & \text{when } \theta \in B \\ 0, & \text{when } \theta \notin B. \end{cases}$$

To get the posterior distribution $dP(\theta, M, \lambda \mid Y_1, \ldots, Y_n)$, Escobar proposed to use a Gibbs sampling approach based on sampling from the appropriate full conditional distributions, $(\theta_j \mid \theta^{(j)}, M, \lambda, Y_1, \ldots, Y_n)$, $(M \mid \theta_1, \ldots, \theta_n, \lambda, Y_1, \ldots, Y_n)$, and $(\lambda \mid \theta_1, \ldots, \theta_n, M, Y_1, \ldots, Y_n)$. The conditional distribution of θ_j given $\theta^{(j)}$, and $Y_1, \ldots, Y_n$ has the following closed form

$$\begin{aligned}
dP(\theta_j \mid \theta^{(j)}, M, \lambda, Y_1, \ldots, Y_n) &= \frac{f(Y_j \mid \theta_j) \left\{ MG_\lambda(d\theta_j) + \sum_{i \neq j} \delta(\theta_i, d\theta_j) \right\}}{\int f(Y_j \mid \theta_j) \left\{ MG_\lambda(d\theta_j) + \sum_{i \neq j} \delta(\theta_i, d\theta_j) \right\}} \quad (1.5.2) \\
&= \frac{Mf(Y_j \mid \theta_j) G_\lambda(d\theta_j) + \sum_{i \neq j} f(Y_j \mid \theta_i) \delta(\theta_i, d\theta_j)}{M \int f(Y_j \mid \theta_j) G_\lambda(d\theta_j) + \sum_{i \neq j} f(Y_j \mid \theta_i)}
\end{aligned}$$

The above distribution follows from the Bayes theorem and the conditional independence of $Y_i \perp\!\!\!\perp \theta_j \mid \theta_i$. The conditional distribution defined in equation (1.5.2) can be sampled according to the following rule:

$$\theta_j \mid \theta^{(j), M, \lambda, Y_1, \ldots, Y_n} \begin{cases} \theta_i \;\; i \neq j, & \text{with probability} \;\; \frac{f(Y_j \mid \theta_i)}{M \int f(Y_j \mid \theta_j) G_\lambda(d\theta_j) + \sum_{i \neq j} f(Y_j \mid \theta_i)} \\ \sim H_j(\theta_j \mid Y_j), & \text{with probability} \;\; \frac{M \int f(Y_j \mid \theta_j) G_\lambda(d\theta_j)}{M \int f(Y_j \mid \theta_j) G_\lambda(d\theta_j) + \sum_{i \neq j} f(Y_j \mid \theta_i)}, \end{cases} \quad (1.5.3)$$

where H_j is the posterior density of θ_j given the data Y_j and the prior distribution G_λ for θ_j. The last algorithm produces and ergodic Markov chain, but the convergence to the posterior distribution may be rather slow, and consequently, sampling under this algorithm may be inefficient. As discussed in Neal (2000), the problem is that there are often groups of observations with high probability that are associated with the same θ. Since the algorithm cannot change the θ for more than one observation simultaneously, a change to the θ values for observations in such a group can occur rarely, as such a change requires passage through a low-probability intermediate state in which observations in the group do not have all the same θ value. Bush and MacEachern (1996) avoided this problem by adding a second stage to the Escobar's Gibbs sampling. In the second stage the cluster locations are moved. Neal (2000) deals with the Escobar's Gibbs sampling problems defining an equivalent model when K (the number of components in a mixture) goes to infinity. The model is given by

$$\begin{aligned} Y_i \mid c_i, \phi &\sim F(\cdot \mid \phi_{c_i}) \\ c_i \mid p &\sim \text{Discrete}(p_1, \ldots, p_K) \\ \phi_c &\sim G_\lambda \\ p &\sim \text{Dirichlet}(M/K, \ldots, M/K) \end{aligned} \tag{1.5.4}$$

Here, c_i indicates which latent class is associated with observations Y_i, with no significance in the numbering of c_i. For each class, c, the parameters ϕ_c determine the distribution of the observations from that class; the collection of all such ϕ_c is denoted by ϕ. The mixing proportions for the classes, $p = (p_1, \ldots, p_K)$, are given by a symmetric Dirichlet prior, with concentration parameter written as M/K, so that it approaches zero as K goes to infinity. Neal (2000) shows that letting $\theta_i = \phi_{c_i}$ model (1.5.4) is equivalent to the Dirichlet process mixture model (1.5.1) when $K \to \infty$. The

problems in the Escobar (1994) algorithm are avoided if Gibbs sampling is applied to the model formulated in (1.5.4), with the mixing proportions, p, integrated out. In the first stage, the algorithm draws the configurations c, then a Gibbs sampling for c_i is based on the following conditional probabilities (with ϕ being the set of ϕ_c currently associated with at least one observation):

$$\begin{aligned} \text{if } c = c_j \text{ for some } j \neq i : P(c_i = c \mid c_{-i}, Y_i, \phi) &= b\frac{n_{-i,c}}{n-1+M} f(Y_i \mid \phi_c) \\ P(c_i \neq c_j \text{for all} j \neq i \mid c_{-i}, Y_i, \phi) &= b\frac{M}{n-1+M} \int f(Y_i \mid \phi) dG_\lambda(\phi) \end{aligned}$$

Here, c_{-i} are all the c_j for $j \neq i$, $n_{-i,c}$ is the number of c_j for $j \neq i$ that are equal to c, b is the appropriate normalizing constant. When Gibbs sampling for c_i chooses a value not equal to any other c_j, a value for ϕ_{c_i} is chosen from H_i, the posterior distribution based on the prior G_λ and the single value Y_i. In the second stage, for all $c \in \{c_1, \ldots, c_n\}$, the algorithm draws a new value of $\phi_c \mid Y_i$ for which $c_i = c$, that is, drawn from the posterior distribution based on the prior G_λ and all the data points associated with latent class c. The above algorithm is essentially the method proposed by Bush and MacEachern (1996). MacEachern (1994) proposed to integrate analytically over the ϕ_c, eliminating them from the algorithm. The state of the Markov chain then consist only of the c_i which are updated in a Gibbs sampling using the following conditional probabilities

$$\begin{aligned} \text{if } c = c_j \text{ for some } j \neq i : P(c_i \mid c_{-i}, Y_i) &= b\frac{n_{-i,c}}{n-1+M} \int f(Y_i \mid \phi) dH_{-i,c}(\phi) \\ P(c_i \neq c_j \text{ for all } j \neq i \mid c_{-i}, Y_i) &= b\frac{M}{n-1+M} \int f(Y_i \mid \phi) dG_\lambda(\phi). \end{aligned}$$

Here, $H_{-i,c}$ is the posterior distribution of ϕ based on the prior G_λ an all observations Y_i for which $j \neq i$ and $c_i = c$.

Jain and Neal (2004) stated that, although the Gibbs sampling approach is straightforward and easily implemented, it could be slow to reach convergence and mix poorly too. In this context, they proposed a split-merge Markov chain algorithm. The split-merge algorithm introduces a new Metropolis-Hastings method that avoids the problems associated with the Gibbs sampling procedure and is suitable for high-dimensional data. Typically, Metropolis-Hastings updates involve simple parametric distributions as the proposal distribution. To split mixtures components, the Jain and Neal (2004) algorithm employs a more complex proposal distribution obtained by using a restricted Gibbs sampling scan for the latent class variables. This method is able to quickly traverse the state space and frequently visit high-probability modes because it splits or merges a group of observations in each update, thereby, bypassing the incremental of the Gibbs sampler. Furthermore, although the proposal distribution used is complex, it does not need to be specially tailored to each model, since the same scheme can be applied to any model with a conjugate prior. For details of its computation refers to Jain and Neal (2004). Finally, Dahl (2005) proposed a split-merge sampler for both conjugate and non-conjugate Dirichlet process mixture models. The sampler borrows ideas from sequential importance sampling. Splits are proposed by sequentially allocating observations to one of two split components using allocations probabilities that condition on previously allocated data. For details of its computation refers to Dahl (2005).

1.6 Statistical Decision Theory

This section provides the basic concepts involved in decision problems. We will use the concepts in Chapter 4, where we deal with the search of optimal information

in an authentication process. Decision theory, as the name implies, is concerned with the problem of making decisions (Berger; 1985). Statistical decision theory is concerned with the making of decisions in the presence of statistical knowledge which sheds light on some of the uncertainties involved in the decision problem. We assume that the uncertainties can be considered as unknown numerical quantities represented by θ. The unknown quantity θ, which affects the decision process, is commonly called the state of nature. The symbol Θ will be used to denote the set of all possible states of nature. Typically, when experiments are performed to obtain information about θ, they are designed so that the observations are distributed according to some probability distribution which has θ as an unknown parameter. In such situations, θ will be called the parameter and Θ the parameter space. In addition to the sample information, two other types of information are typically relevant, these are the knowledge of the possible consequences of the decisions and the prior information about θ. The knowledge of the possible consequences of the decision can be quantified by determining the loss that would be incurred for each possible decision and for the various possible values of θ. Therefore, a key element of decision theory is the loss function. The prior information about θ is the information that arises from past experiences about similar situations involving similar θ. This information often is represented by a probability distribution denoted by $\pi(\theta)$.

Decisions are more commonly called actions in the literature. Particular actions will be denoted by a, while the set of all possible actions under considerations will be denoted $\mathcal{A}$. As mentioned in the last paragraph, a key element of decision theory is the lost function. If a particular action a_1 is taken and θ_1 turns out to be the true state of the nature, then a loss $L(\theta_1, a_1)$ will be incurred. Thus, we will assume a loss

function $L(\theta, a)$, which is defined for all $(\theta, a) \in \Theta \times \mathcal{A}$.

When a statistical investigation is performed to obtain information about θ, the outcome (a random variable) will be denoted as Y. $Y = (Y_1, \ldots, Y_n)$ is often a vector, and Y_i, $i = 1, \ldots, n$ are independent observations from a common distribution, parameterized by θ. That distribution will be denoted by $f(Y \mid \theta)$, commonly named the sample distribution. A particular realization of Y will be denoted by y. The set of possible outcomes is the sample space, and will be denoted $\mathcal{Y}$. When a particular realization of Y is observed, we can update our prior information of $\pi(\theta)$ using the Bayes theorem and obtain the posterior distribution $\pi(\theta \mid y)$.

The incurred loss $L(\theta, a)$, will be never known with certainty (at the time of the decision making). A natural method of proceeding in the face of this uncertainty is to consider the "expected" loss of making a decision, and then choose an "optimal" decision with respect to this expected loss. In Bayesian decision theory, the posterior expected loss of an action a, when the posterior distribution is $\pi(\theta \mid y)$, is

$$\rho(\pi(\theta \mid y), a) = \int_{\Theta} L(\theta, a)\pi(\theta \mid y)d\theta. \tag{1.6.1}$$

The simplicity of the Bayesian approach follows from the fact that and optimal action can be found by simple minimization of (1.6.1). The above concepts are employed in Chapter 4, where we proposed a methodology for finding optimal information in an authentication process.

Chapter 2

Multivariate Bayesian Discrimination for Varietal Authentication of Chilean Red Wine

2.1 Abstract

The process through which food or beverages are verified as complying with its label description is called food authentication. We propose to treat the authentication process as a classification problem. We consider multivariate observations and propose a multivariate Bayesian classifier that extends results from the univariate linear mixed model to the multivariate case. The model allows for correlation between wine samples from the same valley. We apply the proposed model to concentration measurements of nine chemical compounds named anthocyanins in 399 samples of Chilean red wines of the varieties Merlot, Carménère and Cabernet Sauvignon, vintages 2001-2004. We find satisfactory results, with a misclassification error rate based on a leave-one-out cross-validation approach of about 4%. The multivariate extension can be generally applied to authentication of food and beverages, where it is common to have several

dependent measurements per sample unit, and it would not be appropriate to treat these as independent univariate versions of a common model.

Key Words: Bayesian classifier, Gibbs sampling, hierarchical linear models, food authentication.

2.2 Introduction

Consumers increasingly demand reassurance of the origin and content of their food and beverages. The process through which food or beverages are verified as complying with its label description is called food authentication (Winterhalter; 2007). The wine industry has been using the authentication procedure for a long time. Substantial research efforts have been put into this particular topic. von Baer et al. (2005) report that some containers of Chilean red wine have been rejected in Germany because they did not satisfy the parameters applied there to verify wine varieties. These problems have a direct impact on producers and their income. Chilean wine represents an important part of Chile's worldwide exports, which have increased from 52 to 1,256 million U.S. dollars over the period 1997-2007. The main red wine varieties are Merlot, Carménère and Cabernet Sauvignon. Therefore, it is important for sustainable long-term growth to develop a reliable system to verify product authenticity. In this sense, various authors have proposed to differentiate among red wine varieties using their anthocyanin profiles (Eder et al.; 1994; Holbach et al.; 1997; Berente et al.; 2000; Holbach et al.; 2001; Otteneder et al.; 2002, 2004; von Baer et al.; 2005; Revilla et al.; 2001; von Baer et al.; 2007). Anthocyanins are a group of chemical compounds present in red wine, which confer to this beverage its characteristic red color and are transferred from the grape skins to wine during the winemaking process.

Many of the works about wine authentication consider only simple relations between anthocyanins. The method approved by the OIV in 2003 is also based on this principle (OIV; 2003). For a review of exploratory multivariate methods for classification based on anthocyanin profiles and linear discriminant analysis, see von Baer et al. (2007). Other approaches in wine authentication include neural networks (Beltrán et al.; 2005; Kruzlicova et al.; 2009) and similarity index based on mid-infrared spectroscopy data (Bevin et al.; 2006).

Probabilistic modeling for discrimination and authentication purposes was proposed by Brown et al. (1999). In the special case of longitudinal data analysis, Bayesian discrimination has been discussed and used by Brown et al. (2001) and De la Cruz-Mesía and Quintana (2007). Lavine and West (1992) describe Bayesian methods for classification and discrimination using Gibbs sampling. Mallick et al. (2005) discussed Bayesian classification using gene expression data, concluding from their comparison with other methods, that the Bayesian classification approach performed better than other popular alternatives. A similar conclusion was obtained by Rigby (1997) when comparing the Bayesian and classical estimates of P, the probability that a new observation belongs to one of two multivariate normal populations with equal covariance matrices. More recently, Agrawal et al. (2009) consider an incremental framework for feature selection and Bayesian classification for multivariate normal groups.

In the present paper, we extend the univariate Bayesian linear mixed models to the multivariate case, and use this model to build a Bayesian classifier of Chilean red wine varieties using their anthocyanin profiles. In particular, we describe in detail a Bayesian classification strategy based on multivariate hierarchical linear models.

In the context of classical inference, multivariate linear mixed models were proposed by Reinsel (1982) and Reinsel (1984). Our methods are based on a similar model, but using a Bayesian viewpoint. Therefore, our contribution is two-fold in the sense of coherency of the inferential approach, and the novelty of the application of such methods to food authentication problems. In doing so, we treat the classes or groups as predefined and the task is to understand the basis for the classification from a set of labeled samples (training dataset). This information is then used to classify future subjects.

The rest of this paper is organized as follows. We first give a brief description of the dataset in Section 2.3. In Section 2.4.1, we expose a general multivariate Bayesian classification approach. In Section 2.4.2 we present a general multivariate Bayesian linear model for grape variety authentication. In Section 2.4.3 we illustrate the proposed general classifier using data from Chilean anthocyanin profiles of red wine and describe an appropriate posterior simulation scheme based on the Gibbs sampling algorithm. In Section 2.5 we present the results of the selected model application. Finally, Section 2.6 discusses the results.

2.3 The Motivating Dataset

We consider a dataset consisting of concentration measurements of a number of chemical markers in samples of Chilean red wines. For the purpose of this study, we restrict ourselves to measurements of anthocyanins, because these compounds are widely used for red wine authentication, and the methodologies used in their determination are sufficiently accepted and standardized. In addition, we also want to compare the results with other studies carried out with the same data. The dataset includes the

grape variety for each sample *as declared by the producer*, the year of harvest, and the geographic origin or valley. All wine samples came directly from wineries located in the valleys of Aconcagua, Maipo, Rapel, Curicó, Maule, Itata and Bío-Bío. As listed, these valleys are geographically sorted north to south of Chile, and range from 33 to 38 degrees latitude south. The valleys have a wide range of soil types and weather conditions. The largest one is Maule, which is where most of the available samples were taken. The wine samples correspond to the vintages 2001 through 2004. Vinification was made at production scale and samples were taken after malolactic fermentation, but before blending. Anthocyanin determination was made by reverse phase HPLC based on the method described by Holbach et al. (1997), Otteneder et al. (2002) and OIV (2003), with some minor modifications. The response considered for each anthocyanin in a given sample is its log-concentration. More details about anthocyanin determination for the dataset can be found in von Baer et al. (2005) and in von Baer et al. (2007).

The sample size is 399, of which 228 were declared by the producers as Cabernet Sauvignon, 76 as Merlot and 95 samples as Carménère. For later reference, Table 2.1 shows a list of the nine anthocyanins used in the present paper. A brief exploratory analysis of the data uncovered some differences in the anthocyanin log-concentrations across the three grape varieties, and correlations between the nine anthocyanins. These observations support our choice of using the available measurements for discrimination purposes under a multivariate approach, as it would not be reasonable to consider nine separate univariate response models to deal with these data. The multivariate extension we discuss next is thus relevant for the current classification problem.

Anthocyanin	Abbreviation
delphinidin-3-glucoside	DP
cyanidin-3-glucoside	CY
petunidin-3-glucoside	PT
peonidin-3-glucoside	PE
malvidin-3-glucoside	MV
peonidin-3-acetylglucoside	PEAC
malvidin-3-acetylglucoside	MVAC
peonidin-3-coumaroylglucoside	PECU
malvidin-3-coumaroylglucoside	MVCU

Table 2.1: Description of measured anthocyanins.

2.4 Model

We present next the model, discussing some of its properties and implementation issues. The full MCMC details can be found in the Appendix.

2.4.1 Classification Using Multivariate Bayesian Classifier

We assume a classification problem featuring multivariate response observations, and a training dataset comprising n units $\{(y_i, x_i, g_i), i = 1, ..., n\}$. Here $y_i = (y_{i1}, ..., y_{ip})' \in R^p$ represents the observed response vector for the *ith* unit, $x_i = (x_{i1}, ..., x_{iq})'$ is the vector of covariates for the *ith* unit and g_i denotes the known group label for the *ith* unit, $g_i \in \{1, 2, ..., g\}$. Let $y^n = (y_1, ..., y_n, x_1, ..., x_n, g_1, ..., g_n)$ denote the complete data. We adopt a predictive approach for classification. Therefore, we assume an observed data vector $y_{n+1} = (y_{n+1}, x_{n+1})$ for a future unit, for which the corresponding label g_{n+1} is unknown. The primary inferential target is g_{n+1}, i.e. we are interested in estimating $\{p(g_{n+1} = k|y^n, y_{n+1}) : k = 1, \ldots, g\}$. Following De la Cruz-Mesía and Quintana (2007), we consider an augmented model with

marginal prior $P(g_i = k) = \pi_k$ for $k = 1, \ldots, g$. For instance, the π_k probabilities could be taken as the empirical group proportions.

Let θ denote the vector of all possible parameters and hyperparameters. The classification probabilities are obtained by weighting the posterior conditional group probabilities given θ with respect to the posterior distribution $p(\theta|y^n)$. Concretely, the classification probability that a new unit y_{n+1} belongs to the kth group is

$$\begin{aligned}
P(g_{n+1} = k|y_{n+1}, y^n) &= \int \frac{p(g_{n+1} = k, y_{n+1}, y^n, \theta)}{p(y_{n+1}, y^n)} d\theta \\
&= \int \frac{p(g_{n+1} = k|y_{n+1}, y^n, \theta) p(y_{n+1}, y^n, \theta)}{p(y_{n+1}, y^n)} d\theta \\
&= \int p(g_{n+1} = k|y_{n+1}, y^n, \theta) p(\theta|y_{n+1}, y^n) d\theta \\
&= \int p(g_{n+1} = k|y_{n+1}, \theta) p(\theta|y_{n+1}, y^n) d\theta \\
&\propto \int p(g_{n+1} = k|y_{n+1}, \theta) p(\theta|y^n) d\theta \\
&= \int \frac{\pi_k p(y_{n+1}|\theta_k)}{\sum_{l=1}^{g} \pi_l p(y_{n+1}|\theta_l)} p(\theta|y^n) d\theta. \qquad (2.4.1)
\end{aligned}$$

See further details in De la Cruz-Mesía and Quintana (2007). In practice, direct analytical evaluation of (2.4.1) is impossible so we resort to posterior simulation methods. Assuming for now the availability of a sample $\{\theta^{(c)}, c = 1, ..., C\}$ from the posterior distribution $p(\theta \mid y^n)$ (we discuss methods for this later in Section 2.4.2 and in the Appendix), we approximate (2.4.1) by means of (De la Cruz-Mesía and Quintana; 2007)

$$P(g_{n+1} = k|y_{n+1}, y^n) \approx \frac{1}{C} \sum_{c=1}^{C} \frac{\pi_k p(y_{n+1}|\theta_k^{(c)})}{\sum_l \pi_l p(y_{n+1}|\theta_l^{(c)})}. \qquad (2.4.2)$$

We propose classifying an existing unit, i, and a future one, $n+1$, using

$$\hat{g}_i = \arg\max_k P(g_i = k|y^n) \qquad \text{and} \qquad \hat{g}_{n+1} = \arg\max_k P(g_{n+1} = k|y^n, y_{n+1}). \quad (2.4.3)$$

In other words, the unit is classified in the group for which the highest posterior probability is attained, thus minimizing the expected misclassification rate. This is actually the Bayes rule under the zero-one loss function, as discussed in Hastie et al. (2001).

2.4.2 A General multivariate Bayesian Linear Model for Grape Variety Authentication

In practice, the authentication problem can be solved by computing the probability that the product complies with its label description. We propose to do it using the classification approach discussed in Section 2.4.1. To do so, we need a probability model that adequately accounts for all the problem-specific features. We now describe a linear mixed model that is useful for the classification of grape varieties.

We assume that the *ith* response vector is related to the covariates in a linear way. Furthermore, we assume that there are fixed and random effects in the model. The model for the *ith* unit in the *kth* group (grape variety) is thus given by

$$y_i^k = Bx_i^k + Uz_i^k + \epsilon_i^k, \qquad i = 1, \ldots, n \qquad k = 1, \ldots, g \qquad (2.4.4)$$

where y_i^k is the p-dimensional response vector for the *kth* group, x_i^k is the corresponding q-dimensional covariate vector of fixed effects, and z_i^k is the r-dimensional vector of covariates for the random effects. Also, B is a $p \times q$ matrix of regression coefficients for the fixed effects, which we synthetically write as

$$B = [\beta_1, \beta_2, ..., \beta_q]$$

where $\beta_1, ..., \beta_q$ are $p \times 1$ column vectors. In addition, U is a $p \times r$ matrix of random effects which we write as

$$U = [U_1, U_2, ..., U_r]$$

where $U_1, ..., U_r$ are $p \times 1$ column vectors. Finally ϵ_i^k is the p-dimensional error vector.

The formulation of our model is described next. For the top model (2.4.4) we assume ϵ_i^k to be independent with

$$\epsilon_i^k \sim N_p(0, \Sigma_k), \qquad i = 1, \ldots, n, \qquad k = 1, \ldots, g. \tag{2.4.5}$$

As is usual in this context, we assume prior independence for all parameters. The prior distributions for matrices B and U are assumed to be independent by columns, that is $\beta_1, \ldots, \beta_k$ and $U_1, \ldots, U_r$ are mutually independent, with distributions given by

$$\beta_j \sim N_p(\beta_{0j}, \Lambda_0), \qquad j = 1, \ldots, q \tag{2.4.6}$$

$$U_1, \ldots, U_r \sim N_p(0, S) \tag{2.4.7}$$

The prior distribution for the variance-covariance matrices Σ_k, $k = 1, \ldots, g$ and S are given by

$$\Sigma_1, \ldots, \Sigma_g \sim IW(Q_0, \nu_0) \tag{2.4.8}$$

$$S \sim IW(K_0, m_0) \tag{2.4.9}$$

We complete the Bayesian formulation of model (2.4.4) by specifying the prior for hyperparameters $\beta_{01}, \ldots, \beta_{0q}$ and Λ_0 as

$$\beta_{01}, \ldots, \beta_{0q} \sim N_p(\alpha_0, \tau_0) \tag{2.4.10}$$

$$\Lambda_0 \sim IW(L_0, t_0). \tag{2.4.11}$$

The full conditional posterior distributions for the fixed and random effects are normal. The variance-covariance matrices $\Sigma_1, \ldots, \Sigma_g$ and S have full conditional posterior distributions of inverse Wishart type. Finally, the full conditional distribution for hyperparameters Λ_0 and $\beta_{01}, \ldots, \beta_{0q}$ are inverse Wishart and Normal, respectively. Details about the complete set of full conditional distributions are given in the Appendix.

2.4.3 Application to the Wine Dataset

In our application, we have that $n = 399$, $g = 3$, with $g_i = 1$, $g_i = 2$ and $g_i = 3$ indicating Cabernet Sauvignon, Merlot and Carménère, respectively. The label g_i in our example corresponds to the variety *declared by the producer* for each wine sample. This is an important clarification. See the discussion below. We assume that g_i, $i = 1, \ldots, n$ are known and g_{n+1} is unknown, which corresponds to the label of a new sample wine for which we want to verify its authenticity.

We implemented three variations of the general model described in Section 2.4.2:

Model 1: This model has only fixed effects and assumes a common covariance matrix Σ for the three grape varieties. In this model we set $d = 11$, $p = 9$ and the design vector $x_i = (x_{i1}, \ldots, x_{i11})^t$ is given by x_{i1}, x_{i2} and x_{i3}, each one assuming the values 1 or 0 depending on whether the *ith* wine sample corresponds to Carbernet Sauvignon, Merlot or Carménère, respectively. We code x_{i4} as assuming the values $1, \ldots, 4$, depending on whether the year of harvest was 2001, 2002, 2003 or 2004 respectively. This allows us, among other things, to incorporate new data for 2005 that may potentially become available, without having to modify the model if a new sample of harvest 2005, for example, is classified. In

such case we could simply code the year of harvest 2005 as $x_{i4} = 5$. We set $x_{i5} = 1$ if the *ith* sample comes from the Aconcagua valley and 0 otherwise. We define $x_{i6}, \ldots, x_{i11}$ in the same way, to represent samples of the Maipo, Rapel, Curicó, Maule, Itata and Bío-Bío valleys, respectively.

Model 2: This model has both, fixed and random effects and assumes a common covariance matrix Σ for the three grape varieties. In this model we take $d = 4$, $p = 9$, and $r = 7$. The design vector for fixed effects is given by $x_i = (x_{i1}, x_{i2}, x_{i3}, x_{i4})$ where its components were defined exactly as in Model 1. The design vector for the random effects $z_i = (z_{i1}, \ldots, z_{i7})$ represents the valley, where $z_{i1} = 1$ if the *ith* sample comes from the Aconcagua valley and 0 otherwise. We define $z_{i2}, \ldots, z_{i7}$ in the same way, to represent samples of the Maipo, Rapel, Curicó, Maule, Itata and Bío-Bío valleys, respectively. By definition of the z_i matrices, U_1,...,U_7 represent valley-specific random effects and we allow samples that come from the same valleys to be correlated.

Model 3: This model has fixed and random effects and grape variety-specific covariance matrices, Σ_1, Σ_2 and Σ_3. Here, $d = 4$, $p = 9$, $r = 7$, and the design vector for random and fixed effects are the same as in Model 2. The only difference is that we order the data in blocks so we can separate the roles of Σ_1, Σ_2 and Σ_3.

The value of the hyperparameters in (2.4.8) - (2.4.11) for model 1 were taken as $\alpha_0 = (0, 0, 0, 0, 0, 0, 0, 0, 0)^t$, $\tau_0 = 1000 I_9$, $Q_0 = I_9$, $L_0 = I_9$, $\nu_0 = 11$ and $t_0 = 11$. For models 2 and 3 we need the additional choices $K_0 = I_9$ and $m_0 = 11$. The prior means for Σ and S were assumed to be the identity matrix. For the random effects U, we assumed a prior centered at 0, with identity covariance matrix. The selected

hyperparameter values imply proper but vague prior distributions, representing the lack of genuine prior information on the parameters.

The Gibbs sampling algorithm was implemented in a computer program written in FORTRAN. We generated 110,000 iterations. After 10,000 iterations, samples were collected at a spacing of 100 iterations, to obtain independent samples. Finally we totaled $C = 1,000$ samples for calculating posterior quantities of interest. The average time used to run each of the three models above in a standard PC (Intel Core Duo CPU 2.4 Ghz and 2.0 Gb RAM) was 3 hours.

2.5 Results

To evaluate model adequacy and to select among the three models in Section 4.4 we use two model selection criteria, the Conditional Predictive Ordinates (CPO_i) (Chen et al.; 2000) and the Deviance Information Criterion (DIC) (Spiegelhalter et al.; 2002). CPO_i is a useful quantity for model checking, since it is based on how much the *ith* observation supports the model. Large CPO_i values indicate a good fit. It is customary to summarize all the CPOs using the log-pseudo marginal likelihood (LPML) statistic (Geisser and Eddy; 1979), defined as $\text{LPML} = \sum_{i=1}^{n} \log(\text{CPO}_i)$. On the other hand, DIC is an information criterion that was proposed to select Bayesian hierarchical models, where models with smaller values of DIC are preferred. Table 2.2 shows the values of DIC and $\sum_{i=1}^{n} \log(\text{CPO}_i)$ for the three models implemented. Based on both criteria, we select model 2. This suggests that for this particular case of wine data, a model with both, fixed and random effects, is appropriate and that introducing grape variety-specific covariance matrices seems unnecessary. Therefore, in what follows we restrict ourselves to model 2. Figure 2.1 shows the posterior

Criterion	Model 1	Model 2	Model 3
LPML	829.2	834.9	699.6
DIC	-1,682.1	-1,691.5	-1,405.3

Table 2.2: Bayesian Model Adequacy.

distributions of β_1, β_2 and β_3. We clearly see differences across grape varieties for all the anthocyanins. Our results thus support the standard practice of differentiating grape varieties by considering their chemical properties. For example, MVAC presents the same log-concentrations between Carménère and Cabernet Sauvignon, but they differ for Merlot. In terms of classification, the most informative anthocyanins are CY, PE and MV because they yield differences in their log-concentrations between the three grape varieties. This can then be a key element in the classification effort. Figure 2.2 presents the posterior distribution of $U_1, \ldots, U_7$. We see that most of the anthocyanins show differences between valleys, although these are very small in the case of MV, the most abundant anthocyanine in most red wine varieties. For DP the Itata and Bío-Bío valleys behave differently than the rest. The last result was to be expected because the Bío-Bío and Itata valleys have special weather conditions due to their southern geographic location, which implies substantially rainier conditions throughout the year, and generally cooler climate than the northern valleys. Table 2.3 shows the classification results. The total error was 3.0%. We note here that von Baer et al. (2007) quoted an error of 4.22% for the same dataset using classical methods of discrimination. The major error in Table 2.3 is observed for Merlot, whereas for the other varieties the error was very low (0.4 to 2 %). The high error obtained by Merlot with the same dataset was explained by von Baer et al. (2007) as follows: Some years ago, Carménère, which in other countries disappeared due to phylloxera,

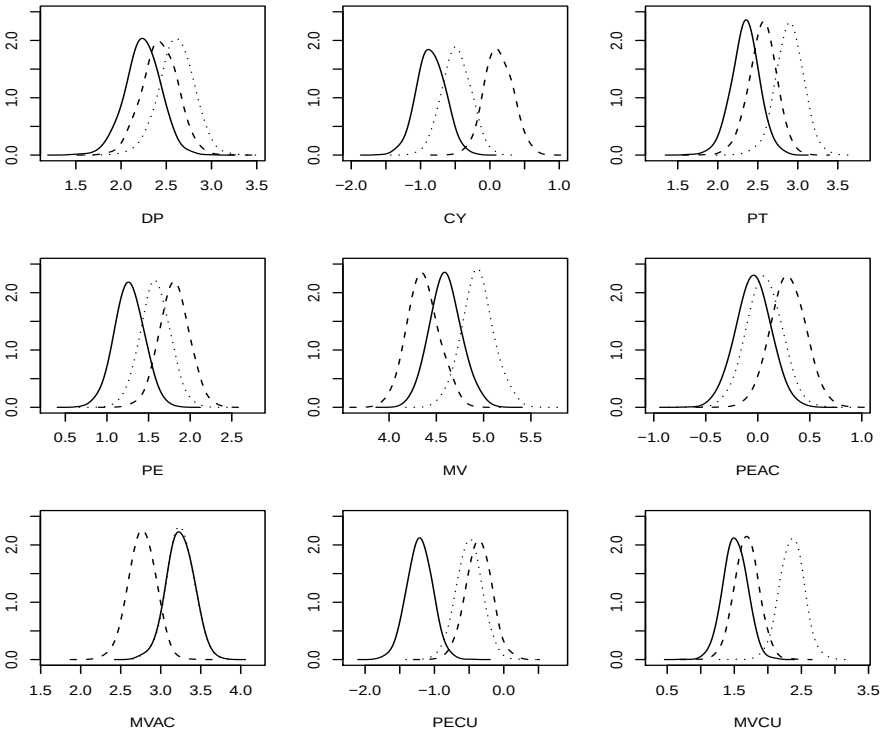

Figure 2.1: Posterior distribution of β_1, β_2 and β_3. For each of the 9 available anthocyanins, the solid line represents β_1 regression coefficients for Cabernet Sauvignon, the dashed line represents β_2 coefficients for Merlot, and the dotted line represents β_3 coefficients for Carménère

was rediscovered in Chile. Formerly, all vineyards planted with this grape variety in Chile were declared as Merlot. Hinrichsen et al. (2001) using SSR DNA markers to confirm the varietal identity, found that from a total of 93 vines of five Chilean vineyards, originally planted as Merlot, four vines matched Carménère. This leads to the conclusion that at the time of collecting wine samples, those vineyards declared as Carménère are correctly identified with high probability, but certain percentage of vineyards declared as Merlot, still correspond to Carménère. It is well known that error rates obtained from applying the classification rule to the same data used to

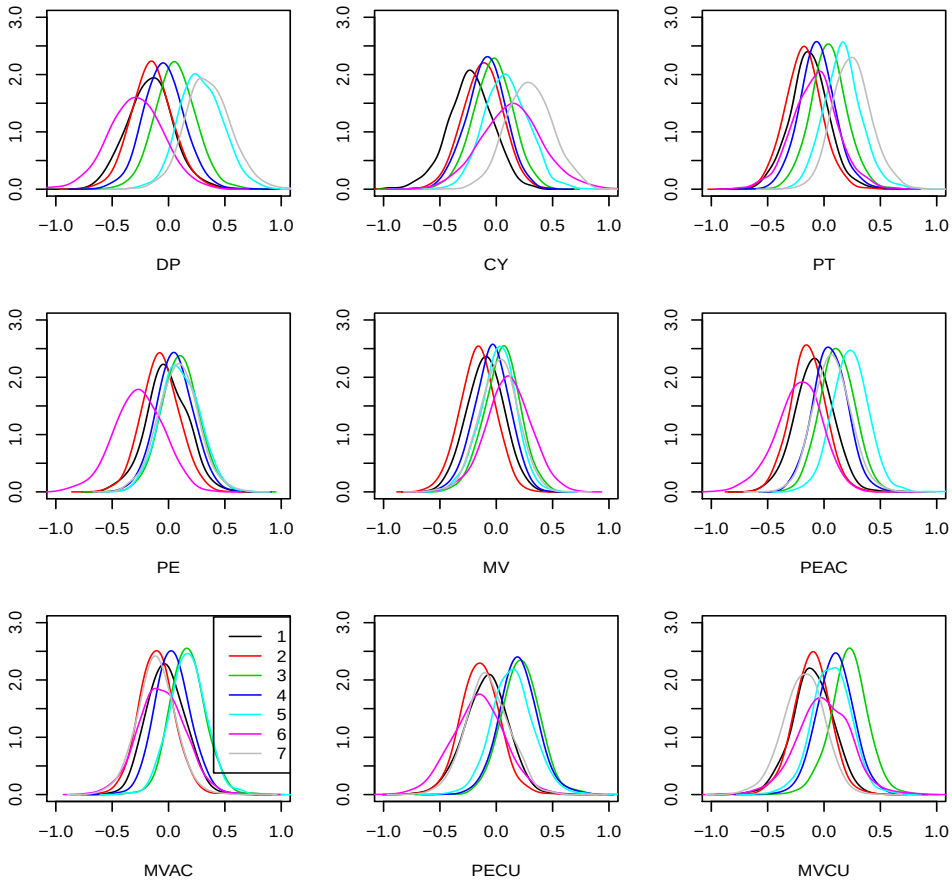

Figure 2.2: Posterior distribution of U_1,...,U_7. 1: Aconcagua, 2: Maipo, 3: Rapel, 4: Curicó, 5: Maule, 6: Itata, 7: Bío-Bío

derive it, tend to be overly optimistic and biased. Several methods are available to solve this problem. For moderately large datasets, we could consider a series of random partitions of the data into two components, one reserved for deriving the classification rule (the training sample) and the other to assessing this rule (the test sample). Under this method, the estimated error rate is the average error rate over all such partitions. For smaller datasets a cross-validation (CV) technique can be used to compensate for the lack of data, which is the road we follow here. Table 2.3 shows the

Variety	Carménère	C. Sauvignon	Merlot	Error
Carménère	93 (92)	1 (1)	1 (2)	2.1% (3.16%)
C. Sauvignon	1 (1)	227 (226)	0 (1)	0.44% (0.88%)
Merlot	9 (9)	0 (0)	67 (67)	11.84% (11.84%)
Total error				3.0% (3.51%)

Table 2.3: Misclassification rate for the three grape varieties. Values within parentheses were obtained using the leave-one-out cross-validation approach.

classification obtained by applying both, the classifier to the same data from which it was computed, and using a leave-one-out CV approach. The latter values are within parentheses. The error rate of 3.51% obtained with leave-one-out CV approach is still quite good when compared to the validated error of 5.3% obtained by von Baer et al. (2005) with classical methods.

2.6 Discussion

This paper proposes a general framework for the classification of multivariate observations from g groups. The underlying models in each group or population are given by linear multivariate models with fixed and random effects. The proposed approach allows to introduce covariates to model the mean responses. This is found to improve the classification when compared to linear or quadratic discriminant analysis, the most popular methods for food authentication. But the proposed method could be used in any situation where the aim is to classify subjects or units into g groups, on the basis of multiple responses as well as covariates.

This approach is particularly appropriate for verifying the authenticity of beverages and food, as it gives us a method to estimate the probability that the food or

beverages comply with the corresponding label description. In most cases, the data collected for authentication purposes have a multivariate structure, because more than one attribute is typically measured by unit sample. As a result, these measurements are not independent and it would not be appropriate to treat them in an univariate way. The proposed multivariate extension allows us to model the multivariate structure in a simple way. For the specific data considered here, we used information about chemical markers which are intrinsic characteristics of the food or beverages that we want to authenticate. In this context, the approach we have presented solves one important problem, as it allows to verify the authenticity of some exports that are subject to heavy regulations prior to admission to the country of destination.

The mixed-effects linear model considered here is quite general and admits several special cases. We compared three of these cases, selecting one of them for the final analysis. One interesting feature of the selected model is that the assumptions on random effects permit us to consider correlation between wine samples from the same valley. This is a reasonable assumption, because the valleys considered here have wide latitudinal variations, and these variations imply different weather and soil conditions.

In our example, we illustrated that anthocyanin profiles are very useful in the process of classifying red wines. Other chemical markers like acid or flavonol concentrations can be used for the same purpose, but we need more research about it. Incorporating information about those markers into the model is a subject currently under study.

2.7 Appendix MCMC

We list all the full conditional distributions below. The specific derivation details are straightforward and therefore omitted. For fixed effect parameters we have that:

$$\beta_j|\text{other parameters and data} \sim N_p(\tilde{\beta}_j, V_j),$$

where

$$\begin{aligned}\tilde{\beta}_j =& V_j[\sum_{k=1}^{g}\{\Sigma_k^{-1}(\sum_{i=1}^{n_k}\{x_{ij}^k y_i^k - x_{ij}^k x_{il_1}^k \beta_{l_1} - \cdots - x_{ij}^k x_{il_q}^k \beta_{l_q} - x_{ij}^k z_{i1}^k U_1 - x_{ij}^k z_{i2}^k U_2 \\ & - \cdots - x_{ij}^k z_{ir}^k U_r\})\} + \Lambda_0^{-1}\beta_{0j}],\end{aligned}$$

and $V_j = [\sum_{k=1}^{g}\{\Sigma_k^{-1}\sum_{i=1}^{n_k}(x_{ij}^k)^2\} + \Lambda_0^{-1}]^{-1}$, where $(l_1, l_2, ..., l_q) \neq j$ for $j = 1, ..., q$.

For the random effect parameters, the full conditional distributions are as follows:

$$U_j|\text{other parameters and data} \sim N_p(\tilde{U}_j, W_j),$$

where

$$\begin{aligned}\tilde{U}_j = W_j[&\sum_{k=1}^{g}\{\Sigma_k^{-1}(\sum_{i=1}^{n_k}\{z_{ij}^k y_i^k - z_{ij}^k x_{i1}^k \beta_1 - z_{ij}^k x_{i2}^k \beta_2 - \cdots - z_{ij}^k x_{iq}^k \beta_q - z_{ij}^k z_{il_1}^k U_{l_1} \\ & - \cdots - z_{ij}^k z_{il_r}^k U_{l_r}\})\}],\end{aligned}$$

and $W_j = [\sum_{k=1}^{g}\{\Sigma_k^{-1}\sum_{i=1}^{n_k}(z_{ij}^k)^2\} + S^{-1}]^{-1}$, for $(l_1, l_2, ..., l_r) \neq j$ and $j = 1, ..., r$.

For the covariance matrices $\Sigma_1, \ldots, \Sigma_g$ the full conditionals are given by

$$\Sigma_k|\text{other parameters and data} \sim IW(H_k, m_k),$$

where

$$H_k = \sum_{i=1}^{n_k} \{(y_i^k - x_{i1}^k\beta_1 - x_{i2}^k\beta_2 - \cdots - x_{iq}^k\beta_q - z_{i1}^k U_1 - z_{i2}^k U_2 - \cdots - z_{ir}^k U_r) \\ \times (y_i^k - x_{i1}^k\beta_1 - x_{i2}^k\beta_2 - \cdots - x_{iq}^k\beta_q - z_{i1}^k U_1 - z_{i2}^k U_2 - \cdots - z_{ir}^k U_r)^t\} + Q_0,$$

and $m_k = n_k + \nu_0$ for $k = 1, \ldots, g$.

For S we get:

$$S|\text{other parameters and data} \sim IW(J, l),$$

where $J = \sum_{j=1}^{r} U_j U_j^t + K_0$ and $l = m_0 + r$.

Next, for the hyperparameters $\beta_{01}, \ldots, \beta_{0q}$ we have:

$$\beta_{0j}|\text{other parameters and data} \sim N_p(\tilde{\beta}_{0j}, D_0),$$

where $\tilde{\beta}_{0j} = D_0[\Lambda_0^{-1}\beta_j + \tau_0\alpha_0]$, for $j = 1, \ldots, q$ and $D_0 = [\Lambda_0^{-1} + \tau_0^{-1}]^{-1}$.

Finally, the full conditional distribution for hyperparameter Λ_0 is given by

$$\Lambda_0|\text{other parameters and data} \sim IW(E, d),$$

where $E = \sum_{j=1}^{q}(\beta_j - \beta_{0j})(\beta_j - \beta_{0j})^t + L_0$ and $d = q + t_0$.

Chapter 3

Multivariate Bayesian Semiparametric Models for Authentication of Food and Beverages

3.1 abstract

Food and beverage authentication is the process by which food or beverages are verified as complying with its label description, e.g., verifying if the denomination of origin of an olive oil bottle is correct or if the variety of a certain bottle of wine matches its label description. The common way to deal with an authentication process is to measure a number of attributes on samples of food and then use these as input for a classification problem. Our motivation stems from data consisting of measurements of nine chemical compounds denominated Anthocyanins, obtained from samples of Chilean red wines of grape varieties Cabernet Sauvignon, Merlot and Carménère. We consider a model-based approach to authentication through a semiparametric multivariate hierarchical linear mixed model for the mean responses, and covariance matrices that are specific to the classification categories. Specifically, we propose a

model of the ANOVA-DDP type, which takes advantage of the fact that the available covariates are discrete in nature. The results suggest that the model performs well compared to other parametric alternatives. This is also corroborated by application to simulated data.

Key Words: Classification, Dependent Dirichlet Process, Wines.

3.2 Introduction

Food and beverage authentication is the process in which food or beverages are verified as complying with its label description (Winterhalter; 2007). From the viewpoint of consumers' acquisition, the mislabeling of foods represents commercial fraud (Mafra et al.; 2008). On the other hand, producers and sellers could have problems if their products are mislabeled. Food authentication is important for foods and beverages of high commercial value, like honey, wines or olive oil, because their prices depend of their quality, variety or origin. It is then important to uncover unscrupulous sellers who decide to increase their profit by adulterating these products with similar but lower quality substances. Misleading labeling might also have negative health implications, especially when the food has undeclared allergenic compounds.

Because of the growing demand from consumers of clarity and certainty in food origins and contents, the importance of food authentication has substantially increased in recent years. Many analytical tools and methods used for authenticity have been consequently developed. In particular, there is a very active area of research on the determination of chemical markers for classification and/or authentication of wines. Anthocyanin profiles are known to be specially useful for the purpose of wine variety

authentication. See, e.g., Eder et al. (1994), Berente et al. (2000), Holbach et al. (2001), Revilla et al. (2001), Otteneder et al. (2004) and von Baer et al. (2007).

Data analysis methods for authentication purposes have been developed mainly outside the statistics fields, and most of them are exploratory techniques designed to deal with multivariate datasets. Probabilistic modeling for discrimination and authentication purposes was proposed by Brown et al. (1999), who used Bayesian methods to discriminate 39 microbiological taxa using their reflectance spectra. More recently, Dean et al. (2006) used a Gaussian mixture model with labeled and unlabeled samples, with application to the authentication of meat samples from five species, and the geographic origin of olive oils. Toher et al. (2007) compared model-based classification methods such as Gaussian mixtures, with partial least squares discriminant analysis, considering samples of pure and adulterated honey.

We propose a model-based procedure to solve the authentication problem of food and beverages. The motivation comes from a dataset consisting of measurements of nine chemical compounds denominated Anthocyanins, obtained from samples of Chilean red wines of grape varieties Cabernet Sauvignon, Merlot and Carménère. We propose a semi-parametric Bayesian model that allows us to define a flexible distribution G for the joint measurements. The model has the advantage of not having to assume any parametric form, which may be particularly difficult to check in multivariate cases. Increased flexibility is added by allowing G to be formulated under the formalism of dependent random probability measures as in De Iorio et al. (2004). A key aspect of the proposed approach is that we formally extend previous univariate semi-parametric models as in De la Cruz et al. (2007b) to the multivariate case.

The rest of the paper is organized as follows. We first present the wine dataset and the related authentication problem in Section 3.3. In Section 3.4 we give a brief theoretical background about Bayesian semi-parametric models and dependent Dirichlet processes, and discuss our approach to the authentication problem. In Section 3.5 we present the model, which is an extension of the univariate semi-parametric Bayesian linear mixed model (Dey et al.; 1998) to the multivariate case. In Section 3.6 we illustrate the performance of the proposed model in a simulated data set. In Section 3.7 we apply the model to authenticate red wines samples based on their anthocyanin profile. The paper concludes in Section 3.8 with a discussion and final remarks.

3.3 The motivating dataset

We consider a dataset consisting of measurements of concentrations of nine anthocyanins on samples of Chilean red wines. Anthocyanins are a group of chemical compounds present in red wine, which confer to this beverage its characteristic red color and are transferred from the grape skins to wine during the winemaking process. The dataset includes the grape variety for each sample *as declared by the producer*, the year of harvest and the geographical origin or valley. The grape varieties in the dataset are Cabernet Sauvignon (228 samples), Carménère (95 samples) and Merlot (76 samples). All wine samples came directly from wineries located in the valleys of Aconcagua, Maipo, Rapel, Curicó, Maule, Itata and Bío-Bío in Chile. They correspond to the vintages 2001, 2002, 2003 and 2004. Anthocyanin determination was made by reverse phase HPLC based on the method described by Holbach et al. (1997), Otteneder et al. (2002) and OIV (2003), with some minor modifications. More details about anthocyanin determination for the dataset can be found

in von Baer et al. (2005) and von Baer et al. (2007). A main concern for the described dataset is the authentication of grape variety using the log-concentrations of the following anthocyanins: delphinidin-3-glucoside (DP), cyanidin-3-glucoside (CY), petunidin-3-glucoside (PT), peonidin-3-glucoside (PE), malvidin-3-glucoside (MV), peonidin-3-acetylglucoside (PEAC), malvidin-3-acetylglucoside (MVAC), peonidin-3-coumaroylglucoside (PECU), and malvidin-3-coumaroylglucoside (MVCU). To do so, we will propose a multivariate linear mixed model in Section 3.5 that attempts to characterize the variability in anthocyanin log-concentrations in terms of variety and valley of origin. We also point out that we will ignore vintage year in our development. The pragmatical reason for this is that by doing so we may easily incorporate data from new years as they become available, without the need to modify the model. In support of this choice, we refer to Gutiérrez et al. (2010) who used the year of harvest as a continuous predictor when proposing a Bayesian parametric model for the same data. The idea was to overcome this very same limitation. Yet, the effect of vintage year was negligible in that context.

3.4 Some Background Material

Semi-parametric models have both, parametric and nonparametric parts, the distinction between these being that the parameters belong to a finite and infinite dimensional space, respectively. Semi- and non-parametric Bayesian models are used mainly to avoid critical dependence on parametric assumptions. An important application of such modeling line is to random effects distributions in hierarchical models, where often little is known about the specific form of such distributions (Müller and Quintana; 2004). To handle the nonparametric part of the model we need to define

a random measure on the space of distribution functions. The most popular such choice is the Dirichlet process (DP) (Ferguson; 1973).

In a food authentication context scenario, we need to build a model that adequately accounts for all the problem-specific features. In the context of our motivating dataset, it is reasonable to think of wines coming from the same valley as being correlated, because soil and weather conditions are similar within a given valley. The usual (and simplest) way to induce a correlation structure is by incorporating random effects or sample specific parameters in a model. Let α_i denote the random effects and let z_i be a categorical covariate with k levels, (e.g. k different regions of origin). We could assume a single nonparametric prior on α_i for all samples, without reference to the levels of z_i. Alternatively, we could consider differences by putting k independent priors on α_i. These two extreme modeling strategies imply that $G_{z_1} = \cdots = G_{z_k}$ for the former and $G_{z_1} \ldots, G_{z_k}$ to be mutually independent for the latter. MacEachern (1999) proposes a modeling strategy, the Dependent Dirichlet Processes (DDP), that allows the set of random effects distributions to be similar but not identical to each other. MacEachern (1999) defines a nonparametric probability model for G_z in such a way that marginally, for each $z = z_j$, $(j = 1, \ldots, k)$, the random measure G_z follows a DP. In this context, the DP representation proposed by Sethuraman (1994) is quite useful. Sethuraman's representation establishes that any $G \sim DP(M, G_0)$ can be represented as an infinite mixture of point masses:

$$\begin{aligned} G(\cdot) &= \sum_{h=1}^{\infty} w_h \delta_{\mu_h}(\cdot), \qquad \mu_h \overset{iid}{\sim} G_0 \\ w_h &= U_h \prod_{j<h}(1 - U_j) \qquad \text{with} \qquad U_h \overset{iid}{\sim} Beta(1, M). \end{aligned} \tag{3.4.1}$$

The key idea behind the DDP is to introduce dependence across the G_z measures

by assuming the distributions of the point masses to be dependent across different levels of z (i.e. μ_{zh}), but still independent across h. If the weights are assumed to be the same across z, the dependent probability measure can be represented as $G_z(\cdot) = \sum_{h=1}^{\infty} w_h \delta_{\mu_{zh}}$. The last idea was used by De Iorio et al. (2004) in the construction of an ANOVA DDP type model. The same approach was used in spatial modeling by Gelfand et al. (2005), who used a Gaussian process for the atoms, Caron et al. (2006) in times series, De la Cruz et al. (2007b) in classification, De Iorio et al. (2009) in survival analysis and recently, by Jara et al. (2010) who proposed a Poisson-Dirichlet process for the analysis of a data set coming from a dental longitudinal study. Griffin and Steel (2006) point out that letting only the atoms to depend on covariate values may lead to certain problems when points in the domain are far from the observed data. They propose an approach that avoids this by locally updating the process and inducing dependence in the weights through distance-based similarities in the ordering of atoms, through viewing the atoms as marks in a point process. Other works where covariate dependence is introduced in the weights are Dunson et al. (2007), and Dunson and Park (2008). Müller et al. (1996) considered a completely different approach for inducing dependence in G. They used a DP mixture of normals for the joint distribution of y and z, and then focused on the implied conditional density of y given z for estimating the mean regression function. A recent reference about nonparametric Bayesian statistics, DDP models and their applications can be found in Hjort et al. (2010).

The almost sure discreteness of the Dirichlet process makes it inappropriate as a model for a continuous quantity y. A standard procedure for overcoming this difficulty

is to introduce an additional convolution so that

$$H(y) = \int f(y \mid \theta) dG(\theta) \qquad \text{with} \qquad G \sim DP(M, G_0). \tag{3.4.2}$$

Such models are known as DP mixtures (DPM) (Antoniak; 1974). The mixture model (3.4.2) can be equivalently written as a hierarchical model by introducing latent variables θ_i and breaking the mixture as

$$y_i \mid \theta_i \sim f(y_i \mid \theta_i), \qquad \theta_i \sim G, \qquad \text{and} \qquad G \sim DP(M, G_0). \tag{3.4.3}$$

For the majority of food authentication problems the responses are continuous multivariate and covariates are discrete. This is the case for the data described in Section 3.3. Thus we will adopt the popular semiparametric modeling strategy that consists of introducing dependence in the random effects distribution and then adding a convolution with a continuous kernel. The ANOVA-DDP approach of De Iorio et al. (2004) is a natural way to build the desired dependence into the model, as will be discussed below in Section 3.5. We remark here that a model that defines dependence in terms of distances would not be appropriate for an authentication problem with categorical covariates, as is our case.

3.5 The model

We first note that due to the multivariate nature of many authentication problems (which is also the case of the wine data), it would not be appropriate to treat the individual responses in an univariate way.

We assume that the *i-th* response vector is related to the covariates in a linear way. Furthermore, we assume that there are fixed and random effects in the model.

The model for the *i-th* unit in the *u-th* group is thus given by

$$\begin{aligned}(y_{iu} \mid x_{iu}, z_{iu}) &\sim N_p(Bx_{iu} + \theta_{iu}, \Sigma_u), \quad i = 1, \ldots, n_u, \quad u = 1, \ldots, g \quad (3.5.1)\\ \theta_{iu} &\sim H_z(\theta_{iu})\\ H_z(\theta) &= \int N(\theta \mid z\alpha, \tau) dG(\alpha)\\ G &\sim DP(M, G_0),\end{aligned}$$

where y_{iu} is a vector of responses in R^p, B is a $p \times q$ matrix of fixed effects, x_{iu} is a vector of covariates in R^q, θ_{iu} is a $p \times 1$ vector of unit-specific random effects, z_{iu} is a $p \times pk$ design matrix for random effects and α_i is a $pk \times 1$ vector of latent variables that define the random effects. The subscript u denotes the group or class in a classification context. Model 3.5.1 implies that $H_z(\theta) = \sum_{h=1}^{\infty} w_h N(\theta \mid z\alpha_h, \tau)$ is an infinite mixture of normal distributions. As usual in mixture models, posterior simulation proceeds by breaking the mixture in (3.5.1) by introducing latent variables α_i:

$$\theta_{iu} = z_{iu}\alpha_i + \eta_i, \quad \alpha_i \sim G, \quad G \sim DP(M, G_0), \quad \text{and} \quad \eta_i \sim N_p(0, \tau). \qquad (3.5.2)$$

By simplicity, we choose a multivariate normal model for the base measure $G_0 \equiv N_{pk}(0, R)$ and as usual in this context, we assume prior independence for all remaining parameters. The prior distribution for matrix $B = [\beta_1, \beta_2, \ldots, \beta_q]$ is assumed to be independent by columns, that is $\beta_1, \beta_2, \ldots, \beta_q$ are mutually independent with distribution given by

$$\beta_1, \ldots, \beta_q \sim N_p(\beta_{0j}, \Lambda), \quad j = 1, \ldots, q. \qquad (3.5.3)$$

The prior distributions for the variance-covariance matrices Σ_u, $u = 1, \ldots, g$, and τ

are given by

$$\Sigma_1, \ldots, \Sigma_g \sim IW_p(\nu_0, Q_0), \qquad \tau \sim IW_p(\gamma_0, \Phi_0). \tag{3.5.4}$$

We complete the Bayesian formulation of model (3.5.1) by specifying the prior for hyperparameters R, $\beta_{01}, \ldots, \beta_{0q}$, Λ and M as

$$R \sim IW_{pk}(r_0, R_0), \quad \beta_{01}, \ldots, \beta_{0q} \sim N_p(\alpha_0, \tau_0) \tag{3.5.5}$$

$$\Lambda \sim IW_p(L_0, t_0), \quad M \sim Ga(a_1, a_2) \tag{3.5.6}$$

The random distribution $H_z(\theta)$ in model 3.5.1 is dependent of the level of covariate z. As such, this is a variation of the model proposed by De Iorio et al. (2004), but our model adds fixed effects and allows us to work with multivariate data. For the wine data analysis later in Section 3.7, we will let the fixed effects be varieties and random effects be the different regions of origin. Matrix R in the model allows for correlation between all components of the vector α_i, which implies correlation between different components of the response vector and between different levels of z. The full conditional posterior distributions and details of the posterior simulation scheme are given in the Appendix section.

Consider now the classification approach. Let $y^n = (y_1, ..., y_n, x_1, ..., x_n, z_1, ..., z_n, g_1, ..., g_n)$ denote the training dataset, where y_i is the response vector, x_i is the vector of covariates for fixed effects, z_i is a vector of covariates for random effects and g_i represents the known group label for the *ith* unit. Consider a new unit for which the response y_{n+1} and covariate vectors x_{n+1} and z_{n+1} are known, but its label g_{n+1} is unknown. We want to assign a label u to the new unit, where $u \in \{1, \ldots, g\}$. Consequently it is necessary to estimate the classification probability $P(g_{n+1} = u \mid y_{n+1}, y^n)$. Following

De la Cruz-Mesía and Quintana (2007) and Gutiérrez et al. (2010) we use

$$P(g_{n+1} = u \mid y_{n+1}, y^n) \approx \frac{1}{C} \sum_{c=1}^{C} \frac{\pi_u p(y_{n+1} \mid \Theta_u^{(c)})}{\sum_l \pi_l p(y_{n+1} \mid \Theta_l^{(c)})}. \tag{3.5.7}$$

In (3.5.7), $\pi_u = P(g_i = u)$ may be taken as the empirical group proportions. We propose classifying an existing unit, i, and a future one, $n+1$, using the zero-one law considered in Hastie et al. (2001)

$$\hat{g}_i = \arg\max_u P(g_i = u \mid y^n) \quad \text{and} \quad \hat{g}_{n+1} = \arg\max_u P(g_{n+1} = u \mid y^n, y_{n+1}), \tag{3.5.8}$$

i.e. assigning the label as the category that maximizes the classification probability (3.5.7).

3.6 Classification performance of the proposed model

To evaluate the classification performance of the proposed model, we simulated a dataset considering $g = 2$, $n = 100$, $p = 2$, $q = 2$, $k = 2$. The dataset was simulated from a mixture of p-variate normal distributions, $\sum_{i=1}^{8} \omega_i N(\mu_i, \Sigma)$, where $\omega_1, \ldots, \omega_8$ are given by (0.25, 0.12, 0.13, 0.1, 0.1, 0.05, 0.12, 0.13) respectively, $\mu_1 = (1.1, 2.3)^t$, $\mu_2 = (0.1, -2)^t$, $\mu_3 = (1.3, 5)^t$, $\mu_4 = (-3, 3.4)^t$, $\mu_5 = (-0.1, 7)^t$, $\mu_6 = (1.8, 5)^t$, $\mu_7 = (-4, 1)^t$, $\mu_8 = (1, -2)^t$ and Σ is given by $\sigma_{11} = 0.932$, $\sigma_{12} = 0.11$ and $\sigma_{22} = 1.632$. Figure 3.1 shows the simulated dataset. Here, $g = 2$ means that we have to classify between two categories and $k = 2$ means that we have two levels for the covariate z. The hyperparameters values were taken as $\beta_0 = (0, 0)^t$, $\tau_0 = 100 I_2$, $Q_0 = I_2$, $L_0 = I_2$, $\nu_0 = 4$, $r_0 = 4$, $t_0 = 4$, $R_0 = I_{pk}$, $\gamma_0 = 4$, $\phi_0 = 0.001 I_p$ and $a_1 = a_2 = 1$. Table 1 shows the classification results of the proposed Bayesian semiparametric model (BSP), comparing with linear discriminant analysis (LDA), which is the usual technique used

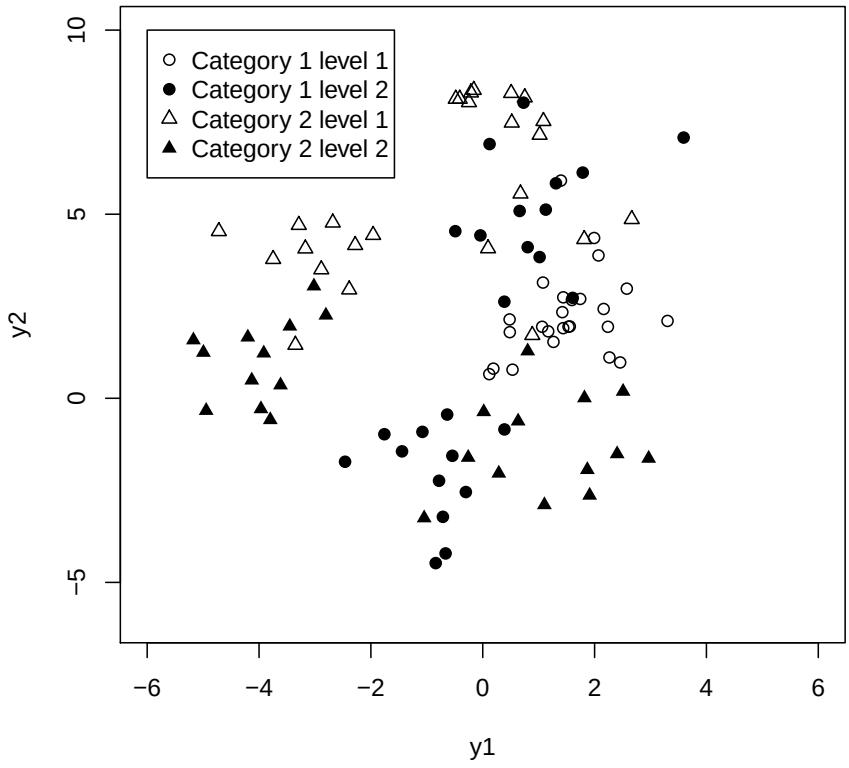

Figure 3.1: Simulated dataset

in the literature for this type of problem, and a parametric (BP) version of model (3.5.1), defined as:

$$\begin{aligned}(y_{iu} \mid x_{iu}, z_{iu}) &\sim N_p(Bx_{iu} + \theta_{iu}, \Sigma_u), \quad i = 1, \ldots, n, \quad u = 1, \ldots, g \quad (3.6.1)\\ \theta_{iu} &= z_{iu}\alpha + \eta_i, \qquad \eta_i \sim N_p(0, \tau)\\ \alpha &\sim N_{pk}(0, R)\end{aligned}$$

Using the proposed BSP model, we obtained a classification error of 7.0% in the

training set and 16% using leave-one-out cross-validation (LOOCV). In contrast, the BP model resulted in a classification error of 12.0% in the training set and 24% under LOOCV, while the corresponding figures for the LDA were 25.0% and 27%, respectively. A common way to assess the performance of classification rules is the Receiver Operating Characteristic curve (ROC) shown in Figure 3.2, which plots the true positive rate against the false positive rate for all the different possible cutpoints. From the ROC curves we also calculated the Area Under ROC curve (AUC) for the three models, with higher values corresponding to models with better discrimination capabilities. We obtained 0.9792 for the BSP model, 0.9334 for the BP model, and 0.7464 for LDA. These results clearly suggest the superiority of the proposed BSP model for wine authentication, compared to the other alternatives.

Another important aspect of the analysis concerns comparing model adequacy of the BP versus our BSP proposal. To this effect we calculated the Conditional Predictive Ordinates (CPO_i) (Chen et al.; 2000), summarized in the log-pseudo marginal likelihood statistic $LPML = \sum_{i=1}^{n} \log(CPO_i)$ (Geisser and Eddy; 1979), and the Deviance Information Criterion (DIC) (Spiegelhalter et al.; 2002). Models with lower DIC and with higher LPML values are to be preferred. The DIC values were 730.0 and 855.2 for the BSP and BP models, respectively. Furthermore, the corresponding LPML values were -370.5 and -427.9. Both criteria consistently point to the superiority of the BSP model compared to the BP one. Overall, the results suggest that the BSP model is more flexible, specially when the data cluster between and within covariate levels.

		BSP		BP		LDA	
		1	2	1	2	1	2
Category	1	46 (43)	4 (7)	47 (35)	3 (15)	42 (42)	8 (8)
	2	3 (9)	47 (41)	9 (9)	41 (41)	17 (19)	33 (31)

Table 3.1: Classification performance. Values within parenthesis were obtained using leave-one-out cross-validation technique

3.7 Performance of the model with wine dataset

We consider now application of the proposed BSP model to the wine dataset. The response vector is formed by the nine anthocyanins listed in Section 3.3. As covariates, we use grape variety (fixed effects) and valleys (random effects). The hyperparameter values were taken as $\beta_0 = (0, 0, 0, 0, 0, 0, 0, 0, 0)^t$, $\tau_0 = 100I_9$, $Q_0 = 0.1I_9$, $L_0 = 0.01I_9$, $\nu_0 = 11$ $r_0 = 65$, $t_0 = 11$, $R_0 = 10I_{pk}$, $\gamma_0 = 11$, $\phi_0 = 0.01I_p$ and $a_1 = a_2 = 1$, where $p = 9$, $q = 3$ and $k = 7$. The resulting prior densities are proper, but the one for B is vague and hence relatively uninformative. The prior density for R is relatively uninformative too. All the variance covariance matrices priors were assumed diagonal.

Table 3.2 shows the classification results, where the values within parenthesis were obtained using a LOOCV approach. The classification error obtained in the training set was 0.5%, and 3.25% under LOOCV. These values are better than those obtained by Gutiérrez et al. (2010) with the same dataset but applying a Bayesian parametric model, namely, 3.0% in the training set and 3.51% using LOOCV.

Table 3.3 shows the AUC values, which were calculated based on separate ROC curves for each grape variety, and for each of the BSP and BP models. All these values are very high, with the BSP model attaining the best performance across the three grape varieties. When comparing the BSP and BP models, the DIC and LPML

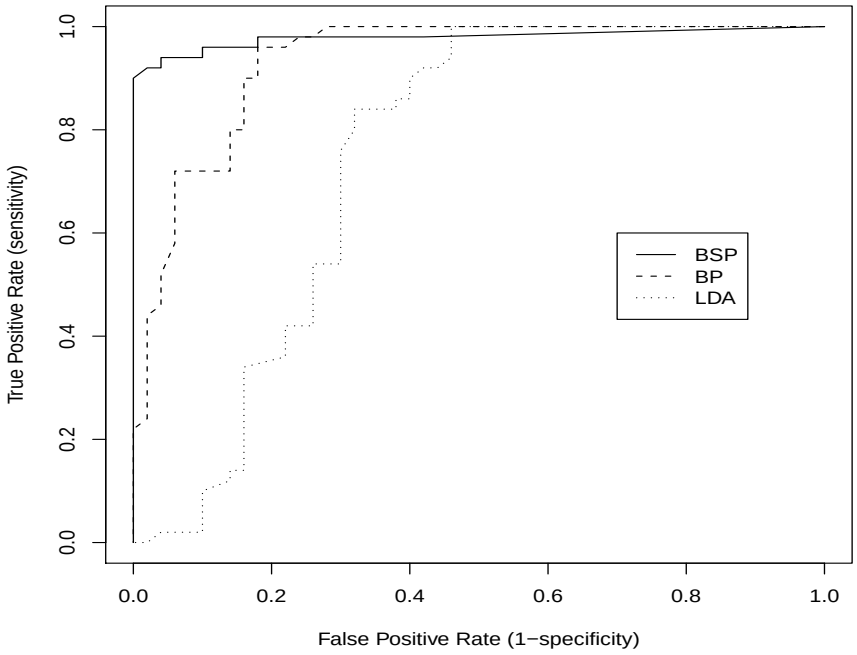

Figure 3.2: ROC curves for classification under Bayesian semiparametric model BSP, Bayesian parametric BP and linear discriminant analysis LDA.

statistics values were -5,901.5 and 2,430.1 for the former, and -5,578.7 and 2,291.8 for the second. Again, these results suggest that the proposed BSP model provides a better fit.

Figure 3.3 displays bivariate posterior predictive distributions for Carménère wines from the valleys of Aconcagua, Maipo, Rapel and Curicó considering anthocyanins PECU and MVCU. The points on the graph are the observed values. We can see the changes in the posterior predictive distribution across valleys. Predictions for the Aconcagua valley shows less variation compared to Maipo valley. Predictions for The Rapel valley show more variability, with some evidence of asymmetry, as dictated by

Variety	Carménère	C. Sauvignon	Merlot	Error
Carménère	94 (93)	1 (1)	0 (1)	1.05% (2.1%)
C. Sauvignon	0 (0)	228 (225)	0 (3)	0.0% (1.31%)
Merlot	1 (8)	0 (0)	75 (68)	1.33% (10.52%)
Total error				0.5% (3.25%)

Table 3.2: Misclassification rate for the three grape varieties

Grape variety	AUC BSM	AUC BPM
Cavernet Sauvignon	0.999999	0.9969221
Merlot	0.999999	0.9967403
Carménère	0.999999	0.9963574

Table 3.3: Area under ROC curve

the observed data, but the model provides a reasonable fit to this behavior. Finally, the Curicó valley also exhibit asymmetry.

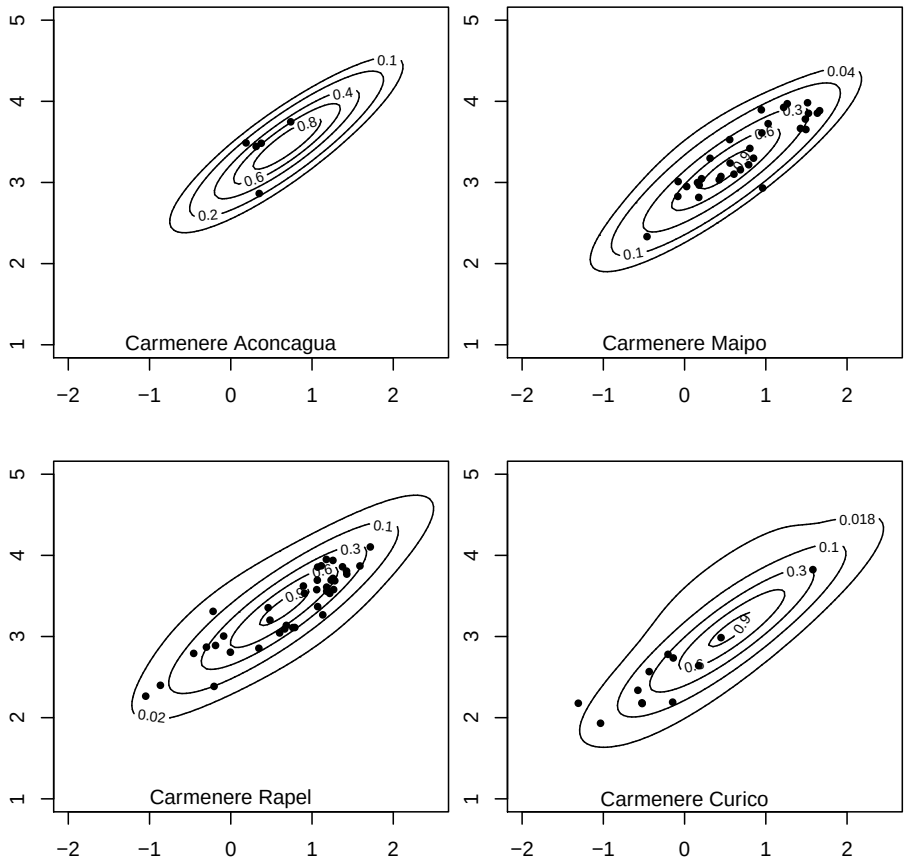

Figure 3.3: Bivariate posterior predictive distributions with BSP model for Carménère wines from the Aconcagua, Maipo, Rapel and Curicó, with points representing observed values. The anthocyanins considered here were PECU and MVCU.

Figure 3.4 shows the bivariate predictive posterior distributions for Cabernet Sauvignon, Carménère and Merlot from Curicó valley considering the DP and CY anthocyanins. This plot is interesting because it shows how informative are DP and CY in terms of the target classification. These two anthocyanins show that some Merlot samples are located near the Carménère ones. This behavior is reasonable because some years ago, Carménère, which in other countries disappeared due to phylloxera, was rediscovered in Chile. Formerly, all vineyards planted with this grape variety in Chile were declared as Merlot. Using SSR DNA markers to confirm varietal identity, Hinrichsen et al. (2001) found that from a total of 93 vines of five Chilean vineyards, originally planted as Merlot, four vines matched Carménère. This leads to the conclusion that at the time of collecting wine samples, those vineyards declared as Carménère are correctly identified with high probability, but certain percentage of vineyards declared as Merlot, still correspond to Carménère.

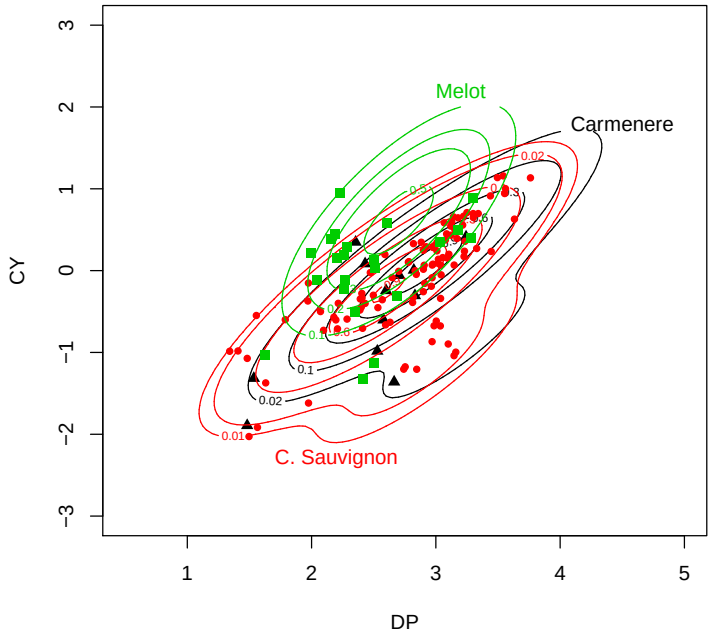

Figure 3.4: Bivariate posterior predictive distributions for Cabernet Sauvignon, Merlot and Carménère wines from the Curicó valley, with points representing the observed values.

3.8 Concluding Remarks

We have proposed a linear mixed effects model for wine authentication, featuring a flexible model for random effects that does not require the restricting ourselves to a given parametric form. We did so by resorting to Dependent Dirichlet Processes, which allow the set of random effects distributions to be similar but not identical to each other, depending on levels of a covariate. For the authentication problem, dependence on covariate levels is important because it is reasonable to think that foods or beverages that come either from the same region of origin, or those which were made with the same technology, could be similar or correlated. The ANOVA-DDP approach was suitable to our purposes, but other types of nonparametric priors could be considered.

The proposed BSP model provided a better fit to the data than a parametric alternative, as we showed in the simulation example and in the application to the wine data. In terms of the target classification, the BSP model also provided slightly better results than other alternatives. Our proposal was motivated by food authentication, but it could be used in any situation where the aim is to classify subjects or units into g groups, on the basis of multiple responses and covariates.

3.9 Appendix

In this section we give the MCMC algorithm that was used for posterior simulation under the proposed model. Because the model is of conjugate type, we use algorithm 2 in Neal (2000). Let $\mathbf{c} = (c_1, \ldots, c_n)$ denote a vector that captures the clustering of α_i and let $\alpha = (\alpha_c : c \in \{c_1, \ldots, c_n\})$. To resample the configurations c_i, we proceed with the following two steps:

Step 1

If $c = c_j$ for some $j \neq i$ we compute the probability that the i-th element in $\mathbf{c}$ equals other element in the same set as

$$P(c_i = c \mid c_{-i}, \theta_i, \alpha) = b\frac{n_{-i,c}}{n-1+M}(2\pi)^{-p/2}|\tau|^{-1/2}\exp\left\{-\frac{1}{2}(\theta_i - z_i\alpha_c)^t\tau^{-1}(\theta_i - z_i\alpha_c)\right\}. \quad (3.9.1)$$

Here $n_{i,c}$ is the number of c_i that are qual to c, c_{-i} are all the c_j for $j \neq i$ and b is such that if $c = c_j$ then $\sum_{j:j\neq i}\{P(c_i = c)\} + P(c_i \neq c_j \forall j \neq i) = 1$. Next, we compute the probability that c_i is different to any other element in $\mathbf{c}$ as

$$P(c_i \neq c_j \text{for all} j \neq i \mid c_{-i}, \theta_i, \alpha) = b\frac{M}{n-1+M}(2\pi)^{-p/2}|\tau|^{-1/2}|R|^{-1/2}|D_i|^{1/2}\times \exp\left\{\frac{-1}{2}[\theta_i^t\tau^{-1}\theta_i - [\theta_i^t\tau^{-1}z_i]D_i[z_i^t\tau^{-1}\theta_i]]\right\}. \quad (3.9.2)$$

If the imputed value of c_i, sampled based on (3.9.1) and (3.9.2), is not associated with any other observation, it is necessary to draw a value of α_{c_i} from H_i, the posterior distribution for α based on the prior G_0 and the single observation θ_i. In our case H_i is given by $H_i \equiv N_{pk}(\tilde{\alpha_i}, D_i)$ where $D_i = [z_i^t\tau^{-1}z_i + R^{-1}]^{-1}$, and $\tilde{\alpha_i} = D_i[z_i^t\tau^{-1}\theta_i]$.

Step 2

In the second step, for all $c \in \{c_1, \ldots, c_n\}$ we draw a new value α_c given all the θ_i for which $c_i = c$, that is, from the posterior distribution based on the prior G_0 and all the data points currently associated with latent class c. In our case, this is given by $N_{pk}(\tilde{\alpha_c}, E)$, where $E = [\sum_{i:c_i=c} z_i^t\tau^{-1}z_i + R^{-1}]^{-1}$ and $\tilde{\alpha_c} = E[\sum_{i:c_i=c} z_i^t\tau^{-1}\theta_i]$.

Now we list all the full conditional distributions for the parametric part of the model. The specific derivation details are straightforward and therefore omitted.

- For fixed effect parameters we have

$\beta_j \mid$ other parameters and data $\sim N_p(\tilde{\beta_j}, V_j)$, where

$$\tilde{\beta_j} = V_j[\sum_{u=1}^{g}\{\Sigma_u^{-1}\sum_{i=1}^{n_u}\{x_{ij}y_i - x_{ij}x_{il_1}\beta_{l_1} - \cdots - x_{ij}x_{il_q}\beta_{l_q} - x_{ij}\theta_i\}\} + \Lambda^{-1}\beta_{0j}], \text{ and}$$

$$V_j = [\sum_{u=1}^{g}\{\sum_{i=1}^{n_u} x_{ij}^2\Sigma_u^{-1}\} + \Lambda^{-1}]^{-1} \quad \text{where} \quad (l_1, l_2, \ldots, l_q) \neq j \qquad j = 1, ..., q$$

- For the random effects parameters $\theta_{1u}, \ldots, \theta_{nu}$, $u = 1, \ldots, g$ we have that:

$$\theta_{iu} \mid \text{other parameters and data} \sim N_p(\tilde{\theta_{iu}}, Q_u), \qquad i = 1, \ldots, n, \text{ where}$$

$$Q_u = [\tau^{-1} + \Sigma_u^{-1}]^{-1} \qquad \text{and} \qquad \tilde{\theta_{iu}} = Q_u[\tau^{-1}z_i\alpha_i + \Sigma_u^{-1}y_i - \Sigma_u^{-1}Bx_i]$$

- For hyperparameters $\beta_{01}, \ldots \beta_{0q}$ we have

$$\beta_{0j} \mid \text{other parameters and data} \sim N_p(\tilde{\beta_{0j}}, D_0), \text{where}$$

$$B_{0j} = D_0[\lambda^{-1}\beta_j + \tau_0^{-1}\beta_0] \qquad j = 1, ..., q \qquad \text{and} \qquad D_0 = [\Lambda^{-1} + \tau_0^{-1}]^{-1}$$

- For hyperparameter Λ we have

$$\Lambda \mid \text{other parameters and data} \sim IW_p(d, E), \text{ where}$$

$$E = \sum_{j=1}^{q}(\beta_j - \beta_{0j})(\beta_j - \beta_{0j})^t + L_0 \qquad \text{and} \qquad d = q + t_0$$

- Finally, for the covariance matrices $\Sigma_1, \ldots, \Sigma_g$, τ and R we have

$$\Sigma_u \mid \text{other parameters and data} \sim IW_p(l_u, H_u), \text{ where}$$

$$H_u = \sum_{i=1}^{n_u}(y_i - Bx_i - \theta_i)(y_i - Bx_i - \theta_i)^t + Q_0 \qquad \text{and} \qquad l_u = n_u + \nu_0$$

$$\tau \mid \text{other parameters and data} \sim IW_p(s, T), \text{ where}$$

$$T = \sum_{i=1}^{n}(\theta_i - z_i\alpha_i)(\theta_i - z_i\alpha_i)^T + \Phi_0 \qquad \text{and} \qquad s = n + \gamma_0$$

$$R \mid \text{other parameters and data} \sim IW_{pk}(f, O), \text{ where}$$

$$O = \sum_{i=1}^{n} \alpha_i \alpha_i^t + R_0 \qquad \text{and} \qquad f = n + r_0$$

Chapter 4

Optimal Information in Authentication of Food and Beverages

4.1 abstract

Food and beverage authentication is the process by which food or beverages are verified as complying with their label descriptions (Winterhalter; 2007). A common way to deal with an authentication process is to measure attributes such as groups of chemical compounds on samples of food, and then use these as input for a classification method. In many applications there may be several types of measurable attributes. An important problem thus consists of determining which of these would provide the best information, in the sense of achieving the highest possible classification accuracy at low cost. We approach the problem under a decision theoretic strategy, by framing it as the selection of an optimal test (Geisser and Johnson; 1992) or as the optimal dichotomization of screening tests variables (Wang and Geisser; 2005), where the "test" is defined through a classification model applied to different groups of chemical compounds. The proposed methodology is applied and compared in the context of a

dataset consisting of measurements of nineteen chemical compounds (anthocyanins, organic acids and flavonols) on samples of Chilean red wines, where the main goal is to determine the combination of chemical compounds that is best for authentication of wine varieties, considering the losses of wrong decision and the cost of determination.

Key Words: Loss function; Classification; Wine

4.2 Introduction

Authentication of food and beverages is the process by which food or beverages are verified to match their label description (Winterhalter; 2007). Authentication problems are typically treated from the viewpoint of classification (Brown et al. (1999); Dean et al. (2006); Toher et al. (2007); Gutiérrez et al. (2010)). The accuracy of a classification model used for authentication depends on the available information. An important issue in this process is to determine what chemical compounds should be analyzed to verify that a given food product complies with its label description. For example, to verify the authenticity of tea varieties and products, different groups of chemical compounds like catechins, total phenolics, theaflavins or caffeine, have been proposed (Engelhardt; 2007).

Motivated by a dataset concerning samples of red wines from different varieties and origins (Gutiérrez et al.; 2010), we address the problem of selecting the compounds that give the best performance. By this we mean that the cost of analyzing the compounds should be low and the accuracy of results good. From a Bayesian viewpoint this is an optimal decision problem (Berger; 1985). A similar problem arises in a biomedical context, when it is necessary to choose between two screening

tests. A possible solution implies the definition of a loss function that combines the penalty associated to a wrong decision with the cost of each test. See for example Geisser and Johnson (1992). A related approach involves the optimal dichotomization of screening test variables, as in e.g., (Wang and Geisser; 2005). See below for a discussion of both methods.

We adapt the methods in Geisser and Johnson (1992) and in Wang and Geisser (2005) to the optimal selection of information for the authentication process. We assume that various types of chemical compounds can be potentially measured, and that additional information leads to increased classification accuracy. Our "test" is a multivariate classification model (Gutiérrez and Quintana; 2010) that can be applied to the different groups of chemical compounds. We consider two populations: one where food samples comply with their label description and the other where they do not. For simplicity, we refer to these as populations having characteristics U or U^c, respectively. The method by Geisser and Johnson (1992) considers the problem of optimally deciding whether a certain characteristic is present, based on one or two screening tests. The authors discuss the relative merits of giving either one or two tests, including the order in which they might be given, as well as their costs. For this method, the input consists of the results of a screening test, e.g. the ELISA test for presence or absence of AIDS. In our case we take the input as the results coming from the classification model, namely, the posterior probability that the sample has characteristic U. To do so, it is necessary to select a threshold for the posterior probability that a given individual is assigned to characteristics U or U^c. On the other hand, the method by Wang and Geisser (2005) considers the problem of finding a most favorable *dichotomizer*, that is, a cut-off value or threshold for which optimal

test performance is obtained. This is so because the accuracy of the screening test often depends on the dichotomization of the test outcome variable. Determination of the optimal dichotomizer is considered under a decision-theoretic Bayesian approach. For this method, the input consists of the outcome test variable values, e.g. in AIDS screening, an ELISA test measuring the level of certain antigens in the blood for ascertaining the presence of the human immunodeficiency virus (HIV) antibodies, and a cut-off value is chosen for dichotomizing the screening outcomes, to indicate the presence or absence of the antibodies (Wittes; 1987). When adapting the Wang and Geisser (2005) method to our case, we take the log-posterior predictive density for a new sample as input. It will be argued that the expected loss function depends on this value, so that we simply proceed to find an "optimal" dichotomizer using minimization techniques. In either case, we consider a loss function that balances the worth of correctly classifying samples with the cost required to measure the chemical compounds. The optimal decision is then the one that minimizes the expected loss.

The rest of the paper is organized as follows. In section 4.3 we introduce the ideas and concepts for defining a loss function and the two approaches for estimating the expected loss. In Section 4.4 we describe the motivating wine dataset, which includes measurements of nineteen chemical compounds: Anthocyanins, Organic Acids and Flavonols. We also briefly describe a classification model that we have found to be particularly useful for authentication in this context (Gutiérrez and Quintana; 2010). We implement and compare the two methods for optimal information selection, considering all possible combinations of groups of compounds that can be used. We conclude in section 4.5, where the the results are compared, and a final discussion of the proposed methodology is given.

4.3 Methodology

4.3.1 Decision-theoretic approach to find optimal information

We assume a classification approach for which a training dataset concerning n experimental units $\{(y_i, x_i, g_i)\}$, $i = 1, \ldots, n$ is available. Here, $y_i = (y_{i1}, \ldots, y_{ip})' \in R^p$ is the observed response vector for the *ith* unit, and $x_i = (x_{i1}, \ldots, x_{iq})$ and $g_i \in E = \{1, \ldots, g\}$ denote the corresponding covariate vector and known group label, respectively. Let $y^n = (y_1, \ldots, y_n, x_1, \ldots, x_n, g_1, \ldots, g_n)$ denote the complete data. Let $y^{n+1} = (y_{n+1}, x_{n+1})$ be the observed data vector for a future unit, for which the corresponding label g_{n+1} is unknown. We adopt a predictive approach for classification, so that the focus is on inference for the g_{n+1} value. Assume a partition of E as $E = U \bigcup U^c$, where $U = \{k\}, k \in E$ and $U^c = \{j \in E \mid j \neq k\}$. Using the above setup, we consider two subpopulations: the units that comply with its label description will be said to have characteristic U, and U^c otherwise. In this context, there are two possible actions, $g_{n+1} = U$ denoted by A, and $g_{n+1} = U^c$ denoted A^c. Assume that we have a generic hierarchical model for the available data, denoted by $\mathcal{M}$, of the form

$$y_i \mid \delta_i, x_i \sim p(y_i \mid \delta_i, x_i), \qquad \delta_i \sim G(\delta_i \mid \phi). \tag{4.3.1}$$

In simple words, the data vector y_i for the *ith* sampling unit are assumed to be sampled from a probability model parameterized by a vector δ_i, where x_i represents a covariate vector. We consider now an additional ingredient of the problem at hand, namely, the dimension p of vector y_i can be changed based on the available information, and on the cost required to obtain that information. For example, in our application, $p = 9$

when we choose to use the anthocyanin compounds, $p = 4$ when we use the Organic acids, $p = 6$ for flavonols, and $p = 19$ when we use a combinations of the three compound groups. See the appendix section for a full list of the mentioned groups of chemical compounds. In all cases the dimension of x_i remains constant, so the covariates are the same for all models. Denote by $\mathcal{M}_{p_j}$ a model of the form (4.3.1), with a corresponding response vector $y_i \in R^{p_j}$, $j = 1, 2, \ldots$. We assume there is a cost c_j associated with model $\mathcal{M}_{p_j}$, and losses in making wrong decisions. Selecting a particular model $\mathcal{M}_{p_j}$ implies selecting the compounds or combinations of them that yield the best performance. By this we mean that the cost c_j of determining the compounds should be low and the accuracy of the results should be good. In our case, we have information on all the different compounds, but we shall take the perspective of identifying the groups or combinations thereof that are most useful for classification. The idea is that, if in the future a producer needs to verify, for example, whether a sample of wine is Cabernet Sauvignon or not, then the analyst will not need to measure all compounds included in the current dataset, but only those providing the best classification for this grape variety at low cost. Therefore we propose a solution that implies the definition of a loss function that combines the penalty associated to a wrong decision with the cost c_j of each model $\mathcal{M}_{p_j}$.

In the case of the actions A and A^c and states U and U^c, a useful loss function is given in Table 4.1. For example, the loss of deciding action A is l_{AU} when the true state is U.

Now, given a decision rule R for model $\mathcal{M}_{p_j}$, the optimal decision is the one

Decision rule outcome	True State U	True State U^c
A	l_{AU}	l_{AU^c}
A^c	l_{A^cU}	$l_{A^cU^c}$

Table 4.1: Loss function

minimizing $E(Loss \mid R)$, given by

$$E(Loss \mid R) = l_{AU}Pr(A,U) + l_{AU^c}Pr(A,U^c) + l_{A^cU}Pr(A^c,U) + l_{A^cU^c}Pr(A^c,U^c) \tag{4.3.2}$$

If the cost associated to model $\mathcal{M}_{p_j}$, c_j, is expressed in the same unit as the losses, then we would minimize $E(Loss \mid R)+c_j$. We can therefore estimate the expected loss for each model under consideration, and select the one yielding the lowest expected loss. To do so, it is necessary to assign values to the losses and the corresponding probabilities as expressed in (4.3.2). The order of magnitude of the quantities in Table 4.1 is crucial for defining the optimal model, and this choice depends on the analyst's viewpoint. In authentication problems, it could be argued that from the viewpoint of a "honest producer", i.e. a producer that says the truth with probability 1,

$$l_{AU} \leq l_{A^cU^c} \leq l_{AU^c} \leq l_{A^cU}. \tag{4.3.3}$$

The worst-case scenario occurs when U is present in the food under authentication but the model estimates this to be not true. A costumer may interpret such model results, as an indication that the producer is committing a fraud, and the losses for the producer could be devastating. A different situation arises when the food under authentication does not have the characteristic U, but the model estimates

that U is present. If so, a customer may think that the producer does not have enough knowledge of their product, which could generate distrust and possible losses. When U is absent from the food under authentication and the model estimates this to be true, the image of the honest producer is strengthened and, probably, no loss is generated. The best scenario is when U is present in the food, and the model estimates this to be true, in which case the honest producer is reliable and most of the time a profit will be made.

4.3.2 Estimation of the expected loss function

Note first that we can rewrite the expected loss function (4.3.2) as

$$\begin{aligned} E(Loss \mid R) &= Pr(U)Pr(A \mid U)(l_{AU} - l_{A^cU}) \\ &+ (1 - Pr(U))Pr(A^c \mid U^c)(l_{A^cU^c} - l_{AU^c}) + Pr(U)l_{A^cU} + (1 - Pr(U))l_{AU^c} \end{aligned} \quad (4.3.4)$$

Denote the probabilities in (4.3.4) as $\pi = Pr(U)$, the probability that a random drawn unit from the population exhibits characteristic U; $\eta = Pr(A \mid U)$, the probability that the model correctly estimates the presence of U (sensitivity); and $\varphi = Pr(A^c \mid U^c)$, the probability that the model correctly estimates the absence of U (specificity).

Conceptually, when all of these quantities are known, we only need to introduce the costs and/or losses, and a few manipulations to determine the optimal decision procedure, given an outcome of the classification model $\mathcal{M}_{p_j}$. In our case, as in many other practical situations, π, η and φ are all unknown.

A simple approach for estimating π, η and φ was proposed by Geisser and Johnson (1992) in the context of a screening test. The method consists of applying the model to n_1 units which are known to have the characteristic U, and also to n_2 units which

are known to be free of U. Assuming that r_1 out of n_1 yield A in the first sample, and r_2 out of n_2 yield A^c in the second, we obtain binomial distributions for both r_1 and r_2, with parameters η and φ, respectively. If π is unknown, we need an additional independent sample of size ν, from which we can count the number t_u of units having U. We obtain another binomial distribution for t_u with parameter π. Let $d = (r_1, n_1, r_2, n_2, t_u, \nu)$. Since the samples are independent, the likelihood function is given by

$$L(\eta, \varphi, \pi \mid d) = L(\eta \mid n_1, r_1)L(\varphi \mid n_2, r_2)L(\pi \mid \nu, t_u). \tag{4.3.5}$$

Under a Bayesian viewpoint it is necessary to assign prior distributions $p(\eta, \varphi, \pi)$ on (η, φ, π), from which the joint posterior density is obtained as

$$p(\eta, \varphi, \pi \mid d) \propto p(\eta, \varphi, \pi)L(d \mid \eta, \varphi, \pi). \tag{4.3.6}$$

We now describe how to obtain the quantities r_1 and r_2 from model $\mathcal{M}_{p_j}$, using the predictive probability $P(g_{n+1} = u|y_{n+1}, y^n)$, which can be approximated as (De la Cruz-Mesía and Quintana; 2007; Gutiérrez et al.; 2010)

$$P(g_{n+1} = u|y_{n+1}, y^n) \approx \frac{1}{C}\sum_{c=1}^{C} \frac{\pi_u p(y_{n+1}|\Theta_u^{(c)})}{\sum_l \pi_l p(y_{n+1}|\Theta_l^{(c)})}. \tag{4.3.7}$$

We use (4.3.7) as follows: take action A if $P(g_{n+1} = u|y_{n+1}, y^n) > p_0$ and A^c otherwise. This rule is of course dependent on the threshold or cut-off value p_0. Therefore, the results depend on the choice of $p_0 \in (0, 1)$, but it is easy to evaluate the expected loss on a suitable grid of values on $(0, 1)$, from which we can select the value of p_0 that gives the minimal expected loss.

A second approach for estimating η and φ, proposed by Wang and Geisser (2005) in the context of dichotomization of screening test variables, consists of assuming that

$l_{AU} = l_{A^cU^c} = 0$ (i.e., no loss for right decisions), $l_{AU^c} = b$ and $l_{A^cU} = a$ with $b \leq a$. Under these assumptions, (4.3.4) simplifies to

$$E(Loss) = b(1-\pi)(1-\varphi) + a\pi(1-\eta). \tag{4.3.8}$$

Wang and Geisser (2005) further assume that $1-\eta$ and φ can be reexpressed in terms of two distribution functions, $\eta = 1 - F_1(\ell)$ and $\varphi = F_2(\ell)$ where ℓ is the result of a classification model $\mathcal{M}_{p_j}$, which in our case corresponds to $\ell = \log(p(y_{n+1} \mid y^n))$. This approach allows us to find the minimum expected loss with respect to ℓ and to find $\ell_0 = \arg\min_\ell L(\ell)$, the optimal dichotomization of the classification model $\mathcal{M}_{p_j}$. Assume that F_i has density function f_i, depending on a parameter ξ_i, $i = 1, 2$. To estimate ξ_1 and ξ_2, it is necessary to fit the model to n_1 units for which U is present, and also to n_2 units for which U is absent. For $i = 1, 2$, let $\ell_{ij} = \{\ell_{ij1}, \ldots, \ell_{ijn_i}\}$ be the values of $\log(p(y_{n+1} \mid y^n))$ with model $\mathcal{M}_{p_j}$ applied to each of the n_i units above. Wang and Geisser (2005) suggest using the predictive distribution

$$\tilde{F_{ij}}(\ell \mid \ell_{ij}) \propto \int F_i(\ell \mid \xi_i) \prod_{m=1}^{n_i} f_i(\ell_{ijm} \mid \xi_i) p_i(\xi_i) d\xi_i \quad i = 1, 2, \tag{4.3.9}$$

from which the expected loss for model $\mathcal{M}_{p_j}$, as a function of ℓ, can be expressed as

$$L_j(\ell) = b(1-\pi)\{1 - \tilde{F}_{2j}(\ell \mid \ell_{2j})\} + a\pi\tilde{F}_{1j}(\ell \mid \ell_{1j}). \tag{4.3.10}$$

The value of π can be inferred just as in the first approach. For simplicity, and to ensure availability of an analytical expression for the posterior predictive distribution, we assume that ℓ, the value of $\log(p(y_{n+1} \mid y^n))$, is distributed as $F_i \sim N(\mu_i, \sigma_i^2)$, $i = 1, 2$ and that the prior distributions for μ_i and σ_i^2 are given by

$$p_i(\mu_i, \sigma_i^2) = N(\mu_i \mid \mu_{i0}, n_{i0}/\sigma_i^2) IG(\sigma_i^2 \mid \alpha_{i0}, \beta_{i0}). \tag{4.3.11}$$

It follows that the posterior predictive distribution follows a Student t distribution $t(\tau_i, \lambda_i, \nu_i)$, with parameters given by

$$\begin{aligned}
\tau_i &= \frac{n_{i0}\mu_{i0} + n_i\bar{\ell_{ij}}}{n_{i0} + n_i} \\
\lambda_i &= \frac{n_i + n_{i0}}{n_i + n_{i0} + 1}(\alpha_{i0} + \frac{1}{2}n_i)\left[\beta_{i0} + \frac{1}{2}(n_i - 1)s_i^2 + \frac{1}{2}\frac{n_{i0}n_i}{n_{i0} + n_i}(\mu_{i0-\bar{\ell}_{ij}})^2\right]^{-1} \\
\nu_i &= 2\alpha_{i0} + n_i
\end{aligned}$$

The value of ℓ_0 can be obtained numerically from Newton-Raphson's method. Given an initial value $\ell_0^{(k=0)}$, we iteratively evaluate

$$\ell_0^{(k)} = \ell_0^{(k-1)} - \frac{L'(\ell_0^{(k-1)})}{L''(\ell_0^{(k-1)})}, \quad k = 1, 2, ...,$$

until convergence is reached. Once ℓ_0 has been computed, we can estimate the minimum expected loss in terms of arbitrary choices of a and b. Under the above assumptions, we have that $L'(\ell)$ and $L''(\ell)$ are given by

$$\begin{aligned}
L'(\ell) &= -b(1-\pi)A_2\left\{1 + \frac{\lambda_2}{\nu_2}(\ell - \tau_2)^2\right\}^{-(\nu_2+1)/2} + a\pi A_1\left\{1 + \frac{\lambda_1}{\nu_1}(\ell - \tau_1)^2\right\}^{-(\nu_1+1)/2} \\
L''(\ell) &= b(1-\pi)A_2\lambda_2\frac{\nu_2 + 1}{\nu_2}(\ell - \tau_2)\left\{1 + \frac{\lambda_2}{\nu_2}(\ell - \tau_2)^2\right\}^{-(\nu_2+3)/2} \\
&- a\pi A_1\lambda_1\frac{\nu_1 + 1}{\nu_1}(\ell - \tau_1)\left\{1 + \frac{\lambda_1}{\tau_1}(\ell - \tau_1)^2\right\}^{-(\nu_1+3)/2},
\end{aligned}$$

where,

$$A_i = \frac{\Gamma\left(\frac{\nu_i+1}{2}\right)}{\Gamma\left(\frac{\nu_i}{2}\right)\Gamma\left(\frac{1}{2}\right)\left(\frac{\lambda_i}{\nu_i}\right)^{1/2}}, \qquad \text{for } i = 1, 2.$$

In the case where the posterior predictive distribution is analytically unavailable, Wang and Geisser (2005) proposed to generate a Markov Chain Monte Carlo (MCMC) sample $\xi_{i1}, \ldots, \xi_{iC}$, $i = 1, 2$. Conditional on each ξ_{il}, we would sample an ℓ_{il}^* from

$F_i(\cdot \mid \xi_{il})$, $i = 1, 2$, $l = 1, \ldots, C$. Then ℓ_0 can be approximated by minimizing

$$b(1-\pi)\left\{1 - \frac{1}{C}\sum_{l=1}^{C} \mathbf{1}_{(-\infty,\ell]}\ell_{2l}^*\right\} + a\pi\frac{1}{C}\sum_{l=1}^{C} \mathbf{1}_{(-\infty,\ell]}\ell_{1l}^*,$$

where

$$\mathbf{1}_{(-\infty,\ell]}\ell_{il}^* = \begin{cases} 1, & \text{if } \ell_{il} \in (-\infty, \ell] \\ 0, & \text{if } \ell_{il} \notin (-\infty, \ell]. \end{cases}$$

4.4 Application to the wine dataset

The wine dataset consists of measurements of concentration of nineteen chemical compounds on 149 samples of Chilean red wines. The grape varieties in the dataset are Cabernet Sauvignon (101 samples), Carménère (29 samples) and Merlot (19 samples). All wine samples come directly from wineries located in the valleys of Aconcagua, Maipo, Rapel, Curicó and Maule. Most of the samples come from 2004 vintage and some of them from 2002 vintage. Our aim is to verify grape authenticity using the decision theoretical approach laid up in Section 4.3. From the nineteen compounds, nine correspond to anthocyanins, four are organic acids and six are flavonols. A full list of the compounds is given in the Appendix. All the compounds have been proposed and used for red wine variety authentication, see e.g. von Baer et al. (2007). Anthocyanins are a group of chemical compounds present on the grape skins, which are transferred to the wine during the winemaking process. They also confer red wines their characteristic color. Anthocyanin determination was made by reverse phase HPLC based on the method described by Holbach et al. (1997), Otteneder et al. (2002) and OIV (2003) with minor modifications. More details about anthocyanin determination can be found in von Baer et al. (2005) and von Baer et al. (2007). Flavonol and Organic acids are antioxidant compounds. Flavonols were determined by

HPLC based on the methodology of McDonald et al. (1998) with minor modifications. Organic acids were determined by a combination of reverse phase and ion exclusion chromatography in series, as described by Holbach et al. (2001) and OIV (2004). More details about Flavonols and Organic acid determination can be found in von Baer et al. (2007).

We apply the methodology developed in Section 4.3 to determine the best combination of chemical compounds for wine authentication. To do so, we consider fitting several models, using the groups of compounds or combinations listed in Table 4.2 as response vector, and grape variety and valley as covariates in all cases.

Model	Information	p_j
$\mathcal{M}_{p_1}$	Anthocyanin	9
$\mathcal{M}_{p_2}$	Organic acids	4
$\mathcal{M}_{p_3}$	Flavonol	6
$\mathcal{M}_{p_4}$	Anthocyanin, Organic acids	13
$\mathcal{M}_{p_5}$	Anthocyanin, Flavonol	15
$\mathcal{M}_{p_6}$	Organic acids, Flavonol	10
$\mathcal{M}_{p_7}$	Anthocyanin, Organic acids, Flavonol	19

Table 4.2: Proposed response vectors for each model

We now need to specify a model for estimating $P(g_{n+1} = u|y_{n+1}, y^n)$ and $\ell = \log(p(y_{n+1} \mid y^n))$, the input quantities in the decision problem under the two approaches described in section 4.3. To this effect, we will use the model proposed by Gutiérrez and Quintana (2010) for food and beverages authentication, which was motivated by the analysis of part of the same dataset. This model turned out to be flexible and useful for classification in that context, outperforming some other competing alternatives. The model considers a semiparametric multivariate hierarchical linear mixed specification for the mean responses, and covariance matrices that are

specific to the classification categories. The model considers a flexible distribution for the random effects, using the formalism of dependent random probability measures as in De Iorio et al. (2004). Concretely, we assume

$$\begin{aligned}
(y_{iu} \mid x_{iu}, z_{iu}) &\sim N_p(Bx_{iu} + \theta_{iu}, \Sigma_u), \quad i = 1, \ldots, n_u, \quad u = 1, \ldots, g \quad (4.4.1)\\
\theta_{iu} &\sim H_z(\theta_{iu}), \quad H_z(\theta) = \int N(\theta \mid z\alpha, \tau) dG(\alpha)\\
G \mid M, G_0 &\sim DP(M, G_0), \quad G_0 \equiv N_{pk}(0, R)\\
\Sigma_1, \ldots, \Sigma_g &\sim IW_p(\nu_0, Q_0), \quad \tau \sim IW_p(\gamma_0, \Phi_0), \quad R \sim IW_{pk}(r_0, R_0)\\
\beta_{01}, \ldots, \beta_{0q} &\sim N_p(\alpha_0, \tau_0), \quad \Lambda \sim IW_p(L_0, t_0), \quad M \sim Ga(a_1, a_2).
\end{aligned}$$

In model (4.4.1), y_{iu} is a vector in R^p, B is a $p \times q$ matrix of fixed effects, x_{iu} is a vector of covariates in R^q, θ_{iu} is a $p \times 1$ vector of unit-specific random effects, z_{iu} is a $p \times pk$ design matrix for random effects, and α_i is a $pk \times 1$ vector of latent variables that define the random effects. The subscript u denotes the group or class in a classification context. Basically, the random effects in model (4.4.1) are modeled from an infinite mixture, where the mixing distribution G depends on the level of covariate z through an ANOVA DDP process (De Iorio et al.; 2004). This is a natural approach to introduce dependence on factors. More details can be found in Gutiérrez and Quintana (2010).

The hyperparameter values in model (4.4.1) were taken as $\beta_0 = (0, \ldots, 0)^t$, $\tau_0 = 100I_p$, $Q_0 = 0.1I_p$, $L_0 = 0.01I_p$, $\nu_0 = p + 2$, $r_0 = pk + 2$, $t_0 = p + 2$, $R_0 = 10I_{pk}$, $\gamma_0 = p + 2$, $\phi_0 = 0.01I_p$ and $a_1 = a_2 = 1$. The resulting prior densities are proper, but the one for B is vague and hence relatively uninformative. The prior density for R is relatively uninformative too. All the prior variance-covariance matrices were assumed diagonal.

Table 4.3 shows two model adequacy measures, LPML and DIC. LPML (Geisser and Eddy; 1979) is the log-pseudo marginal likelihood, defined as $LPML = \sum_{i=1}^{n} \log(CPO_i)$, where CPO_i are the Conditional Predictive Ordinates (Chen et al.; 2000). Models with higher LPML are preferred. DIC is the Deviance Information Criterion proposed by Spiegelhalter et al. (2002), and models with the smallest DIC values are preferred. From Table 4.3 we can generally conclude that models including more information perform better.

Model	LPML	DIC
$\mathcal{M}_{p_1}$	1,095.7	-2,492.3
$\mathcal{M}_{p_2}$	163.2	-381.1
$\mathcal{M}_{p_3}$	294.2	-1,103.6
$\mathcal{M}_{p_4}$	1,348.7	-3,459.9
$\mathcal{M}_{p_5}$	1,833.7	-4,560.6
$\mathcal{M}_{p_6}$	665.2	-2,134.1
$\mathcal{M}_{p_7}$	2,097.3	-5,759.9

Table 4.3: Model adequacy measures

In our application, U is the grape variety of wine. Recall that the grape varieties and sample sizes are Cabernet Sauvignon (101 samples), Merlot (19 samples) and Carménère (29 samples). Then, when U = Cabernet Sauvignon, each model in Table 4.2 was applied to $n_{11} = 101$ samples that are Cabernet Sauvignon, and $n_{21} = 48$ samples where U is absent, corresponding to the 29 Carménère plus 19 Merlot samples. Similarly, for Merlot we apply the models to $n_{12} = 19$ samples (so $n_{22} = 130$), and for Carménère we have $n_{13} = 29$ and $n_{23} = 120$. With these samples we obtained the values of r_{ijm} and ℓ_{ijm}, for $i = 1, 2$, $j = 1, 2, \ldots, 7$, and $m = 1, 2, 3$ with i denoting population, j denoting model $\mathcal{M}_{pj}$, and m denoting the grape variety.

To estimate π we used an additional independent sample of size $\nu = 100$, where the

number of Cabernet Sauvignon samples (as declared by the producer) was $t_{u1} = 54$, the number of Merlot was $t_{u2} = 20$ and the number of Carménère was $t_{u3} = 26$. For the first approach, we completed the Bayesian formulation assuming independent beta prior distributions for π, η and φ:

$$(\eta_i) \sim Beta(1,1), \quad (\varphi_i) \sim Beta(1,1) \quad i = 1,2,3$$
$$(\pi_1) \sim Beta(2,2), \quad (\pi_2) \sim Beta(1,3), \quad (\pi_3) \sim Beta(1,5)$$

The prior distribution for η_i and φ_i are proper and uninformative. The prior for $\pi_1 = Pr(U = \text{Cabernet Sauvignon})$, $\pi_2 = Pr(U = \text{Merlot})$ and $\pi_3 = Pr(U = \text{Carménère})$ were assigned using information about wine production (thousand of liters by grape variety) supplied by the National Statistics Institute of Chile (INE; 2008).

From the discussion leading to (4.3.3), we choose $l_{AU} = 0$ US\$, $l_{A^cU^c} = 0$ US\$ (i.e., no loss for right decisions), $l_{A^cU} = 10,000$ US\$, and $l_{AU^c} = 4,000$ US\$. We note that the actual costs for wrong decisions depend on additional information which we do not have, such as the number of rejected bottles, transportation, publicity, etc. Nevertheless, the above values were chosen having in mind that our goal is to select a model, and that the expected loss for a particular model is not important in itself, but in relative terms. Additionally, all models assume the same loss, so that what varies between models is the cost of collecting data c_j. The cost of an anthocyanin analysis for wines in a lab in Chile is about US \$ 73.7, an organic acid analysis costs US \$ 81.9, and a flavonol analysis costs US \$ 102.4. Therefore the cost of collecting data for the seven models were: $c_1 = 73.7$, $c_2 = 81.9$, $c_3 = 102.4$, $c_4 = 155.6$, $c_5 = 176.1$, $c_6 = 184.3$ and $c_7 = 258$ all in US\$ (von Baer; 2010).

With the losses and costs described above we estimated the expected loss (4.3.4) as a function of the threshold p_0. The expected loss for Cabernet Sauvignon for each

of the seven models is given in Figure 4.1. For almost all values of p_0, $\mathcal{M}_{p_1}$ is the

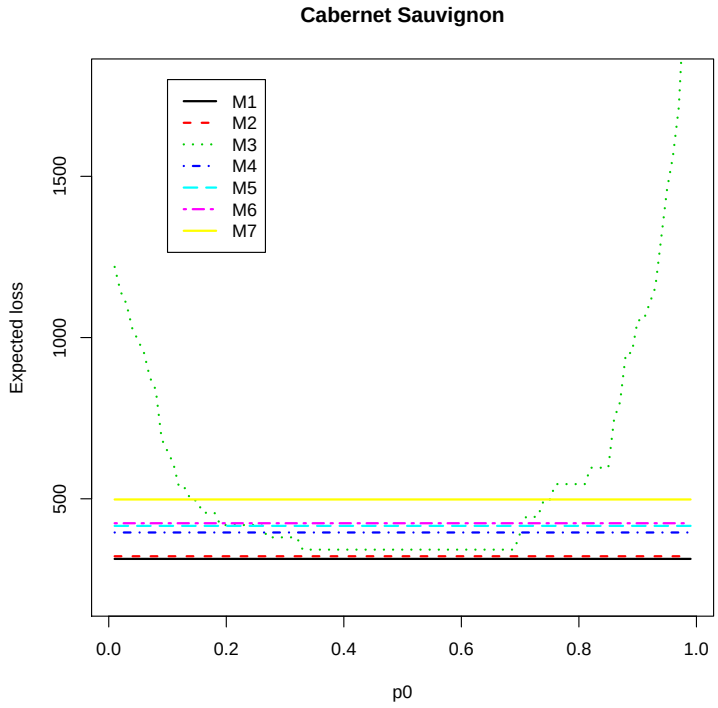

Figure 4.1: Expected loss for Cabernet Sauvignon as a function of p_0

best model. $\mathcal{M}_{p_2}$ has similar expected loss than $\mathcal{M}_{p_1}$. Therefore, measurements of Anthocyanins or Organic Acid are most useful when a producer wants to verify that a sample of wine is Cabernet Sauvignon.

Figure 4.2 shows the expected loss function for Merlot. In this case, $\mathcal{M}_{p_1}$ yields good results but not for all range of p_0 values, as $\mathcal{M}_{p_3}$ is better than $\mathcal{M}_{p_1}$ when p_0 is near 1. Although the expected losses of $\mathcal{M}_{p_4}$ or $\mathcal{M}_{p_5}$ are both bigger, they appear almost invariant to the choice of p_0. Therefore if a producer wants to verify that a sample of wine is Merlot, measurements of Anthocyanin and Flavonol are needed.

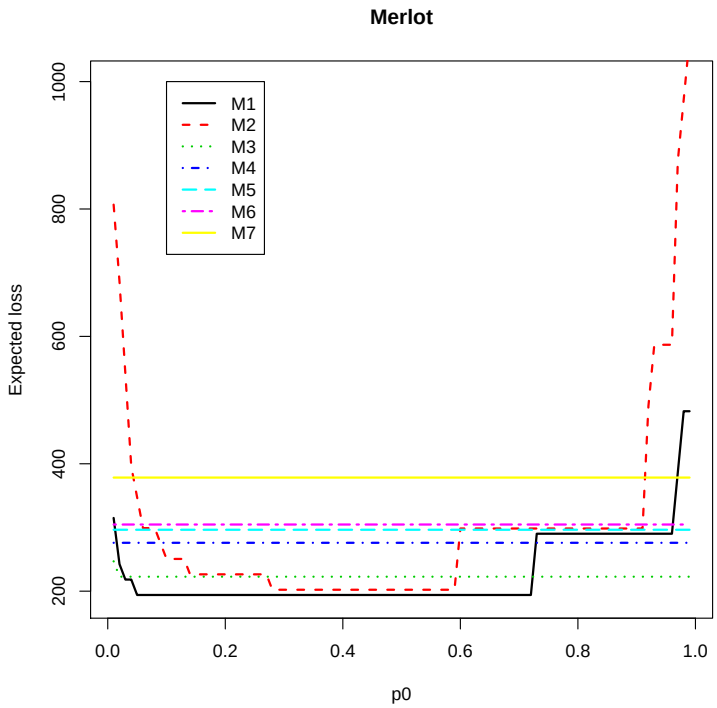

Figure 4.2: Expected loss for Merlot as function of p_0

Finally, Figure 4.3 shows the expected loss function for Carménère. We find that $\mathcal{M}_{p_1}$ is the best over a wide range of p_0. When p_0 is near 0.5, $\mathcal{M}_{p_2}$ has a similar performance than $\mathcal{M}_{p_1}$. On the other hand, $\mathcal{M}_{p_4}$ implies a bigger loss but it is almost invariant to the choice of p_0. Therefore if a producer wants to verify that a sample of wine is Carménère, measurements of Anthocyanins and Organic acids are needed.

For the second approach described in Section 4.3, we selected the prior distribution parameters as $\mu_{i0} = 0$, $n_{i0} = 1$, $\alpha_{i0} = 3$, $\beta_{i0} = 1$ for the three grape varieties. After a minimization process we obtained ℓ_0, the optimal value of ℓ, and evaluated the

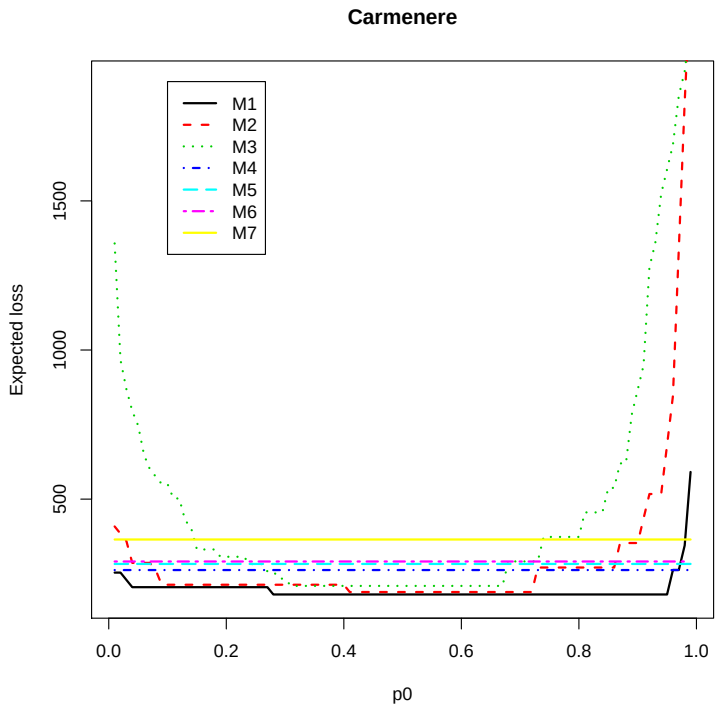

Figure 4.3: Expected loss for Carménère as a function of p_0

expected loss as a function of losses a and b. For a we evaluated the expected loss over the range from 50 (small loss) to $20,000$ US$ (big loss), keeping b fixed at 1 US$. For b the expected loss was calculated between 1 and $7,000$ US$, keeping a fixed at 10,000 US$. These choices were motivated by inequality (4.3.3). Again, the losses of wrong decisions are the same for all models and the cost of data collecting c_j varies across models. The ranges of losses were selected in order to obtain a broad view of the minimum expected loss under different scenarios. The results are shown in Figure 4.4. Figure 4.4 shows, for grape variety Cabernet Sauvignon, that $\mathcal{M}_{p_1}$ attains the minimal expected loss; A similar performance was obtained by $\mathcal{M}_{p_2}$. For

b, $\mathcal{M}_{p_2}$ attains the minimal expected loss uniformly over the whole range. In the case of Merlot the minimum expected loss is attained by $\mathcal{M}_{p_1}$ as function of a and $\mathcal{M}_{p_6}$ as function of b, especially when b increases. For Carménère, $\mathcal{M}_{p_1}$ reached the minimum loss; as function of b, $\mathcal{M}_{p_2}$ attains the minimum loss for the same grape variety. These results are consistent with those obtained with the first approach, and with the results obtained by von Baer et al. (2007) with part of the same data, but using linear discriminant analysis applied to the same groups of compounds. Additionally, we performed a sensibility analysis for different values of prevalence (fixing π in $0.1, 0.2, \ldots, 0.8$ for each grape variety). From this analysis we found that the prevalence affects the expected loss, but for all values of prevalence, the conclusions for each grape variety were not affected.

In summary, the two approaches lead to identical conclusions: (i) to verify whether a wine sample is Cabernet Sauvignon or not, anthocyanin or organic acid measurements are more appropriate than flavonols; (ii) to verify whether a wine sample is Merlot or not, flavonols or anthocyanins are more appropriate than organic acids; and (iii) to verify whether a wine sample is Carménère, organic acids or anthocyanins are appropriate.

4.5 Concluding remarks

The applied methodology allows us to select the optimal information when we need to verify the grape authenticity of red wines. The methodology could be applied to any authentication problem where more than one group of chemical markers could be used for authentication. In the case of red wines, many chemical markers have been proposed for authentication purposes, but as we can see in the results, different groups of chemical markers provide different information. For instance, if we want to verify whether a sample of wine is Cabernet Sauvignon or not, anthocyanin or organic acid measurements are more appropriate than flavonols. The methodology allows us to incorporate the cost of chemical determination, so an analyst can decide the best combination of chemical compounds to use when verifying the authenticity of each sample.

In our application we used a semiparametric Bayesian model, but the model could be parametric as well, and there is no constrain about it. The focus is on the information that the model uses, and as suggested by the adequacy measurements DIC and LPLM, the more information we add to the model, the better fit we get. But improving the fit might be too expensive, and so our approach balances the precision we get with the cost we have to pay for using the additional information. In that sense, the conclusions we draw can be useful to producers and consumers, as they allow to focus their effort on the appropriate chemicals to consider for each wine variety.

4.6 Appendix

Anthocyanin	Organic Acids	Flavonol
Delphinidin-3-glucoside	Tartaric	Myricetin
Cyanidin-3-glucoside	Shikimic	Quercetin
Petunidin-3-glucoside	Lactic	Total myricetin
Peonidin-3-glucoside	Acetic	Total quercetin
Malvidin-3-glucoside		Conjugate myricetin
Peonidin-3-acetylglucoside		Conjugate quercetin
Malvidin-3-acetylglucoside		
Peonidin-3-coumaroylglucoside		
Malvidin-3-coumaroylglucoside		

Table 4.4: Measured compounds

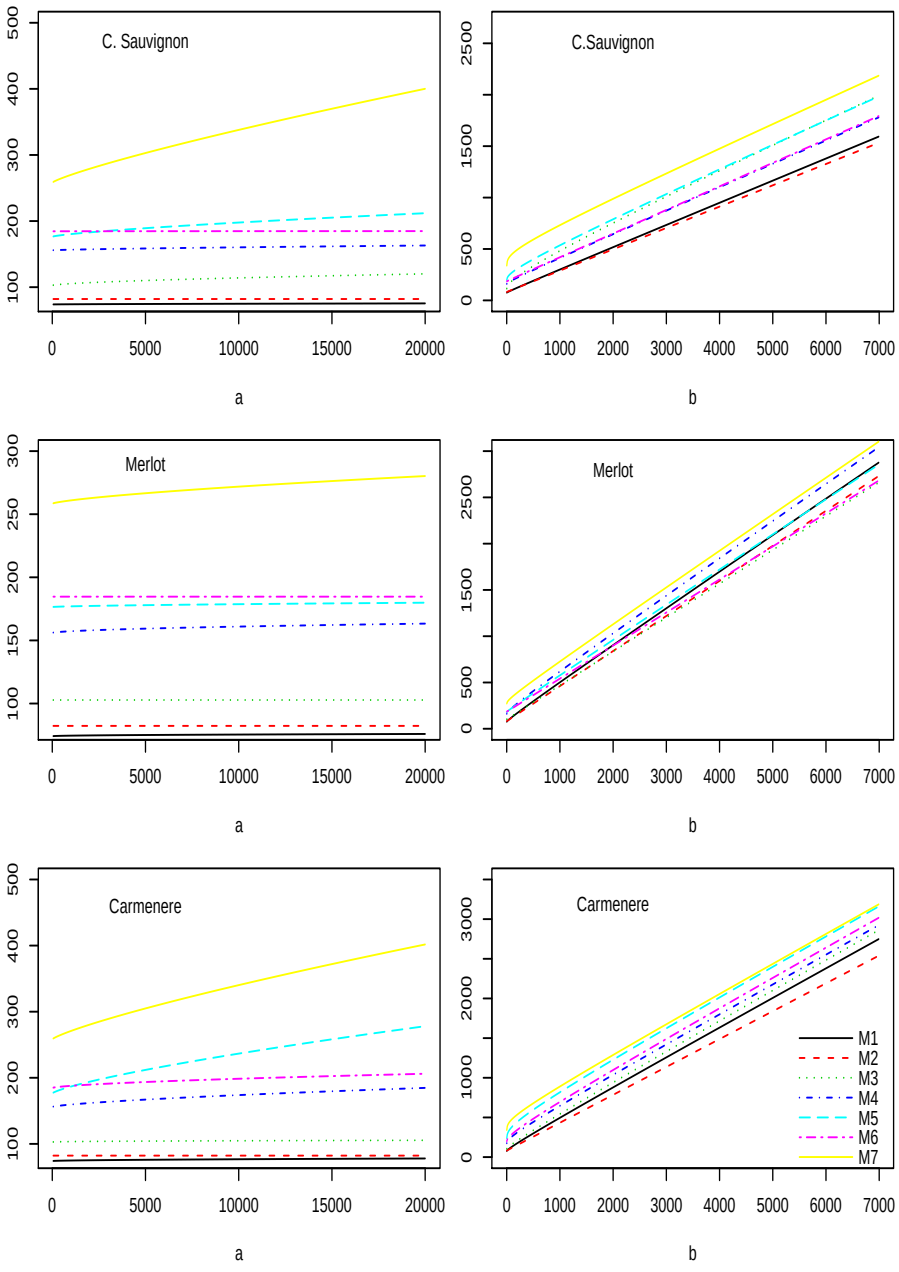

Figure 4.4: Minimum expected loss as function of losses a and b

Chapter 5

Further Research

In this chapter we consider possible future research directions.

5.1 Motivated by the wine dataset

In Chapter 1 we used the vintage year as a continuous predictor when proposing a Bayesian parametric model. The pragmatical reason for this was that by doing so we may easily incorporate data from new years as they become available, without the need to modify the model. The effect of vintage year was negligible in that context. Therefore in Chapter 2, we ignored vintage year in our development. It is well known that the vintage year could affect the wine properties, because the weather conditions could change. In future work, we want to explore effective modeling strategies that incorporate the vintage effect without the need to modify the model each year.

5.2 Motivated by near-infrared spectroscopic measurements

Many analytical chemistry techniques are used in food authenticity studies, including high performance gas chromatography (HPLC), mass spectroscopy and vibrational spectroscopy techniques. All of these techniques have been shown to be capable of discriminating between certain sets of similar biological materials. Spectroscopy is the study of the interaction between radiation and matter as a function of wavelength. Spectrometry is the spectroscopic technique used to assess the concentration or amount of a given chemical (atomic, molecular, or ionic) species. Downey (1996) and Reid et al. (2006) provide reviews of food authenticity studies with emphasis on spectroscopic methods. Near infrared (NIR) spectroscopy is the spectroscopic technique that deals with the infrared region of the electromagnetic spectrum. Near infrared (NIR) spectroscopy provides a quick and efficient method of collecting data for use in food authenticity studies (Downey; 1996). It is particularly useful because it requires very little sample preparation and is nondestructive to the samples being tested.

Murphy et al. (2010) presents two food authenticity data sets which consist of combined visible and near-infrared spectroscopic measurements from food samples of different types. In the first dataset the aim is to classify meats into species (Beef, Chicken, Lamb, Pork, Turkey). In the second, the aim is to classify olive oils into geographic origin (Crete, Peloponese, other). In both studies, combined visible and near infrared spectra were collected in reflectance mode using an NIRSystem 6500 instrument over the wavelength range 400-2498 nm at 2 nm intervals. Hence, the values

collected for each food sample consist of 1050 reflectance values. The reflectance values in the visible and near-infrared region are produced by vibrations in the chemical bonds in the substance being analyzed. This type of data exhibits high correlation due to the presence of a large number of overlapping broad peaks in this region of the electromagnetic spectrum and the presence of combinations and overtones. Most of the time more variables than observation are available, that is, large p, small n ($n << p$). Problems where ($n << p$) were described by West (2003) in a context of factor regression models and Murphy et al. (2010) in model-based discriminant analysis for high dimensional data. The Murphy et al. (2010) proposal is an adaptation of the model based clustering with variable selection method of Raftery and Dean (2006), where the basic idea is to recast the variable selection problem as one of comparing competing models for all of the variables initially considered. This comparison is made using approximate Bayes factor.

We think that more research is needed with near-infrared spectroscopic measurements, especially under different viewpoints, because this type of data are quickly obtained, and there is the possibility of developing portable sensor for food authenticity. In fact, portable sensors have been development on a commercial basis for the authentication of Scottish whiskeys (Connoly; 2006). The data used in Murphy et al. (2010) are available online as supplementary material, and can be used as a motivating example to adopt a different modeling approach.

Bibliography

Agrawal, R., Bala, M. and Bala, R. (2009). Incremental Framework for Feature Selection and Bayesian Classification for Multivariate Normal Distribution, *Advance Computing Conference, 2009. IACC 2009. IEEE International*, pp. 1469 –1474.

Aleixandre, J., Lizana, V., Alvarez, I. and Garcia, J. (2002). Varietal differentiation of red wines in the Valencian region (Spain), *Journal of Food Agricultural and Food Chemistry* **50**: 751–755.

Antoniak, C. (1974). Mixtures of Dirichlet process with applications to Bayesian nonparametric problems, *The Annals of Statistics* **2**(6): 1152–1174.

Beltrán, N., Duarte-Mermound, M., Salah, S., Bustos, M., Peña-Neira, A., Loyola, E. and Jalocha, J. (2005). Feature extraction and classification of Chilean wines, *Journal of Food Engineering* **67**: 483–490.

Berente, B., De la Calle Garcia, D., Reichenbächer, M. and Danzer, K. (2000). Method development for the determination of anthocyanins in red wines by high-performance liquid chromatography and classification of German red wines by means of multivariate statistical methods, *Journal of Chromatography A* **871**: 95–103.

Berger, J. (1985). *Statistical decision theory and bayesian analysis*, Springer, New York.

Bevin, C., Fergusson, A., Perry, W., Janik, L. and Cozzolino, D. (2006). Development of a rapid fingerprinting system for wine authenticity by mid-infrared Spectroscopy, *Journal of Agricultural and Food Chemistry* **54**: 9713–9718.

Binder, D. (1978). Bayesian cluster analysis, *Biometrika* **65**: 31–38.

Blackwell, D. and MacQueen, J. B. (1973). Ferguson distributions via Pólya urn schemes, *Annals of Statistics* **1**(353–355).

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

Brown, P., Kenward, M. and Bassett, E. (2001). Bayesian discrimination with longitudinal data, *Biostatistics* **2**: 417–432.

Bush, C. and MacEachern, S. (1996). A semiparametric Bayesian model for randomised block designs, *Biometrika* **83**(2): 275–285.

Caron, F., Davy, M., Doucet, A., Duflos, E. and Vanheeghe, P. (2006). Bayesian inference for dynamic models with Dirichlet process mixtures, *In International Conference on Information Fusion*, Florence Italy.

Chen, M., Shao, Q. and Ibrahim, J. (2000). *Monte carlo methods in Bayesian computation*, Springer Series in Statistics. Springer-Verlag, New York.

Chung, Y. and Dunson, D. (2011). The Local Dirichlet process, *Annals of the Institute of Statistical Mathematics* . Online First.

Connoly, C. (2006). Spectroscopy and analytical developments Ltd fingerprints brand spirits with ultraviolet spectrophotometry, *Sensor Review* **26**: 94–97.

Dahl, D. (2005). Sequentially-Allocated Merge-Split sampler for conjugate and nonconjugate Dirichlet process mixture models. Technical Report, Texas AM University, USA.

De Iorio, M., Johnson, W., Müller, P. and Rosner, G. (2009). Bayesian nonparametric nonproportional hazards survival modeling, *Biometrics* **65**: 762–771.

De Iorio, M., Müller, P., Rosner, G. and MacEachern, S. (2004). An Anova model for dependent random measures, *Journal of the American Statistical Association* **99**(465): 205–215.

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, R. (2008). Bayesian non-linear regression models with skew-elliptical erros: Applications to the classification of longitudinal profiles, *Computational Statistics and Data Analysis* **53**: 436–449.

De la Cruz, R., Quintana, F. and Marshall, G. (2008b). Model-based clustering for longitudinal data, *Computational Statistics and Data Analysis* **52**: 1441–1457.

De la Cruz, R., Quintana, F. and Müller, P. (2007b). Semiparametric Bayesian Classification with Longitudinal Markers, *Journal of the Royal Statistical Society, Series C* **56**(2): 119–137.

de Villiers, A., Vanhoenacker, G., Mejek, P. and Sandra, P. J. (2005). Determination of anthocyanins in wine by direct injection liquid Chromatography diode array detection mass spectrometry and classification of wines using discriminant analysis, *Journal of Chromatography A* **1054**: 195–204.

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.

Dey, D., Müller, P. and Sinha, D. (1998). *Practical nonparametric and semiparametric Bayesian statistics*, Lecture notes in statistics, Springer.

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

Dunson, D. and Park, J. (2008). Kernel stick-breaking processes, *Biometrika* **95**(2): 307–323.

Dunson, D. and Peddada, S. (2008). Bayesian nonparametric inference on stochastic ordering, *Biometrika* **95**: 859–874.

Dunson, D., Pillai, N. and Park, J. (2007). Bayesian density regression, *Journal of the Royal Statistical Society,Series B* **69**(2): 163–183.

Eder, R., Wendelin, S. and Barna, J. (1994). Classification of red wine cultivars by means of anthocyanin analysis. 1st Report: application of multivariate statistical methods for differentiation of grape samples, *Mitteilungen Klosterneubug* **44**: 201–212.

Engelhardt, U. (2007). *Authenticity of tea (C. sinensis) and tea products*, ACS SYMPOSIUM SERIES 952 American Chemical Society Washington DC.

Escobar, M. (1994). Estimating normal means with Dirichlet process prior, *Journal of the American Statistical Association* **89**(425): 268–277.

Escobar, M. and West, M. (1995). Bayesian density estimation and inference using mixtures, *Journal of the American Statistical Association* **90**(430): 577–588.

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

Fischerleitner, E., Korntheuer, K., Wendelin, S. and Eder, R. (2005). Über die Eignung des Gehalts an Shikimisäure im Wein als Authentizitätsparameter, *Mitteilungen Klosterneuburg* **54**: 234–238.

Geisser, S. and Eddy, W. F. (1979). A predictive approach to model selection, *Journal of the American Statistical Association* **74**(365): 153–160.

Geisser, S. and Johnson, W. (1992). Optimal administration of dual screening test for detecting a characteristic with special reference to low prevalence diseases, *Biometrics* **48**: 839–852.

Gelfand, A., Kottas, A. and MacEachern, S. (2005). Bayesian nonparametric spatial modeling with Dirichlet process mixing, *Journal of the American Statistical Association* **100**(471): 1021–1035.

Griffin, J. and Steel, M. (2006). Order-based dependent Dirichlet processes, *Journal of the American Statistical Association* **101**(473): 179–194.

Gutiérrez, L. and Quintana, F. (2010). Multivariate Bayesian semiparametric models for authentication of food and beverages. Submitted to The Annals of Applied Statistics.

Gutiérrez, L., Quintana, F., von Baer, D. and Mardones, C. (2010). Multivariate Bayesian discrimination for varietal authentication of Chilean red wine. Accepted in Journal of Applied Statistics.

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

Hinrichsen, P., Narvaez, C., Bowers, J., Boursiquot, J., Valenzuela, J., Muñoz, C. and Meredith, C. (2001). Distinguishing Carmenère from similar cultivars by DNA typing, *American Journal of Enology and Viticulture* **52**: 396–399.

Hjort, N., Holmes, C., Müller, P. and Walker, S. (2010). *Bayesian Nonparametrics*, Cambrigde, Series in Statistical and Probabilistic Mathematics.

Holbach, B., Marx, R. and Ackerman, M. (2001). Bedeutung der shikimisäure und des anthocyanspek-trums für die charakterisierung von rebsorten, *Lebensmittelchenie* **55**: 32–34.

Holbach, B., Marx, R. and Ackermann, M. (1997). Bestimmung der anthocyanzusammenset-zung von rotwein mittels hochdruckflüssig chromatographi, *Lebensmittelchemie* **51**: 78–80.

INE (2008). Enfoque Estadístico: Vinos Atraen Millones de Dólares Para Chile @ONLINE.
URL: *http://www.ine.cl/canales/menu/boletines/enfoques/2008/septiembre/viticolap.pdf*

Jain, S. and Neal, R. (2004). A Split-Merge Markov Chain Monte Carlo procedure for Dirichlet process mixture model, *Journal of Computational and Graphical Statistics* **13**(1): 158–182.

Jara, A., Lesaffre, E., Iorio, M. D. and Quintana, F. (2010). Bayesian semiparametric inference for multivariate doubly-interval-censored data, *The Annals of applied statistics to appear* .

Kottas, A. and Krnjajić, M. (2009). Bayesian semiparametric modelling in quantile regression, *Scandinavian Journal of Statistics* **36**: 297–319.

Kruzlicova, D., Mocak, J., Balla, B., Petka, J., Farkova, M. and Havel, J. (2009). Classification of Slovak white wines using artificial neural networks and discriminant techniques, *Food Chemistry* **112**: 1046–1052.

Lavine, M. and West, M. (1992). A Bayesian method for classification and discrimination, *The Canadian Journal of Statisitics* **20**: 451–461.

MacEachern, S. (1994). Estimating normal means with a conjugate style Dirichlet process prior, *Communications in Statistics: Simulation and Computation* **23**: 727–741.

MacEachern, S. (1999). Dependent nonparametric processes, *Proc. Bayesian Statistical Science. Amer. Statistic. Assoc., Alexandria, VA.* pp. 50–55.

Mafra, I., Isabel, M. P., Ferreira, P., Beatriz, M. and Oliveira, P. (2008). Food authentication by PCR-based methods, *European Food Research Technology* **277**: 649–665.

Mallick, B., Ghosh, D. and Ghosh, M. (2005). Bayesian classification of tumours by using gene expression data, *Journal of the Royal Statisitcal Society* **67**: 219–234.

McClachlan, G. and Peel, D. (2000). *Finite mixture models*, Wiley seires in probability and statistics.

McDonald, M., Hughes, M., Burns, J., Lean, M., Matthews, D. and Crozier, D. (1998). Survey of the free and conjugated Myricetin and Quercetin content of red wines of different geographical origins, *Journal of Agricultural and Food Chemistry* **46**: 368–375.

Müller, P., Erkanli, A. and West, M. (1996). Bayesian curve fitting using multivariate normal mixtures, *Biometrika* **83**: 67–79.

Müller, P. and Quintana, F. (2004). Nonparametric Bayesian data analysis, *Statistical Science* **19**(1): 95–110.

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.

Neal, R. (2000). Markov chain sampling for Dirichlet process mixture models, *Journal of Computational and Graphical Statistics* **9**(2): 249–265.

OIV (2003). *Resolution OENO 22/2003*, International Organization of Vine and Wine, Paris.

OIV (2004). *Resolution OENO 33/2004*, International Organization of Vine and Wine, Paris.

Otteneder, H., Holbach, B., Marx, R. and Zimmer, M. (2002). Rebsortenbestimmung in Rotwein anhand der Anthocyanspektren, *Mitteilungen Klosterneuburg* **52**: 187–194.

Otteneder, H., Marx, R. and Zimmer, M. (2004). Analysis of anthocyanin composition of Cabernet Sauvignon and Portugieser wines provides an objective assessment of the grape varieties, *Journal of Grape Wine Research* **10**: 3–7.

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

Reid, L., O'Donnell, C. and Downey, G. (2006). Recent technological advances in the determination of food authenticity, *Trends in Food Science and Technology* **17**: 344–353.

Reinsel, G. (1982). Multivariate repeated-measurement or growth curve models with multivariate random-effects covariance structure, *Journal of the American Statistical Association* **77**(377): 190–195.

Reinsel, G. (1984). Estimation and prediction in a multivariate random effects generalized linear model, *Journal of the American Statistical Association* **79**(386): 406–414.

Revilla, E., Garcia-Beneytez, E., Cabello, F., Martin-Ortega, G. and Ryan, J. (2001). Value of high-performance liquid chromatographic analysis of anthocyanins in the differentiation of red grape cultivars and red wines made from them, *Journal of Chromatography A* **915**: 53–60.

Rigby, R. (1997). Bayesian discrimination between two multivariate normal populations with equal covariance matrices, *Journal of the American Statistical Association* **92**: 1151–1154.

Rodriguez, A. and ter Horst, E. (2008). Bayesian dynamic density estimation, *Bayesian Analysis* **3**(2): 339–366.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors, *Statistica Sinica* **4**: 639–650.

Spiegelhalter, D., Best, N., Carlin, B. and van der Linde, A. (2002). Bayesian measures of model complexity and fit, *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **64**(4): 583–639.

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.

von Baer, D. (2010). Personal communication.

von Baer, D., Mardones, C., Gutiérrez, L., Hofmann, G., Becerra, J., Hitschfeld, A. and Vergara, C. (2005). Varietal authenticity verification of Cabernet sauvignon, Merlot and Carmenère wines produced in Chile by their Anthocyanin, Flavonol and Shikimic acid profiles, *Le Bulletin de L'OIV* **78**: 45–57.

von Baer, D., Mardones, C., Gutiérrez, L., Hofmann, G., Becerra, J., Hitschfeld, A. and Vergara, C. (2007). *Anthocyanin, Flavonol, and Shikimic acid profiles as a tool to verify varietal authenticity in red wines produced in Chile*, ACS SYMPOSIUM SERIES 952 American Chemical Society Washington DC.

Wang, M. and Geisser, S. (2005). Optimal dichotomization of screening test variables, *Journal of statistical planing and inference* **131**: 191–206.

West, M. (2003). Bayesian factor regression models in the "large p, small n" paradigm, *Bayesian Statistics* **7**: 723–732. Oxford Univ. Press, Oxford. MR20033537.

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

Wittes, J. (1987). Comment on the statistical precision of medical screening test by J.L. Gastwirth, *Statistical Science* **2**: 228–230.