Multivariate Bayesian Discrimination for Varietal Authentication of Chilean Red Wine

Luis Gutiérrez Fernando A. Quintana* Dietrich von Baer
Claudia Mardones †

March 19, 2010

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

*Departamento de Estadística, Facultad de Matemáticas, Pontificia Universidad Católica de Chile, e-mail:{llgutier,quintana}@mat.puc.cl

†Departamento de Análisis Instrumental, Facultad de Farmacia, Universidad de Concepción, Chile, e-mail:{dvonbaer,cmardone}@udec.cl

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.

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). 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.

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 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) and De la Cruz-Mesía and Quintana (2007). 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 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. In Section 3.1, we expose a general multivariate Bayesian classification approach. In Section 3.2 we present a general multivariate Bayesian linear model for grape variety authentication. In Section 3.3 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 4 we present the results of the selected model application. Finally, Section 5 discusses the results.

2 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 proportion, relative to the sum of all of the corresponding concentrations across the same sample. This is an important observation,

because the concentration of each anthocyanin depends on the processing technology and maceration time, whereas the proportions between the anthocyanins in a sample provides a correction for this variation and can show more efficiently different patterns between varieties due to their different genetic identity. 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 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-proportions 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.

3 Model

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

3.1 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

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

See further details in De la Cruz-Mesía and Quintana (2007). In practice, direct analytical evaluation of (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 3.2 and in the Appendix), we approximate (1) by means of (De la Cruz-Mesía and Quintana; 2007)

$$P(g_{n+1} = k|y_{n+1}, y^n) \approx \frac{1}{C} \sum_{c=1}^{C} \frac{\pi_k p(y_{n+1}|\theta_k^{(c)})}{\sum_l \pi_l p(y_{n+1}|\theta_l^{(c)})}. \tag{2}$$

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

$$\hat{g}_i = \arg\max_k P(g_i = k|y^n) \qquad \text{and} \qquad \hat{g}_{n+1} = \arg\max_k P(g_{n+1} = k|y^n, y_{n+1}). \tag{3}$$

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).

3.2 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 3.1. 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

$$y_i^k = Bx_i^k + Uz_i^k + \epsilon_i^k, \qquad i = 1, \ldots, n \qquad k = 1, \ldots, g \tag{4}$$

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 (4) we assume ϵ_i^k to be independent with

$$\epsilon_i^k \sim N_p(0, \Sigma_k), \qquad i = 1, \ldots, n, \qquad k = 1, \ldots, g. \tag{5}$$

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

$$\beta_j \sim N_p(\beta_{0j}, \Lambda_0), \qquad j = 1, \ldots, q \tag{6}$$

$$U_1, \ldots, U_r \sim N_p(0, S) \tag{7}$$

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

given by

$$\Sigma_1, \ldots, \Sigma_g \sim IW(Q_0, \nu_0) \quad (8)$$

$$S \sim IW(K_0, m_0) \quad (9)$$

We complete the Bayesian formulation of model (4) by specifying the prior for hyperparameters $\beta_{01}, \ldots, \beta_{0q}$ and Λ_0 as

$$\beta_{01}, \ldots, \beta_{0q} \sim N_p(\alpha_0, \tau_0) \quad (10)$$

$$\Lambda_0 \sim IW(L_0, t_0). \quad (11)$$

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.

3.3 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 3.2:

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 covari-

ance 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 (8) - (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 1,000,000 iterations. After 10,000 iterations, samples were collected at a spacing of 990 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 last models in a standard PC (Intel Core Duo CPU 2.4 Ghz and 2.0 Gb RAM) was 8 hours.

4 Results

To evaluate model adequacy and to select among the three models in Section 3.3 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. DIC is an information criterion that was proposed to select Bayesian hierarchical models, where

models with smaller values of DIC are preferred. Table 2 shows the values of DIC and $\sum_{i=1}^{n} CPO_i$ for the three models implemented. Based on both criterion, 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.

[Table 2 here]

Figure 1 shows the posterior 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, DP presents the same log-proportions between Carménère and Cabernet Sauvignon, but they differ for Merlot. In terms of classification, the most informative anthocyanins are PEAC, PECU and MVCU because they yield differences in their proportions between the three grape varieties. This can then be a key element in the classification effort.

[Figure 1 here]

Figure 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 MVAC 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.

[Figure 2 here]

Table 3 shows the classification results. The total error was 3.26%. 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 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, 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.

[Table 3 here]

It is well known that error rates obtained from applying the classification rule to the same data used to 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 3 shows the 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 4.01% 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.

5 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.

Acknowledgments

The authors thank the Associate Editor and two anonymous referees for their valuable comments. The fist author thanks the Comisión Nacional de Investigación Científica y Tecnológica - CONICYT, for supporting his Ph.D. studies at the Pontificia Universidad Católica de Chile. The second author was partially funded by grant FONDECYT 1060729. The anthocyanin profiles were obtained in the frame of the FONDEF grant D00I1138.

6 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

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

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 $j = 1, ..., q$. [1]

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

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

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

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

[1] Chequear! Cambié $l1$ por l_1, etc. en todas partes.

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 $d = q + t_0$.

References

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.

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.

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.

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.

Chen, M., Shao, Q. and Ibrahim, J. (2000). *Monte Carlo Methods in Bayesian Computation*, Springer Series in Statistics. Springer-Verlag, New York.

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 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.

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.

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.

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.

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.

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.

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.

OIV (2003). *Resolution OENO 22/2003*, 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.

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.

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.

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.

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

Figure 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.

Figure 2. Posterior distribution of U_1,...,U_7. --- Aconcagua, --- Maipo, --- Rapel, --- Curicó, --- Maule, --- Itata, --- Bío-Bío.

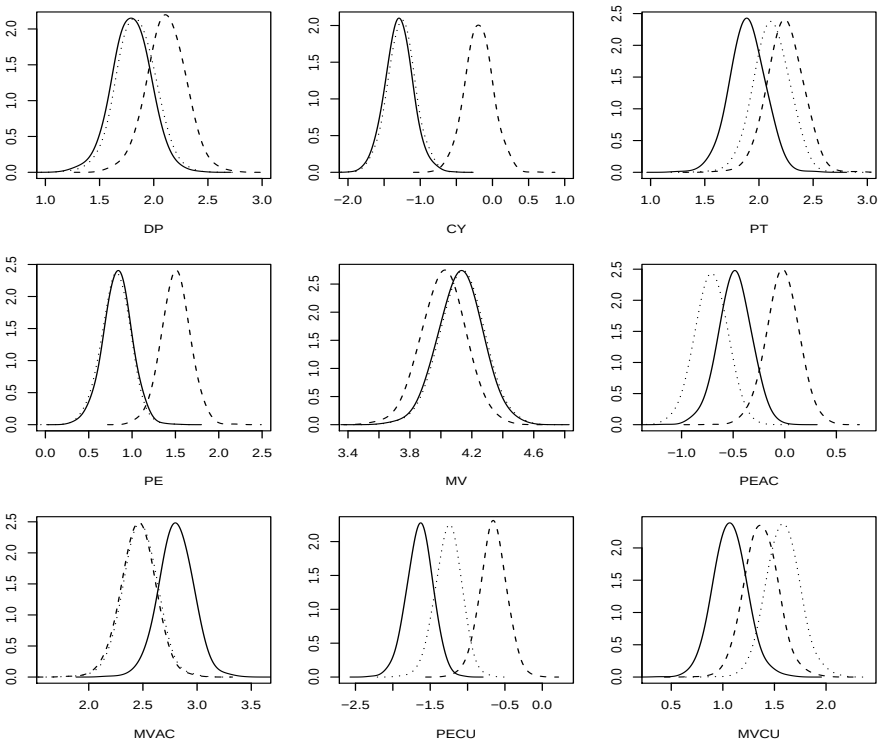

Figure 1:

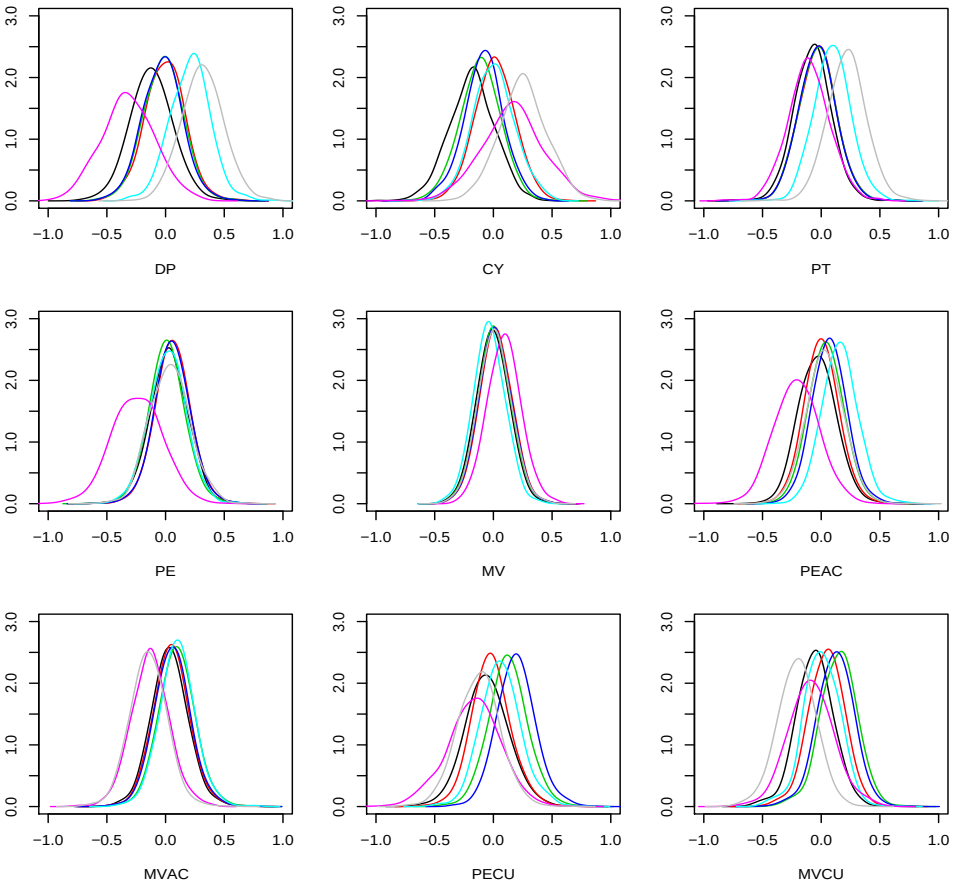

Figure 2:

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: Description of measured anthocyanins

Criterion	Model 1	Model 2	Model 3
CPO	193,027	193,879	103,721
DIC	-3,620.175	-3,628.691	-3,268.616

Table 2: Bayesian Model Adequacy

[2]

[2]Luis: estos números para el CPO se ven muy grandes. Lo que se suele reportar es $\sum_{i=1}^{n} \log(CPO_i)$, donde cada CPO_i se calcula como en Chao, Chen & Ibrahim (2000). ¿Qué calculaste exactamente?

Variety	Carménère	C. Sauvignon	Merlot	Error
Carménère	93 (92)	1 (1)	1 (2)	2.1% (3.16%)
C. Sauvignon	1 (2)	227 (225)	0 (1)	0.44% (1.32%)
Merlot	10 (10)	0 (0)	66 (66)	13.16% (13.16%)
Total error				3.26% (4.01%)

Table 3: Misclassification rate for the three grape varieties. Values within parentheses were obtained using leave-one-out cross-validation approach