# Multivariate Bayesian Semiparametric Models for Authentication of Food and Beverages

Luis Gutiérrez Fernando A. Quintana[*]

November 5, 2010

## 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

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

parametric alternatives. This is also corroborated by application to simulated data.

**Key Words:** Classification, Dependent Dirichlet Process, Wines.

## 1 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 2. In Section 3 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 4 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 5 we illustrate the performance of the proposed model in a simulated data set. In Section 6 we apply the model to authenticate red wines samples based on their anthocyanin profile. The paper concludes in Section 7 with a discussion and final remarks.

## 2 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-proportion 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 4 that attempts to characterize the variability in anthocyanin log-proportions 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 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 $\alpha_i$ 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 $\alpha_i$ for all samples, without reference to the levels of $z_i$. Alternatively, we could consider differences by putting $k$ independent priors on $\alpha_i$. 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:

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

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. $\mu_{zh}$), 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

$$H(y) = \int f(y \mid \theta) dG(\theta) \qquad \text{with} \qquad G \sim DP(M, G_0). \tag{2}$$

Such models are known as DP mixtures (DPM) (Antoniak; 1974). The mixture model (2) can be equivalently written as a hierarchical model by introducing latent variables $\theta_i$ and breaking the mixture as

$$y_i \mid \theta_i \sim f(y_i \mid \theta_i), \qquad \theta_i \sim G, \qquad \text{and} \qquad G \sim DP(M, G_0). \tag{3}$$

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 2.

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 4. 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.

## 4 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

$$\begin{aligned}(y_{iu} \mid x_{iu}, z_{iu}) &\sim N_p(Bx_{iu} + \theta_{iu}, \Sigma_u), \qquad i = 1, \ldots, n_u, \qquad u = 1, \ldots, g \qquad (4)\\ \theta_{iu} &\sim H_z(\theta_{iu})\\ H_z(\theta) &= \int N(\theta \mid z\alpha, \tau) dG(\alpha)\\ G &\sim DP(M, G_0),\end{aligned}$$

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 $R^q$, $\theta_{iu}$ is a $p \times 1$ vector of unit-specific random effects, $z_{iu}$ is a $p \times pk$ design matrix for random effects and $\alpha_i$ 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 4 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 (4) by introducing latent variables $\alpha_i$:

$$\theta_{iu} = z_{iu}\alpha_i + \eta_i, \qquad \alpha_i \sim G, \qquad G \sim DP(M, G_0), \qquad \text{and} \qquad \eta_i \sim N_p(0, \tau). \tag{5}$$

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

$$\beta_1, \ldots, \beta_q \;\sim\; N_p(\beta_{0j}, \Lambda), \qquad j = 1, \ldots, q. \tag{6}$$

The prior distributions for the variance-covariance matrices $\Sigma_u$, $u = 1, \ldots, g$, and $\tau$ are given by

$$\Sigma_1, \ldots, \Sigma_g \sim IW_p(\nu_0, Q_0), \qquad \tau \sim IW_p(\gamma_0, \Phi_0). \tag{7}$$

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

$$R \sim IW_{pk}(r_0, R_0), \quad \beta_{01}, \ldots, \beta_{0q} \sim N_p(\alpha_0, \tau_0) \tag{8}$$

$$\Lambda \sim IW_p(L_0, t_0), \quad M \sim Ga(a_1, a_2) \tag{9}$$

The random distribution $H_z(\theta)$ in model 4 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 6, 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 $\alpha_i$, 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

$$P(g_{n+1} = u \mid y_{n+1}, y^n) \approx \frac{1}{C} \sum_{c=1}^{C} \frac{\pi_u p(y_{n+1} \mid \Theta_u^{(c)})}{\sum_l \pi_l p(y_{n+1} \mid \Theta_l^{(c)})}. \tag{10}$$

In (10), $\pi_u = P(g_i = u)$ 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)

$$\hat{g}_i = \arg\max_u P(g_i = u \mid y^n) \qquad \text{and} \qquad \hat{g}_{n+1} = \arg\max_u P(g_{n+1} = u \mid y^n, y_{n+1}), \tag{11}$$

i.e. assigning the label as the category that maximizes the classification probability (10).

## 5 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 $\Sigma$ is given by $\sigma_{11} = 0.932$, $\sigma_{12} = 0.11$ and $\sigma_{22} = 1.632$. Figure 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 $\beta_0 = (0, 0)^t$, $\tau_0 = 100I_2$, $Q_0 = I_2$, $L_0 = I_2$, $\nu_0 = 4$, $r_0 = 4$, $t_0 = 4$, $R_0 = I_{pk}$, $\gamma_0 = 4$, $\phi_0 = 0.001I_p$ 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 in the literature for this type of problem, and a parametric (BP) version of model (4), defined as:

$$\begin{aligned}
(y_{iu} \mid x_{iu}, z_{iu}) &\sim N_p(Bx_{iu} + \theta_{iu}, \Sigma_u), \qquad i = 1, \ldots, n, \qquad u = 1, \ldots, g \qquad (12)\\
\theta_{iu} &= z_{iu}\alpha + \eta_i, \qquad \eta_i \sim N_p(0, \tau)\\
\alpha &\sim N_{pk}(0, R)
\end{aligned}$$

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

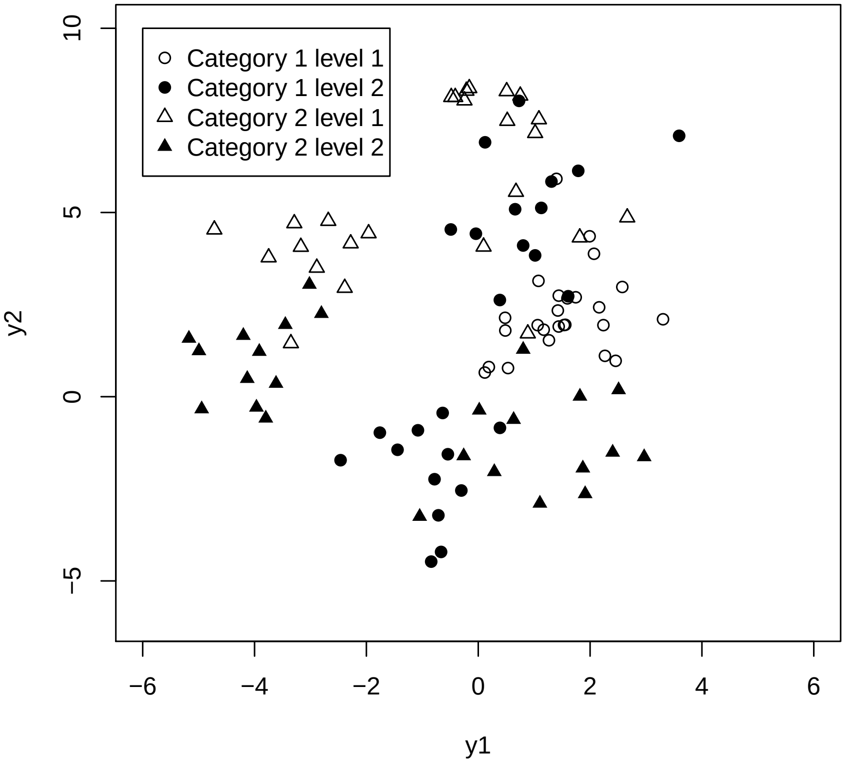

Figure 1: Simulated dataset

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 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

|  |  | 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 1: Classification performance. Values within parenthesis were obtained using leave-one-out cross-validation technique

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.

## 6 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 2. 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

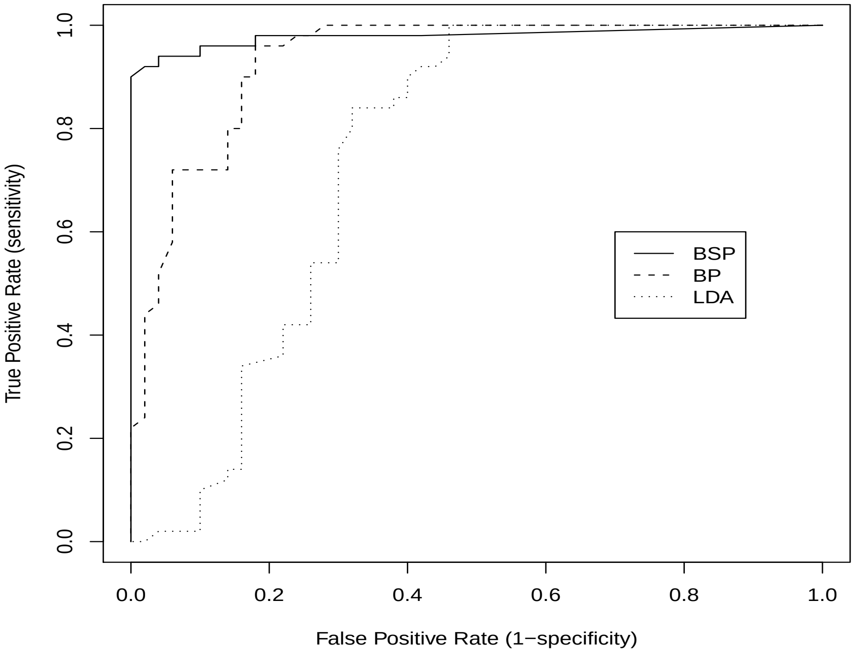

Figure 2: ROC curves for classification under Bayesian semiparametric model BSP, Bayesian parametric BP and linear discriminant analysis LDA.

$p = 9$, $q = 3$ 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 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 1.5%, and 3.76% 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, 3.26% in the training set and 4.01% using LOOCV. Table 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

| Variety | Carménère | C. Sauvignon | Merlot | Error |
|---|---|---|---|---|
| Carménère | 95 (91) | 0 (2) | 0 (2) | 0.00% (4.21%) |
| C. Sauvignon | 0 (1) | 228 (225) | 0 (2) | 0.00% (1.32%) |
| Merlot | 6 (7) | 0 (1) | 70 (68) | 7.9% (10.53%) |
| Total error | | | | 1.5% (3.76%) |

Table 2: Misclassification rate for the three grape varieties

the BSP model attaining the best performance across the three grape varieties. When comparing the BSP and BP models, the DIC and LPML statistics values were -6473.6 and 2987 for the former, and -5493.7 and 2425.2 for the second. Again, these results suggest that the proposed BSP model provides a better fit.

| Grape variety | AUC BSP | AUC BP |
|---|---|---|
| Cavernet Sauvignon | 0.9999999 | 0.9969221 |
| Merlot | 0.9990223 | 0.9867403 |
| Carménère | 0.9991690 | 0.9863574 |

Table 3: Area under ROC curve

Figure 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 and Maipo valleys are of similar form, with some evidence of asymmetry in both cases. Predictions for The Rapel valley show more variability, as dictated by the observed data, but the model provides a reasonable fit to this behavior. Finally, the Curicó valley also exhibit asymmetry, but in a different direction than the others we have displayed.

Figure 4 shows the bivariate predictive posterior distributions for Cabernet Sauvignon, Carménère and Merlot from Rapel valley considering the PEAC and MVAC

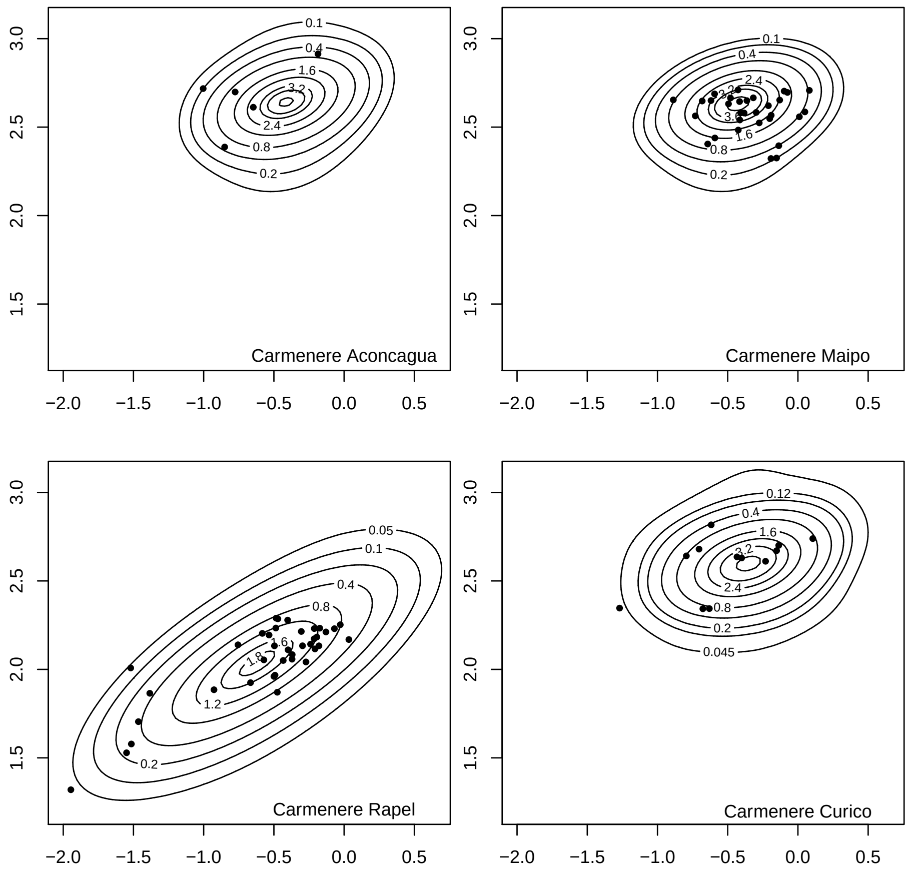

Figure 3: Bivariate posterior predictive distributions with BSP model for Carménère wines from the Aconcagua, Maipo, Rapel and Curicó, with points representing observed values. The anthocyanins considered here were PECU and MVCU.

anthocyanins. This plot is interesting because it shows how informative are PEAC and MVAC in terms of the target classification. These two anthocyanins show a good separation between Cabernet Sauvignon and the rest of the grape varieties, but it is clear 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.

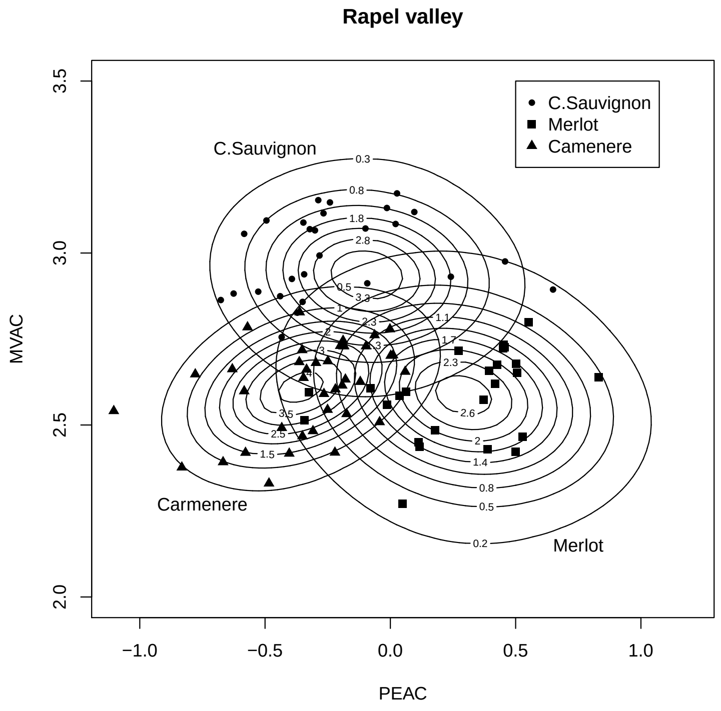

Figure 4: Bivariate posterior predictive distributions for Cabernet Sauvignon, Merlot and Carménère wines from the Rapel valley, with points representing the observed values.

## 7 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.

## Acknowledgments

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 1100010. The anthocyanin profiles were obtained in the frame of the FONDEF grant D00I1138.

## 8 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 $\alpha_i$

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

$$\begin{aligned} P(c_i = c \mid c_{-i}, \theta_i, \alpha) \\ = b\frac{n_{-i,c}}{n-1+M}(2\pi)^{-p/2}|\tau|^{-1/2} \exp\left\{-\frac{1}{2}(\theta_i - z_i\alpha_c)^t\tau^{-1}(\theta_i - z_i\alpha_c)\right\}. \end{aligned} \quad (13)$$

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

$$\begin{aligned} P(c_i \neq c_j \text{for all} j \neq i \mid c_{-i}, \theta_i, \alpha) = b\frac{M}{n-1+M}(2\pi)^{-p/2}|\tau|^{-1/2}|R|^{-1/2}|D_i|^{1/2}\times \\ \exp\left\{\frac{-1}{2}[\theta_i^t\tau^{-1}\theta_i - [\theta_i^t\tau^{-1}z_i]D_i[z_i^t\tau^{-1}\theta_i]]\right\}. \end{aligned} \quad (14)$$

If the imputed value of $c_i$, sampled based on (13) and (14), is not associated with any other observation, it is necessary to draw a value of $\alpha_{c_i}$ from $H_i$, the posterior distribution for $\alpha$ based on the prior $G_0$ and the single observation $\theta_i$. 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 $\alpha_c$ given all the $\theta_i$ 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

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

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

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

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

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

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

- 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 $\Lambda$ 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 $\Sigma_1, \ldots, \Sigma_g$, $\tau$ and $R$ we have

$$\Sigma_u \mid \text{other parameters and data} \sim IW_p(l_u, H_u), \text{ where}$$
$$H_u = \sum_{i=1}^{n_u} (y_i - Bx_i - \theta_i)(y_i - Bx_i - \theta_i)^t + Q_0 \qquad \text{and} \qquad l_u = n_u + \nu_0$$

$$\tau \mid \text{other parameters and data} \sim IW_p(s, T), \text{ where}$$
$$T = \sum_{i=1}^{n} (\theta_i - z_i\alpha_i)(\theta_i - z_i\alpha_i)^T + \Phi_0 \qquad \text{and} \qquad s = n + \gamma_0$$

$$R \mid \text{other parameters and data} \sim IW_{pk}(f, O), \text{ where}$$
$$O = \sum_{i=1}^{n} \alpha_i\alpha_i^t + R_0 \qquad \text{and} \qquad f = n + r_0$$

## References

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

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.

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

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.

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., Quintana, F. and Müller, P. (2007b). Semiparametric bayesian classification with longitudinal markers, *Journal of the Royal Statistical Society, Series C* **56**(2): 119–137.

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

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

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

Dunson, D., Pillai, N. and Park, J. (2007). Bayesian density regression, *Journal of the Royal 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.

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

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

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., Quintana, F., von Baer, D. and Mardones, C. (2010). Multivariate bayesian discrimination for varietal authentication of chilean red wine. Conditionally 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.

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

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

Mafra, I., Ferreira, I. M. P. L. O. and Oliveira, M. B. P. P. (2008). Food authentication by pcr-based methods, *European Food Research Technology* **277**: 649–665.

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.

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.

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.

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.

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., 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.