Optimal Information in Authentication of Food and Beverages

Luis Gutiérrez * Fernando A. Quintana†

July 24, 2013

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 motivated by data consisting of measurements of nineteen

*Escuela de Salud Pública, Facultad de Medicina, Universidad de Chile, e-mail:luisgutierrez@med.uchile.cl

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

chemical compounds (Anthocyanins, Organic Acids and Flavonols) on samples of Chilean red wines. The main goal is to determine the combination of chemical compounds that provides the best information for authentication of wine varieties, considering the losses associated to wrong decisions and the cost of the chemical analysis. The proposed methodology performs well on simulated data, where the best combination of responses is known beforehand.

Key Words: Loss function; Classification; Wine

1 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.; 2011). 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.; 2011), in this paper 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 can be seen as a 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 involves 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 and Section 2 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, but at a higher cost. Our "test" is a multivariate classification model (Gutiérrez and Quintana; 2011) 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 di-

chotomization 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 our classification approach, we model a response vector $y \in R^p$ as function of covariates $x \in R^q$. We deal specifically with the case where the dimension p of y can be changed based on the available information, while the dimension q of x remains constant. This differs from most traditional approaches, where the response vector dimension remains constant and the focus is on covariate selection. Furthermore, we take into account the cost c_j required to obtain information, and so it is natural to consider the problem of optimally selecting information. The basic idea can be summarized as follows. Let $j = 1, 2, \ldots$ index the different combinations of chemical compounds to be considered, yielding a response vector y of dimension p_j to which we fit a classification model $\mathcal{M}_{p_j}$. We also define a loss function that balances the worth of correctly classifying these samples, with the cost c_j required to measure the chemical compounds. The optimal group of compounds to use is then determined as the one minimizing the expected loss function, i.e. the one giving the best classification results at the lowest possible cost. Calculations are based on adapted versions of the methods by Geisser and Johnson (1992) and Wang and Geisser (2005). We compare these methods and

show that they ultimately lead to the same decisions for our problem.

The rest of the paper is organized as follows. In section 2 we introduce the ideas and concepts for defining a loss function and the two approaches for estimating the expected loss. In Section 3 we apply the proposed methodology to a simulated data set. We also briefly describe a classification model that we have found to be particularly useful for authentication in this context (Gutiérrez and Quintana; 2011). In Section 4 we describe the motivating wine dataset, which includes measurements of nineteen chemical compounds: Anthocyanins, Organic Acids and Flavonols. 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 5, where the results are compared, and a final discussion of the proposed methodology is given.

2 Methodology

2.1 A decision-theoretic approach to find an optimal information subset

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 $g_i \in E = \{1, \ldots, m\}$ 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 g_{n+1}. 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 sub-populations: one consists of those units that comply with its label description (which will be referred to as having characteristic U), and the other formed by those units that do not, which we denote as U^c. In this context, there are two possible actions, $g_{n+1} = U$, and $g_{n+1} = U^c$, denoted respectively by A and A^c. Here, U and U^c define a partition of the set E as defined above. In other words, our actions are based on classification predictions that result from a certain model. Concretely, assume now that we have a generic hierarchical model, denoted by $\mathcal{M}$, for the available responses, covariates and group labels, of the form

$$y_i \mid \theta_i, x_i \sim p(y_i \mid \theta_i, x_i), \qquad \theta_i \sim G(\theta_i \mid \phi). \tag{1}$$

In simple words, the data vector y_i for the *ith* sampling unit is assumed to be sampled from a probability model parameterized by a vector θ_i, in turn modeled by a distribution G that depends on hyperparameters ϕ. Our main motivation and focus is on the problem of computing predictions when the dimension p of 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 Anthocyanins, $p = 4$ when we use the Organic Acids, $p = 6$ for Flavonols, and $p = 19$ when we use all of the available compounds. See a full list of the mentioned groups of chemical compounds in the Appendix. In all cases the dimension of x_i remains constant, so the covariates are the same for all models. For the wine data set, the covariates are the grape variety and valley for all models. Denote by $\mathcal{M}_{p_j}$ a model of the form (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 classification predictions 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 collecting the data for each model $\mathcal{M}_{p_j}$.

In the case of actions A and A^c and states U and U^c, a useful loss function is given in Table 1. For example, the loss of deciding action A is l_{AU} when the true state is U.

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

Table 1: Loss function

Now, following Geisser and Johnson (1992), given a decision rule R for model $\mathcal{M}_{p_j}$, the optimal decision is the one minimizing $E(Loss \mid R)$, given by

$$E(Loss \mid R) = l_{AU}Pr(A,U) + l_{AU^c}Pr(A,U^c) + l_{A^cU}Pr(A^c,U) + l_{A^cU^c}Pr(A^c,U^c). \quad (2)$$

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

$$f(E(Loss \mid R), c_j) = E(Loss \mid R) + c_j. \quad (3)$$

We can therefore estimate (3) 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 (2). The order of magnitude of the quantities in Table 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,

$$l_{AU} \leq l_{A^cU^c} \leq l_{AU^c} \leq l_{A^cU}. \tag{4}$$

The worst-case scenario occurs when U is present in the food under authentication but the model estimates this to be not true. A customer 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 her 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.

2.2 Estimation of the expected loss function

Note first that we can rewrite the expected loss function (2) as

$$E(Loss \mid R) = Pr(U)Pr(A \mid U)(l_{AU} - l_{A^cU}) + (1 - Pr(U))Pr(A^c \mid U^c)(l_{A^cU^c} - l_{AU^c}) + Pr(U)l_{A^cU} + (1 - Pr(U))l_{AU^c}. \quad (5)$$

Denote the probabilities in (5) as $\pi = Pr(U)$, the probability that a randomly 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.

We describe now two different approaches for estimating the expected loss function (5).

2.2.1 Geisser and Johnson Approach

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

$$L(\eta, \varphi, \pi \mid d) = L(\eta \mid n_1, r_1)L(\varphi \mid n_2, r_2)L(\pi \mid \nu, t_u). \tag{6}$$

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

$$p(\eta, \varphi, \pi \mid d) \propto p(\eta, \varphi, \pi)L(d \mid \eta, \varphi, \pi). \tag{7}$$

We will discuss specific choices below.

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.; 2011)

$$P(g_{n+1} = u|y^{n+1}, y^n) \approx \frac{1}{C}\sum_{c=1}^{C} \frac{\upsilon_u p(y^{n+1}|\Theta_u^{(c)})}{\sum_{l=1}^{m} \upsilon_l p(y^{n+1}|\Theta_l^{(c)})}, \tag{8}$$

where $\{\Theta^{(c)}, c = 1, \ldots, C\}$ denote a sample of size C from the posterior distribution $p(\theta \mid y^n)$ under the classification model. Details of this model will be given in Section 3. We note that a conventional procedure consists of choosing action A (i.e., declare feature U to be present in the sample) or A^c (i.e., feature U is absent), according to the zero-one law (Hastie et al.; 2001):

$$\hat{g}_i = \arg\max_u P(g_i = u \mid y^n) \qquad \text{and} \qquad \hat{g}_{n+1} = \arg\max_u P(g_{n+1} = u \mid y^n, y_{n+1}), \tag{9}$$

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

(8). Instead, we use (8) 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.

2.2.2 Wang and Geisser Approach

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, (5) simplifies to

$$E(Loss) = b(1 - \pi)(1 - \varphi) + a\pi(1 - \eta). \tag{10}$$

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 part of the output of a classification model $\mathcal{M}_{p_j}$. We select $\ell = \log(p(y^{n+1} \mid y^n))$, because the posterior predictive density is the key element in a Bayesian classifier. In fact, the Monte Carlo approximation in (8) is the average of posterior predictive odds for category u. Thus, the logarithm of posterior predictive density is a very natural choice as an optimization variable. This approach allows us to find the minimum expected loss with respect to ℓ and to find $\ell_0 = \arg\min_\ell Loss(\ell)$, 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. We refer to these as sub-

populations $i = 1, 2$, respectively. For $i = 1, 2$, let $\ell_{ij} = \{\ell_{ij1}, \ldots, \ell_{ijn_i}\}$ denote the values of $\log(p(y^{n+1} \mid y^n))$ obtained when model $\mathcal{M}_{p_j}$ is applied to each of the n_i units above, and where j indexes the various groups or combinations of chemical compounds to be considered. Wang and Geisser (2005) suggest using the predictive distribution

$$\tilde{F_{ij}}(\ell \mid \ell_{ij}) \propto \int F_i(\ell \mid \xi_i) \prod_{m=1}^{n_i} f_i(\ell_{ijm} \mid \xi_i) p_i(\xi_i) d\xi_i \quad i = 1, 2, \tag{11}$$

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

$$Loss_j(\ell) = b(1-\pi)\{1 - \tilde{F}_{2j}(\ell \mid \ell_{2j})\} + a\pi \tilde{F}_{1j}(\ell \mid \ell_{1j}). \tag{12}$$

The value of π can be inferred just as in Section 2.2.1. To simplify calculations, ensuring the availability of an analytical expression for the posterior predictive distribution, we assume, as an approximation, 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

$$p_i(\mu_i, \sigma_i^2) = p_i(\mu_i \mid \sigma_i^2) p_i(\sigma_i^2) = N(\mu_i \mid \mu_{i0}, n_{i0}/\sigma_i^2) IG(\sigma_i^2 \mid \alpha_{i0}, \beta_{i0}), \tag{13}$$

Here, n_{i0} is the hyperparameter that controls our prior knowledge about μ_i. The above assumptions imply that the posterior predictive distribution follows a Student t distribution (Wang and Geisser; 2005) $t(\tau_i, \lambda_i, \nu_i)$, with parameters given by

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

Here, $\bar{\ell}_{ij}$ is the mean of $\{\ell_{ij1}, \ldots \ell_{ijn_i}\}$ and s_{ij} its sample variance. 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

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

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 $Loss'(\ell)$ and $Loss''(\ell)$ are given by

$$Loss'(\ell) = -b(1-\pi)A_2 \left\{1 + \frac{\lambda_2}{\nu_2}(\ell - \tau_2)^2\right\}^{-(\nu_2+1)/2} + a\pi A_1 \left\{1 + \frac{\lambda_1}{\nu_1}(\ell - \tau_1)^2\right\}^{-(\nu_1+1)/2} \quad (14)$$

$$\begin{aligned} Loss''(\ell) = {} & b(1-\pi)A_2\lambda_2\frac{\nu_2+1}{\nu_2}(\ell - \tau_2)\left\{1 + \frac{\lambda_2}{\nu_2}(\ell - \tau_2)^2\right\}^{-(\nu_2+3)/2} \\ & -a\pi A_1\lambda_1\frac{\nu_1+1}{\nu_1}(\ell - \tau_1)\left\{1 + \frac{\lambda_1}{\tau_1}(\ell - \tau_1)^2\right\}^{-(\nu_1+3)/2}, \end{aligned} \quad (15)$$

where,

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

Alternatively, we could try other approximations based on distributional assumptions for ℓ, such as a student t or a mixture of normals. For some choices, however, the corresponding posterior predictive distribution is analytically unavailable. In those cases, Wang and Geisser (2005) proposed using Markov Chain Monte Carlo (MCMC) to generate a posterior sample $\xi_{i1}, \ldots, \xi_{iC}$, for sub-population $i = 1, 2$. 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

$$b(1-\pi)\left\{1-\frac{1}{C}\sum_{l=1}^{C}\mathbf{1}_{(-\infty,\ell]}\ell_{2l}^{*}\right\}+a\pi\frac{1}{C}\sum_{l=1}^{C}\mathbf{1}_{(-\infty,\ell]}\ell_{1l}^{*}, \tag{16}$$

where

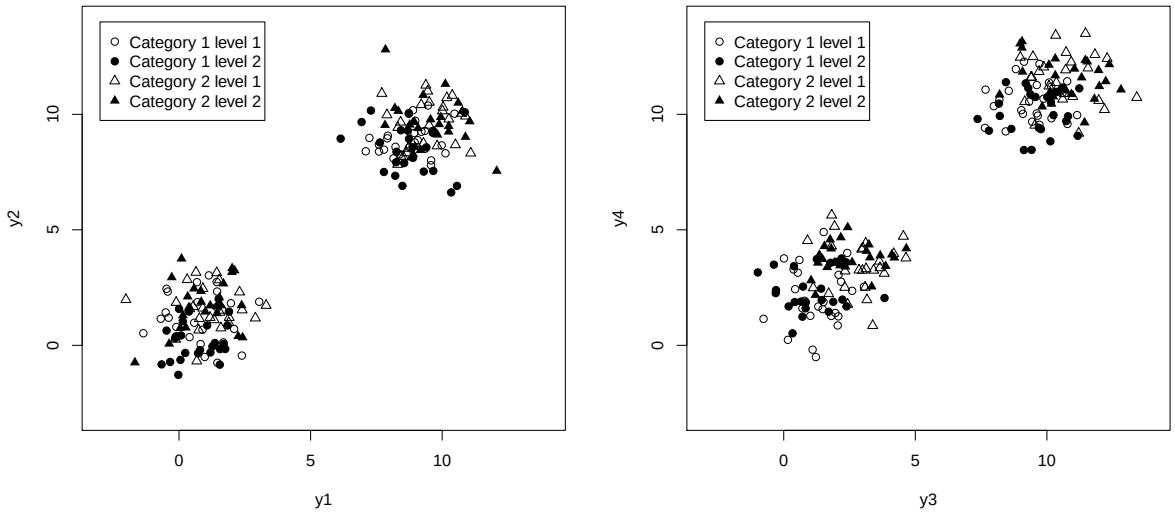
$$\mathbf{1}_{(-\infty,\ell]}\ell_{il}^{*}=\begin{cases}1, & \text{if } \ell_{il}^{*}\in(-\infty,\ell]\\ 0, & \text{if } \ell_{il}^{*}\notin(-\infty,\ell].\end{cases}$$

Having the value of ℓ_0 available, we estimate the minimum expected loss as a function of losses a and b which can vary on an arbitrary range. The final decision consists of selecting the model that yields the minimum expected loss over the range of values for a and b. An additional advantage of this approach is that we can evaluate the sensitivity of conclusions to the choices of a and b.

3 Application to a simulated dataset

To illustrate the use of the proposed methodology we simulate a data set considering $m = 2$, $p = 4$, $k = 2$ and $n = 200$. Here, $m = 2$ means that we have to classify between two categories; $p = 4$ is the dimension of multivariate normal components; $k = 2$ means that we have a categorical covariate z with two levels; and finally, $n = 200$ is the sample size, where $n_1 = 100$ are from category (sub-population) 1 and $n_2 = 100$ come from category 2. Given the simulation scenario, we will also assume the prevalence to be known as $\pi = 0.5$.

The observations were simulated from a mixture distribution, with components given by p-variate normal distributions. Specifically, we consider a four-component mixture, $\sum_{i=1}^{4}\omega_i N(\mu_i,\Sigma)$, where $\mu_1 = (0.8, 0.6, 1.4, 2.2)^t$ and $\mu_2 = (8.8, 8.6, 9.4, 10.2)^t$ are the



(a) (b)

Figure 1: Simulated data. (a) components y_1 and y_2, (b) components y_3 and y_4.

means for category 1, levels 1 and 2 of the covariate, with weights $\omega_1 = \omega_2 = 0.25$; $\mu_3 = (1.2, 1.4, 2.6, 3.8)^t$ and $\mu_4 = (9.2, 9.4, 10.6, 11.8)^t$, are the means for category 2, levels 1 and 2 of the covariate, with weights $\omega_3 = \omega_4 = 0.25$, $\Sigma = I_p$. Figure 1, panel (a), shows scatter plots of the first two dimensions of the four dimensional dataset (y_1 and y_2), while the third and fourth dimensions (y_3 and y_4) are shown in panel (b). Our aim here is to correctly classify circles and triangles, which represent the fixed covariate x in model (17). Also, z is a discrete random covariate, indicated as solid/empty symbols. Furthermore, it is clear that the data in Figure 1 are clustered in two groups, which justifies using a flexible modeling approach.

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 2. To this effect, we will use the model proposed by Gutiérrez and

Quintana (2011) for food and beverages authentication, which was motivated by the analysis of part of the wine dataset to be described in the next section. 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 also considers a flexible distribution for the random effects, using the formalism of dependent random probability measures as in De Iorio et al. (2004). Concretely, the model assumes

$$\begin{aligned}
(y_{iu} \mid x_{iu}, z_{iu}) &\sim N_p(Bx_{iu} + \theta_{iu}, \Sigma_u), \quad i = 1, \ldots, n_u, \quad u = 1, \ldots, m \qquad (17)\\
\theta_{iu} &\sim H_z(\theta_{iu})\\
H_z(\theta) &= \int N(\theta \mid z\alpha, \tau) dG(\alpha)\\
G &\sim DP(M, G_0),
\end{aligned}$$

where $y_{iu} = (y_i : g_i = u)$, $u = 1, \ldots, m$ is the response vector for the ith unit in the uth group, and g_i is the label for the ith unit. The subscript u denotes the group or class in the classification context, B is a $p \times q$ matrix of fixed effects, with columns give by $B = [\beta_1, \beta_2, \ldots, \beta_q]$. $x_{iu} = (x_i : g_i = u)$ is a vector of covariates in R^q for fixed effects, θ_{iu} is a $p \times 1$ vector of unit-specific random effects, $z_{iu} = (z_i : g_i = u)$ is a $p \times pk$ design matrix for random effects, α is a $pk \times 1$ vector of latent variables that define the random effects, and $DP(M, G_0)$ denotes the Dirichlet process prior (Ferguson; 1973) with total mass parameter M and centering distribution G_0. Model (17) 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 (17) via

the introduction of latent variables α_i:

$$\theta_{iu} = z_{iu}\alpha_i + \eta_i, \quad \alpha_i \sim G, \quad G \sim DP(M, G_0), \quad \text{and} \quad \eta_i \sim N_p(0, \tau). \tag{18}$$

We choose a multivariate normal model for the base measure, $G_0 := N_{pk}(0, \Omega)$. Matrix Ω in the model allows for correlation between all components of vector α_i, which implies correlation between different components of the response vector and between different levels of z. Also, τ is the matrix of covariance for θ_{iu}. The Bayesian formulation of the model is completed with a prior specified as follows. For matrix B we assume column-wise independence, that is, $\beta_1, \beta_2, \ldots, \beta_q$ are mutually independent with $\beta_j \sim N_p(\beta_{0j}, \Lambda)$ for $j = 1, \ldots, q$. The prior distributions for the variance-covariance matrices Σ_u, $u = 1, \ldots, m$, and τ are given by $\Sigma_1, \ldots, \Sigma_m \sim IW_p(\nu_0, Q_0)$ and $\tau \sim IW_p(\gamma_0, \Phi_0)$, where $IW_p(\nu, Q)$ indicates the Inverse Wishart distribution on p dimensional positive definite matrices, with ν degrees of freedom and mean $(\nu - p - 1)^{-1}Q$. We complete the Bayesian formulation of model (17) by assuming $\Omega \sim IW_{pk}(r_0, R_0)$, $\beta_{01}, \ldots, \beta_{0q} \sim N_p(\alpha_0, \tau_0)$, $\Lambda \sim IW_p(L_0, t_0)$, and $M \sim Ga(a_1, a_2)$, the Gamma distribution with mean a_1/a_2. More details about properties and performance of the model and a suitable posterior simulation scheme can be found in Gutiérrez and Quintana (2011). To illustrate the methodology developed in section 2 we consider the models listed in Table 2.

Model	Coordinate	p
$\mathcal{M}_{1s}$	y_1, y_2	2
$\mathcal{M}_{2s}$	y_3, y_4	2
$\mathcal{M}_{3s}$	y_1, y_2, y_3, y_4	4

Table 2: Proposed response vector for each model

The hyperparameter values in model (17) were taken as $\beta_0 = (0, \ldots, 0)^t$, $\tau_0 = 100I_p$,

$Q_0 = I_p$, $L_0 = I_p$, $\nu_0 = p+2$, $r_0 = pk+2$, $t_0 = p+2$, $R_0 = 1000I_{pk}$, $\gamma_0 = p+2$, $\phi_0 = 0.1I_p$ 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 Ω is relatively uninformative too. All the prior variance-covariance matrices were assumed diagonal. Table 3 shows the classification results obtained for the three models using the zero-one law, as described in (9). Sorting the models in decreasing order by their classification performance we have $\mathcal{M}_{3s}$, followed by $\mathcal{M}_{2s}$, and finally $\mathcal{M}_{1s}$.

		$\mathcal{M}_{1s}$		$\mathcal{M}_{2s}$		$\mathcal{M}_{3s}$	
		1	2	1	2	1	2
Category	1	90	10	97	3	99	1
	2	13	87	3	97	0	100

Table 3: Classification performance for the simulated dataset. The real state of nature is represented by columns (1 and 2). The classification results are represented by rows.

Letting U denote category 1, each model in Table 2 was applied to the data simulated as described earlier. From each model we estimated the quantities $P(g_{n+1} = u|y^{n+1}, y^n)$, and $\ell_{ij} = \log(p(y^{n+1} \mid y^n))$, where as before, i indexes sub-populations with characteristic U $(i = 1)$ and U^c $(i = 2)$ and j refers to model $\mathcal{M}_{p_j}$. Recall also that quantity $P(g_{n+1} = u \mid y^{n+1}, y^n)$ is used to obtain r_{1j} and r_{2j}, the number of samples that yield A and A^c from sub-populations 1 and 2, respectively, using model j.

For the first approach in Section 2.2.1 we complete the Bayesian formulation assuming independent beta prior distributions for η and φ:

$$(\eta) \sim Beta(1, 1), \quad (\varphi) \sim Beta(1, 1).$$

Recall also that we assume π to be known and fixed at 0.5. From the discussion leading to (4), 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 also assume that the cost of collecting data for these models were $c_1 = 200$, $c_2 = 50$ and $c_3 = 250$ all in US$. These values, though arbitrary, depict a scenario where measuring variables y_1 and y_2 to apply $\mathcal{M}_{1s}$ is more expensive than measuring coordinates y_3 and y_4 for model $\mathcal{M}_{2s}$. Note also that the cost of $\mathcal{M}_{3s}$ is $c_3 = c_1 + c_2$ because that model uses all four coordinates y_1, y_2, y_3, y_4. With the losses and costs described above we estimated the expected loss (5) as a function of the threshold p_0. The expected loss for each of the three models is given in Figure 2.

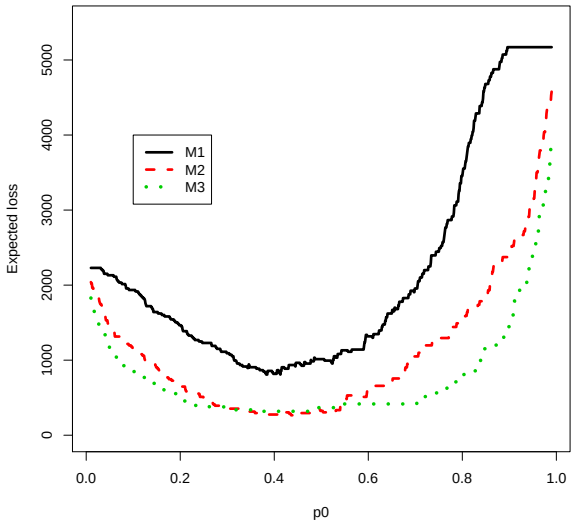
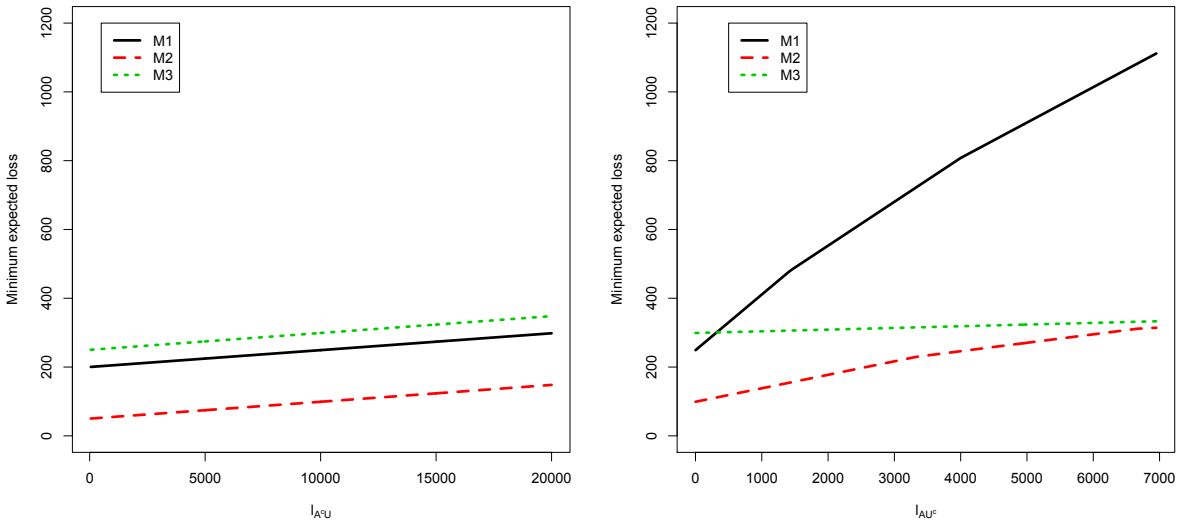


Figure 2: Expected loss as a function of p_0

From Figure 2 we can see that the minimal expected loss for all models was reached for values of p_0 in the range of 0.4 to 0.5. In the above range, $\mathcal{M}_{2s}$ and $\mathcal{M}_{3s}$ obtained the same performance and they are better than $\mathcal{M}_{1s}$. Because models $\mathcal{M}_{2s}$ and $\mathcal{M}_{3s}$ have similar expected loss and $\mathcal{M}_{2s}$ is cheaper than $\mathcal{M}_{3s}$, under this approach $\mathcal{M}_{2s}$ is preferred. To evaluate the sensitivity of the conclusions to the choices of l_{A^cU} and

l_{AU^c} in Figure 3 we present the minimum expected loss. That is, the expected loss at the optimal p_0 value. The optimal p_0 value was selected using a discrete grid in the interval (0,1) following the ideas in Greiner (1996). For l_{A^cU} we evaluated the minimum expected loss over the range from 50 (small loss) to 20,000 US$, keeping l_{AU^c} fixed at 1 US$. For l_{AU^c} the minimum expected loss was calculated between 1 and 7,000 US$, keeping l_{A^cU} fixed at 10,000 US$. These choices were motivated by inequality (4). The optimal values of p_0 vary between 0.34 and 0.41 for $\mathcal{M}_{1s}$, 0.41 to 0.53 for $\mathcal{M}_{2s}$, and there is a unique p_0 value for $\mathcal{M}_{3s}$, because, as we can see from Figure 3 (panel (a) and (b)) $\mathcal{M}_{3s}$ shows a linear behavior of the minimum expected loss. Thus, from Figure 2 and 3 we conclude that the group of variables formed by y_3 and y_4 provides the optimal information.

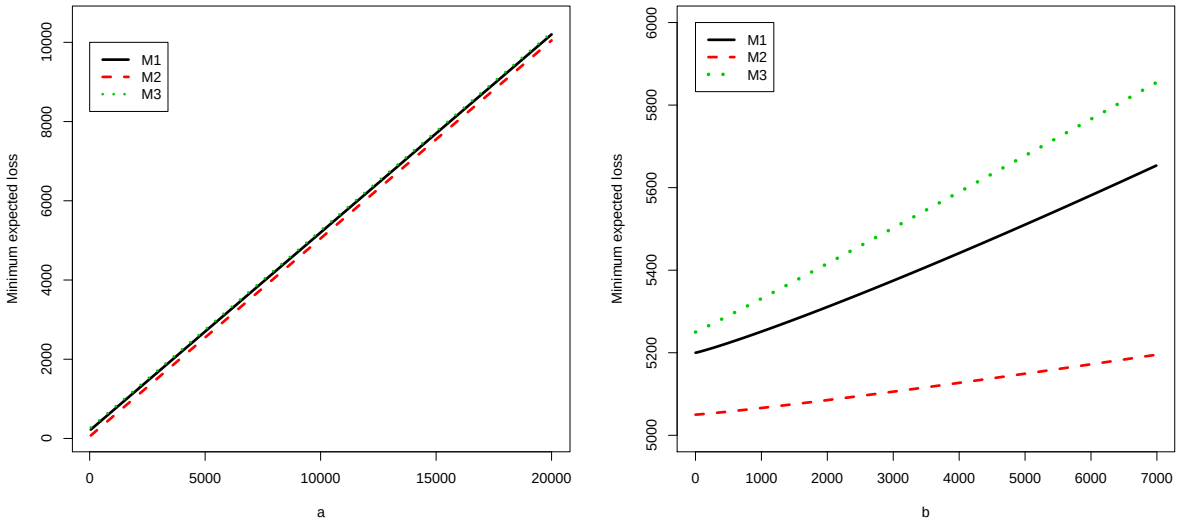


(a) (b)

Figure 3: Minimum expected loss as a function of losses l_{A^cU} (a) and l_{AU^c} (b)

For the second approach, described in Section 2.2.2, we selected the prior distribu-

tion parameters as $\mu_{i0} = 0$, $n_{i0} = 1$, $\alpha_{i0} = 3$, $\beta_{i0} = 1$. After a minimization process we obtained ℓ_0, the optimal value of ℓ, and evaluated the expected loss as a function of losses $l_{A^cU} = a$ and $l_{AU^c} = b$ using the values of a and b as in the first approach. Figure 4 shows the minimal expected loss as a function of loss a, panel (a), and the minimal expected loss as a function of loss b, panel (b).



(a) (b)

Figure 4: Minimum expected loss as a function of losses a and b

From Figure 4, panel (a), we can see that $\mathcal{M}_{2s}$ yields the minimum expected loss as a function of loss a; and from panel (b) the same model ($\mathcal{M}_{2s}$) produces the minimum expected loss as a function of loss b. The above results agree with those obtained under the previous approach. Furthermore, Figures 3 and 4 show that for all considered values of a and b the conclusions are the same, therefore, these are invariant to the choices of the values of losses in Table 1. Thus, variables y_3 and y_4 provide the optimal information. It is interesting to point out that in this simulation example the best

classification results are obtained using model $\mathcal{M}_{3s}$. But because the cost associated with variables y_1 and y_2 is high, the model that uses all the available information is not the optimal one. Of course, in this application, we deliberately simulated coordinates y_3 and y_4 to be more informative for classification purposes than coordinates y_1 and y_2. In real applications though, we only have some intuition about the quality of information for authentication problems, and it is in this aspect that the proposed methodology could be useful.

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. These samples form a data set with mixed wine types, representing the most abundant grape varieties cultivated in Chile across different valleys. Our aim is to verify grape authenticity using the decision theoretical approach laid up in Section 2. 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 High Performance Liquid Chromatography (HPLC), a chromatography technique that can separate a mixture of compounds and is

used in analytical chemistry to identify, quantify and purify the individual components of complex mixtures, like wines and other beverages or foods. The analytical chemistry procedure was based on the method described by Holbach et al. (1997), Otteneder et al. (2002) and by the International Organization of Vine and Wine (OIV), as described in 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). Additionally, 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 2 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 as response vector, and grape variety and valley as covariates in all cases. For further discussion of these covariates, see Gutiérrez and Quintana (2011).

Model	Information	# variables 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: Proposed response vectors for each model

The hyperparameter values in model (17) were taken as $\beta_0 = (0, \ldots, 0)^t$, $\tau_0 = 100 I_p$,

$Q_0 = 0.1I_p$, $L_0 = 0.01I_p$, $\nu_0 = p + 2$, $r_0 = pk + 2$, $t_0 = p + 2$, $R_0 = 10I_{pk}$, $\gamma_0 = p + 2$, $\phi_0 = 0.01I_p$ and $a_1 = a_2 = 1$. The selected hyperparameter values imply proper but vague prior distributions, representing the lack of genuine prior information on the parameters. All the prior covariance matrices were assumed of diagonal form.

We fitted each of the seven models in Table 4, and in particular, evaluated the classification performance using the wine data set described earlier. Quite remarkably, all models yielded perfect classification (i.e., 100% accuracy) with the zero-one law over the observed data (training set). To explore possible differences among these models, we computed some model adequacy measures (a full leave-one-out cross-validation study of each of the models is unnecessary for our purpose). Table 5 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 the CPO_i's 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. We specifically compute DIC_1 (Celeux et al.; 2006). For all models, the effective dimension p_D as described in Celeux et al. (2006) was positive. From Table 5 we can generally conclude that models including more information perform better.

Model	LPML	DIC_1
$\mathcal{M}_{p_1}$	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 5: Model adequacy measures

In our application, U represents that the grape variety under consideration is correctly classified using the model described earlier. We therefore take the view of an individual who wants to learn the best combination of chemical compounds to determine whether the wine variety under consideration is indeed as indicated in the bottle label. Thus, when U = Cabernet Sauvignon, each model in Table 4 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 $m = 1, 2, 3$ where i denotes sub-population, j denotes model $\mathcal{M}_{p_j}$, and m denotes 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 $t_{u3} = 26$. The above sample was taken from part of the wine data that was not included in this application, because of only partial availability of measurements for all the chemical compounds included here.

Under the first approach, model specification is completed by assuming independent beta prior distributions for π, η and φ:

$$(\eta_m) \sim Beta(1,1), \quad (\varphi_m) \sim Beta(1,1) \quad m = 1,2,3$$
$$(\pi_1) \sim Beta(2,2), \quad (\pi_2) \sim Beta(1,3), \quad (\pi_3) \sim Beta(1,5).$$

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 nation-wide production (thousands of liters by grape variety) supplied by the National Statistics Institute of Chile (INE; 2008).

As part of routine procedures related to wine exports, a sample of bottles is taken upon arrival to the corresponding customs point, and chemical analysis of the samples is performed to verify authenticity. Specifically, the analysis may include measuring concentrations for some of the chemical compounds, including those listed in Table 6. The bottles in the sample are then representative of the whole set of bottles in the container or batch. We therefore think of the loss as associated to a batch. From the discussion leading to (4), 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 of a batch depend on additional information which we do not have, such as the batch size, number of rejected bottles, transportation costs, 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 ordinal terms. In fact, all models assume the same loss, so 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 expressed in US\$ as of January 2011 (von Baer; 2010).

With the losses and costs described above we estimated the expected loss (5) as a function of the threshold p_0. The expected loss for Cabernet Sauvignon for each of the seven models is given in Figure 5.

For almost all values of p_0, $\mathcal{M}_{p_1}$ is the best model. $\mathcal{M}_{p_2}$ has similar expected loss than $\mathcal{M}_{p_1}$. Therefore, measurements of Anthocyanins or Organic Acids are most useful

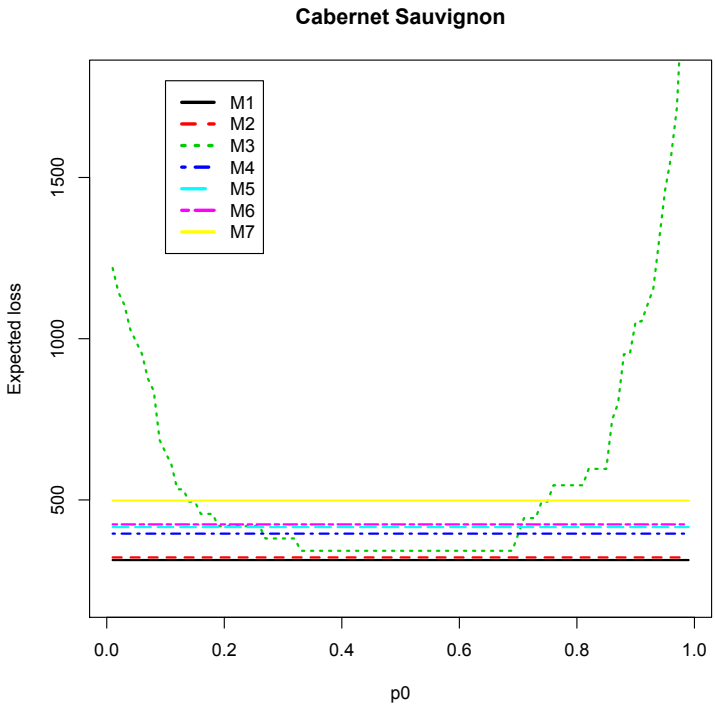


Figure 5: Expected loss for Cabernet Sauvignon as a function of p_0

when a producer wants to verify that a sample of wine is Cabernet Sauvignon. All the expected loss functions in Figure 5 look similar, which is due to the similar classification of the seven models of Table 4. When examining the estimated probabilities $\widehat{P}(g_{n+1} = u|y^{n+1}, y^n)$ for Cabernet Sauvignon samples, we realized that these values were unusually high, e.g. about 0.9999. Thus the proposed model is very flexible for classifying these data, to the point of rendering the expected loss not sensitive to the value of p_0. We also point out that this was not at all the case for the simulated data example, indeed the behavior of the expected loss function was quite different, because the simulated classification pattern was more complex. This explains why the expected loss functions in Figure 5 are so flat. The zero-one law gives the same classification, but when varying the value of p_0, the expected loss for model 3, $\mathcal{M}_{p_3}$ in Figure 5,

exhibits some differences, because Flavonol concentrations are less informative than Anthocyanins and Oganics Acids to authenticate Cabernet Sauvignon (von Baer et al.; 2007). Figure 6 shows the expected loss function for Merlot. In this case, $\mathcal{M}_{p_1}$ yields

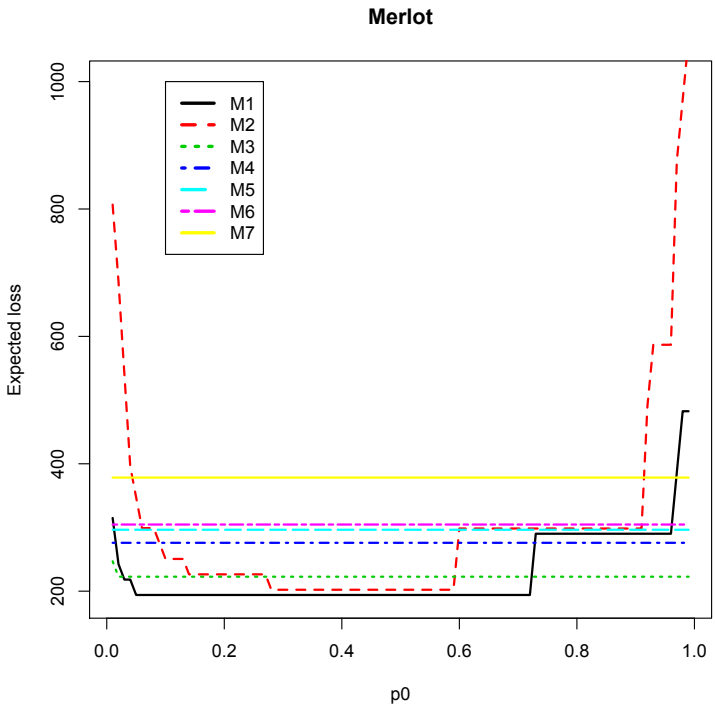


Figure 6: Expected loss for Merlot as function of p_0

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 loss function for $\mathcal{M}_{p_2}$ is not invariant to the choice of p_0, for values of p_0 in the range of 0.3 to 0.6 this model presents lower losses than $\mathcal{M}_{p_3}$, and, moreover, it is cheaper. Therefore if a producer wants to verify that a sample of wine is Merlot, measurements of Anthocyanins and Organic Acids are suggested.

Finally, Figure 7 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

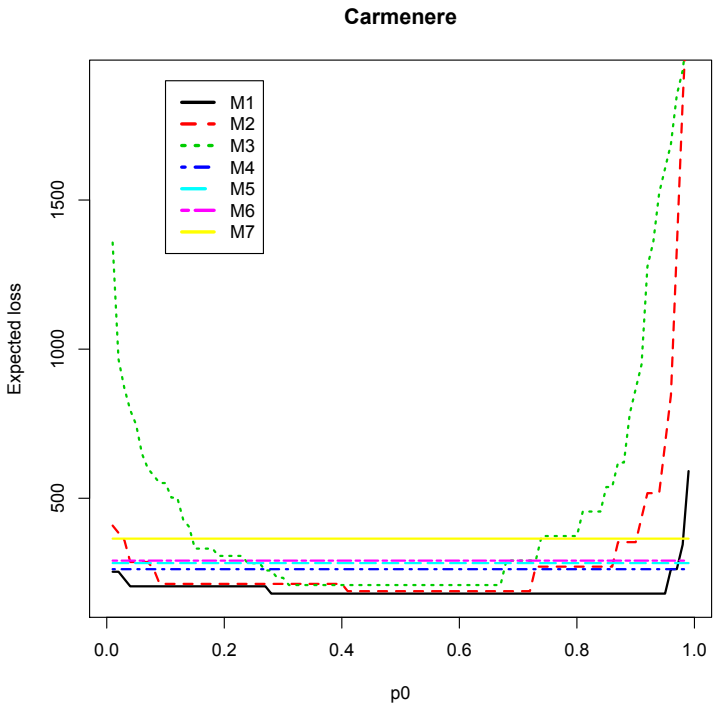


Figure 7: Expected loss for Carménère as a function of p_0

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 the best choice. Following the same scheme of analysis of the simulated data example, we evaluated the sensitivity of the conclusions to the choices of l_{A^cU} and l_{AU^c}. Here, we used the same range of values for l_{A^cU} and l_{AU^c} employed in the simulated data example. Figure 8 shows the minimal expected loss (that is, the expected loss at the optimal p_0 value) as function of l_{A^cU} (left panels) and l_{AU^c} (right panels) for the three grape varieties. The minimum expected loss shows a linear behavior due to the flat shape of the expected losses of the Figures 5, 6 and 7, due to the linear behavior there are unique optimal values for p_0. The optimal p_0 value for Cabernet Sauvignon was 0.5 for $\mathcal{M}_{p_1}$, $\mathcal{M}_{p_2}$, $\mathcal{M}_{p_4}$, $\mathcal{M}_{p_5}$, $\mathcal{M}_{p_6}$ and $\mathcal{M}_{p_7}$; for $\mathcal{M}_{p_3}$, p_0 was 0.51. For Merlot, p_0 was 0.5 in the case of

$\mathcal{M}_{p_3}$, $\mathcal{M}_{p_4}$, $\mathcal{M}_{p_5}$, $\mathcal{M}_{p_6}$ and $\mathcal{M}_{p_7}$; for $\mathcal{M}_{p_1}$ p_0 was 0.38 and 0.43 for $\mathcal{M}_{p_2}$. In the case of Carménère, the optimal p_0 was 0.5 for $\mathcal{M}_{p_4}$, $\mathcal{M}_{p_5}$, $\mathcal{M}_{p_6}$ and $\mathcal{M}_{p_7}$; for the other models the values were 0.61 ($\mathcal{M}_{p_1}$), 0.56 ($\mathcal{M}_{p_2}$) and 0.49 ($\mathcal{M}_{p_3}$). The results from Figure 8 are concordant with the results obtained in Figures 5, 6 and 7. In conclusion, under this approach, measurements of Anthocyanins and Organic Acids are the best choice for the three grape variety.

For the second approach described in Section 2.2.2, 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 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). Again, the losses of wrong decisions are the same for all models and the cost of data collecting c_j varies across models. The loss ranges were selected so as to obtain a broad view of the minimum expected loss under different scenarios. Under this approach we can see how sensitive our conclusions are, regarding the choice of groups of chemical compounds, to the choice of values in Table 1. The results are shown in Figure 9.

Figure 9 shows, for grape variety Cabernet Sauvignon, that $\mathcal{M}_{p_1}$ attains the minimal expected loss for all values of a. 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 $\mathcal{M}_{p_3}$ 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.

Additionally, we performed a sensitivity analysis for different values of prevalence

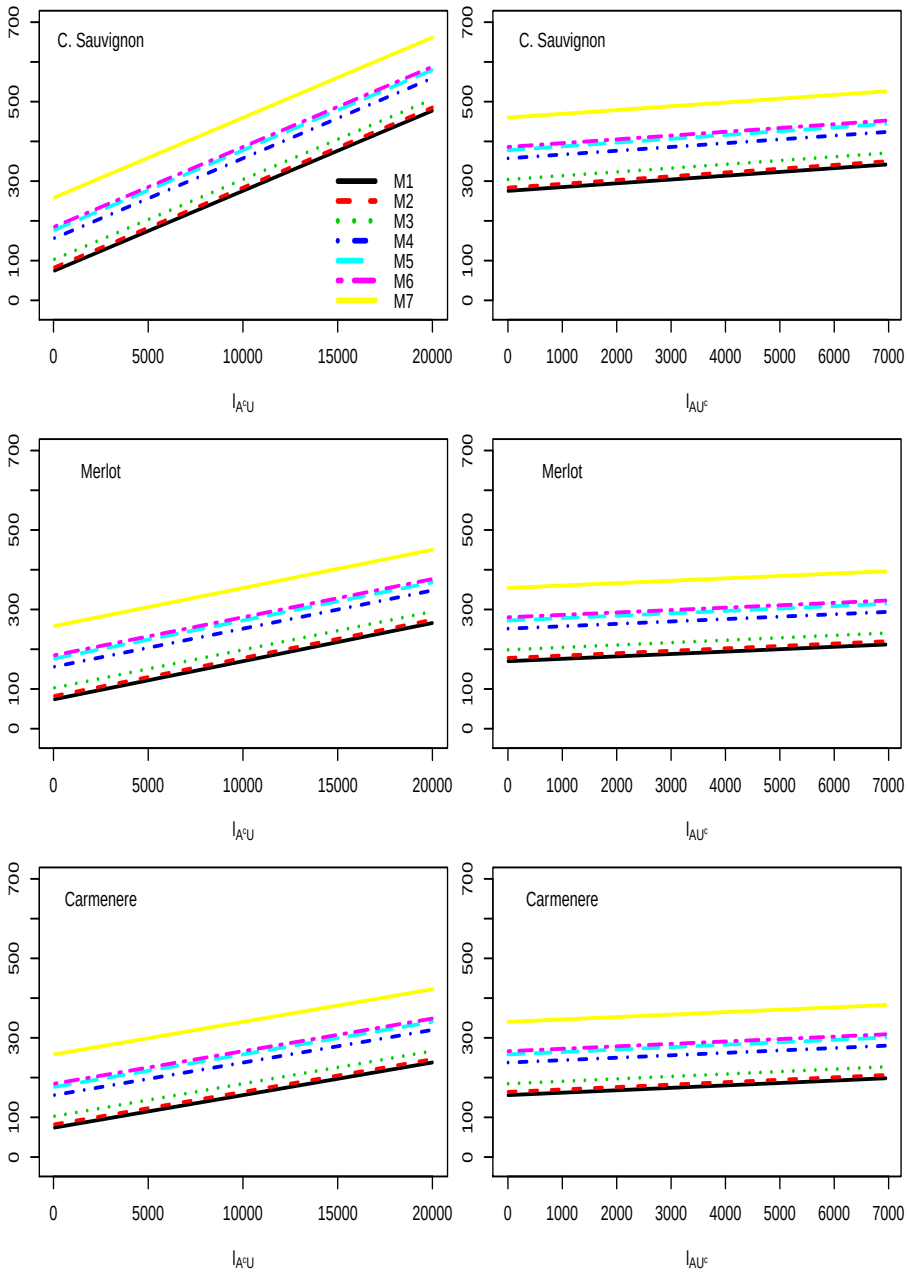


Figure 8: Minimum expected loss as a function of losses l_{A^cU} (left panels) and l_{AU^c} (right panels).

(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

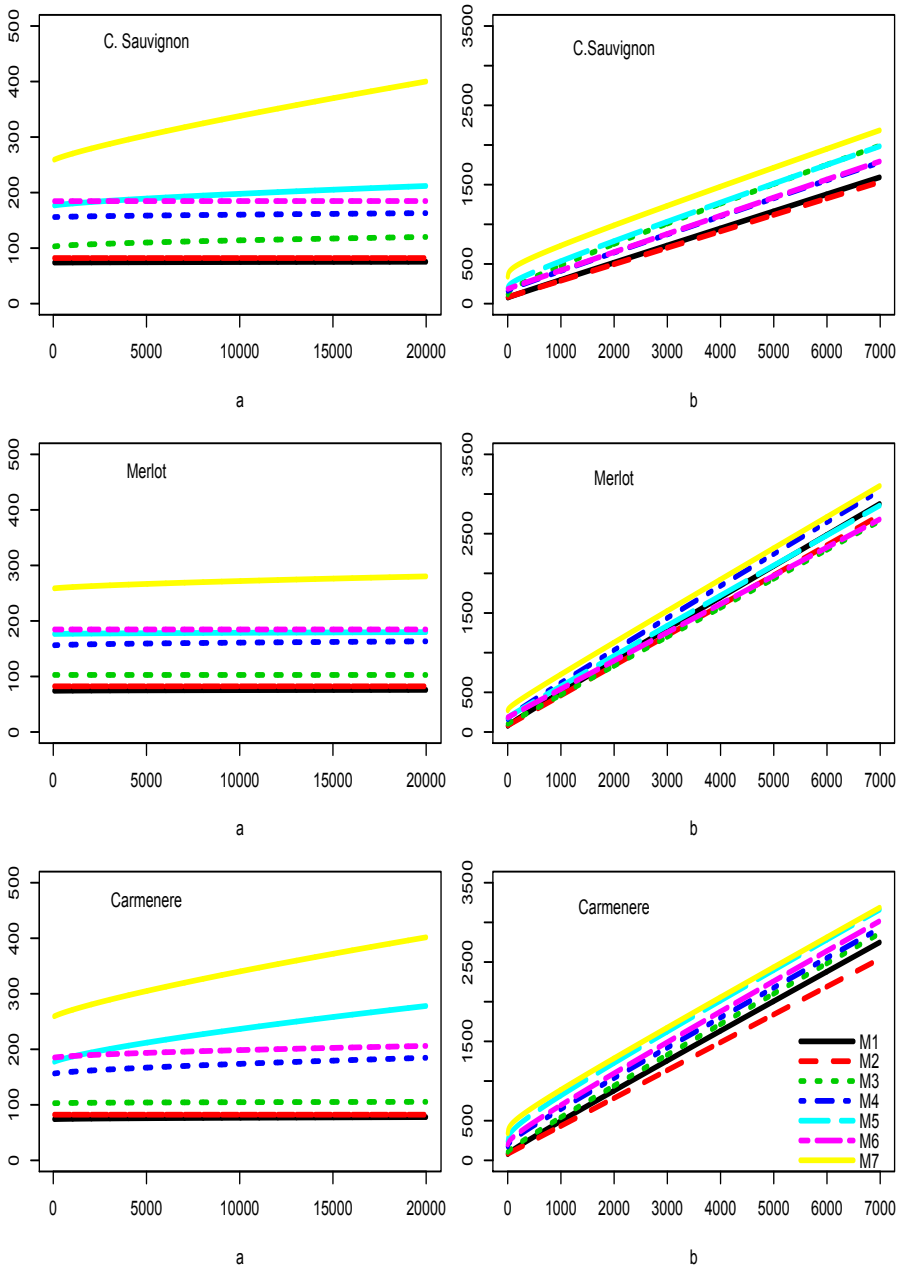


Figure 9: Minimum expected loss as function of losses a and b

for each grape variety were not affected.

In summary, the two approaches lead to the following 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, Anthocyanins or Flavonols are more appropriate than Organic Acids; and (iii) to verify whether a wine sample is Carménère, Organic Acids or Anthocyanins are appropriate. Figures 8 and 9 show that the conclusions are invariant to the values in Table 1 for a broad range of loss values.

5 Concluding remarks

The methodology discussed allows the user to select the optimal information needed to verify authenticity of red wine varieties. In our examples, the conclusions are invariant to the choice of values in Table 1 under the constraint in (4). The methodology could be applied to any authentication problem where more than one group of chemical markers are available for the analysis. 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, generally the better fit we get. But improving the fit might be too expensive, and so our approach balances the achieved

precision with the cost required to use the additional information. In that sense, the conclusions we draw can be useful to producers and consumers, as they allow to focus their efforts on the most appropriate combination of chemicals to consider for each wine variety.

Acknowledgements

The authors thank the Associate Editor and two anonymous referees for their valuable comments. The first author was partially funded by Program U-INICIA VID 2011, grant U-INICIA 02/12A; University of Chile. The second author was partially funded by grant FONDECYT 1100010.

6 Appendix

Anthocyanins	Organic Acids	Flavonols
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 6: Measured compounds

References

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

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

Celeux, G., Forbes, F., Robert, C. and Titterington, D. (2006). Deviance information criteria for missing data models, *Bayesian Analysis* **1**(4): 651–674.

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., 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.

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.

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

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.

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.

Greiner, M. (1996). Two-graph receiver operating characteristic (TG-ROC): update version supports optimisation of cut-off values that minimise overall misclassification cots, *Journal of Immunological Methods* **191**: 93–94.

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

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

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

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*

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.

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.

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 Carménère 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.

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.