Available online at www.sciencedirect.com
ScienceDirect
Chemometrics and Intelligent Laboratory Systems 87 (2007) 200–207
Chemometrics and intelligent laboratory systems
www.elsevier.com/locate/chemolab

Multivariate prototype approach for authentication of food products

S. Preys [a,*], E. Vigneau [b], G. Mazerolles [a], V. Cheynier [a], D. Bertrand [b]

[a] *UMR Sciences pour l'Oenologie, INRA, 2 Place Viala, 34060 Montpellier, France*
[b] *Unité de Sensométrie et de Chimiométrie, ENITIAA/INRA, La Géraudière, BP 82225, 44322 Nantes Cedex, France*

Received 28 July 2006; received in revised form 17 January 2007; accepted 22 January 2007
Available online 31 January 2007

Abstract

Authentication basically consists in deciding if a given unknown product belongs or not to a group of interest, defined by producers or regulators. More often, in order to demonstrate the authentication ability of a given instrumental analysis, several other groups are arbitrarily chosen. Then a Factorial or Linear Discriminant Analysis (FDA or LDA) or a Partial Least Squares Discriminant Analysis (PLS-DA) is usually performed; the model therefore depends on the nature of all observed groups of the study. The aim of this paper was to investigate an approach, named "prototype approach", based on a model built up only using the group of products of interest. Such an approach has the advantage not to depend on the whole complementary data of the study.

Prototype approach is inspired by Multivariate Statistical Process Control and Hotelling T^2 statistic and consists in buiding up the assignment model according to the group of interest. Then, authentication step of new data is performed. Prototype approach and FDA were compared on a case study (authentication of Beaujolais red wines using their polyphenolic composition). False negative (#FN) and false positive (#FP) numbers were estimated by bootstrapping procedures for both methods.

Compared to FDA, the prototype approach gave higher #FP with larger variability and lower #FN with lower variability. Wines produced with the same grape variety as AOC Beaujolais but in other regions were poorly authenticated. The prototype approach appears to be more flexible than FDA. The user can adjust the theoretical α risk in relation to its strategy, making that decision tool an alternative to discriminant analyses for authentication.

Keywords: Authentication; Multivariate Statistical Process Control (MSPC); Prototype; Factorial/Linear Discriminant Analysis (FDA/LDA); Food; Wine; Polyphenols

1. Introduction

Authentication is the ability to assign an unknown product to a known class of products by means of its physico-chemical or even sensorial characterization and a learning model. In the food industry, the authentication of products is an important need in the scope of traceability, food safety and quality control [1,2]. Authentication tools can also be used for marketing purposes, especially in order to build commercial brands including very well differentiated products for consumers. In this scope, authenticating quality marks such as 'AOC' (*Appellation d'Origine Contrôlée*=Protected Denomination of Origin) are of prime interest. This is more and more often achieved by characterizing products in a multivariate way, rather than analyzing independently one or few markers [3].

Many studies have dealt with differentiation or authentication of food products such as wines. The wines were differentiated in relation to their variety using markers such as volatile compounds [4,5], to their vintage by analyzing stable isotopes of minerals [6], to their geographical origin by means of trace element measurements [7,8,9], or to the wine-making process by analyzing amino-acids [10,11]. Some authors explored the discriminative potential of some polyphenolic compounds, which are secondary metabolites of the grape berry mainly responsible for wine color and astringency. Anthocyanin composition was used to differentiate red wines made from different grape varieties in various regions [12–15]. Some other polyphenolic compounds, i.e. flavonols [16,17] or phenolic acids [17,18], were analyzed to discriminate wines from various varieties, regions and technologies.

* Corresponding author. Fax: +33 4 99 61 26 83.
E-mail address: spreys@ondalys.fr (S. Preys).

0169-7439/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.chemolab.2007.01.003

Discriminant analyses are commonly carried out in such studies. Most of the time, the authors made use of FDA or LDA (Factorial or Linear Discriminant Analysis) [5,6,8,10,15,18], and more rarely SIMCA (Soft Independent Modelling of Class Analogy) [7]. More recently, PLS-DA (Partial Least Squares Discriminant Analysis) [19,20,21] appeared to be an interesting tool. Non-parametric methods, e.g. *k*-NN (*k*-Nearest Neighbors) [7,22], and neural networks [22] were also used.

In many situations, the very purpose of authentication studies is to separate a single "group of interest", from other groups. When using discriminant analyses, it is thus necessary to build up groups of observations representing the group of interest and also complementary groups, including products, which do not belong to the group of interest. As it is almost impossible to study all the existing groups of products, the resulting model thus depends on the nature of these complementary groups. Moreover, if a group has a particular importance, it seems reasonable to make principal use of it in building up the model.

The objectives of this work were (i) to investigate an authentication approach, named the "prototype approach", where only the knowledge of the group of interest, called the "reference group", is used to build up a set of decision rules; and (ii) to compare this prototype approach to FDA, which is usually used in authentication problems. The performances of the two methods will be discussed on an illustrative example, dealing with the authentication of AOC commercial red wines using their polyphenolic composition.

2. Statistical methods

2.1. Prototype approach

The prototype approach only requires that the reference group has been well defined previously. The presented methodology was inspired by Multivariate Statistical Process Control (MSPC) [23–26]. The main difference between the proposed prototype approach and MSPC methodology lies in the fact that the notion of time-series in MSPC, with observations repeated at every time point of a continuous process, is no longer appropriate in authentication studies. However, the rationale of the method is the same: once having defined a model giving a description of the reference products, new observations are considered and assessed to be compatible, or not, with the reference.

In MSPC, when p multinormal variables are measured on each observation, discrepancy from the in-control or reference situation is evaluated by using the Hotelling T^2 statistic [27]:

$$T^2 = \left(\mathbf{x} - \hat{\boldsymbol{\mu}}\right)' \hat{\boldsymbol{\Sigma}}^{-1} \left(\mathbf{x} - \hat{\boldsymbol{\mu}}\right) \tag{1}$$

where $\mathbf{x}$ is the $(p \times 1)$ measurement vector for one particular observation (or a sample of size one) and $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$ are respectively the mean vector and variance–covariance matrix estimated under the in-control situation. This T^2 statistic is actually the squared Mahalanobis distance between each multidimensional observation and the centroid of all observations involved in the estimation of the parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ [28]. The training set of n observations, used for estimating $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, is supposed to be representative of the reference situation.

In rather common cases, the number p of measured variables is large, and may present a high level of colinearity. This results in a variance–covariance matrix $\boldsymbol{\Sigma}$ that is nearly singular. A procedure for reducing the dimensionality of the variable space is to use Principal Components (PC) or PLS components [26]. If we consider the PC $\mathbf{t}_k$, processed after a Principal Component Analysis (PCA), organized in decreasing order of their variance λ_k (for $k = 1,\ldots, \min(n-1, p)$), the T^2 statistic can be expressed as:

$$T^2 = \sum_{k=1}^{\min(n-1,p)} \frac{\mathbf{t}_k^2}{\lambda_k} = T_A^2 + \widetilde{T}^2 \tag{2}$$

where

$$T_A^2 = \sum_{k=1}^{A} \frac{\mathbf{t}_k^2}{\lambda_k} \tag{3}$$

T_A^2 is the T^2 value estimated from the A first PC. $\widetilde{T}^2$ is thus a residual value, representing the deviation from the PCA model.

From Eq. (2), it clearly appears that the last PC, associated with the smallest eigenvalues λ, can play a main role in the statistic value. Thus, an alternative is to consider only the A first PC and to make use of the statistic T_A^2 (Eq. (3)). Nevertheless, as the T_A^2 values do not take into account possible new phenomena, which are not expressed in the principal space formed by the A first PC of the training set, another statistic is also considered. This statistic, denoted RSPE (Root Squared Prediction Error), is the square root of the variance of the residuals, obtained after projection into the principal space.

For each of these statistics, a decision rule is built up, the null hypothesis being associated with the reference situation. For authentication purpose, the parameters of the model under the null hypothesis are estimated on the basis of a training set of reference observations. The critical values of the tests, named Upper Control Limit (UCL) in MSPC, are defined [24,29] as follows:

$$\mathrm{UCL}_{T_A^2} = \frac{A(n-1)(n+1)}{n(n-A)} F_{1-\alpha,A,n-A} \tag{4}$$

$$\mathrm{UCL}_{\mathrm{RSPE}} = \sqrt{\left(\frac{v}{2n}\right)\chi^2_{1-\alpha,\frac{2\eta^2}{v}}} \tag{5}$$

where n is the size of the training reference set and A the number of PC retained. η and v are respectively the mean and the variance of the SPE (Squared Prediction Error), i.e. the variance of residuals, obtained for the training set. F and χ^2 hold for Fisher and Chi-squared distributions, and α is the chosen significance level.