Analytica Chimica Acta, 162 (1984) 241—251
Elsevier Science Publishers B.V., Amsterdam — Printed in The Netherlands

PREDICTION OF WINE QUALITY AND GEOGRAPHIC ORIGIN FROM CHEMICAL MEASUREMENTS BY PARTIAL LEAST-SQUARES REGRESSION MODELING

I. E. FRANK and BRUCE R. KOWALSKI*

Laboratory for Chemometrics, Department of Chemistry BG-10, University of Washington, Seattle, WA 98195 (U.S.A.)

(Received 23rd February 1984)

SUMMARY

A multivariate regression method, PLS, was applied to model the relationship between objective chemical measurements and subjective sensory evaluation of Pinot Noir wine samples. Descriptive and predictive models were calculated according to preset pathways in order to classify the wines according to their geographic origin and to predict several organoleptic characteristics. The importance of inorganic elements in these prediction problems was investigated.

Product quality control is a very important problem in all industrial processes. In most cases, the quality requirements are well defined and there is a narrow range of various measurable characteristics that a "good" product must match. In the food industry, it is a more complex problem. The quality of certain products cannot easily be defined by objective measurements and analytical chemical methods cannot fully replace organoleptic examinations. Wine is one of those interesting chemical mixtures, the overall sensory impression of which is defined by its many organic and inorganic composites in a very complex way.

Recently, a regression method called PLS (partial least squares) was developed [1], which models the relationship among information sources of multiple measurements according to a preset causal path. This method not only describes the connection between variables segregated into blocks, but gives a predictive model for measurements or observations in a response block.

Several studies have been done to explore the relationship between chemical measurements and sensory scores of wine samples [2—5]. Another important question is to identify the geographic origin of wines revealing counterfeits, instability of different types, labels [6—8]. In some other studies [9—11], in which pattern recognition was applied, only one source of information was investigated at a time or all the chemical measurements coming from different analytical methods were treated together, regardless of their natural separation.

0003-2670/84/$03.00 © 1984 Elsevier Science Publishers B.V.

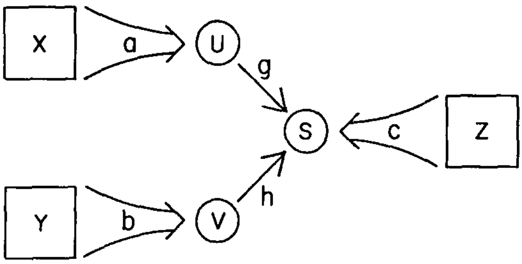


Fig. 1. PLS model with two predictor blocks and a multivariate response block.

In this study, the relationship among elemental, organic composition, sensory evaluation and geographic origin of 40 Vitis Vinifera cv. Pinot Noir from France and the United States is modeled by the PLS method. The PLS model provides useful information about which elements and organic compounds are responsible mainly for the good and bad characteristics of the different wines, about which chemical components differ in the French and American wine regions and about how the overall sensory score is composed of 13 individual organoleptic characteristics of the wine samples.

METHOD

The method applied here to build linear regression models on the data of four blocks is called PLS, which stands for partial least squares referring to the solution of the regression.

The idea behind this method is to describe each block of variables (type of measurement) by a set of latent variables (underlying components) which are linear functions of the original variables in the block. The role of these latent variables is not only to describe their own block (outer relationship), but also to relate to each other according to the preset causal pathway (inner relationship).

The first application in physical sciences [12] used the so-called MODE B algorithm, calculating only one latent variable for each block by multiple regression. The disadvantage of this algorithm is that it does not give a predictive model for the response block. The MODE A algorithm, which calculates several latent variables by single regression to build a predictive model, was applied in several two block chemical systems [13—16]. In the present study the MODE A algorithm was also used and was extended to more than two blocks.

This PLS algorithm describes more than the relationship among the various blocks. By explaining the influence of one block on another, it can predict all the variables in the response block (the block at the end of the causal path) and can establish which predictor block (type of measurement) is the most important as well as which predictor variables have the largest effect on the response variables. A PLS model with two predictor blocks, X and Y, and one response block, Z, is illustrated in Fig. 1. The relationship between the latent variables and the original variables of the same block (outer relationship) is analogous to an eigenvector expansion. Lesser numbers of latent variables, which are mutually orthogonal, span most of the variance associated with the response block in a predictor block with a higher number of original variables

$$
\begin{aligned}
u_i^m &= \sum_j a_j^m x_{ij} & i &= 1 \ldots I \\
& & j &= 1 \ldots J \\
v_i^m &= \sum_k b_k^m y_{ik} & k &= 1 \ldots K \\
& & l &= 1 \ldots L \\
s_i^m &= \sum_l c_l^m z_{il} & m &= 1 \ldots M
\end{aligned}
\tag{1}
$$

where i is the sample index, j, k and l are the indices of the variables, m is the index of components, u, v and s are the latent variables, and a, b and c are coefficients calculated by single regressions.

However, there is another criterion in the calculation of these latent variables, namely that they must be maximally correlated to each other, i.e., u and v must explain the maximum variance in s. The relationship among the latent variables is called the inner relationship

$$s_i^m = g^m u_i^m + h^m v_i^m \tag{2}$$

The optimal number (M) of the latent variables for a PLS model of the highest predictive power is determined by Stone-Geisser's cross-validation test [17]. This method estimates prediction error as a function of the number of latent variables by the leave-one-out technique. The model with least predictive error is then chosen. For faster computation, instead of leaving one sample out for prediction from the model and repeating the procedure I times, one-fourth of the samples were omitted and the procedure was repeated four times.

A special case of the PLS model is also used here where there is no Y block and there is only one variable in the Z block. PLS gives a suboptimal solution for the underdetermined system. It mitigates the collinearity problem by regressing the response variable on orthogonal latent variables. In this respect, the PLS regression is similar to the principal component regression. Increasing the number of latent variables to the number of original variables in the predictor block, the PLS solution converges to multiple regression by ordinary least squares. However, by removing some of the later latent variables, which similarly to the principal components describe only variance caused by noise, automatic noise filtering can be achieved. In this case, although the fit of the model is worse than that of the multiple regression solution, the predictive power of the model is increased, as variance caused by random noise is omitted from the model. The coefficients of the resulting linear model can easily be calculated from the loadings of the latent variables (a_j^m) and the inner relationship coefficients (g^m). The PLS method gives biased estimates for the regression coefficients in contrast to the ordinary least squares solution. However, the variance of PLS estimates is generally smaller than those for least squares, so that often the expected squared error is smaller. This is especially true when $J > I$.

DATA

The 40 Pinot Noir wine samples used in this study, and previously examined by various pattern recognition methods, have been listed elsewhere [9]. The contents of 17 elements were determined by atomic emission spectrometry [9]. The organic components were separated by gas chromatography (g.c.); the mathematical characterization of the chromatograms contained 137 integrated peak areas of organic acids and neutral components [10]. The sensory evaluation was given in terms of 13 individual characteristics and one overall quality score [18]. The geographic categories and their three-digit binary codes are as follows

wines from the Pacific Northwest	1	0	0
wines from California	0	1	0
wines from France	0	0	1

The total of 171 (17 + 137 + 14 + 3) measurements, scores and binary codes, called variables, were segregated into four blocks, each describing a different type of measurement. The data used here are summarized in Table 1. As identification of the 137 chromatographic peaks was not available for this study, the whole chromatogram was used directly as a 137-dimensional measurement vector.

Because of the different scale in each measurement, all variables were scaled to zero mean and unit variance.

RESULTS AND DISCUSSION

The effect of individual sensory parameters on the overall score for wine quality

A PLS regression model was calculated on the sensory evaluation block, trying to explain the variance in the overall quality score from the individual sensory parameters. Three models were built: one on all 40 samples, one on the 26 American wines and one on the 14 French wines. The optimal number of latent variables from prediction point of view was determined by cross-validation. The results are summarized in Table 2. In the overall model, two orthogonal components were found; both describe the "goodness" of the wine (positive correlation to the overall quality). The first, most important component is composed of aroma character and intensity, flavor character and intensity, body and color as positive contributors and undesirable odor

TABLE 1

Data for the wine study

Wine type	Block I Category (3)	Block II Elemental content (17)	Block III G.c. (137)	Block IV Sensory scores (14)
17 from Pacific Northwest	1 0 0	1 Cd	Peaks of organic acids and neutral components	1 Clarity
		2 Mo		2 Color
		3 Mn		3 Aroma intensity
		4 Ni		4 Aroma character
9 from California	0 1 0	5 Cu		5 Undesirable odor
		6 Al		6 Acidity
		7 Ba		7 Sugar
14 from France	0 0 1	8 Cr		8 Body
		9 Sr		9 Flavor intensity
		10 Pb		10 Flavor character
		11 B		11 Oakiness
		12 Mg		12 Astringency
		13 Si		13 Undesirable taste
		14 Na		14 Overall quality
		15 Ca		
		16 P		
		17 K		

and taste as negative contributors. The second, weaker component has positive loadings from clarity, acidity and negative loadings from color, sugar and oakiness. Astringency turned out to be an unimportant parameter according to the panel for predicting the overall quality. If the model is calculated with only one type of wine, cross-validation finds only one predictive component. When the regression coefficients of the American and French model are compared, significant differences can be noted. Although in both models, the highest positive contributors are the aroma and flavor characters and body and the highest negative contributors are the undesirable taste and odor, strangely in the French model, the most negative parameter is clarity; the acidity got a high negative loading while oakiness and sugar became positive in both models. This means that the judges' impression of the overall quality as a function of individual characteristics differs from one type of wine to another. Certain parameters are ignored (low loadings) in the overall and the American model, while in the French model, all the individual scores have high loadings.

The PLS method gives a model with a good fit and prediction accuracy for the overall score, considering the high noise level in sensory evaluation data. However, the regression coefficients differ from those of the Davis scorecard and of the Modified scorecard [18].

TABLE 2

Magnitude of regression coefficients in three different sensory models

Estimated regression coefficients	American and French	American	French	
	Flavor char. Aroma char. Body Flavor int. Color Aroma int.	Flavor char. Aroma char. Body Color Sugar Oakiness Flavor int.	Flavor char. Aroma char. Body Color Sugar Aroma int. Oakiness Astringency	+
	Oakiness Acidity Sugar Astringency Clarity	Aroma int. Clarity Acidity		0
	Undesirable odor Undesirable taste	Astringency Undesirable odor Undesirable taste	Acidity Undesirable odor Undesirable taste Clarity	–
Error in fit (%)	6	8	4	
No. of latent variables	2	1	1	
Error in prediction (%)	8	10	12	

Prediction of geographic origin

An attempt was made to predict the geographic origin of the wine samples on the basis of one or more information sources by the PLS method. The results are listed in Tables 3 and 4. In previous pattern recognition analyses [9—11] only selected variables were included in the model in order to maintain a high enough sample/variable ratio. With the PLS method, information relevant to the prediction problem from all the variables can be used. Also, in the previous studies, only the goodness of fit of the models was investigated and no result was reported referring to the prediction power of the models calculated by different methods. It is very important in PLS models (and, in general, in any regression model) to choose the number of parameters (variables or components) by cross-validation, which selects the optimal model for prediction.

TABLE 3

Parameters of PLS models predicting geographic origin

Predictor block(s)	Models			
	Sensory	Elemental	G.c.	Elemental + g.c.
No. of components determined by cross-validation	1	4	5	4
Sum of the squared residuals in cross-validation	2.80	0.97	2.11	0.82
Average absolute errors in fit	0.456 0.302 0.253	0.115 0.134 0.116	0.134 0.138 0.176	0.069 0.094 0.081
Misclassification fit: prediction (out of 40 samples)	25:36	0:7	0:10	0:6

TABLE 4

Regression matrix for predicting geographic origin from elemental analysis

	Cd	Mo	Mn	Ni	Cu	Al	Ba	Cr	Sr
Pacific Northwest	1.526	0.381	−0.105	0.722	0.054	0.185	1.333	−5.422	0.105
California	0.760	−0.298	0.110	0.175	−0.250	−0.293	0.112	3.516	0.071
France	−2.287	−0.084	−0.005	−0.898	0.197	0.108	−1.444	1.906	−0.177
	Pb	B	Mg	Si	Na	Ca	P	K	
Pacific Northwest	0.142	−0.107	0.003	−0.003	0.000	0.005	−0.003	−0.000	
California	−0.074	0.057	−0.000	0.013	−0.001	−0.000	0.002	0.000	
France	−0.068	0.049	−0.003	−0.010	0.000	−0.004	0.001	−0.000	

One solution to calculate regression models for category response variables is to introduce as many dummy variables as there are categories to be separated. Each of these dummy variables has two values: 1 if the samples belong to the category described by the variable and 0 if it does not. With the two-block PLS method, all the dummy variables can be incorporated in the model. This means that parallel separation of all three wine types (French, California and Pacific Northwest) can be achieved.

The sensory data do not provide enough information to separate the three wine regions. These data separate only the French wine samples, and the model has practically no predictive power. Both the organic and inorganic compositions provide ample information to separate completely the three wine categories (see Table 3 misclassification in fit), but the prediction power of the model based on the elemental analysis is much better (see Table 3 misclassification in prediction). Combination of the two chemical blocks as two predictor blocks gives, by far, the best fit (average absolute error in fit), the reported values should be compared on a relative basis with the actual values: 17/40 = 0.43, 9/40 = 0.23 and 14/40 = 0.35.

From the regression coefficient matrix, which in the case of multiple response variables is analogous to the regression coefficient vector, the significance of the predictor variables in separating one particular wine type can be revealed. The regression matrix for the elemental block is given in Table 4. For Californian wines the most characteristic elements are chromium and cadmium; in wines from the Pacific Northwest cadmium, barium and nickel; and in French wines chromium, aluminum and copper. Barium, Cd, Ni and Sr are the best elements in distinguishing between American and French wines (high positive coefficients for the American wines and high negative coefficients for the French wines), while for the separation between the two American types, Al, Cr, Mo and Mn are the most important elements.

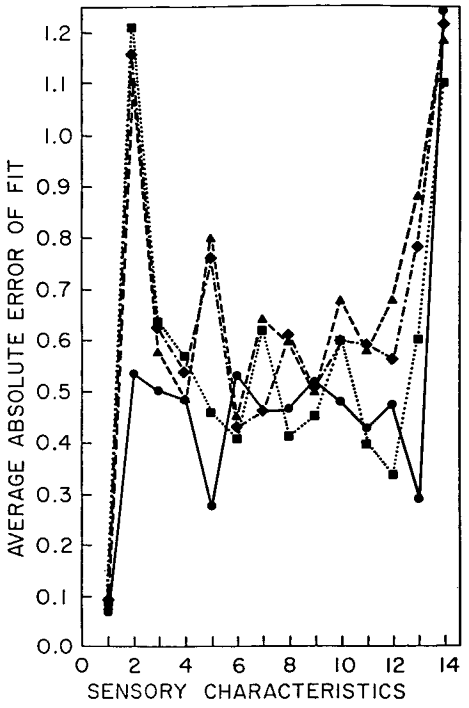
In a previous study [9], in two-category separations barium and calcium were selected as variables best separating American and French wines, while aluminum and potassium were best for separating wines from the Pacific Northwest and California.

Prediction of all sensory parameters from chemical measurements

In a previous study [11], an attempt was made to correlate the chemical measurements to the sensory evaluation data. Only two peaks of the gas chromatogram were included in the stepwise regression analysis in order to avoid overfitting, however, many other variables had high correlation to the overall quality score. In the same paper, the first three principal components of the chemical and sensory variables (in the former only g.c. peaks were reported having high loadings) did not show high correlations with one another, so this model failed to capture the relationship between the objective measurements and organoleptic evaluations. The following results show that PLS gives an applicable model, where other multivariate methods failed. With the general two- and three-block PLS method, it is possible to calculate the regression model to predict all the individual and the overall sensory scores together from the chemical composition of the wines. Three causal paths were examined: sensory data prediction from elemental analysis, from gas chromatograms and from both elemental and organic composition. In the first case, three models were built: on all wines, only on American wines and only on French wines.

The American wine model consists of one component of "goodness": positive loadings of positive characteristics and negative loadings of negative ones. The French wine model has one component of "badness": negative loadings of positive characteristics. In the overall model, both components appear. In the French wines, Cd, Mo, Cu and Cr, and in the American wines, Cd, Pb, Ca, Mo, Ni have a positive influence on the quality. Manganese, Cr, Sr, Mg, K and Ba have negative effects in both regions. The slight differences between the effect of the French and American elemental composition is probably due to the different types of soil that can change the matrix of the elements.

In Fig. 2 goodness of fit for the three models is compared. The values on the horizontal axis are the identification numbers of the sensory characteristics. On average, the French model fits better than the American or the overall model; for most of the sensory parameters the average absolute error is around 10% of the mean of the actual value. A specially bad fit is observed for color (2), undesirable odor (5) and undesirable taste (13). The explanation might be that the undesirable characteristics develop because of organic reactions during the storage time, so that the elemental analysis does not have predictive information about these characteristics.



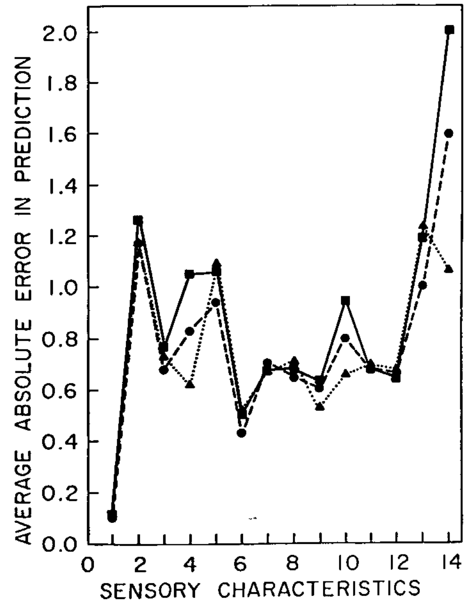


Fig. 2. Goodness of fit of PLS models predicting all the sensory characteristics from elemental analysis. (·—·) All 40 samples; (——) 10% of the mean of the actual value; (– – –) American wine samples; (· · ·) French wine samples.

Fig. 3. Goodness of prediction of PLS models predicting all sensory characteristics from different chemical measurements. (——) G.c.; (– – –) g.c. and elemental analysis; (· · ·) elemental analysis.

As the next step, a predictor model for sensory data on the basis of the chromatographic data was calculated and finally both elemental and organic composition were included in the two predictor block model. The prediction powers of the three models (elemental, organic and elemental and organic) on all 40 samples are compared in Fig. 3.

From the inner relationship coefficients (0.75 of elemental and 0.53 of g.c. block), it is clear that elemental analysis contains more information relevant to the organoleptic characteristics than the organic composition. For certain characteristics, like clarity (1), color (2), acidity (6) and sugar (7), the two chemical blocks have the same predictive power. The prediction based on elemental analysis is better for aroma intensity (3) and character (4), flavor character (9) and intensity (10) and overall judgment (14), while g.c. data give more information about body (8), oakiness (11), astringency (12) and undesirable taste (13) and odor (5). Most of the prediction errors are between 10% and 20%, which is a good result considering the high noise level in sensory evaluation data. Cross-validation chose two components for the elemental analysis and the g.c. plus elemental analysis models and one component for the g.c. model. Because the predictive information from the inorganic and organic composition blocks are highly correlated, the prediction power cannot be increased by using both chemical information sources.

The relationship between the two chemical blocks can be described by a five-component PLS model explaining, for example, the variance in the elemental analysis block from the chromatographic data. The percent error of fit for each element is shown in Table 5. Except for magnesium and copper, all the elements have less than 50% error; the best fit by the PLS model is given for K, Ca, P and Ba.

Conclusion

The aim of this study was to show how the PLS method can be applied to calculate a descriptive and predictive model of the relationship between objective chemical measurements and sensory evaluation data of wine samples. It was demonstrated that the chemical data contain sufficient information to predict the geographic origin, the individual sensory parameters and the overall quality of wines. By segregating chemical variables into blocks coming

TABLE 5

PLS model predicting elemental analysis from organic composition

Element	Cd	Mo	Mn	Ni	Cu	Al	Ba	Cr	Sr
Error in fit (%)	20	31	21	48	82	22	18	25	24
Element	Pb	B	Mg	Si	Na	Ca	P	K	
Error in fit (%)	38	22	120	35	47	11	15	8	

from different analytical measurements, not only the importance of single variables, but the relevance of each block in the prediction problem can be investigated. The composition of latent variables (components) in the PLS model gives quantitative information about which chemical measurements and individual sensory parameters are associated with the quality of the wines.

Because the overall quality score is a different combination of individual parameters in each type of wine, even by the same panel of judges, it is important to develop a model which is able to predict not only the overall sensory judgment, but all the individual quality parameters as well. PLS models predicting response blocks of several variables were calculated successfully to connect inorganic and organic composition with the various sensory parameters.

The PLS method can handle several blocks of multiple measurements, extracting significant components to predict many response variables together. Therefore, it is desirable to include results from several multivariate analytical methods to enable PLS to integrate information from all the relevant sources.

The authors thank the 3M Company for supporting the PLS program development.

REFERENCES

1 K. G. Joreskog and H. Wold (Eds.), Systems Under Indirect Observation, Parts I and II, North Holland, Amsterdam, 1982.

2 R. R. Nelson, T. E. Acree and R. M. Butts, J. Agric. Food Chem., 27 (1979) 1188.

3 A. C. Noble, R. A. Flath and R. R. Forrey, J. Agric. Food Chem., 28 (1980) 346.

4 P. X. Etievant, J. Agric. Food Chem., 29 (1981) 65.

5 H. R. Buser, C. Zainer and H. Tanner, J. Agric. Food Chem., 30 (1982) 359.

6 I. Moret, G. Scarponi, G. Capodaglio, S. Zanin, G. Camaiani and A. Toniolo, Am. J. Enol. Vitic., 31 (1980) 245.

7 G. Scarponi, I. Moret and G. Capodaglio, Riv. Vitic. Enol., 34 (1981) 254.

8 G. Scarponi, I. Moret, G. Capodaglio and P. Cescon, J. Agric. Food Chem., 30 (1982) 1135.

9 W. Kwan, B. R. Kowalski and R. K. Skogerboe, J. Agric. Food Chem., 27 (1979) 1321.

10 W. Kwan and B. R. Kowalski, J. Agric. Food Chem., 28 (1980) 356.

11 W. Kwan and B. R. Kowalski, Anal. Chim. Acta, 122 (1980) 215.

12 R. W. Gerlach, B. R. Kowalski and H. A. Wold, Anal. Chim. Acta, 112 (1979) 417.

13 W. Lindberg, J. Persson and S. Wold, Anal. Chem., 55 (1983) 643.

14 I. E. Frank, J. H. Kalivas and B. R. Kowalski, Anal. Chem., 55 (1983) 1500.

15 M. L. Bisani, D. Faraone, S. Clementi, K. H. Esbensen and S. Wold, Anal. Chim. Acta, 150 (1983) 129.

16 M. Sjöström, S. Wold, W. Lindberg, J. Persson and H. Martens, Anal. Chim. Acta, 150 (1983) 61.

17 M. Stone, J. R. Stat. Soc., Ser. B, 36 (1974) 111.

18 W. Kwan and B. R. Kowalski, J. Food Sci., 45 (1980) 213.