AUSTRIAN JOURNAL OF STATISTICS
Volume 35 (2006), Number 1, 45–55

Application of Interactive Regularized Discriminant Analysis to Wine Data

Ute Römisch[1], Dimitar Vandev[2] and Katrin Zur[1]
[1]Technical University, Berlin
[2]St. Kl. Ohridski University, Sofia

Abstract: Testing the possibility of determining the geographical origin (country) of wines on the base of chemico-analytical parameters was the aim of the European project "Establishing of a wine data bank for analytical parameters for wines from Third countries (G6RD-CT-2001-00646-WINE DB)" supported by the European Commission. Therefore a data base containing 400 samples of commercial and authentic wines from Hungary, Czech Republic, Romania and South Africa was created. For each of those samples around 100 analytical parameters, among them rare earth elements and isotopic ratios were measured.

Besides other multivariate statistical methods of discrimination and classification the method of regularized discriminant analysis (RDA) was used to distinguish the wines of the different countries on the base of a minimal number of the most important parameters. A MATLAB-program, developed by Vandev (2004) which allows an interactive stepwise discriminant model building on the base of an optimal choice of the "nonlinearity" parameter alpha was used. This program will be described shortly and models for commercial wines with corresponding classification and prediction error rates will be given.

As a result of using RDA it was possible to reduce the number of analytical parameters to the eight to infer the geographical origin of these commercial wines.

Zusammenfassung: Das Prüfen der Möglichkeit der geographischen Herkunftsbestimmung von Weinen auf der Basis chemisch-analytischer Parameter war das Ziel des von der Europäischen Kommission unterstützten Europäischen Projektes „Errichtung einer Weindatenbank für analytische Parameter von Weinen aus Drittländern (G6RD-CT-2001-00646-WINE DB)". Hierfür wurde eine Datenbasis, die 400 kommerzielle und authentische Weinproben aus Ungarn, Tschechien, Rumänien und Süd Afrika enthält, erhoben. Für jede dieser Proben wurden ca. 100 analytische Parameter gemessen, unter ihnen seltene Erden und Isotopendaten.

Neben weiteren multivariaten Methoden der Diskriminierung und Klassifikation wurde die Regularisierte Diskriminanzanalyse (RDA) verwendet, um die Weine verschiedener Länder mit minimaler Anzahl der wichtigsten Parameter zu unterscheiden. Ein von Vandev (2004) entwickeltes MATLAB-Programm, das eine interaktive schrittweise Diskriminanzmodellbildung bei optimaler Wahl des „Nichtlinearitäts-Parameters" α gestattet, fand hierbei

Anwendung. Dieses Programm wird kurz beschrieben und es werden Modelle für kommerzielle Weine mit den entsprechenden Klassifikations- und Vorhersagefehlern angegeben.

Als Ergebnis der Anwendung der RDA konnte die Anzahl der analytischen Parameter auf die für die Unterscheidung nach ihrer geographischen Herkunft (Land) wichtigsten acht reduziert werden.

Keywords: Regularization, Classification.

1 Introduction

The responsible wine controlling authorities are often confronted with products which are not correctly marked with regard to their origin, vintages and quality parameters. To find out such adulterations of wines, the identification of the geographical origin of wines is of great interest to wine consumers and producers (Römisch et al., 2001). This was the background for creating a data base of wines from Hungary, Czech Republic, Romania and South Africa over a period of three years (2001-2003) in the scope of a European project.

Every year 400 commercial and authentic wine samples were collected based on a sample plan. Commercial wines were purchased directly from the wine producers of the respective countries, whereas authentic wines were produced under standardized conditions in a laboratory. For each of these samples of the first year around 100 chemical parameters were analyzed. After these first analyzes, taking the experiences of involved oenologists into consideration, it was possible to reduce this number to 63: regular 58 parameters plus 5 rare earth ratios, the chemists suggested to include.

Data management included data handling of missing and censored data, log-transformations of 90% of the data and the identification of univariate and multivariate outliers. Then descriptive and inferential univariate methods, variance and correlation analyzes and multivariate classification and projection methods were applied to all wine data and separately to authentic as well as commercial red and white wines.

For the case of commercial wines some results of linear, quadratic and regularized discriminant analyzes will be presented.

2 Discriminant Analysis

Discriminant analysis is used to analyze differences of two or more groups with respect to a set of variables measured on the objects of these groups. Two questions are to be answered:

1. Which variables are the most important to discriminate between the groups? (discrimination problem)
2. In which groups objects (elements, cases), whose group membership is unknown, will be classified based on their variable values? Which correct classification rates can be found with the estimated discriminant model? (classification and prediction problem)

The influence of the independent variables on the groups is to be investigated. Discriminant functions, which contain significant variables, are estimated and objects will be classified on the base of the estimated discriminant model. Different methods of discrimination (e.g. linear, quadratic, regularized, nonparametric, ...) can be used. "Good" discriminant models contain the most important variables for separating groups with minimal misclassification rates.

Here we restrict to presenting the results of an one-parameter based regularized discriminant analysis, including the linear and quadratic case.

2.1 Classification

Methods of discriminant analysis (McLachlan, 1992; Fahrmeir et al., 1996) allow assigning objects to one of K ($K \geq 2$) distinct groups on the base of a feature vector $x = (x_1, \ldots, x_p)$, containing the measurements from each object. Moreover, the separability of groups in the feature space will be analyzed.

Let the categorical variable Y denote the group membership of the object, where $Y = k$ implies that it belongs to the group with index k ($k = 1, \ldots, K$). Moreover, each object is characterized by the p-dimensional feature vector X. Let $p_k = P(Y = k)$ be the prior probabilities, that an object belongs to the group with index k and $f(x|k)$ be the conditional density of X given $Y = k$. The unconditional distribution of X is then

$$f(x) = \sum_{k=1}^{K} p_k f(x|k) \,.$$

Of special interest for classification problems is the posterior probability $p(k|x)$, i.e. the probability, that an object with observed feature vector x belongs to group k. Then according to the formula of Bayes this conditional probability of Y given $X = x$ can be written as

$$p(k|x) = P(Y = k|X = x) = \frac{p_k f(x|k)}{f(x)} \,.$$

Two well known allocation rules can be derived:

- Allocation rule of Bayes

$$p(\hat{k}|x) \geq p(j|x) \,, \quad \text{resp.} \quad p_{\hat{k}} f(x|\hat{k}) \geq p_j f(x|j) \,, \quad j = 1, \ldots, K \,, \tag{1}$$

- Maximum Likelihood allocation rule for the special case that $p_k = p$, $\forall k$,

$$f(x|\hat{k}) \geq f(x|j) \,, \quad j = 1, \ldots, K \,.$$

That is, an object with feature vector x will be assigned to that group with index $\hat{k}$ which has the largest posterior probability. The Bayes rule achieves minimal misclassification risk among all possible rules. All allocation rules considered have the general structure

$$d_{\hat{k}}(x) \geq d_j(x) \,, \quad j = 1, \ldots, K \,, \tag{2}$$

where $d_j(x)$ are called discriminant functions.

In practice the conditional densities $f(x|k)$ and sometimes also the prior probabilities p_k are unknown and have to be estimated on the base of a learning sample. For this purpose an assumption about the group distribution can be used for example.

2.2 Linear, Quadratic, and Regularized Discriminant Analysis

We assume normality for the p-dimensional feature vector X_k in group k

$$X_k \sim \mathrm{N}(\mu_k, \Sigma_k)\,, \quad k = 1, \ldots, K\,,$$

where $\mu_k$denote the group mean and Σ_k the group covariance matrix. Then the conditional distribution of X given $Y = k$ can be described by the density of the normal distribution

$$f(x|k) = (2\pi)^{-p/2}|\Sigma_k|^{-1/2} \exp\left(-1/2(x-\mu_k)'\Sigma_k^{-1}(x-\mu_k)\right)\,, \quad k = 1, \ldots, K\,. \quad (3)$$

Substituting equation (3) into $d_k(x) = f(x|k)p_k$ (see (1) and (2))and taking the logarithm leads to the discriminant function of the form

$$d_k(x) = -\frac{1}{2}\left((x-\mu_k)'\Sigma_k^{-1}(x-\mu_k) + \log|\Sigma_k|\right) + \log p_k\,, \quad k = 1, \ldots, K\,. \quad (4)$$

Using allocation rule (2) with equation (4) minimizes the misclassification risk and is called Quadratic Discriminant Analysis (QDA), since it separates the disjoint regions of the feature space corresponding to each group assignment by quadratic boundaries.

The Linear Discriminant Analysis (LDA) is used if the group covariance matrices are identical, i.e., $\Sigma_k = \Sigma, \forall k$. In this case the rule that minimizes the misclassification risk leads to a linear separation of the groups. The quadratic term in the discriminant function for all groups then is the same and can be eliminated. Whether LDA or QDA should be preferred depends on the structure of the data. If we consider real data, the parameters μ_k and Σ_k are unknown and have to be estimated ($\hat{\mu}_k$ and $\hat{\Sigma}_k$) from a given training sample. In practice, often LDA leads to better classification results than QDA, even when the true group covariance matrices are not equal, because less model parameters have to be estimated and LDA is more robust against violations of its basic assumptions.

Regularization techniques are successfully used in solving ill- and poorly posed problems. If the number of parameters to be estimated is comparable or even larger than the sample size, the parameter estimates can be highly unstable. Friedman (1989) has proposed the Regularized Discriminant Analysis (RDA) as a compromise between linear and quadratic discriminant analyzes. He has proposed two steps of regularization. First, the estimated group covariance matrix $\hat{\Sigma}_k$ should be regularized by

$$\hat{\Sigma}_k(\lambda) = (1-\lambda)\hat{\Sigma}_k + \lambda\hat{\Sigma} = \frac{(1-\lambda)(n_k-1)S_k + \lambda(n-K)S}{(1-\lambda)(n_k-1) + \lambda(n-K)}\,,$$

where S_k and S are the sample-based covariance matrix estimates and n_k and n the corresponding sample sizes. The regularization parameter $\lambda \in [0, 1]$ controls the degree of shrinkage of the group covariance matrix estimates toward the pooled estimate. If n is less than or comparable to p, the estimate of Σ_k should be regularized further by

$$\hat{\Sigma}_k(\lambda, \gamma) = (1-\gamma)\hat{\Sigma}_k(\lambda) + \gamma c_k I_p\,,$$

where I_p is the $p \times p$ identity matrix, and $c_k = \mathrm{tr}\left(\hat{\Sigma}_k(\lambda)\right)/p$. For a given value of $\lambda \in [0, 1]$, the additional regularization parameter $\gamma \in [0, 1]$ controls shrinkage toward a

multiple of the identity matrix. The multiplier c_k is the average value of the eigenvalues of $\hat{\Sigma}_k(\lambda)$. This shrinkage has the effect of decreasing the larger eigenvalues and increasing the smaller ones of $\hat{\Sigma}_k(\lambda)$, thereby counteracting the bias of the estimates. In Vandev (2004) the covariance matrices are stabilized by one parameter α, i.e.,

$$\hat{\Sigma}_k(\alpha) = \alpha\hat{\Sigma}_k + (1-\alpha)\hat{\Sigma}\,.$$

This parameter $\alpha \in [0,1]$ corresponds to $(1-\lambda)$ above. The limiting cases correspond to LDA ($\alpha = 0$) and QDA ($\alpha = 1$). To determine the optimal value of this parameter α, the error rate estimation has to be minimized during the model building process. As error rate estimations often resubstitution, cross validation or simulation methods are used. The methods we have used are described in Section 3.

3 The MATLAB-Program "ldagui"

The MATLAB-program "ldagui" is described in detail in Vandev (2004). It can be used by means of menus, shortcuts and listboxes. The main window of the program shows Figure 1.

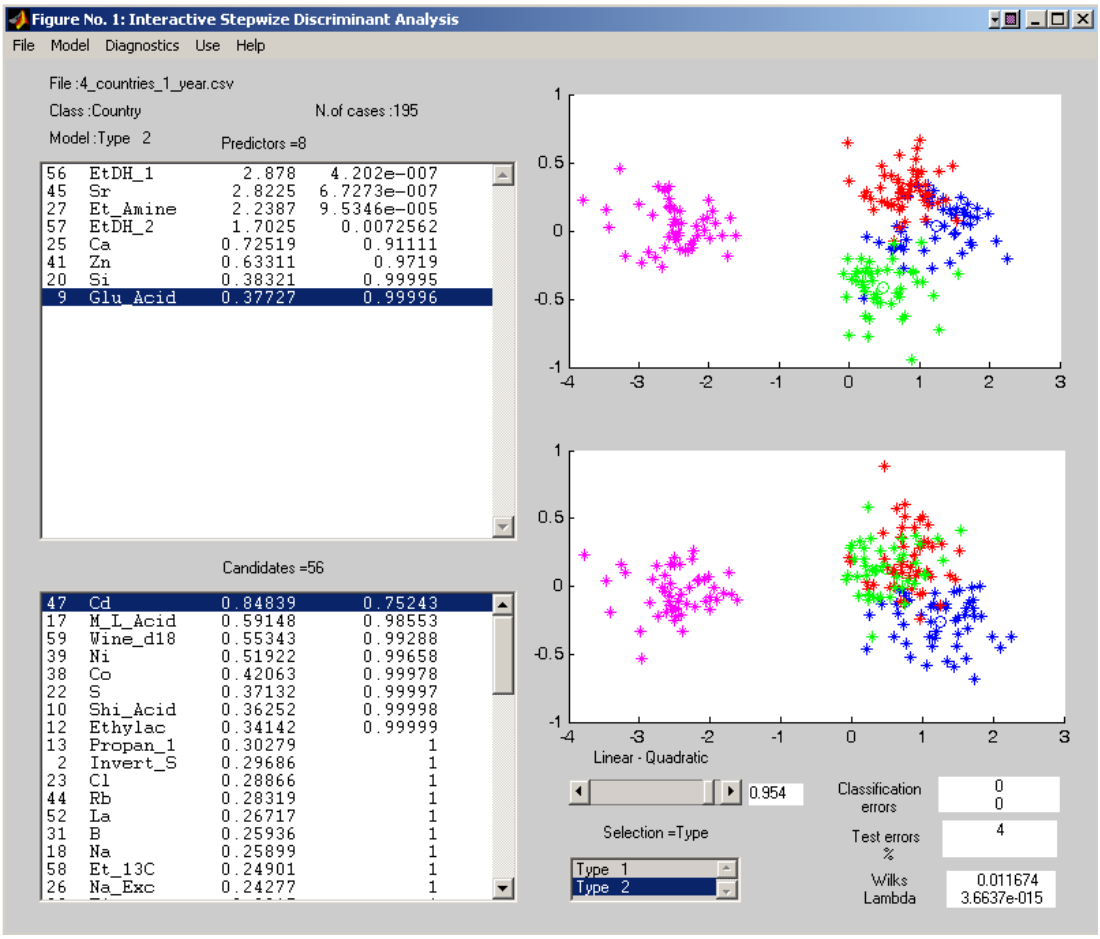

Figure 1: Main window of "ldagui"

3.1 Menus

Five menus **File**, **Model**, **Diagnostics**, **Use** and **Help** can be activated.

- In **File** a csv-data file can be loaded and by choosing a classification and selection variable missing data will be replaced with group means.
- **Model** allows to build interactively a model in dependence on a minimal classification and test error and an optimal choice of the regularization parameter $\alpha \in [0, 1]$.
- **Diagnostics** contains three tools for making adequate decisions:
  - Test: A small random test sample with 600 observations for each group will be produced according estimated group means and covariance matrices and will be classified.
  - "Leave-one-out" (LOO – special case of cross validation):
    Classical: For each observation in the training sample a model with the same variables will be built but without that particular observation. Then each removed observation will be classified with this model, all misclassifications are counted and the LOO-error will be estimated.
    Modification: Not only the one removed, but all observations from the training sample will be classified, all misclassifications are counted and LOO-error will be estimated.
  - Plot: Second and third canonical variables will be plotted against the first.
- In **Use** other (csv)-data files can be loaded for testing the model ("Hold-out" method).

More detailed results are printed in the MATLAB command window, e.g.

- Ordered variables in model with their F- and p-values,
- Wilk's Λ- and p-value,
- Results of error estimation of the training sample by methods of resubstitution, simulation (test and theoretical error) and cross validation (classical and modified LOO), including number and cases of misclassifications and cases classified with probability < 0.8. The theoretical error was estimated in the same way as the test error, but by using a large (6000 per group) simulated data sample and the LOO error was obtained as proportion of all errors to the size of training sample.

The algorithms are based on papers of Jennrich (1977) and Einslein et al. (1977).

4 Results of Applying RDA to Wine Data

4.1 Overview about Models for Commercial Wines

Several models for commercial wines obtained by using RDA are presented in Table 1. Here we have used the following strategy: At first we have looked for our „best" model (Model 1) by choosing the optimal parameter α manually so that the model has 0 or only a small number of classification and test errors. Then we have considered the same model for $\alpha = 0$ (LDA) and $\alpha = 1$ (QDA). In a next step we wanted to find a better linear and quadratic model and we have considered some other acceptable models for different α. Classification and prediction errors and misclassified samples will be given.

Table 1: Model results for commercial wines ($N = 195$)

	RDA-M. 1	LDA-M. 1	QDA-M. 1	LDA-M. 2	QDA-M. 2	RDA-M. 2	RDA-M. 3
Parameter α	0.7 0.95	0.0	1.0	0.0	1.0	0.8	0.8
No. of variables	8	8	8	14	9	7	11
Invert Sugar							•
Gluconic Acid	•	•	•	•	•		•
Shikimic Acid				•			•
2-Methylbutanol							•
Malic-L. Acid				•			
Sodium				•			
Silicon	•	•	•	•	•		
Calcium	•	•	•	•	•	•	
Ethanolamine	•	•	•	•	•	•	
Putrescine				•			•
Lithium							
Boron							•
Titanium				•			•
Cromium							•
Nickel							•
Copper							•
Zinc	•	•	•	•	•	•	
Strontium	•	•	•	•	•	•	
Cadmium				•	•		
Ethanol (D/H)1	•	•	•		•	•	
Ethanol (D/H)2	•	•	•	•	•	•	
Wine δ^{18}O				•		•	•
Class. error (Resubstitution)	0	6	1	0	0	2	1
(No. and %)	0	3.08	0.51	0	0	1.03	0.51
Incorrectly classified samples (ID-No.) (Resubstitution)		100006 100032 100068 100085 100114 100148	100060			100119 100120	100099
No. of cases with post. prob. < 0.8	16 13	22	12	5	10	10	18
Theor. error (%)	4.4 3.7	4.8	3.4	1.7	2.4	4.1	4.4
LOO error (class.)	10 7	9	9	5	6	10	21
(No. and %)	5.13 3.6	4.62	4.62	2.56	3.08	5.13	10.77
LOO error (modif.)	78 16	195	152	8	12	195	195
(No.)*	0.59 0.1	6.02	0.85	0.05	0.06	2.15	1.84
(No. and %)**	0.3 0.05	3.09	0.43	0.03	0.03	1.10	0.94

*No. of LOO-cases leading to one or more misclassifications of cases of the whole training sample

**LOO-mean error of misclassifications over the whole training sample

4.2 Description of RDA-Model 1

On the base of the print results of "ldagui" our preferred model (RDA-Model 1 for $\alpha = 0.95$) will be described in more detail. Table 2 contains the variables as result of interactive model building and Figure 2 illustrates this model.

- Wilk's Λ: 0.0117; p-value (tail): 0.0000

Table 2: Interactive model building (variables in model: 8)

No.	Name	F-value	p-value
55	Ethanol (D/H)1	2.9535	2.40E-07
44	Sr	2.8966	3.88E-07
26	Ethanolamine	2.2975	6.14E-05
56	Ethanol (D/H)2	1.7472	0.00531
24	Ca	0.7442	0.89076
40	Zn	0.6497	0.96337
19	Si	0.3933	0.99991
8	Gluconic Acid	0.3872	0.99993

- Method of error estimation: Resubstitution. No. of classification errors: 0
  Cases classified with probability below 0.8: 100006, 100016, 100019, 100020, 100027, 100030, 100060, 100074, 100085, 100115, 100120, 100140, 100143.

- Method of error estimation: Theoretical error by simulation (6000 per group)

Table 3: Classification matrix

	% Correct	Hungary	Romania	Czech Rep.	South Africa	Total
Hungary	94.73	**5684**	18	298	0	6000
Romania	94.87	62	**5692**	246	0	6000
Czech Rep.	95.67	212	48	**5740**	0	6000
South Africa	100.00	0	0	0	**6000**	6000
Total	**96.32**	5958	5758	6284	6000	

Rows: Observed classifications, Columns: Predicted classifications

- Method of error estimation:
  1. LOO (classical) error (No. and %): 7; 3.59%
  Misclassified LOO-cases (ID-No.): 100020, 100030, 100060, 100068, 100085, 100115, 100120.

  2. LOO (modif.) mean error (No. and %): 0.097; 0.05%
  No. of LOO-cases, which lead to misclassifications: 16
  Two misclassified cases for leaving out case (ID-No.): 100030, 100039, 100050.
  One misclassified case for leaving out case (ID-No.): 100013, 100020, 100049, 100058, 100060, 100061, 100062, 100068, 100072, 100085, 100103, 100115, 100120.

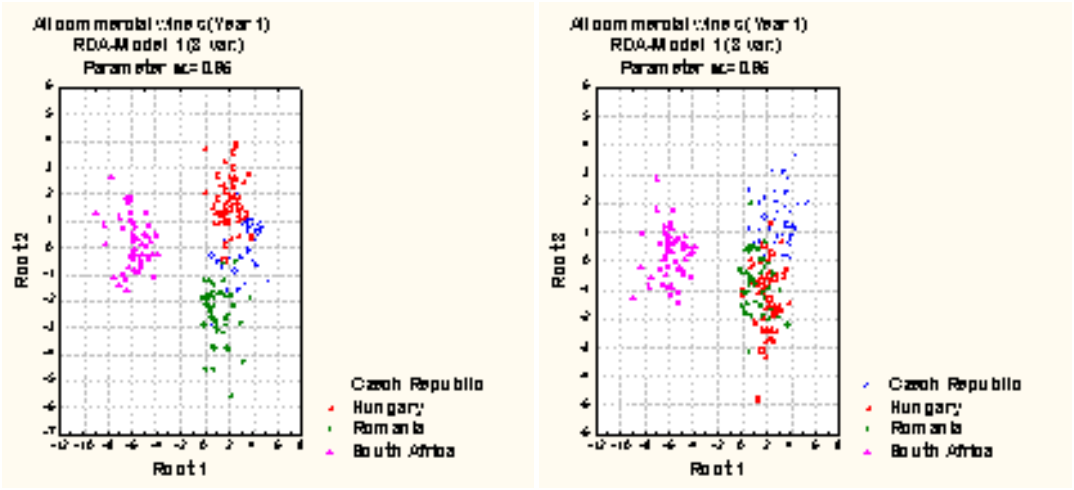

Figure 2: Discriminating plots for commercial wines concerning the 4 countries (RDA).

4.3 Models for White and Red Wines

Table 4 contains our preferred RDA-models for white and red wines. In both cases only six variables were selected as being important to separate the four countries. By simulating 6000 samples per group very small "theoretical" error rates could be obtained.

5 Conclusions

The classical methods of discriminant analysis are suitable for distinguishing wines from different countries. Discriminant models containing most important parameters and allowing minimal misclassification rates can be given. Particularly, methods of regularized discriminant analysis led to good results in our case of investigating commercial wines.

Using our preferred model 1 of RDA for all commercial wines, which is much better than the corresponding one of LDA and comparable with that of QDA, all 195 wines could be classified correctly by resubstitution method. Wilk's Λ near 0 shows a high discriminating power of the chosen model. Only 13 wines were classified with posterior probability < 0.8. By simulating 6000 wine samples per country a "theoretical" correct classification rate of 96.32% could be obtained. Using "Leave-One-Out" method led to correct classification rates between 96.4% (classical LOO) and 99.95% (modified LOO).

The eight most important variables are: the isotopic ratios **Ethanol (D/H)1** and **Ethanol (D/H)2**, the trace elements **Strontium** and **Zinc**, the macroelements **Calcium** and **Silicon** and the biogenic amine **Ethanolamine** and the classical parameter **Gluconic Acid**. Figure 2 shows the well separation of the countries by model 1.

As expected the South African wines could be separated very easily from those of the other countries. Only the isotopic ratios could be identified as being important and sufficient parameters in the discriminant model.

Considering only white respectively red commercial wines, RDA-models with six variables led to very good results of discriminating the wines of the four countries.

Table 4: Model results for white and red commercial wines.

	White wines ($N = 136$)	Red wines ($N = 59$)	
	RDA-Model 1	RDA-Model 1	RDA-Model 2
Parameter α	0.9	0.9	0.7
No. of Variables	6	6	5
Malic-L. Acid	•	•	•
Ca	•		
Ethanolamine	•		
Li		•	•
B		•	•
Al		•	
Ti			•
Cu		•	
Sr	•		
Ethanol (D/H)1		•	•
Ethanol (D/H)2	•		
Wine δ^{18}O	•		
No. Class. error (Resubstitution)	0	0	0
No. of cases with post. prob. < 0.8	2	2	2
Theor. error (%)	2.4	1.5	2.4
LOO (class.)	4	6	5
(No. and %)	2.94	10.17	8.46
LOO (modif.)	6	6	7
(No.)*	0.04	0.11	0.11
(No. and %)**	0.03	0.20	0.20

*No. of LOO-cases which lead to one or more misclassifications of cases of the whole training sample
**LOO-mean error of misclassifications over the whole training sample

Acknowledgements

Project Steering Committee: R. Wittkowksi BfR, Germany, P. Brereton CSL, United Kingdom, E. Jamin Eurofins, France, X. Capron VUB, Belgium, C. Guillou JRC, Italy, M. Forina UGOA, Italy, U. Römisch TUB, Germany, V. Cotea UIASI.VPWT.LO, Romania, E. Kocsi NIWQ, Hungary, R. Schoula CTL, Czech Republic.

References

Einslein, K., Ralston, A., and Wilf, H. S. (1977). *Statistical Methods for Digital Computers*. New York: J. Wiley & Sons.

Fahrmeir, L., Hamerle, A., and Tutz, G. (1996). *Multivariate statistische Verfahren*. Berlin: W. deGruyter.

Friedman, J. H. (1989). Regularized discriminant analysis. *Journal of the American Statistical Association*, *84*, 165-175.

Jennrich, R. I. (1977). Stepwise discriminant analysis. In K. Einslein, A. Ralston, and

H. S. Wilf (Eds.), *Statistical Methods for Digital Computers* (p. 76-95). New York: J. Wiley & Sons.

McLachlan, G. J. (1992). *Discriminant Analysis and Statistical Pattern Recognition*. New York: J. Wiley & Sons.

Römisch, U., Vandev, D., Klimmek, A., and Wittkowski, R. (2001). *Determination of the Geographical Origin of Wines from East European Countries by Methods of Multivariate Data Analysis.* (Proceedings of the ROeS Seminar, Mayrhofen/Austria, 24.-27.09.2001)

Vandev, D. (2004). Interactive stepwise discriminant analysis in MATLAB. *Pliska Stud. Math. Bulg.*, *16*, 291-298.

Authors' addresses:

Ute Römisch and Katrin Zur
Faculty of Process Engineering
Department of Informatics
Technical University Berlin
Gustav-Meyer-Allee 25
D-13355 Berlin

E-mail : ute.roemisch@tu-berlin.de
Homepage: http://www.tu-berlin.de/fak3/staff/roemisch/homepage1.html

Prof. Dimitar Vandev passed away in Sept., 2004. This joined work is dedicated to him.