PLISKA
STUDIA MATHEMATICA BULGARICA
ПЛИСКА
БЪЛГАРСКИ МАТЕМАТИЧЕСКИ СТУДИИ

For further information on
Pliska Studia Mathematica Bulgarica
visit the website of the journal http://www.math.bas.bg/~pliska/
or contact: Editorial Office
Pliska Studia Mathematica Bulgarica
Institute of Mathematics and Informatics
Bulgarian Academy of Sciences
Telephone: (+359-2)9792818, FAX:(+359-2)971-36-49
e-mail: pliska@math.bas.bg

Pliska Stud. Math. Bulgar. **18** (2007), 327–339

PLISKA
STUDIA MATHEMATICA
BULGARICA

APPLICATION OF REGULARIZED DISCRIMINANT ANALYSIS

Ute Roemisch Henry Jäger Dimitar Vandev

The method of regularized discriminant analysis (RDA) was used for identifying the geographical origin of wines on the base of chemical-analytical parameters in the scope of a European project "WINE DB"1. A data base with 63 measured parameters of 250 authentic wine samples from five countries of the vintage 2003 was taken as a basis for classifying and discriminating wines. Uni- and multivariate methods of data analysis were applied. By using a Matlab-program, which allows an interactive stepwise discriminant model building, some different models for authentic wines with corresponding classification and prediction error rates (resubstitution, classical and modified "Leave-one-out", simulation and test) will be presented. The goodness of our preferred model was analysed by classifying a test sample that was created by splitting the data set based on Duplex-algorithm of Snee. 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. Rmisch TUB, Germany, V. Cotea UIASI.VPWT.LO, Romania, E. Kocsi NIWQ, Hungary, R. Schoula CTL, Czech Republic.

2000 *Mathematics Subject Classification:* 62H30, 62P99

Key words: Discrimination of wines; Regularization; Classification.

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. Roemisch TUB, Germany, V. Cotea UIASI.VPWT.LO, Romania, E. Kocsi NIWQ, Hungary, R. Schoula CTL, Czech Republic.

The method of regularized discriminant analysis (RDA) was used for identifying the geographical origin of wines on the base of chemical-analytical parameters in the scope of a European project "WINE DB". A data base with 63 measured parameters of 250 authentic wine samples from five countries of the vintage 2003 was taken as a basis for classifying and discriminating wines. Uni- and multivariate methods of data analysis were applied. By using a Matlab-program, which allows an interactive stepwise discriminant model building, some different models for authentic wines with corresponding classification and prediction error rates (resubstitution, classical and modified "Leave-one-out", simulation and test) will be presented. The goodness of our preferred model was analysed by classifying a test sample that was created by splitting the data set based on Duplex-algorithm of Snee.

1. Introduction

The determination of the geographical origin of wines is very important for identifying wines, which come not up to European quality standards. That's why a wine data base, containing 600 authentic and 600 commercial wines from Hungary, Czech Republic, Romania and South Africa, was built over a period of three years (2001-2004) in the scope of a European project. The second-year-data could be extended by fifty authentic wine samples from Australia.

The sampling strategy for collecting wines was to obtain a statistical sample that is proportional to the production of wines and that is representative for the wine regions and for the wine varieties. For each sample 63 chemical parameters were considered.

The statistical data analysis was starting with: Data Management (data control, data handling of missing and censored data, log-transformations of 90% of the data and identification of uni- and multivariate outliers), Descriptive Statistics and Analysis of Correlations, One- and Multifactor-Variance Analyses and Principal Component Analyses. Analysing these results, the data set could be reduced to 244 authentic wines. Then multivariate classification and projection methods as Cluster Analyses, Projection Pursuit methods, Partial Least Square methods (PLS-UVE), Classification and Regression Trees (CART), Class modelling techniques (SIMCA) as well as Linear, Quadratic and Regularized Discriminant analyses were used for classifying and discriminating the wines of the different countries.

After presenting first results of applying regularized discriminant analysis for commercial wine data of the first year in Roemisch, et al. (2006), now results for authentic wines of the second year will be presented, including some improvements of the used Matlab-program.

2. Regularized Discriminant Analysis

Mc Lachlan (1992) and Fahrmeir, et al. (1996) give an overview about methods of discriminant analysis. These methods 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 analysed.

Let the categorical variable Y denote the group membership of the object, where $Y = k$ implies that it belongs to the group with index $k (k = 1, , K)$. Each object is characterized by the p-dimensional feature vector X.

Let $P(Y = k) = p_k,\ k = 1, \ldots, K$, be the prior probabilities, that an object belongs to the group with index k and $f(x|k),\ k = 1, \ldots, K$, be the conditional distribution density of X given for $Y = k$. The distribution of X is then

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

For classification problems the posterior probability $p(k|x)$, i.e. the probability, that an object with observed feature vector x belongs to the k^{th} group, is very important. According to the formula of Bayes this conditional probability of Y given by $X = x$ is

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

The allocation rule of Bayes, which achieves minimal misclassification risk among all possible rules, can be derived

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

For the special case that $p_k = p\ \forall k$, the Maximum Likelihood allocation rule is used

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

That means, an object with feature vector x will be assigned to that group with index $\hat{k}$, which has the largest posterior probability.

These allocation rules have the general structure

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

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

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

Let us assume normality for the p-dimensional feature vector X_k in the k^{th} group

$$X_k \sim N(\mu_k, \Sigma_k), \qquad k = 1, \ldots, K,$$

where μ_k denote the group means and Σ_k the group covariance matrices.

Then the conditional distribution of X given for $Y = k$ can be described by the density of the normal distribution

$$f(x|k) = (2\pi)^{-p/2}|\Sigma_k|^{-1/2}\exp\{-\frac{1}{2}(x-\mu)'\Sigma_k^{-1}(x-\mu)\},\ k = 1, \ldots, K.$$

Substituting equation (7) into $d_k(x) = f(x|k)p_k$ (see (3) and (5)) and taking the logarithm leads to the discriminant function of the form

$$d_k(x) = -\frac{1}{2}[(x-\mu_k)'\Sigma_k^{-1}(x-\mu_k) + \ln|\Sigma_k|] + \ln p_k, \qquad k = 1, \ldots, K.$$

Using allocation rule (5) with equation (8) 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.

If the group covariance matrices are identical, i.e., $\Sigma_k = \Sigma\ \forall k,\ (k = 1, \ldots, K)$, the Linear Discriminant Analysis (LDA) can be used, because the rule that minimizes the misclassification risk leads to a linear separation of the groups.

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

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

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.

In the case that n is less than or comparable to p, the estimate of Σ_k should be regularized further by a second parameter γ

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

where I_p is the $p \times p$ identity matrix, and $c_k = (tr\hat{\Sigma}_k(\lambda))/p$. 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.

Vandev (2004) has stabilized the covariance matrices by only one parameter α, which corresponds to $(1 - \lambda)$ of Friedman

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

For the case of $(\alpha = 0)$ the RDA corresponds to LDA and for the case of $(\alpha = 1)$ to QDA. To determine the optimal value of the parameter, the error rate estimation has to be minimized during the model building process. Our preferred methods of error estimation are described in section 3.

3. The Matlab-program "ldagui"

The first version of the Matlab-program "ldagui" is described in detail in Vandev (2004). During the process of applying the program to the wine data improvements were necessary, which can be found in Rmisch et al. (2006). Mateev (2006) has continued improving the user-convenience and has supplemented some print results of the error estimations.

In the main window (figure 1) of the program five menus: *File, Model, Diagnostics, Use* and *Help* can be activated. A model can be built interactively in dependence on a minimal classification (resubstitution) and simulation error (simulation of a small test sample with 600 samples by group) and an optimal choice of the regularization parameter $\alpha \in [0, 1]$. The second and the third canonical variables can be plotted against the first.

More detailed results can be printed in the Matlab- command window, such as results of different error rate estimations.

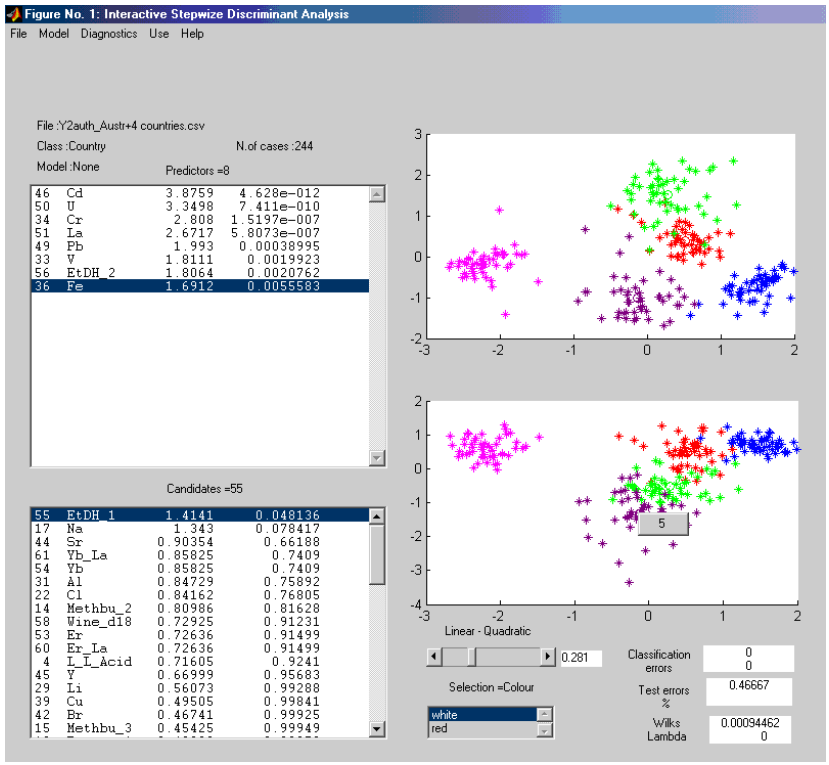

Figure 1: Main window of "ldagui"

The following methods of error rate estimation can be used:

∘ **Resubstitution ↦ Classification error (classification table)** Misclassified samples are counted and identified (ID-No.) and the classification error will be estimated. Cases classified with posterior probability ¡0.8 are given.

∘ **Cross validation ↦ "Leave-one-out"-error**

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 identified 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 identified and LOO-error will be estimated.

∘ **Simulation ↦ Simulation error (classification table)** In the main window

a small test random sample with 600 observations for each group will be produced according estimated group means and covariance matrices and will be classified. Error rates are given. Together with resubstitution error this simulation error is used as quick error estimation method for model building. In a second step a greater random sample with 6000 observations for each group can be produced and classified in an analogous way and then misclassifications for each country can be given.

∘ **Test ↦ Test error (classification table)** The wine data set will be split by Duplex-algorithm of Snee (1977) into a learning (2/3 of the data) and a test (1/3 of the data) sample. Then models are built based on the learning sample and objects of the test sample are classified. A test error will be estimated and misclassified samples are identified.

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

4. RDA-results for authentic wine data

4.1. Models for authentic wines

Several RDA-models for all authentic wines are presented in table 1. We have used the following strategy: At first we have looked for our "best" RDA-model (model 1) by choosing the optimal parameter α manually so that the model has 0 or only a small number of classification and simulation (small test sample) errors. Then we have considered the same model for $\alpha = 0$ (LDA) and $\alpha = 1$ (QDA). In a next step we have tried to find a better linear and quadratic model and at last we wanted to find some other acceptable RDA-models, containing also different variables than model 1 for different α .Classification and prediction errors and misclassified samples will be given.

4.2. Description of RDA-Model 1 ($\alpha = 0.28$)

The variables of our preferred RDA-model for $\alpha = 0.28$ as a result of our interactive stepwise model building contains table 2 and figure 2 illustrates this model. The different error rate estimations can be compared. Wilk's λ near 0 shows a high discriminating power of the chosen model.

Wilk's λ, P-value(tail): 0,00094 0,000000
Method of error estimation: Resubstitution
No. of classification errors: 0
No. of cases classified with probability below 0.8: 5
ID-No. of the sample: 100792, 100793, 100808, 100827 and 100828.
Method of error estimation: "Simulation error" (6000 per group)

The diagonal of the classification matrix contains the correct classified simulated wines.
Method of error estimation: Leave-One-Out (LOO)
1. LOO (classical) error [No. and %]: 3 ; 1.23 % (ID-No.: 100827, 100828 and 100213)
2. LOO (modif.) means error [No. and %]: 0.02 ; 0.084 %
No. of LOO-cases which lead to one misclassification: 5 ID-No.: 100808, 100813, 100827, 100828, 100213.

4.3. Error estimation with split data

The whole wine data set was split into two independent data sets, the learning and the test data set by using the Duplex-algorithm of Snee (1977). The learning data set containing 185 authentic samples was used for building the RDA models, whereas the test data set consisting of 59 authentic wines was used for estimating an unbiased error rate as result of this classification and for testing the predictive ability of the RDA-models.

The results of classifying test data are summarized as "Test error" in table 4. The "best" RDA-Model 1 we had found based on the whole data set, proved to be again an excellent model based on the learning data set (see the corresponding error rates). Testing the goodness of the model by the test sample didn't lead to any misclassifications of that data. Some other comparable good models as former QDA-Model 2 and RDA-Model 5, but also a new RDA-Model 8 can be found in table 8.

For the group of white and red wines in the same way models with corresponding error rates can be found.

5. Conclusions

- Using regularized discriminant analysis (RDA) interactively by determining the optimal value of the parameter α for minimal error rate estimates is a successful strategy to obtain good models for discriminating the wines of the five countries. For all authentic wines we could find acceptable models with 7–9 variables. For the group of authentic white wines models with 5–8 variables can be given and for the group of authentic red wines models with 4–5 variables.

- Our preferred RDA-Model 1 showed a high stability and very small error rates in comparison of all different methods of error estimation we have used.

	RDA-M. 1+2	LDA-M. 1	QDA-M. 1	LDA-M. 2	QDA-M. 2	RDA-M. 3+4	RDA-M. 5	RDA-M. 6+7
Parameter	0.28	0.0	1.0	0.0	1.0	0,28	0.8	0,5
No. of Vars	8	8	8	10	8	8	7	9
Tartaric Acid								•
Sodium	•			•		•		•
Silicon					•		•	
Chlorine								•
Potassium						•		
Vanadium	• •	•	•	•	•	•		• •
Chromium	• •	•	•	•	•	• •	•	• •
Iron	• •	•	•	•	•	•		• •
Copper						•		
Strontium				•	•	•		•
Arsenic								
Cadmium	• •	•	•	•	•	• •	•	• •
Lead	•	•	•	•	•	• •	•	• •
Uranium	• •	•	•	•		• •	•	• •
Lanthanum	• •	•	•	•		• •	•	•
Ethanol D/H-1					•		•	
Ethanol D/H-2	• •	•	•	•				•
Class. error - Resubstitution (No. and %)	0 0	3 1.2	3 1.2	0 0	0 0	0 0	0 0	0 0
Incorr. class. samples (ID-No.)		100827 100828 100214	100861 100872 100827					
No. of cases with post. prob. ≤ 0.8	5 6	9	5	5	1	4 5	3	1 2
Theor. error (%)	0.7 1.1	0.6	0.4	0.3	0.9	0.7 1.0	0.7	0.5 0.4
Leave-one-Out error(classical)	3 4	8	11	5	5	4 7	5	3 5
No. and %	1.23 1.6	3.3	4.5	2.05	2.05	1.64 2.87	2.05	1.23 2.05
LOO error (modif.)	5 22	244	244	10	5	5 17	11	5 5
No.*	0.02 0.09	3.0	3.0	0.05	0.02	0.02 0.07	0.05	0.02 0.02
No. and % **	0.01 0.04	1.2	1.2	0.02	0.008	0.01 0.03	0.02	0.01 0.01

* No. of Leave-one-Out -cases which lead to one or more misclassifications of cases from the whole training sample.

** Leave-one-Out -mean error of misclassifications over the whole training sample.

Table 1: Model results for authentic wines (N=244)

No.	Name	F-value	p-value
47	Cd	3.7845	9.2384e-012
51	U	3.2709	1.3776e-009
35	Cr	2.7418	2.5932e-007
52	La	2.6087	9.6744e-007
50	Pb	1.946	0.00056522
34	V	1.7684	0.0027626
57	EtDH2	1.7638	0.0028756
37	Fe	1.6514	0.0074707

Table 2: Interactive model building (RDA-Model 1 with 8 variables in model)

	Correct (%)	CR	HU	RO	SA	AU	Total
Czech Rep.	99.17	5950	50	0	0	0	6000
Hungary	98.68	9	5921	60	0	10	6000
Romania	98.80	0	67	5928	0	5	6000
S. Africa	100.00	0	0	0	6000	0	6000
Australia	99.90	0	4	1	1	5994	6000
Total	99,31	5905	6037	5994	6001	6013	

Rows: Observed classifications;
Columns: Predicted classifications.

Table 3: Classification Matrix

	RDA-M. 1***	QDA-M. 2	RDA-M. 5	RDA-M. 8
Parameter	0.28	1.0	0,8	0.58
No. of Vars	8	8	7	8
Silicon		•	•	•
Chlorine				•
Vanadium	•	•		•
Chromium	•	•	•	•
Iron	•	•		
Strontium		•		•
Cadmium	•	•	•	•
Lead	•	•	•	•
Uranium	•		•	
Lanthanum	•		•	
Ethanol D/H-1		•	•	
Ethanol D/H-2	•			•
Class. error - Resubstitution (No. and %)	1 0.54	0 0	1 0.54	0 0
Incorr. class. samples (ID-No.)	100827		100792	
No. of cases with post. prob. ≤ 0.8	2	1	1	6
Theor. error (%)	0.60	0.66	0.56	1.05
Leave-one-Out error(classical) No. and %	3 1.62	9 4.86	3 1.62	5 2.70
LOO error (modif.) No.* No. and % **	173 0.94 0.51	26 0.15 0.08	184 1.0 0.54	18 0.11 0.06
Test error No. and %	0 0	1 1.69	1 1.69	2 3.39

* No. of Leave-one-Out-cases which lead to one or more misclassifications of cases from the whole training sample.
** Leave-one-Out-mean error of misclassifications over the whole training sample.
*** Best model from the discrimination analyses of the complete data set

Table 4: Model results for authentic wines (Split data set: learning data with N=185, test data with N = 59)

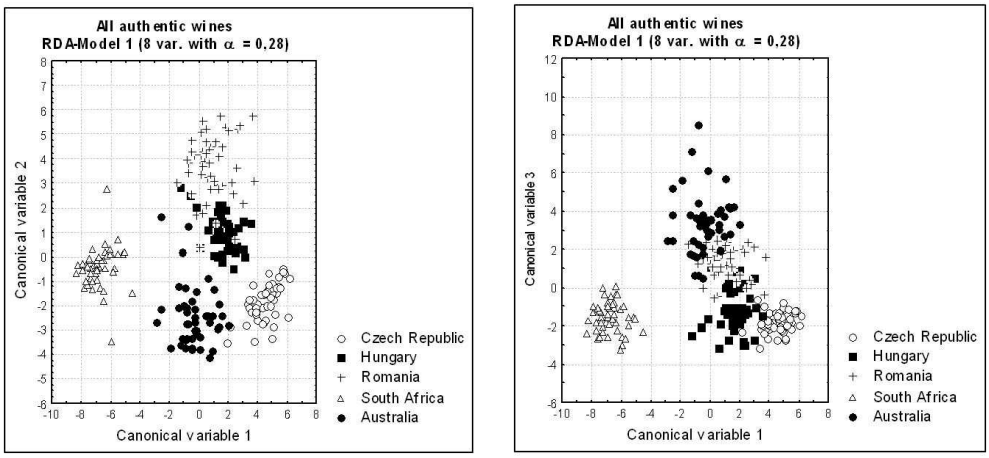

Figure 2: Discriminating plots for authentic wines concerning the 5 countries (RDA)

- The following variables had a very high or high (in brackets) discriminating power:
  - for all authentic wines: (Na), V, Cr, Fe, (Sr), Cd, Pb, U, La, (Ethanol D/H 2 or Ethanol D/H 1);
  - for authentic white wines: V, (Cr), Fe, Cd, (Pb), U, La, (Wine_d18);
  - for authentic red wines: V, (Fe), Cd, (Pb, U), EtDH_2;
- The discrimination of South African and Australian wines was not difficult and also Czech and Romanian wines could be separated very well. The simulation results and the graphics show some overlap of wines between Hungary and Czech Republic and Hungary and Romania.

REFERENCES

[1] K. Einslein, A. Ralston, H. S. Wilf. Statistical Methods for Digital Computers. J. Wiley & Sons, New York, 1977.

[2] L. Fahrmeir, A. Hamerle, G. Tutz. Multivariate statistische Verfahren. W. de Gruyter, Berlin, 1996

[3] J. H. Friedman. Regularized discriminant analysis. J. Amer. Statist. Assoc. 84 (1989), 165-175

[4] R. I. Jennrich. Stepwise Discriminant Analysis. In K. Einslein, A. Ralston and H.S. Wilf (Eds.), Statistical Methods for Digital Computers, pages 76-95, J. Wiley & Sons, New York, 1977.

[5] G. J. McLachlan. Discriminant Analysis and Statistical Pattern Recognition. J. Wiley & Sons, New York, 1992.

[6] P. Mateev. Error rate estimations in the Matlab-program "ldagui", Manuscript, Berlin, 2006.

[7] U. Roemisch, D. Vandev, A. Klimmek, R. Wittkowski. Determination of the Geographical Origin of Wines from East European Countries by Methods of Multivariate Data Analysis. *Proc. of the RoeS Sem.* Mayrhofen, 24-27.09.01.

[8] U. Roemisch, D. Vandev, K. Zur. Application of Interactive Regularized Discriminant Analysis to Wine Data *Austr. J. of Stat.* **351** (2006), 45–55.

[9] R. D. Snee. Validation of regression models: method and examples, *Technometrics* **19** (1977), 415–428.

[10] D. Vandev. Interactive Stepwise Discriminant Analysis in MATLAB. *Pliska Stud. Math. Bulg.* **16** (2004), 291–298.

[11] D. Vandev, U. Roemisch. Comparing several Methods of Discriminant Analysis on the Case of Wine Data. *Pliska Stud. Math. Bulg.* **16** (2004), 299–308.

Dr. Ute Roemisch and Henry Jaeger
Faculty of Process Engineering
Department of Informatics
Technical University Berlin
Gustav-Meyer-Allee 25 D- 13355
Berlin, Germany
e-mail: ute.roemisch@tu-berlin.de
http://www.tu-berlin.de/fak3/staff/roemisch/homepage1.html