Prediction Optimal Classification of Business Phases

K. Luebke * C. Weihs

October 2005

Universität Dortmund
Fachbereich Statistik

Abstract

Linear Discriminant Analysis (LDA) performs well for classification of business phases – even though the premises of an LDA are not met. As the variables are highly correlated there are numerical as well as interpretational shortcomings. By transforming the classification problem to a regression setting both problems can be addressed by a computer-intensive prediction oriented method which also improves the classification performance.

1 Introduction

A vast amount of methods have been developed for classification. Despite its age the Linear Discriminant Analysis developed by R.A. Fisher in 1936 does perform well even in situations, where the underlying premises like normally distributed data with constant covariance matrices over all classes are not met. So it is not exceptional for the problem at hand, classification of business phases, that Weihs and Garczarek (2002) show the good performance of LDA.

*e-mail: luebke@statistik.uni-dortmund.de

As LDA, however, includes matrix inversion it may run into problems in high dimensional situations or when the variables are highly correlated like macroeconomic variables. To overcome this problem, Hastie et.al. (1995) utilize the fact that LDA is equivalent to canonical correlation analysis and optimal scoring and therefore can be transformed into a regression problem. They use a penalty term for the within-covariance matrix as well as smoothing of the estimates. As such a penalty term is not directly linked to the classification problem, Luebke and Weihs (2003) propose a projection on latent factors optimized for classification. By using latent factors, Luebke and Weihs (2004a) also obtain better predictions than e.g. Partial Least Squares in a linear regression model. In the present paper, both ideas are combined to obtain a Prediction Optimal Classification (POC) criterion to evaluate projection matrices for predictive classification.

The optimal solution, i.e. the corresponding projection matrix, is found by Simulated Annealing, the flexibility of which allows to include cost terms for deviations of estimators from zero so variable selection or measurement of importance can be included in the method.

The paper is organized as follows: In the next section we introduce the underlying scoring function of the classification problem which is minimized by Simulated Annealing. In section 3 the implementation of the Simulated Annealing Algorithm is described. A performance criterion for time related data is proposed in section 4. The data and the results of the new method are shown section 5. After that an outlook on future work will be given as well as some concluding remarks.

2 Optimal Scoring with Latent Factors

Linear Discriminant Analysis is a statistical method for classification. In LDA the classification is based on the calculation of the posterior probabilities of a trial point. The class with the highest posterior probability is chosen. To calculate the posterior probability it is assumed that the data comes from a multivariate normal distribution where the classes share a common covariance matrix but have different mean vectors. Hastie et.al. (1995) show that LDA is equivalent to canonical correlation analysis and optimal scoring. So LDA can be seen as a special linear regression. One of the problems in LDA is that the estimated covariance matrix of the data points has to be inverted for the classification. Especially in a high dimensional problem or when the

variables are highly correlated this can cause numerical problems. In the paper of Hastie et.al. (1995) they try to overcome possible numerical problems by using a penalty term. This may not be optimal as the covariance matrix is transformed away from singularity without using information in the data. In the new method the data is first projected on (few) latent factors guaranteeing the invertibility of the optimal scoring matrices.

Assume that there are n observations with p variables in the predictor space and k classes. Let

- $X \in I\!R^{n \times p}$: Predictor variables.
- $Y \in I\!R^{n \times k}$: Indicator matrix of the classes.

The basic idea is as follows: Assign $l \leq k-1$ scores to the classes and regress these scores on X. We are looking for scores (of the k classes) and a suitable regression of these scores on the predictor variables so that the residuals are small for the true class and large for the wrong. The average squared residual function is:

$$ASR(H, M) = \frac{1}{n}\|YH - XM\|^2, \tag{1}$$

where

- $H \in I\!R^{k \times l}$ is the score matrix of the classes,
- $M \in I\!R^{p \times l}$ is the regression parameter matrix, and
- $\|\cdot\|^2$ is the Frobenius matrix norm.

To avoid trivial solutions the constraint

$$H'(Y'Y/n)H = I_l \tag{2}$$

is used.
To tackle the numerical problems in calculating M so called latent factors are derived. Latent factors are linear combinations of the original predictor variables $Z = XG$ with $G \in I\!R^{p \times l}, l < p$. It can be shown that in a latent factor model an optimal solution can be found with at most l latent factors if l is the number of response variables (Luebke and Weihs, 2004b).

In order to avoid numerical problems these latent factors must fulfill the side-condition that they are orthonormal, i.e.

$$Z'Z = (XG)'(XG) = I_l. \tag{3}$$

In the regression context latent factors (i.e. Reduced Rank Regression) turned out to be quite an improvement over the ordinary least squares regression (see for example Frank and Friedman (1993)).

With latent factors Z instead of the original X the ordinary least squares estimator of the regression coefficient $\tilde{M}$ on Z is:

$$\hat{\tilde{M}} = (Z'Z)^{-1}Z'YH = Z'YH. \tag{4}$$

Note that we made use of the side-condition (3) guaranteeing the invertibility of $(Z'Z)$. With $Z = XG$ we would like to have $XM = Z\tilde{M}$ and thus the estimator of M is $\hat{M} = G\hat{\tilde{M}}$. So equation (1) leads to:

$$ASR(H, G) = \frac{1}{n}\|YH - X(G(XG)'YH)\|^2. \tag{5}$$

As in general $\|AB\| \neq \|A\|\,\|B\|$ it is necessary in minimizing (5) to optimize G and H together.

Suppose that the regression matrix M and the score matrix H are estimated on a training set X, Y and that it is crucial how these estimators will perform on n_0 future values X_0. The point prediction of the future response values is:

$$\widehat{Y_0H}_{X,Y} = X_0\hat{M}_{X,Y}. \tag{6}$$

With a known test set X_0, Y_0 the loss in n_0 (new) observations can be measured by

$$L = \frac{1}{n_0}\|Y_0\hat{H} - \widehat{Y_0H}\|^2. \tag{7}$$

Note that instead of the Frobenius norm also a more robust norm can be used.

Usually one is not only interested in the performance of the estimator for some observations but also in the "general" or average performance. The corresponding mean loss (mean squared error of prediction) is defined as:

$$\begin{aligned} MSEP &= \frac{1}{n_0}E_{Y|X}E_{Y_0|X_0}\|Y_0\hat{H} - \widehat{Y_0H}\|^2 \\ &= \frac{1}{n_0}E_{Y|X}E_{Y_0|X_0}\|Y_0\hat{H}_{X,Y} - X_0\hat{M}_{X,Y}\|^2 \\ &= \frac{1}{n_0}E_{Y|X}E_{Y_0|X_0}\|Y_0\hat{H}_{X,Y} - X_0(\hat{G}_{X,Y}\hat{G}'_{X,Y}X'Y)\hat{H}_{X,Y}\|^2 \end{aligned} \tag{8}$$

Equation (8) shows that the *MSEP* can be seen again as a function of the projection matrix G and scoring matrix H. So by using different G and H – and taking care of the side-condition – different *MSEP* can be achieved. Estimation of *MSEP* can be done by bootstrap methods or cross-validation.

After the calculation of G and H the classification can then take place by means of the linear map of the data X:

$$\eta(x) = XM, \quad \eta(X) \in R^{n \times l} \tag{9}$$

Let $\bar{\eta}^k$ be the mean of the linear map of observations from class k. Then the assigning of observations is obtained by

$$\hat{c} = \arg\min \sum_{i=1}^{l} w_i (\eta(x)_i - \bar{\eta}_i^k)^2, \tag{10}$$

where η_i is the i-th column of η and w_i is the weight corresponding to the i-th dimension of the linear map space. If different a-priori probabilities of the classes are given, equation (10) is adapted, for example by subtracting $-2log\pi_k$ with π_k as the a-priori class probability.
Hastie et.al. (1995) show that if the weight is calculated as

$$w_i = \frac{1}{r_i^2(1 - r_i^2)} \tag{11}$$

with r_i^2 being the mean squared residual of the i-th optimally scored fit, then the distance in (10) is proportional to the Mahalanobis distance in the original feature space X.

3 Optimizing Prediction Criteria in the Latent Factor Model

As described in the previous section the objective functions ASR (5) and therefore L (7) and *MSEP* (8) can be described as functions of G and H, with G fulfilling the side-condition (3) and H the side-condition (2). In Prediction Optimal Classification one of these cost functions is minimized by applying a simulated annealing (SA) algorithm to the vectorized projection and scoring matrices $(vec(G)', vec(H)')'$. The choice of the cost function depends on the problem: If (5) is used the prediction purpose is neglected, with (8) a general

prediction purpose is evaluated but a time related structure in the data is not used which can be incorporated in (7).

Simulated annealing was already applied successfully to a Reduced Rank Regression in order to get optimal predictions (Luebke and Weihs, 2004a). To fulfill the side-conditions, new trial points are adapted to the requirements of the side-condition. This is done by a QR decomposition of the appropriate matrix product. (In a QR decomposition a matrix A is decomposed into $A = QR$ with Q being orthonormal and R a triangle matrix.) As with the QR decomposition the image space is unchanged and just an orthonormal basis (Q) is constructed, it is a suitable tool for the given problem. Thus in order to fulfill (2) a new trial point $\tilde{H}$ (generated by the transition function in the SA algorithm) is updated to a trial point H by

$$(Y'Y/n)^{\frac{1}{2}}\tilde{H} = Q_Y R_Y \quad (12)$$

$$H = \tilde{H}R_Y^{-1}, \quad (13)$$

The same procedure is applied to a trial point $\tilde{G}$ to fulfill (3).

The implementation of the SA algorithm is based on the one described in Press et.al. (1992). So the algorithm is a stochastic version of the well-known Nelder-Mead Simplex method (Nelder and Mead, 1965). The Nelder-Mead (or downhill simplex) method is transforming a simplex of $m+1$ points in an m dimensional problem. The functional values are calculated and the worst point is reflected through the opposite face of the simplex. If this trial point is best the new simplex is expanded further out. If the function value is worse than the second highest point the simplex is contracted. If no improvement at all is found the simplex is shrunk towards the best point. This procedure terminates when the differences in the function values between the best and worst points are small.
The implemented simulated annealing algorithm can be summarized as follows:

1. Build a random start simplex on $(vec(G)', vec(H)')'$. Set the start temperature t_0 to e.g. 1.

2. Add to the function values $f = ASR \vee L \vee MSEP$ (5, 7, 8) of the points in the simplex a random number so that $f_{temp}(simplex) = f(simplex) + t|log(u)|$, where u is uniformly distributed over $(0, 1)$. So the (simulated) deterioration of performance is random and proportional to t.

3. According to the Nelder-Mead transition function generate a trial point using the temporary function values f_{temp}.

4. Adapt the trial point to the side conditions by a QR decomposition.

5. Accept the new trial point according to Nelder-Mead with the function value $f_{temp}(trial) = f(trial) - t|log(u)|$ of the trial point. So a better trial point is always accepted and a worse trial point is accepted with a certain probability.

6. Repeat step 2-5 sufficiently often (e.g. 100 times). Reduce the temperature according to the cooling scheme, e.g. $t_{new} = 0.8t_{old}$.

7. Repeat step 2-6 sufficiently often, for example 50 times.

4 Error Rates for Time Related Data

In business phase classification one is interested in a reliable result of the predicted classes for e.g. the next 6 quarters. Therefore in order to evaluate a classification method it is necessary to look at the so-called "Ex-Post-Ante" which measures the retrospective prediction performance:

$$epa(t; pre) = \frac{\sum_{i=t}^{\min(t+pre,T)} I_{\{c_i \neq \hat{c}_i^t\}}}{\min(pre, T-t)} \tag{14}$$

where

- $c_i, \hat{c}_i$ are the true and estimated class of observation number i.
- pre is the number of successive observations of which the classes should be predicted.
- t is the last observation from which the classification model is estimated.
- T is the last observation from which the class is given.

An estimator for the misclassification rate for the next pre observation is then given by

$$\widehat{err}_{t_0} = \sum_{t=t_0}^{T-1} w(t) epa(t; pre). \tag{15}$$

At least two different weights $w(t)$ are possible:

Method	$\widehat{err}_C$	$\widehat{err}_T$
LDA	0.222	0.150
POC	0.196	0.132

Table 1: Estimated Error Rates on B3 Data

1. Constant with the number of training observations: $w(t)_C = \frac{1}{T-t_0}$
2. Increasing with the number of training observations: $w(t)_T = \frac{t}{\sum_{t=t_0}^{T-1} t}$

where t_0 is the first observation where the Ex-Post-Ante error rate is calculated. $w(\cdot)_T$ gives more weight to error rates which are calculated recently so it may be more suitable for the problem at hand.

5 Application to Business Phase Classification

The data set consists of 13 economic variables with quarterly observations from 1961/2 to 2004/2 (see Heilemann and Münch (1996)) of the German business cycle. The German business cycle is classified in a four phase scheme: upswing, upper turning point, downswing and lower turning point. The classes are given until $T = 2002/4$, so estimation of the Ex-Post-Ante error (14) can only be done until $t = 2002/3$. A prediction interval length of $pre = 6$ quarters was used and we started for LDA at $t_0 = 10$ which is 1963/3.

As POC is splitting the training data into a training and a test set to evaluate the loss it needs more data: The first estimation of optimal projection and score matrices was possible at 1967/2 (see also Figure 1). For Estimation of the error rates of LDA and POC t_0 in (15) was set to 1967/2, so both error series have the same length. Table 1 shows that POC is outperforming LDA with a constant weight ($\widehat{err}_C$) as well as with a weight that increases with the amount of training data ($\widehat{err}_T$). The time series of the "Ex-Post-Ante" rates (Figure 1) shows very interesting results: One can see the reunification of Germany (1990) which of course changes the business cycle so the phases could not be predicted by the classification methods – especially the yearly changes can be considered as outliers. Also the start of

the oil-crisis (oil price increases after 1971) and the second oil-crisis (1979) causes problems for classification methods.

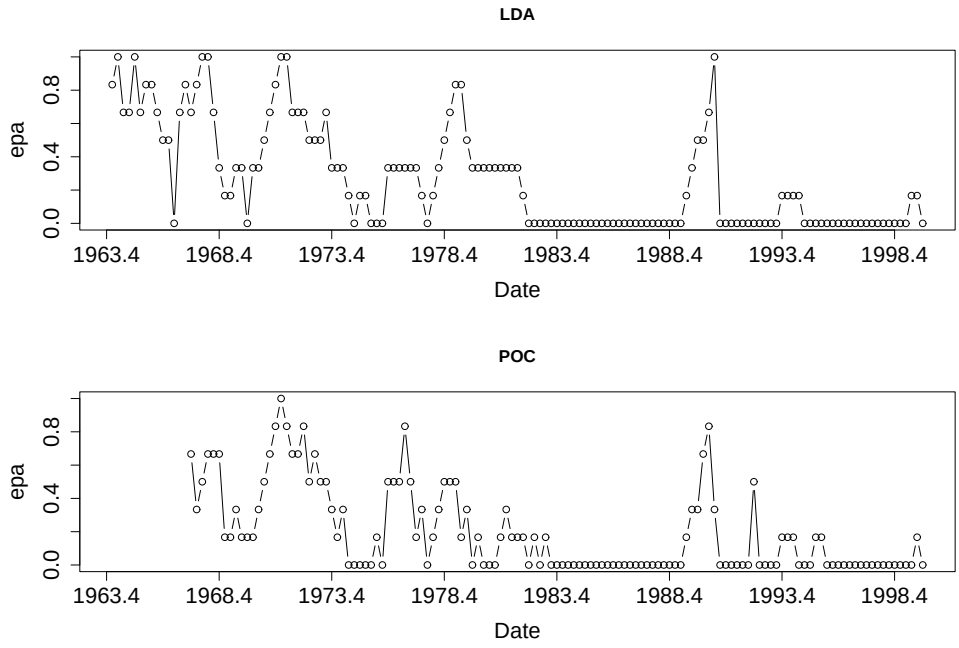

Figure 1: Comparison of Ex-Post-Ante Error Rates

6 Conclusion and outlook

The new prediction pursuit projection method for classification (POC) based on a simulated annealing algorithm outperforms the classical LDA (and therefore a lot of other classification methods) in business phase classification. Also the proposed error estimation by "Ex-Post-Ante" rates produces time series of error rates which agree with the economic knowledge and can be interpreted.

The flexibility of simulated annealing can be used by adding cost terms or using a more robust matrix norm for residual evaluation – both options are included in our program. So a lot of tuning is possible by these meta-parameters. On the other hand there is no theoretic knowledge about the

choice of these parameters so more research in this area is necessary.

Acknowledgment

This work has been supported by the Collaborative Research Center 'Reduction of Complexity in Multivariate Data Structures' (SFB 475) of the German Research Foundation (DFG).

References

I. E. FRANK and J. H. FRIEDMAN (1993): A statistical view of some Chemometrics regression tools. *Technometrics, 35(2): 199–209.*

T. HASTIE, A. BUJA, and R. TIBSHIRANI (1995): Penalized discriminant analysis. *The Annals of Statistics, 23(1):73–102.*

U. HEILEMANN and J. M. MÜNCH (1996): West german business cycles 1963-1994: A multivariate discriminant analysis. *CIRET-Conference in Singapore, CIRET-Studien 50.*

K. LUEBKE and C. WEIHS (2003): Testing a Simulated Annealing Algorithm in a Classification Problem, in: *Stochastic Algorithms: Foundations and Applications.* Lecture Notes in Computer Science, 2827:61–70.

K. LUEBKE and C. WEIHS (2004a): Generation of prediction optimal projection on latent factors by a stochastic search algorithm. *Computational Statistics and Data Analysis, 47(2):297–310.*

K. LUEBKE and C. WEIHS (2004b): A Note on the Dimension of the Projection Space in a Latent Factor Regression Model with Application to Business Cycle Classification. *Technical Report 29/2004, SFB 475, Universität Dortmund.*

J.A. Nelder and R. MEAD (1965): A Simplex Method for Functional Minimization. *Computer Journal, 7:308–313.*

W. H. PRESS, B.P. FLANNERY, S. A. TEUKOLSKY and W. T. VETTERLING (1992): *Numerical Recepies in C.* Cambridge University Press, Cambridge.

C. WEIHS and U. GARCZAREK (2002): Stability of multivariate representation of business cycles over time. *Technical Report 20/2002, SFB 475, Universität Dortmund.*