Optimal vs. Classical Linear Dimension Reduction

Michael C. Röhl, Claus Weihs

Lehrstuhl für Computergestützte Statistik,
Universität Dortmund, D-44221 Dortmund, Germany

Abstract: We describe a computer intensive method for linear dimension reduction which minimizes the classification error directly. Simulated annealing (Bohachevsky et al. (1986)) is used to solve this problem. The classification error is determined by an exact integration. We avoid distance or scatter measures which are only surrogates to circumvent the classification error. Simulations (in two dimensions) and analytical approximations demonstrate the superiority of optimal classification opposite to the classical procedures. We compare our procedure to the well–known canonical discriminant analysis (homoscedastic case) as described in Mc Lachlan (1992) and to a method by Young et al. (1987) for the heteroscedastic case. Special emphasis is put on the case when the distance based methods collapse. The computer intensive algorithm always achieves minimal classification error.

1 Introduction

Classification deals with the allocation of objects to g predetermined groups $G = \{1, 2, \ldots, g\}$, say. The goal is to minimize the misclassification rate over all possible future allocations, characterized by the conditional densities $p_i(x)$ $(i = 1, 2, \ldots, g)$. The minimal error is the so–called *Bayes error* (Mc Lachlan (1992)). Often we want to reduce the dimension of the classification problem to one or two dimensions in order to support human imagination without significantly increasing the misclassification rate. This article deals with linear combinations of the original variables to achieve this goal: Linear Dimension Reduction. The next section reviews the classical approach based on distance measures and presents the idea of Young et al. (1987) in a way that facilitates such a distance formulation. Section 3 introduces computerintensive dimension reduction and simulated annealing. Section 4 compares the classical and the computerintensive method.

2 Classical Linear Dimension Reduction

The intuitive idea is to project the data in a way that maximizes the distance between the groups (hopefully this will also minimize the misclassification rate). The distance measure relates the between–group scatter matrix

$$S_B = \sum_{i=1}^{g} p(i)(\mu_i - \bar{\mu})(\mu_i - \bar{\mu})' \quad ; \quad \bar{\mu} = \sum_{i=1}^{g} p(i)\mu_i \tag{1}$$

to the pooled within–group scatter matrix

$$S_W = \sum_{i=1}^{g} p(i)\Sigma_i, \tag{2}$$

where $p(i)$ $(i = 1, 2, \ldots, g)$ denotes the apriori probability of the different groups, μ_i their means and Σ_i their covariance matrices. The maximal rank of S_B is $g-1$. If we project on direction a, we have to maximize the quotient

$$\frac{a'S_B a}{a'S_W a} \tag{3}$$

by variation of a. The maximum is attained at the eigenvector v_1 corresponding to the largest eigenvalue λ_1 of $S_W^{-1}S_B$.
Formula (3) is equivalent to

$$\Delta_1 := \frac{a'S_B a}{a'S_W a} = \frac{1}{g-1}\sum_{i=1}^{g}(\nu_i - \bar{\nu})^2, \tag{4}$$

where

$$\nu_i = \frac{a'\mu_i}{(a'S_W a)^{1/2}} \quad \text{and} \quad \bar{\nu} = \frac{a'\bar{\mu}}{(a'S_W a)^{1/2}}. \tag{5}$$

This expression is easier to analyze.
An idea of Young et al. (1987) incorporates different covariance matrices for different groups. First we build the matrix

$$M = [\mu_2 - \mu_1 | \cdots | \mu_g - \mu_1 | \Sigma_2 - \Sigma_1 | \cdots | \Sigma_g - \Sigma_1], \tag{6}$$

where $M \in \mathbb{R}^{d \times s}$, $s = (g-1)(d+1)$ by juxtaposition of the vectors and matrices. We assume $\Sigma_i \neq \Sigma_1$ for at least one $i \in G$.
The analogon to (4) is

$$\Delta_2 := a'MM'a = \underbrace{\sum_{i=2}^{g}(a'(\mu_i - \mu_1))^2}_{\text{mean portion}} + \underbrace{\sum_{i=2}^{g}\sum_{j=1}^{d}(a'(\Sigma_i^j - \Sigma_1^j))^2}_{\text{covariance portion}}, \tag{7}$$

where Σ_i^j denotes the jth column vector of the ith covariance matrix. The term is divided into a pure mean portion and a pure covariance portion.
We get a similiar result to Δ_1 (apart from a different origin: $\bar{\mu}$ in Δ_1 and μ_1 in Δ_2) if all covariance matrices are identical and the variables are transformed in such a way that $\Sigma_i = I_d$ (identity matrix). This corresponds to

$$M = [\mu_2 - \mu_1 | \cdots | \mu_g - \mu_1]. \tag{8}$$

Further directions can be calculated by means of the eigenvectors of $S_W^{-1} S_B$ and MM' respectively.

3 Computerintensive Dimension Reduction and Optimization

3.1 Computerintensive Dimension Reduction

This section applies simulated annealing to the linear dimension reduction. The algorithm optimizes the entries in the projection matrix. The optimization problem is therefore

$$\begin{aligned} \text{Minimize} f: \quad \mathbb{R}^{dimred \times dim} &\rightarrow \quad \mathbb{R}^{+} \\ \text{projection matrix} &\mapsto \text{error rate}, \end{aligned} \tag{9}$$

where *dimred* and *dim* denote the dimension of the lower dimensional space and the original one, respectively.
We now sketch the simulated annealing algorithm used as an optimization tool.
Simulated annealing does not need derivatives, a great advantage compared to gradient methods. It can also be used if the function values are discrete. On the other hand you need more function evaluations than common gradient algorithms.
The computerintensive method achieves minimal misclassification error if adequately implemented.

3.2 Simulated Annealing

The freezing and crystallizing of liquids overcomes local energy minima. This physical strategy serves as the prototype for a computer program: *Simulated Annealing* (Bohachevsky (1986)). To model the natural procedure, we need a configuration space (a discrete or continous domain), a mechanism which describes how to get from one configuration to another and a cooling schedule describing how to decrease the temperature T ($T_0 \rightarrow T_1 \rightarrow \ldots \rightarrow T_n \rightarrow \ldots$). At each temperature – beginning at an optional configuration x_0 – we start a markov chain. Each trial point x_p is accepted with probability $\exp(-(f(x_p) - f(x_0))/T)$. After a number of steps in the markov chain, the temperature will be decreased, for example $T_n = \alpha T_{n-1}$ $(0 < \alpha < 1)$, and a new chain will be created (the starting point of the new chain is the end point of the last one, see Figure 1). In a concrete optimization, the temperature T is not a physical quantity but an abstract parameter which controls the optimization.

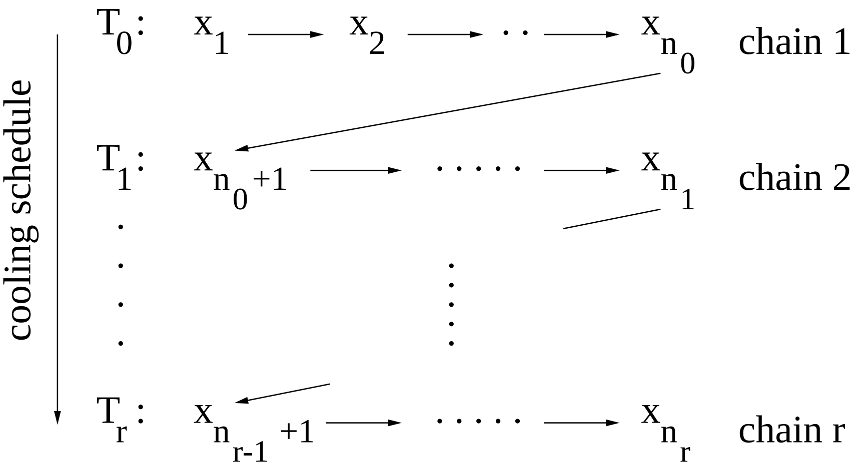

Figure 1: Flow chart of the simulated annealing algorithm.

In our application of simulated annealing, the function to be optimized is the misclassification rate. In each optimization step we calculate the error by exact integration using the conditional densities, that is for each group $i \in G$, we determine the regions where at least one of the other conditional densities is greater. We integrate $p_i(x)$ over these regions and get the misclassification error conditional on this group i. The total error is calculated as an average over all groups weighted by their apriori probability.

4 Comparison with the Classical Approach

The optimization algorithm introduced in section 3 is now compared to the classical approach. The classical procedures do not provide a direct link to the misclassification rate (that is, from a small perturbation of the direction a, you can not analytically derive the corresponding variation in the misclassification rate). In fact, in some special cases (depending on constellation of the groups, form of the covariance matrices), a significant difference between the two procedures can be detected. Apart from the pure comparison, emphasis is put on the question when distance based "analytical" methods collapse. In these cases only the algorithm in section 3 supplies valid results.

4.1 Equal Covariance Matrices and $g = 3$ Groups

In formula (4), assume

$$|\mu_i - \bar{\mu}| \gg |\mu_j - \bar{\mu}| \quad \forall j \neq i, \tag{10}$$

for one i. Then the sum has one dominant term which is maximized at the cost of the other summands, because the distances in (4) are squared. Therefore we get the approximation

$$\Delta_1 := \frac{a' S_B a}{a' S_W a} = \sum_{i=1}^{g} (\nu_i - \bar{\nu})^2 \approx \frac{1}{a' \Sigma a} ((\mu_i - \bar{\mu})' a)^2. \tag{11}$$

Maximization yields the value

$$(\mu_i - \bar{\mu})' \Sigma^{-1} (\mu_i - \bar{\mu}) \quad \text{attained at} \quad a \propto \Sigma^{-1} (\mu_i - \bar{\mu}). \tag{12}$$

Henceforth, we project on a direction that is dominated by μ_i. The other means are only incorporated by $\bar{\mu}$. This behaviour leads to suboptimality.

To get a better understanding, we conduct some simulations. First, we transform the common covariance matrix Σ by the transformation $x_{new} := \Sigma^{-1} x_{old}$ to the identity matrix I_d. This does not increase the misclassification rate. Because of the symmetry induced by three groups, it suffices to take $d = 2$. Therefore we set

$$\mu_1 = (0,0)', \quad \mu_2 = (2,0)' \quad \text{and} \quad \mu_3 = (x,y)'. \tag{13}$$

Mean μ_1 only determines the origin and μ_2 is somewhat arbitrary. A variation of μ_2 would only alter the misclassification level, not the qualitative conclusion. The third mean contains two variables x and y. This two dimensional surface can be conveniently plotted. Once again because of the symmetry of the constellation, it is enough to regard the positive quadrant. We take the range $0 \leq x \leq 2.5$ and $0 \leq y \leq 2.5$.

Figures 2 and 3 show the misclassification rates of the classical and the optimized procedure, respectively (simulated annealing given the means and the covariance matrix). Note the different scales of the two graphs.

The results of the classical procedure are qualitatively similiar in the "front" range ($0 \leq x \leq 2.5$ and $0 \leq y \leq 1.7$), whereas there is a significant difference in the "back". We now analyze the reason of the "mountain ridge" in the classical case in more detail. To achieve this goal, we calculate S_B.

A special situation arises, if the means of the three groups constitute a regular triangle. For that reason, we reparametrize the third mean: $\mu_3 = (1 + \delta x, \sqrt{3} + \delta y)$. Then we have

$$S_B = \frac{2}{9} \begin{pmatrix} \delta^2 x + 3 & \sqrt{3}\delta x + \delta x \delta y \\ \sqrt{3}\delta x + \delta x \delta y & \delta^2 y + 3 + 2\sqrt{3}\delta y \end{pmatrix}. \tag{14}$$

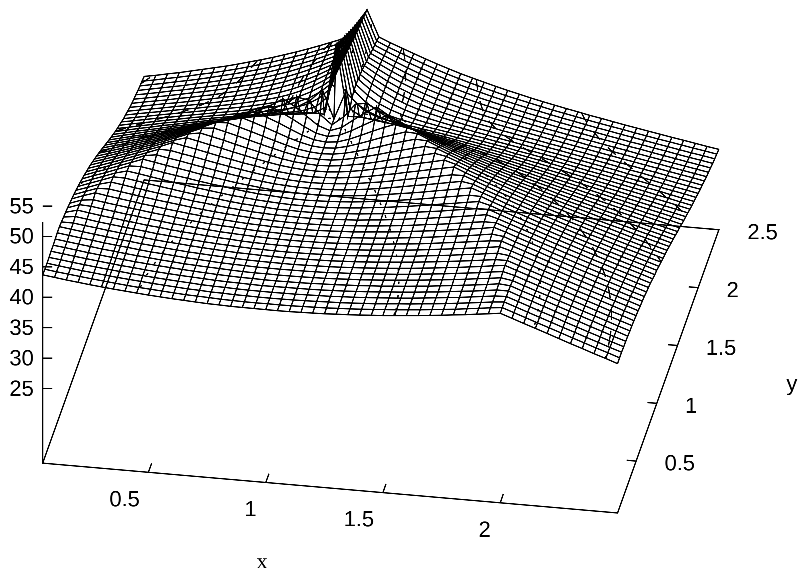

Figure 2: Misclassification rate classical procedure.

The special case $\delta x = 0$ yields

$$\frac{a'S_B a}{a'S_W a} = \frac{2}{9}\left\{1 - a_1^2\left((\delta y + \sqrt{3})^2 - 3\right)\right\}, \tag{15}$$

and after maximization we get the following distinction of cases:

$$\begin{aligned}
(\delta y + \sqrt{3})^2 > 3 \quad &\Leftrightarrow \quad \delta y > \sqrt{3} \quad \Rightarrow \quad a_1 = 0, a_2 = 1 \\
(\delta y + \sqrt{3})^2 < 3 \quad &\Leftrightarrow \quad 0 \le \delta y < \sqrt{3} \quad \Rightarrow \quad a_1 = 1, a_2 = 0 \\
(\delta y + \sqrt{3})^2 = 3 \quad &\Leftrightarrow \quad \delta y = 0 \quad \Rightarrow \quad a_1, a_2 \text{ arbitrary.}
\end{aligned} \tag{16}$$

The mean $\mu_3 = (1, \sqrt{3})'$ results in a singularity (projection vector $a = (a_1, a_2)'$ not defined). But this mean is realized with probability zero by the empirical mean value and is therefore unimportant. But important is the fact that the projection behaviour "turns over" at this value. Up to $\delta y < \sqrt{3}$, the projection is onto the x–axis (like the optimized procedure), then onto the y–axis. This causes a higher misclassification rate compared to the optimized procedure, because the projected first group coincides with the second one, while the optimized method still projects onto the x–axis.
The classical approach even more often fails for more than $g = 3$ groups, because there are more critical constellations.

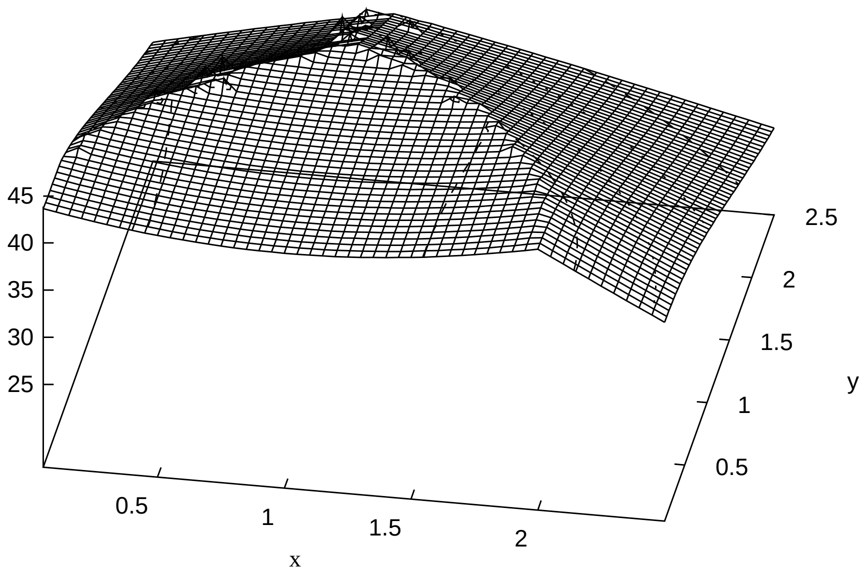

Figure 3: Misclassification rate optimized procedure.

4.2 Unequal Covariance Matrices and $g = 3$ Groups

The central formula (7)

$$\Delta_2 := a'MM'a = \sum_{i=2}^{g}(a'(\mu_i - \mu_1))^2 + \sum_{i=2}^{g}\sum_{j=1}^{d}(a'(\Sigma_i^j - \Sigma_1^j))^2 \tag{17}$$

yields more possibilities for dominant terms than (4).
For example, assume

$$|\Sigma_i^j - \Sigma_1^j| \gg |\Sigma_{i'}^{j'} - \Sigma_1^{j'}| \qquad \forall \quad (i', j') \neq (i, j), \tag{18}$$

true for one i, then only the jth column in the ith group differs from the pendant in the first group. Thus we get

$$\Delta_2 \approx (a'(\Sigma_i^j - \Sigma_1^j))^2. \tag{19}$$

The inequality of Schwarz supplies the maximum at

$$a \propto (\Sigma_i^j - \Sigma_1^j). \tag{20}$$

This solution uses only a small part of the available information and we therefore get – once again – a difference between the optimized and classical solution.
A small simulation study in two dimensions demonstrates the key issue. The means are now fixed at

$$\mu_1 = (0,0)', \quad \mu_2 = (2,0)' \quad \text{and} \quad \mu_3 = (1,1)'. \tag{21}$$

The covariance matrices are

$$\Sigma_1 = I_2, \quad \Sigma_2 = I_2 \quad \text{and} \quad \Sigma_3 = \mathrm{diag}(1+x, 1+y). \tag{22}$$

We take again the range $0 \leq x \leq 2.5$ and $0 \leq y \leq 2.5$. This time, the graphical representation does not show the constellation of the groups, but in a more abstract manner the variance of the third group. The figures 4 and 5 plot the misclassification rate of the classical and optimized procedure.

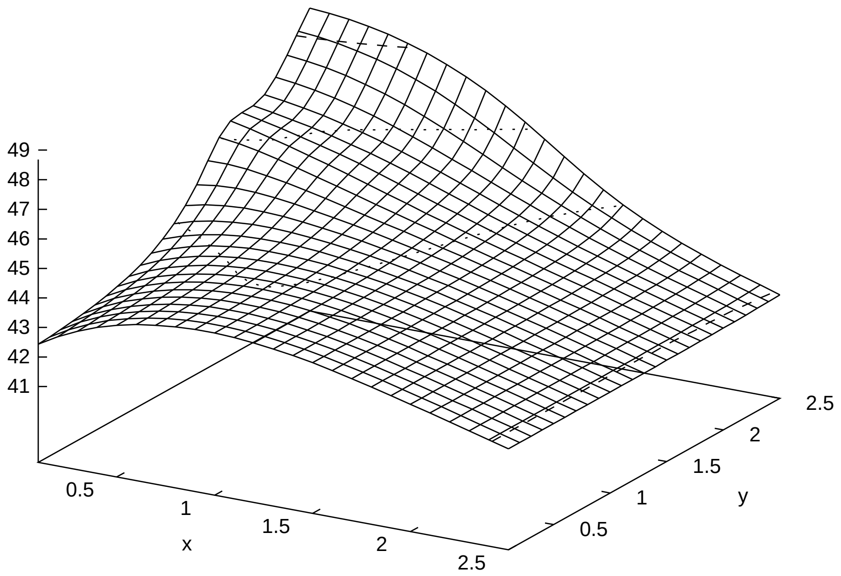

Figure 4: Misclassification rate classical procedure.

The differences are significant, especially if y is large and x small. In this case, the classical method projects onto the y–axis and the first and second group collapse. The optimized procedure still projects onto the x–axis.

In a concrete application, it is useful to compare the classical procedure with the optimized method in one dimension. If the results differ significantly, we have to use the optimized approach in higher dimensions (even if the computational burden is higher), otherwise we use the idea of Young et al. (1987), if the covariance matrices are unequal (especially if $d' > 2$).

5 Conclusions

After we introduced the classical discriminant analysis based on scatter matrices, we discussed a less well-known approach of Young et al. (1987) which we have reformulated using a distance measure.

These classical procedures were compared to an optimized procedure based on simulated annealing by means of simulations and analytical approximations. The differences and drawbacks of the classical approach were discussed in detail. The differences for more than two groups can be severe. It is exactly this case that is mainly ignored in the literature.

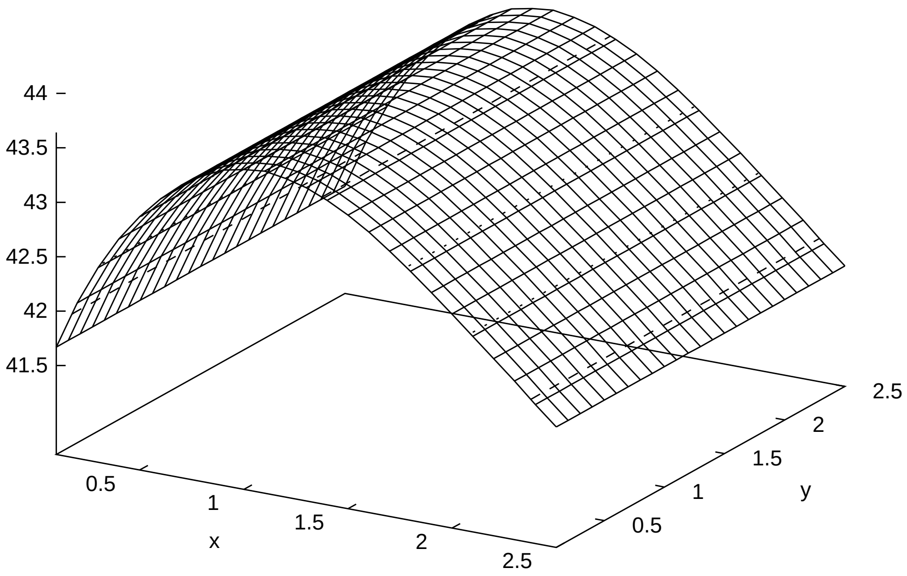

Figure 5: Misclassification rate optimized procedure.

This article clearly demonstrates the power of computerintensive methods. They help the statistician to concentrate on the real problem at hand: here the minimization of the misclassification rate.

Acknowledgment

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

References

BOHACHEVSKY, I.O., JOHNSON, M.E., STEIN, M.L. (1986):
Generalized Simulated Annealing for Function Optimization, *Technometrics, 28, 209-217.*

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

YOUNG, D.M., MARCO, V.R., ODELL, P.L. (1987): Quadratic Discrimination: Some Results on Optimal Low–Dimensional Representation, *J. Statist. Planning and Inference, 17, 307-319.*