A Note on the Dimension of the Projection Space in a Latent Factor Regression Model with Application to Business Cycle Classification

K. Luebke * C. Weihs

April 2004

Universität Dortmund
Fachbereich Statistik

Abstract

In this paper it is shown that the number of latent factors in a multiple multivariate regression model need not be larger than the number of the response variables in order to achieve an optimal prediction. The practical importance of this lemma is outlined and an application of such a projection on latent factors in a classification example is given.

Keywords
Latent Factor Models, Projection Matrix, Regression, Classification

1 Introduction

It is known that predictions in a multiple, multivariate linear regression are rather poor when the explanatory variables are collinear or the number of

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

parameters to estimate is not much larger then the number of observations (Helland and Almøy, 1994). This may, for example, be caused by overfitting or unstable estimates. To tackle these problems a reduced rank regression (RRR) method can be used (Reinsel and Velu, 1998). In a reduced rank regression the explanatory variables are projected on (few) so-called latent factors which are used as regressors for the response variables.
The different techniques like Partial Least Squares, Canonical Correlation Analysis or Redundancy Analysis differ only in the way they project the original variables on latent factors (see Schmidli (1995), page 61). To achieve a prediction optimal projection, the mean squared error of prediction (MSEP) can be written as a function of the projection matrix (see Weihs and Hothorn (2002), page 6). All possible projection matrices must fulfill the side-condition that the latent factors are orthonormal (see Schmidli (1995), page 55). To do a computer intensive minimization of the MSEP it is therefore useful to know the space of possible solutions of the side-condition. The general solution space of such a model is given in Groß et al. (2002). Within this solution space the MSEP can be minimized for example by means of simulated annealing (Luebke and Weihs, 2003a,b).
We show that if the objective is to find a prediction optimal projection the number of latent factors need not be larger than the number of response variables. So the minimization of the MSEP can be made much faster as the number of parameters to estimate is reduced compared to the general model – in cases where the number of response variables is small compared to the number of predictor variables.
This paper is organized as follows: In section 2 the reduced rank regression model is briefly introduced. Section 3 presents and proves a lemma on the necessary dimension of the projection matrix on latent factors in a regression context. In section 4 this lemma is applied to the classification problem. In a real world example phases of the German Business Cycles are classified (Section 5).

2 The Latent Factor Model

The basic multiple, multivariate linear model looks as follows:

$$Y = 1_n\mu + XM + E, \tag{1}$$

where

$Y \in \mathbb{R}^{n \times q}$	data of response variables,
$\mu \in \mathbb{R}^{q}$	mean column vector of responses,
$X \in \mathbb{R}^{n \times p}$	data of explanatory variables (X is mean centered),
$M \in \mathbb{R}^{p \times q}$	unknown regression coefficient matrix,
$E \in \mathbb{R}^{n \times q}$	matrix of errors.

Instead of the original explanatory variables X in this work a projection of these (possible) high dimensional variables on (few) variables Z is used. This may be important because of numerical reasons (collinearity or overfitting) or because of some model assumptions, e.g. that the response variables Y depend on some underlying latent factors. So in a latent factor model instead of the variables X in model (1) latent variables Z under the side-condition $Z'Z = I_r$ ($r \leq p$) with $Z = XG$ are used (I_r is the r-dimensional identity matrix). The model with latent factors is:

$$Y = 1_n\mu + XM + E = 1_n\mu + (XG)B + E. \tag{2}$$

with the side condition

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

Given estimates $\hat{G}$ of G, fulfilling the side-condition $(X\hat{G})'(X\hat{G}) = I_r$, and $\hat{\mu}$ of μ it is assumed that in the latent factor model the estimate of B is the usual least square estimate of Y on $Z = XG$.

$$\hat{B} = [(X\hat{G})'(X\hat{G})]^{-1}(X\hat{G})'(Y - 1_n\hat{\mu}) = (X\hat{G})'(Y - 1_n\hat{\mu}). \tag{4}$$

The ordinary least squares estimator for B is used by all reduced rank regression techniques like Partial Least Squares, Principal Component Regression, Canonical Correlation Analysis and Redundancy Analysis.

3 Lemma on the Number of Latent Factors in a Regression Problem

In this section it is shown that the dimension of the projected space $Z = XG$ need not be larger then the dimension of Y in the regression context. Let:

- $\tilde{q} = rank(Y) \leq q$,
- $\tilde{p} = rank(X) \leq p$.

Define $\tilde{r} = \min(\tilde{q}, \tilde{p})$. So $\tilde{r} \leq q$ and $\tilde{r} \leq p$. To simplify the notation assume that Y is mean centered ($\mu = 0$).

Lemma

1. If $\tilde{q} \geq \tilde{p}$ there exists no $r > \tilde{r}$ with $(XG)'(XG) = I_r$ with $G \in I\!R^{p \times r}$.
2. If $\tilde{q} < \tilde{p}$ there is for every $G \in I\!R^{p \times r}$ with $r > \tilde{r}$ and $(XG)'(XG) = I_r$ there is a $\tilde{G} \in I\!R^{p \times \tilde{r}}$ with $(X\tilde{G})'(X\tilde{G}) = I_{\tilde{r}}$ and $\tilde{M} = \tilde{G}(X\tilde{G})'Y = GB = M$.

Proof

1. $\tilde{r} = \min(\tilde{p}, \tilde{q}) = \tilde{p}$. Assume $r > \tilde{r}$. But from $(XG)'(XG) = I_r$ it follows that $r \leq \tilde{p} = \tilde{r}$. This contradicts the assumption.
2. From $(XG)'(XG) = I_r$ it follows that $r \leq \tilde{p}$. From $B = (XG)'Y$ and $Y = XGB$ it follows that $rank(B) = \tilde{r}$ because of

   - $rank(B) \leq \min(r, \tilde{q}) \leq \min(\tilde{p}, \tilde{q}) = \tilde{r}$.
   - $rank(B) \geq rank(Y) = \tilde{q} \geq \min(\tilde{q}, \tilde{p}) = \tilde{r}$.

   $B =: UV'$, $U \in I\!R^{r \times \tilde{r}}$ and $V \in I\!R^{q \times \tilde{r}}$ and $U'U = I_{\tilde{r}}$. By the Singular Value Decomposition and the fact that $rank(B) = \tilde{r}$ there is a U so that $B = UV'$ and $U'U = I_{\tilde{r}}$ (see e.g. Harville (1997), page 550).
   Let: $\tilde{G} = GU$. Now:
   $(X\tilde{G})'(X\tilde{G}) = U'(XG)'(XG)U = U'I_rU = I_{\tilde{r}}$ and
   $\tilde{B} = \tilde{G}'X'Y = U'G'X'Y = U'B = U'UV' = V'$ so
   $\tilde{M} = \tilde{G}\tilde{B} = GUV' = GB = M$.

□

So if for example MSEP is to be minimized as a function of G (Luebke and Weihs, 2003a) it is only necessary to minimize it for $\tilde{r}$ latent factors. In situations when there is only on response variable ($\tilde{q} = q = 1$) only one latent factor is needed to achieve an optimal projection concerning prediction. In order to understand the data it may be necessary to obtain and visualize more latent factors but for a prediction optimal regression only $\tilde{r}$ factors are needed. As to find the optimal G $p\tilde{r}$ parameters are optimized a relevant decrease in the number of parameters can be achieved. So far up to p^2 parameters in the matrix G are estimated (Luebke and Weihs, 2003a).

4 Application: Classification via Regression

A regression with the use of latent factors can be applied in a classification problem.[1].
Linear Discriminant Analysis (LDA) is a statistical method for classification. In LDA the classification is based on the calculation of the posteriori probabilities of a trial point. The class with the highest posteriori probability is chosen. To calculate the posteriori 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 and linked to regression via Average Squared Residual (ASR). 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. Thus the following Average Squared Residual is to be minimized. The Average Squared Residual function is (see (Hastie et al. (2001), p. 392):

$$ASR(H, M) = \frac{1}{n}||YH - XM||^2, \tag{5}$$

where

- Y is an indicator or dummy matrix of the classes,
- $H \in I\!R^{k \times l}$ is the score matrix of the classes,
- $M \in I\!R^{p \times l}$ is the regression parameter matrix, and
- $||\cdot||$ is the Frobenius Norm of a matrix.

To avoid trivial solutions the constraint

$$H'(Y'Y/n)H = I_l, \tag{6}$$

is applied. The ASR with latent factors for the regression is:

$$ASR(H, Z) = \frac{1}{n}||YH - ZZ'YH||^2, \tag{7}$$

[1]Part of this work is taken from Luebke and Weihs (2003c)

subject to (6) and (3) with $Z = XG$. This is equivalent to:

$$ASR(H, G) = \frac{1}{n}||YH - XG(XG)'YH||^2 \quad (8)$$

$$= \frac{1}{n}||(I_n - XG(XG)')YH||^2, \quad (9)$$

subject to (6) and (3).
In general $||AB|| \neq ||A||\,||B||$ (see for example: $||I_2 I_2|| = ||I_2|| = \sqrt{2} \neq 2 = ||I_2||\,||I_2||$). Therefore it is necessary in minimizing (8) to optimize G and H together and not to optimize H after the optimization of G.
The Lemma of this paper on the number of latent factors can be applied to such a classification problem, as in a classification problem there are often only $k = 2$ or $k = 4$ classes and then the rank of the indicator matrix Y is $k - 1$ which in many situations is much smaller than p.
After the calculation of G and H the classification can then take place in the linear map of the data X:

$$\eta(x) = XM, \quad \eta(X) \in R^{n \times l} \quad (10)$$

Let $\bar{\eta}^k$ be the mean of the linear map of observations from class k. Assigning of observations to classes is done by

$$\hat{k} = argmin \sum_{i=1}^{l} w_i(\eta(x)_i - \bar{\eta}_i^k)^2, \quad (11)$$

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 (11) 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

$$w_i = \frac{1}{s_i^2(1 - s_i^2)} \quad (12)$$

with s_i^2 being the mean squared residual of the i-th optimally scored fit, then the weight is proportional to the Mahalanobis distance in the original feature space X. As this equivalence is based on the way they calculate the scoring and regression matrix it may not be usable here. Another problem is that the weight is symmetric to $\frac{1}{2}$. That means, that if the squared residual in

dimension a is very good, for example $s^2_{i_a} = 0.05$, it gets the same weight as a dimension in which the prediction is very bad $s^2_{i_b} = 0.95$.
Let

$$w_i = n(\sum_{g=1}^{k} \sum_{x_j \in class g} (\eta(x_j)_i - \bar{\eta}_i^g)^2)^{-1}. \tag{13}$$

Here the weight is reciprocal to the squared sum of distances of the observations from the mean of the class (in the projected space).
The implementation of a Simulated Annealing algorithm to minimize (8) is described in Luebke and Weihs (2003c).
In the following the classification performance of this Classification Pursuit Projection (ClPP) is compared to LDA in a real world problem.

5 Example: Business Cycle Classification

The data set consists of 13 economic variables with 157 quarterly observations from 1955/4 to 1994/4 (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. With help of the Lemma the Simulated Annealing optimization only must be done in a $3 \cdot 13 = 39$-dimensional space instead of $13^2 = 169$-dimensional space which is less than $\frac{1}{4}$ of maximum dimensions.

There were 6 complete cycles in the time period. The prediction ability was tested by the leave-one-cycle out validation: One cycle was left out as a validation set, the other 5 cycles are used to train the method and then the misclassification rate was estimated on the validation set. It is shown in Weihs and Garczarek (2002) that in general LDA is among the best classifiers for this classification task. Despite the fact that the observed group sizes vary the a-priori group probabilities are set equal. As it turned out that 'unit labor costs' (LC) and 'wage and salary earners' (L) are the most stable economic indicators for business-cycle classification LDA and ClPP were also compared using only these two variables. The results are shown in Table 1.

Table 1 shows that ClPP is slightly performing better than LDA. This was also found in a simulation study in Luebke and Weihs (2003c).

Table 1: Estimated Error Rates in German Business-Cycle Classification

var	LDA	ClPP
all	0.49	0.45
L,LC	0.36	0.35

Acknowledgment

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

References

Jürgen Groß, Karsten Luebke, and Claus Weihs. A note on the general solution for a projection matrix in latent factor models. Technical Report 28, Sonderforschungsbereich 475, Universität Dortmund, 2002.

David A. Harville. *Matrix Algebra From a Statisticians's Perspective*. Springer, 1997.

Trevor Hastie, Andreas Buja, and Robert Tibshirani. Penalized discriminant analysis. *The Annals of Statistics*, 23(1):73–102, 1995.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer, 2001.

Ulrich Heilemann and H.J. Münch. West german business cycles 1963-1994: A multivariate discriminant analysis. In *CIRET-Conference in Singapore*, CIRET-Studien 50, 1996.

Inge S. Helland and Trygve Almøy. Comparison of prediction methods when only a few components are relevant. *Journal of the American Statistical Association*, 89(426):583–591, 1994.

Karsten Luebke and Claus Weihs. Generation of prediction optimal projection on latent factors by a stochastic search algorithm. *Computational Statistics & Data Analysis*, pages ??–??, 2003a. accepted for publication.

Karsten Luebke and Claus Weihs. Prediction optimal data analysis by means of stochastic search. In Martin Schader, Wolfgang Gaul, and Maurizio Vichi, editors, *Between Data Science and Applied Data Analysis*, pages 305–312. Springer, 2003b.

Karsten Luebke and Claus Weihs. Testing a simulated annealing algorithm in a classification problem. In Andreas Albrecht and Kathleen Steinhoefel, editors, *Stochastic Algorithms: Foundations and Applications*, volume 2827 of *Lecture Notes in Computer Science*, pages 61–70. Springer, 2003c.

Gregory C. Reinsel and Raja P. Velu. *Multivariate Reduced-Rank Regression, Theory and Applications*. Springer, 1998.

Heinz Schmidli. *Reduced Rank Regression*. Physica Verlag, 1995.

Claus Weihs and Ursula Garczarek. Stability of multivariate representation of business cylces over time. Technical Report 20, Sonderforschungsbereich 475, Universität Dortmund, 2002.

Claus Weihs and Torsten Hothorn. Determination of optimal prediction oriented multivariate latent factor models using loss functions. Technical Report 15, Sonderforschungsbereich 475, Universität Dortmund, 2002.