This is a preprint of an article published in *Journal of Chemometrics* 2001; **15**: 413–426
(http://www.interscience.wiley.com/jpages/0886-9383/).

Understanding the collinearity problem in regression and discriminant analysis.

Tormod Næs and Bjørn-Helge Mevik

MATFORSK, Oslovegen 1, 1430 Ås, Norway
and University of Oslo, Blindern, Oslo, Norway

Abstract

This paper presents a discussion of the collinearity problem in regression and discriminant analysis. The paper describes reasons why the collinearity is a problem for the prediction ability and classification ability of the classical methods. The discussion is based on established formulae for prediction errors. Special emphasis is put on differences and similarities between regression and classification. Some typical ways of handling the collinearity problems based on PCA will be described. The theoretical discussion will be accompanied by empirical illustrations.

Key words: Regression, classification, discriminant analysis, collinearity, PCR, PCA.

1. Introduction

Multivariate regression and discriminant analysis are among the most used and useful techniques in modern applied statistics and chemometrics. These techniques, or rather classes of techniques, are used in a number of different areas and applications, ranging from chemical spectroscopy to medicine and social sciences.

One of the main problems when applying some of the classical techniques is the collinearity among the variables used in the models. Such collinearity problems can sometimes lead to serious stability problems when the methods are applied (Weisberg(1985), Martens and Næs(1989)). A number of different methods can be used for diagnosing collinearity. The most used are the condition index and the variance inflation factor (Weisberg(1985)).

A number of different techniques for solving the collinearity problem have also been developed. These range from simple methods based on principal components to more specialised techniques for regularisation (see e.g. Næs and Indahl(1998)). The most frequently used methods for collinear data in regression and classification resemble each other strongly and are based on similar principles.

Often, the collinearity problem is described in terms of instability of the small eigenvalues and the effect that this may have on the empirical inverse covariance matrix which is involved both in regression and classification. This explanation is relevant for the regression coefficients and classification criteria themselves, but does not explain why and in which way the collinearity is a problem for the prediction and classification ability of the methods.

The present paper presents a discussion of the reasons why the collinearity is a problem in regression and classification. The discussion is focussed on prediction and classification ability of the methods. Some simple ways of handling the problem will also be mentioned and illustrated by examples. The methods presented are not selected because they are optimal in all possible cases, but because they are closely linked to how the problem is formulated and therefore well suited for discussing possible ways of solving the problem. Other and sometimes more efficient methods will also be referenced. In particular, we will describe similarities and differences of the effect that the collinearity has in the regression and classification situations respectively. Computations on real and simulated data will be used for illustration.

2. The effect of collinearity in linear regression

2.1 Least squares (LS) regression

Assume that there are *N* observations of a vector $(\mathbf{x}^t, y)$ and the purpose is to build a predictor for the scalar *y* based on the *K*-dimensional vector **x**. Say that **x** is easier or cheaper to measure than *y*. The data used for regression can be collected in the matrix **X** and the vector **y**. Assume that the relationship between **X** and **y** is linear. Without loss of generality we assume that **X** is centred. We also assume that **X** has full rank. The model can then be written as

$$\mathbf{y} = \mathbf{1}b_0 + \mathbf{X}\mathbf{b} + \mathbf{e} \qquad (1)$$

The main problem is to estimate the regression vector **b** in order to obtain a predictor

$$\hat{y} = \bar{y} + \mathbf{x}^t\hat{\mathbf{b}}\,, \qquad (2)$$

which gives as good predictions of unknown *y*'s as possible. Another possible application is interpretation of **b**, but here we will focus on prediction. A measure of prediction accuracy, which is much used, is mean square error (MSE) defined by

$$\mathrm{MSE}(\hat{y}) = \mathrm{E}(\hat{y} - y)^2 \qquad (3)$$

The most frequently used method of estimation for the regression vector is least squares (LS). The sum of squared residuals is minimised over the space of **b** values. The LS estimator is convenient to work with from a mathematical perspective and has a very nice closed form solution

$$\hat{\mathbf{b}} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{y} \qquad (4)$$

The covariance matrix of $\hat{\mathbf{b}}$ is equal to

$$\mathrm{COV}(\hat{\mathbf{b}}) = \sigma^2(\mathbf{X}^t\mathbf{X})^{-1} \qquad (5)$$

This can also be written as

$$\mathrm{COV}(\hat{\mathbf{b}}) = \sigma^2\sum_{k=1}^{K}\mathbf{p}_k(1/\lambda_k)\mathbf{p}_k^t \qquad (6)$$

where the **p**'s are the eigenvectors of $\mathbf{X}^t\mathbf{X}$ and the λ's are the corresponding eigenvalues.

The predictor $\hat{y}$ using the LS estimator $\hat{\mathbf{b}}$ is unbiased with MSE equal to

$$\mathrm{MSE}(\hat{y}) = \sigma^2/N + \sigma^2\mathbf{x}^t(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{x} + \sigma^2 \qquad (7)$$

The first term comes from the contribution of the estimated intercept, which is the average when **X** is centred. The last term σ^2 is due to the noise in *y* for the prediction sample. Even a perfect predictor will have this error if compared to a measured *y* value.

Using the eigenvectors and eigenvalues decomposition of $\mathbf{X}^t\mathbf{X}$, the MSE can be written as

$$\mathrm{MSE}(\hat{y}) = \sigma^2/N + \sigma^2\sum_{k=1}^{K}t_k^2/\lambda_k + \sigma^2 \qquad (8)$$

Here t_k is the score of **x** along eigenvector *k*, i.e. $t_k = \mathbf{x}^t\mathbf{p}_k$.

2.2 The effect of collinearity in the X-data

A common situation in many applications of linear models is that there are linear or near-linear relationships among the **x**-variables. If the linear relations are exact, the inverse of $\mathbf{X}^t\mathbf{X}$ does not exist and no unique $\hat{\mathbf{b}}$ can be produced. In this paper, however, we will concentrate on the case when **X** has full rank, i.e situations where a unique mathematical solution exists for the LS problem.

It is easy to see from (6) that if some of the eigenvalues are very small, the variances of the regression coefficients become very large.

For the prediction case, however, the situation is somewhat different. Directions with "small eigenvalues" will not necessarily give large MSE's. As can be seen from equation (8), the score values t_k relative to the eigenvalues λ_k are the important quantities for the size of the $MSE(\hat{y})$. In other words, what matters is how well the new sample fits into the range of variability of the calibration samples along the different eigenvector axes. As will be seen below, this fit has a tendency of being poorer for the eigenvectors with small eigenvalue than for those with larger eigenvalue.

In the following we will use the term prediction leverage for the quantities t_k^2 / λ_k because of their similarity with the leverages used for **x**-outlier detection (see Weisberg(1985)). Note that there is a prediction leverage for each factor *k*. Note also that the MSE of the LS predictor is essentially a sum of prediction leverages for the new sample plus two constant terms.

2.3 Principal component regression (PCR) used to solve the collinearity problem.

One of the simplest ways that the collinearity problem is solved in practice is by the use of principal component regression (PCR). Experience has shown that this usually gives much better results than LS for prediction purposes. Note that PCR is not selected because it is optimal, but because it links easily to the problem discussion above and also makes it clear in which direction solutions to the problem should be sought. Other and more sophisticated solutions may sometimes give better results (see e.g. Martens and Næs(1989)).

The singular value decomposition (SVD) of **X**, gives the equation

$$\mathbf{X} = \mathbf{USP}^t \qquad (9)$$

The column vectors **u** of **U** have sum of squares equal to 1 and are orthogonal. They are linked to the principal component score matrix **T** by **T = US**. The **S** matrix is a diagonal matrix with elements equal to the square root of λ (the singular values *s*). The **P** is defined as above, i.e. as the matrix of eigenvectors of $\mathbf{X}^t\mathbf{X}$.

A regression model for **y** given **U** can be written as

$$\mathbf{y} = \mathbf{1}\alpha_0 + \mathbf{U}\boldsymbol{\alpha} + \mathbf{e} \qquad (10)$$

Since **U** is a linear transformation of **X**, the model (10) is equivalent to the model (1) in the sense that the two will provide the same LS fit and predicted values. The α_0 is equal to b_0 above and the $\boldsymbol{\alpha}$ can be transformed linearly into **b**. Alternatively, the equation is sometimes transformed into a regression equation based on the scores **T** = **US**, i.e.

$$\mathbf{y} = \mathbf{1}\alpha_0 + \mathbf{T}\boldsymbol{\gamma} + \mathbf{e} \qquad (11)$$

The models (10) and (11) give the same fit as model (1), so the error term **e** is identical in all three models.

The PCR is defined as regression of **y** onto a subset (usually those which correspond to the larger eigenvalues, λ) of the components/columns of **U** (or **T** in (11)). The idea is to avoid those dimensions, i.e. those columns of **U**, which cause the instability. Let the matrix $\mathbf{U}_A$ be defined as the columns of **U** corresponding to the *A* largest eigenvalues of $\mathbf{X}^t\mathbf{X}$. The PCR is then defined as the regression of **y** onto $\mathbf{U}_A$.

$$\mathbf{y} = \mathbf{1}\alpha_0 + \mathbf{U}_A\boldsymbol{\alpha}_A + \mathbf{f} \qquad (12)$$

Here **f** is generally different from the error term **e** above. The estimates of the α's in $\boldsymbol{\alpha}_A$ are found by LS. The PCR predictor $\hat{y}_{\mathrm{PCR}}$ is obtained as

$$\hat{y}_{\mathrm{PCR}} = \bar{y} + \mathbf{u}_A^t\hat{\boldsymbol{\alpha}}_A \qquad (13)$$

The value of $\mathbf{u_A}$ for a new sample is found from projecting **x** onto the *A* first principal components and by diving the score/projection, *t*, by the square root of the eigenvalues. Note that for *A* = *K*, the PCR predictor becomes identical to the LS predictor $\hat{y}$. In practice, the best choice of A is usually determined by cross-validation or prediction testing. The predictor (13) can also be presented as an explicit function of **x**.

Some researches like to think of the model (12) as regression on so-called latent variables $\mathbf{U_A}$. The standardised scores $\mathbf{U_A}$ are then thought of as underlying latent variables describing the main variability of **X**. More information about this way of viewing the problem and also other approaches based on the latent variable approach can be found in Kvalheim(1987) and Burnham, et al(1996).

See Joliffe(1986) for other ways of selecting eigenvectors for regression. Jackson(1991) discusses several important aspects of using principal components.

2.4. Properties of the PCR predictor.

The variance, bias and the MSE of the predictor $\hat{y}_{\mathrm{PCR}}$ are

$$\mathrm{Var}(\hat{y}_{\mathrm{PCR}}) = \sigma^2/N + \sigma^2\sum_{k=1}^{A} t_k^2/\lambda_k \qquad (14)$$

$$\mathrm{bias}(\hat{y}_{\mathrm{PCR}}) = -\sum_{k=A+1}^{K}\left(t_k/\sqrt{\lambda_k}\right)\alpha_k \qquad (15)$$

$$\mathrm{MSE}(\hat{y}_{\mathrm{PCR}}) = \mathrm{Var}(\hat{y}_{\mathrm{PCR}}) + \mathrm{bias}(\hat{y}_{\mathrm{PCR}})^2 + \sigma^2 = \sigma^2/N + \sigma^2\sum_{k=1}^{A} t_k^2/\lambda_k + \left(-\sum_{k=A+1}^{K}\left(t_k/\sqrt{\lambda_k}\right)\alpha_k\right)^2 + \sigma^2$$

(16)

Note again that the σ^2 contribution in (16) comes from the error in *y* for the prediction sample. Note also that the PCR predictor is biased as opposed to the LS predictor above. The only difference between the MSE for LS and PCR is the contribution along the eigenvectors with small eigenvalue (a=A+1,….,K).

In many practical situations with collinear data, the PCR predictor performs much better than LS from a MSE point of view. Comparing the MSE formulae for the two predictors, one can see that the reason for this must lie in the replacement of the variance contribution for the LS predictor along the small eigenvalue directions ($\sigma^2 \sum_{k=A+1}^{K} t_k^2 / \lambda_k$) by the bias contribution from the PCR along the same directions $(-\sum_{k=A+1}^{K} \left(t_k / \sqrt{\lambda_k}\right)\alpha_k)^2$. In other words, a large variance contribution (for LS) is replaced by a smaller bias contribution (for PCR).

2.5 Empirical illustration.

In the following we illustrate the above aspects for near infrared (NIR) data which are usually highly collinear. The data are from calibration of protein in wheat. The number of samples is 267 (133 calibration, 134 test), and the dimension of the **x**-vector is 20 (reduced from 100 by averaging). The α's for the different factors and the contribution of the prediction leverages along the different eigenvectors are plotted in Figures 1 and 2. The prediction ability of PCR for different number of components is presented in Figure 3. As can be seen from the latter, the error is high for small values of *A*, then it decreases to a flat level before it increases again. The LS predictor is obtained as the PCR for 20 components. As can be seen, much better results than LS are obtained for PCR using for instance 10 components.

It is clear from this illustration that

1) The regression coefficients ($\hat{\alpha}'s$) for the smaller eigenvalues are very small (not significant, Figure 1).
2) The prediction leverage contribution for the smaller eigenvalues is larger than for the directions with large eigenvalues (Figure 2).

The first point means that the bias for PCR (with for instance 10 components) in this application is small. The second point shows that the poor performance of LS comes from the large prediction leverages for the smaller eigenvector directions.

Exactly the same phenomena (1 and 2) were observed in Næs and Martens(1988).

2.6. Discussion.

These two points are also easy to argue for intuitively. The reason for the first point (1) comes from the fact that the t's, not the u's, are in the same scale as the original measurements, **x** . In other words, if the different directions in **X**-space are comparable in importance for their influence on y, the γ's in model (11) will be comparable in size. Since u is obtained by dividing the score t by the singular value s, the α is identical to γ multiplied by s. Since s is very small for the smaller eigenvalues, the corresponding α's must also be very small (see also Frank and Friedman(1993)). In other words, the regression coefficients of the smaller eigenvalues will become small because of the small variation along the corresponding axes. A possible effect that comes on top of this is of course that the "smaller eigenvectors" may be less interesting than the rest, i.e. that the main information about y is in the eigenvector directions with large eigenvalue. In many, but not all, reasonably planned and conducted experiments, the smallest eigenvalue directions will be of less interest than the rest.

The other point (2) can be argued for by using formulae for the eigenvector stability (Mardia et al(1979)). It is clear from these formulae that the eigenvector directions with small variability are less stable than the rest. This means that their directions can change strongly from dataset to dataset taken from the same population. This will cause a tendency for larger prediction leverages.

Note that another consequence of the arguments above is that it usually matters little what is done to a few large eigenvectors with small values of α. The prediction leverages along these eigenvectors will be small or moderate and therefore their contribution to the MSE will typically be relatively small. Similarly, it is clear that the "small" eigenvector directions will always be difficult to handle. If such eigenvectors have large values of α, the prediction ability of any regression method will probably be poor. The most interesting discussions about differences among possible regression methods (for instance PCR, PLS, RR etc.) for solving the collinearity problem should therefore relate to how they handle of the eigenvectors with intermediate size of the eigenvalues.

3 The effect of collinearity in discriminant analysis

3.1. QDA/LDA

Assume that there are a number of vector-valued observations (**x**, dimension K) available for a number of groups, C. The purpose is to use these data in order to build a classification rule that can be used to classify future samples into one of the groups. The training data matrix for group j will be denoted by $\mathbf{X}_j$. The number of training samples in group j is denoted by N_j. The total number of samples is denoted by N and is defined as the sum of the N_j's.

One of the most used methods of discrimination assumes that the populations (groups) are normally distributed and assumes that there is a probability π_j attached to each group. This probability indicates that prior to the observation of **x** is taken there is a probability π_j that an unknown object comes from group j.

The so-called quadratic discriminant analysis (QDA) method, which uses the Bayes rule, maximises the posterior probability that a sample belongs to a group, given the observation of the vector **x**. The discrimination rule results in the following criterion if the distributions within all groups are normal with known means and covariance matrices: Allocate a new sample (with measurement **x**) to the group (*j*) with the smallest value of the criterion

$$L_j = (\mathbf{x} - \boldsymbol{\mu}_j)^t \boldsymbol{\Sigma}_j^{-1} (\mathbf{x} - \boldsymbol{\mu}_j) + \log\left|\boldsymbol{\Sigma}_j\right| - 2\log \pi_j \qquad (17)$$

In practice, the means and covariances for the groups are unknown and must be estimated from the data. Usually one uses the empirical mean vectors $\bar{\mathbf{x}}_j$ and the empirical covariance matrices $\hat{\boldsymbol{\Sigma}}_j = (\mathbf{X}_j - \mathbf{1}\bar{\mathbf{x}}_j^t)^t (\mathbf{X}_j - \mathbf{1}\bar{\mathbf{x}}_j^t) / N_j$. Then we obtain $\hat{L}_j$ as the direct plug-in estimate for L_j by replacing all parameters by their estimates. $\hat{L}_j$ can be written as

$$\hat{L}_j = (\mathbf{x} - \bar{\mathbf{x}}_j)^t \hat{\boldsymbol{\Sigma}}_j^{-1} (\mathbf{x} - \bar{\mathbf{x}}_j) + \log\left|\hat{\boldsymbol{\Sigma}}_j\right| - 2\log \pi_j \qquad (18)$$

As can be seen, $\hat{L}_j$ is a squared Mahalanobis distance plus a contribution from the log of the covariance matrix minus a contribution from the prior probabilities. Note that when prior probabilities are equal, the last term vanishes. Note also that when covariance matrices are equal, the second term vanishes.

If a pooled covariance matrix is used, which is natural when the covariance structure of the two groups are similar, the method (18) reduces to a linear discriminant function. The method is called linear discriminant analysis (LDA). This is the method which will be focused in the computations to follow.

As can be seen, in the same way as for the regression method above, the estimated criterion $\hat{L}_j$ contains an estimated inverse covariance matrix ($\hat{\boldsymbol{\Sigma}}^{-1}$).

3.2 The effect of collinearity in discriminant analysis.

The criterion $\hat{L}_j$ can also be written as

$$\hat{L}_j = \sum_{k=1}^{K} (t_{jk})^2 / \lambda_{jk} + \log\left|\sum_{k=1}^{K} \mathbf{p}_{jk} (\lambda_{jk}) (\mathbf{p}_{jk})^t\right| - 2\log \pi_j \qquad (19)$$

where $\mathbf{p}_{jk}$ is the k'th eigenvector of the empirical covariance matrix $\hat{\Sigma}_j$ for group *j* and λ_{jk} is the corresponding eigenvalue. The t_{jk} is the score/coordinate for the new sample along eigenvector *k* in group *j*.

The smallest eigenvalues and their corresponding eigenvectors may be very unstable (Mardia et al(1979)) and since the eigenvalues are inverted, they will have a very large influence on $\hat{L}_j$. The variance of the criterion $\hat{L}_j$ as an estimate of L_j will then obviously be very large. Note the similarity between this and the first part of the MSE formula for the LS predictor.

The most important aspect, however, is usually not the instability of the criterion itself (either $\hat{L}_j$ in classification or $\hat{\mathbf{b}}$ in regression), but rather the performance of the method when applied to prediction/classification. The formula (19) shows that it is not the instability of the criterion itself that matters for the classification power of the method, but merely the relative size of the scores t (computed for a new sample) compared to the corresponding eigenvalues of the training set. These quantities were above named the prediction leverages.

What matters for a classification rule like (19) is that at least some of the principal component directions with corresponding scores, t, distinguish between the groups. If directions with no predictive power are known exactly (as they are in (17)), the non-informative directions will vanish from the criterion. If they are estimated imprecisely (as they may be in (19)), they will, however, represent noise. Since the small eigenvectors and eigenvalues may be very imprecise estimates for their population analogues (Mardia et al(1979)), the noise effect may be more serious for the smaller eigenvalues than for the larger. If the small population eigenvectors have little predictive power, their estimates may therefor weaken the classification power of $\hat{L}_j$ in equation (19) substantially.

3.3. Modifications of QDA/LDA based on PCA

An obvious suggestion for improvement indicated by the PCR method above is then to use only a few principal components with large eigenvalue in the classification rule instead of the original variables themselves. This results in the reduced QDA criterion given by

$$\hat{L}_j^{\mathrm{PC}} = \sum_{k=1}^{A} (t_{jk})^2 / \lambda_{jk} + \log\left|\sum_{k=1}^{A} \mathbf{p}_{jk} (\lambda_{jk}) (\mathbf{p}_{jk})^t\right| - 2\log \pi_j \qquad (20)$$

Note that this is identical to (19) except that the sums in the squared Mahalanobis distance and the determinant is up to A instead of up to K. Thus, $\hat{L}_j$ is reduced to a criterion that is a sum of contributions along the eigenvectors with largest variance. In other words, the suggestion (20) solves the large variance problem of the criterion $\hat{L}_j$.

Note that in the same way as $\hat{L}_j$ is an estimate of L_j, $\hat{L}_j^{\mathrm{PC}}$ is an estimate of a population analogue (here denoted by L_j^{PC}).

Again, we will stress that the choice of illustration method is here made based on its close relation to the problem formulation above. It is a good method and indicates clearly in which directions solutions should be sought, but in some cases other and more sophisticated methods can sometimes perform better (see e.g. Indahl et al(1999) for a comparison of various classification methods on NIR data).

In the formula (20), the PCA is run for each group separately (QDA). If a joint covariance matrix is used, the PCA is run on the pooled covariance matrix (LDA) instead.

3.4. Properties of the PCA based approach and some alternatives.

The above modification of QDA obviously solves the large variance problem since the effects of the small and unstable eigenvalue directions are eliminated. Therefore, in cases where the "small eigenvector" directions are unimportant for classification, $\hat{L}_j^{\mathrm{PC}}$ will clearly be a good method. Such a situation is depicted in Figure 4a.

In regression, directions with small eigenvalue will always be difficult to handle. In the following, we will argue that this is not necessarily the case for classification. The situation we have in mind is the one depicted in Figure 4b. This is a situation where the directions of small variability for the different groups is responsible for the difference between the groups. What should then be done with such directions? They are important for classification, but their estimates are uncertain and can lead to very unstable classification rules.

One possible way of handling this problem is to use the Euclidean distance within the space orthogonal to the first stable eigenvectors with large eigenvalues. This technique avoids the instability problem of the individual eigenvectors since the Euclidean distance does not divide each direction by its eigenvalue. The instability is solved at the same time as the "small eigenvectors" are not left out of the criterion.

This technique is used for the well known SIMCA (Wold(1976)) and also for the method DASCO (Frank and Friedman(1989)).

Another option for solving the collinearity problem which should be mentioned, is the following: Compute the PCA on the whole data set and use LDA or QDA on the joint components. By doing this, one transforms the small eigenvector directions in Figue 4b into directions with substantial variability. Small eigenvector directions for a particular subgroup are turned into directions with substantial variability for the whole data set. These directions will have discrimination power and will not represent any problem with respect to stability.

3.5. Empirical illustration

Two simulated examples will be used to illustrate the ideas presented.

For both data sets, we generated two groups ($C = 2$) with different means and with the same covariance matrix. There were 20 **X**-variables in both cases. The generation of the **X**-variables is done by using linear combinations of 5 loading spectra from a NIR example plus random noise. Each simulated NIR spectrum is generated according to a factor model

$$\mathbf{x} = \mathbf{L}\mathrm{t} + \boldsymbol{\varepsilon} \qquad (20)$$

Where **L** is the matrix of 5 estimated NIR spectra, the **t** consists of uncorrelated Gaussian variables with variances (8, 4, 1, 0.5, 0.1). The random noise ε has uncorrelated, normally distributed components with variance 0.0025. The difference between the two examples is the way the differences between the two groups are generated.

In both cases we generated 10 spectra in each group for training, and 40 for test set validation (20 from each group).

Note that model (20) is a latent variable model used to generate collinearity. See Section 2.3. and the references given there for a discussion of the relation between collinearity and latent variable models.

Example 1.

In the first example, the **t** variables have different means in the two groups, namely (–1, 0.5, 1, 2, 1) for group one and (1, 0.5, 1, 1, 0.5) for group two. As can be seen, the difference between the groups is related to components 1, 4 and 5, i.e. in the space generated by eigenvectors with “large eigenvalue” (see Figure 5 for an illustration of the empirical eigenvalue structure). As can be seen the spectra are highly collinear. The situation here corresponds to the conceptual situation in Figure 4a.

Example 2.

Here the groups are different with respect to the orthogonal compliment to the 5 “NIR loadings”. This is achieved in the following way: The constant 0.18 is multiplied by a 6’th loading vector (orthogonal to the other 5) and added to the second group. Both groups had initially the same means as group one in the first example. The situation corresponds to Figure 4b. Again the data are highly collinear.

Results.

The two examples were approached by using LDA based on different ways of computing and selecting components. The methods compared are the following:

a) LDA based on the first 5 and 10 components computed from the pooled covariance matrix. This corresponds to $\hat{L}_j^{\mathrm{PC}}$ above based on 5 and 10 components.
b) LDA based on the components 6–10 from the pooled matrix. This corresponds to the using $\hat{L}_j^{\mathrm{PC}}$ for those components only.
c) The Euclidean distance for the space orthogonal to the 5 first components.
d) LDA based on principal components computed from the full data set (10 components).

The quality of the different methods is evaluated by the percentage of wrong classifications (error rate, %) obtained in a prediction testing. The results are given in Table 1.

As can be seen for Example 1, using the 5 first components gives the best results, i.e $\hat{L}_j^{\mathrm{PC}}$ based on 5 components is best. This was to be expected, since the 5 most dominant dimensions are responsible for the discrimination. The components 6–10 give very bad results. It is also interesting to note that the components 1–10 solution performs poorer than the first 5 components, showing again that the components 6–10 only introduce instability. The Euclidean distance was not able to improve the results in this case, since the main discrimination power lies in the first few components. The solution for the joint PCA followed by LDA gives the same results as obtained by the $\hat{L}_j^{\mathrm{PC}}$ with 5 components.

For example 2 the 5 first components give very bad results. The same is true for the components 6–10 and for 1–10 if Mahalanobis distance is used. All this is to be expected from the discussion above. The Euclidean distance in the space orthogonal to the first 5 components, however, gives good results. This also supports the discussion above, namely that even situations as generated in Example 2 can be handled if data are used properly (Figure 4b). It is also worth noting that also in this case, the joint PCA followed by LDA gives good results.

3.6 Discussion.

As has been described above, if larger eigenvectors are important and the small eigenvector directions are irrelevant, the $\hat{L}_j^{\mathrm{PC}}$ based on the first few components gives good results. It was also shown that if this is not the case, the problem is difficult to solve within the framework of LDA. A better approach in such cases is to use Euclidean distance in the orthogonal compliment to the first few components. The main reason for this lies in the lack of stability of the eigenvectors with the smallest eigenvalue.

In practice, the Euclidean distance may possibly be accompanied by a Mahalanobis distance criterion in the space of main variability. This is the case for both SIMCA and

DASCO. Another possibility is to use PCA first on the whole space and use LDA on the joint components (Figure 4c).

In the discussion for regression it was argued that the importance (bias) of the eigenvectors often decreases with the size of the eigenvalues (α's decrease). On top of this mathematical aspect, there may also be reasons to believe that for most well planned and conducted experiments the main information is in the larger and intermediate eigenvector directions. For the classification situation, it is hard to come up with a similar argument. There is no reason as far as we can see for assuming that the smaller eigenvalue directions (for a class), neither in the population or in the formula (19), should be less important than the others in the classification rule.

4. General discussion and conclusion

Collinearity is a problem both for regression and for classification when standard methods are applied. In both cases, this is related to instability of information along the small eigenvector directions.

If the relevant information is gathered in the "larger eigenvectors", the problem can be solved by using only the first few components from the covariance matrix. This is true for both regression and classification.

If, however, this is not the case, this approach will give poor results. For regression, such directions will always be difficult to handle, but for discriminant analysis they can be handled if used properly. If the "small eigenvector" space is handled as the orthogonal compliment of the "larger eigenvectors" and if a Euclidean distance is used instead of the Mahalanobis distance, the problem can be solved.

It should be stressed again that the methods selected for illustration here are not necessarily optimal. They are good candidates, and selected primarily because of their close connection to how the problem of collinearity is described. In some cases, other and more sophisticated (but still closely related) methods can do even better. An obvious and probably the most used candidate for solving the collinearity problem in regression is PLS regression (see e.g. Martens and Næs(1989)). This is also a method based on regressing y onto well selected linear combinations of **x.** It has been shown that in some cases, PLS may be more efficient than PCR in extracting the most relevant information in x by as few components as possible. This has to do with the ability that PLS has in discarding components with little relation to *y*. For classification, PLS is also much used. Instead of using a continuous *y* variable as regressand, dummy variables representing the different classes are regressed onto the spectral data **x.** Again linear combinations with good ability to distinguish between the groups are extracted. In Indahl et al(1999) it was shown that a compromise between this so-called PLS discriminant analysis and the LDA described above can give even better results. PLS is used to generate relevant components, and LDA is applied for these components only. Note that this method is very similar to the method described above where principal components are extracted before an LDA is used on the extracted components. The only difference is the way the components are extracted.

In some situations, like for instance in NIR spectroscopy, the problem of collinearity can sometimes be reduced by using carefully selected transforms. Examples here are the MSC method proposed by Geladi et al(1985) and the OSC method suggested by Wold et al(1998). These are methods which remove some of the uninteresting variability which is causing much of the collinearity in the spectral data. These transforms should always be considered before the regression or classification is performed.

References.

Burnham, A, Viveros, R. and MacGregor, J.F. (1996). Frameworks for latent variable multivariate regression. J. Chemometrics, 10, 31-45.

Frank, I. and Friedman, J.(1989). Classification: Old-timers and newcomers. J Chemometrics, 3, 463-475.

Frank, I. and Friedman, J (1993). A statistical view of some chemometric regression tools. Technometrics, 35, 2, 109-135.

Geladi, P., McDougall, D. and H. Martens (1985). Linearisation and scatter correction for near infrared reflectance spectra of meat", Appl. Spectrosc. 39, 491.

Indahl, U.G., Sahni, N.S., Kirkhus, B. and Næs, T. (1999). Multivariate strategies for classification based on NIR spectra – with applications to mayonnaise. Chemometrics and int. lab. systems. 49, 19-31.

Jackson, J.E. (1991). A user's guide to principal components. Wiley, NY.

Joliffe, I. T. (1986). Principal component analysis. Springer Verlag, New York.

Kvalheim, O.M. (1987) Latent-structure decomposition projections of multivariate data. Chemometrics and int. lab. systems, 2, 283-290.

Næs, T. and Martens, H. (1988). Principal component regression in NIR analysis: viewpoints, background details and selection of components. J. Chemometrics, 2, 155-167.

Næs, T. and Indahl, U. (1998). A unified description of classical classification methods for multicollinear data. J. Chemometrics. 12, 205-220.

Mardia, K.V., Kent, J.T. and Bibby, J.M. (1979). Multivariate analysis. Academic Press, London.

Martens, H. and Næs, T.(1989). Multivariate Calibration. J. Wiley and Sons, Chichester, UK.

Weisberg, S. (1985). Applied Linear Regression. J. Wiley and Sons, NY.

Wold, S. (1976). Pattern recognition by means of disjoint principal components models. Pattern recognition, 8, 127-139.

Wold, S. Antii, H. Lindgren, F. and Öhman, J. (1998). Orthogonal signal correction of near-infrared spectra. *Chemometrics and Int. Lab. Systems*. **44**, 175 - 185, 1998)

Figure captions.

Figure 1. Regression coefficients and 95% confidence limit for their t-tests for the different principal components.

Figure 2. Prediction leverage for the different components

Figure 3. Root mean square error of prediction (RMSEP) as a function of the 20 principal components.

Figure 4. Different classification situations. In a) is depicted a situation whete the first PC is the most important direction for discrimination. In b) the second PC is the most important. In c) both directions are important for discrimination.

Figure 5. Eigenvalues for the 7 first principal components.

Table 1. Error rates (in %) for the different methods used for example 1 and example 2. Prediction testing is used for validation.

	1–5 Mah	6–10 Mah	1–10 Mah	6–> Euclid	1–10 Joint PCA
Example 1	12.5	45.0	17.5	32.5	12.5
Example 2	55	25.0	30.0	10.0	17.5

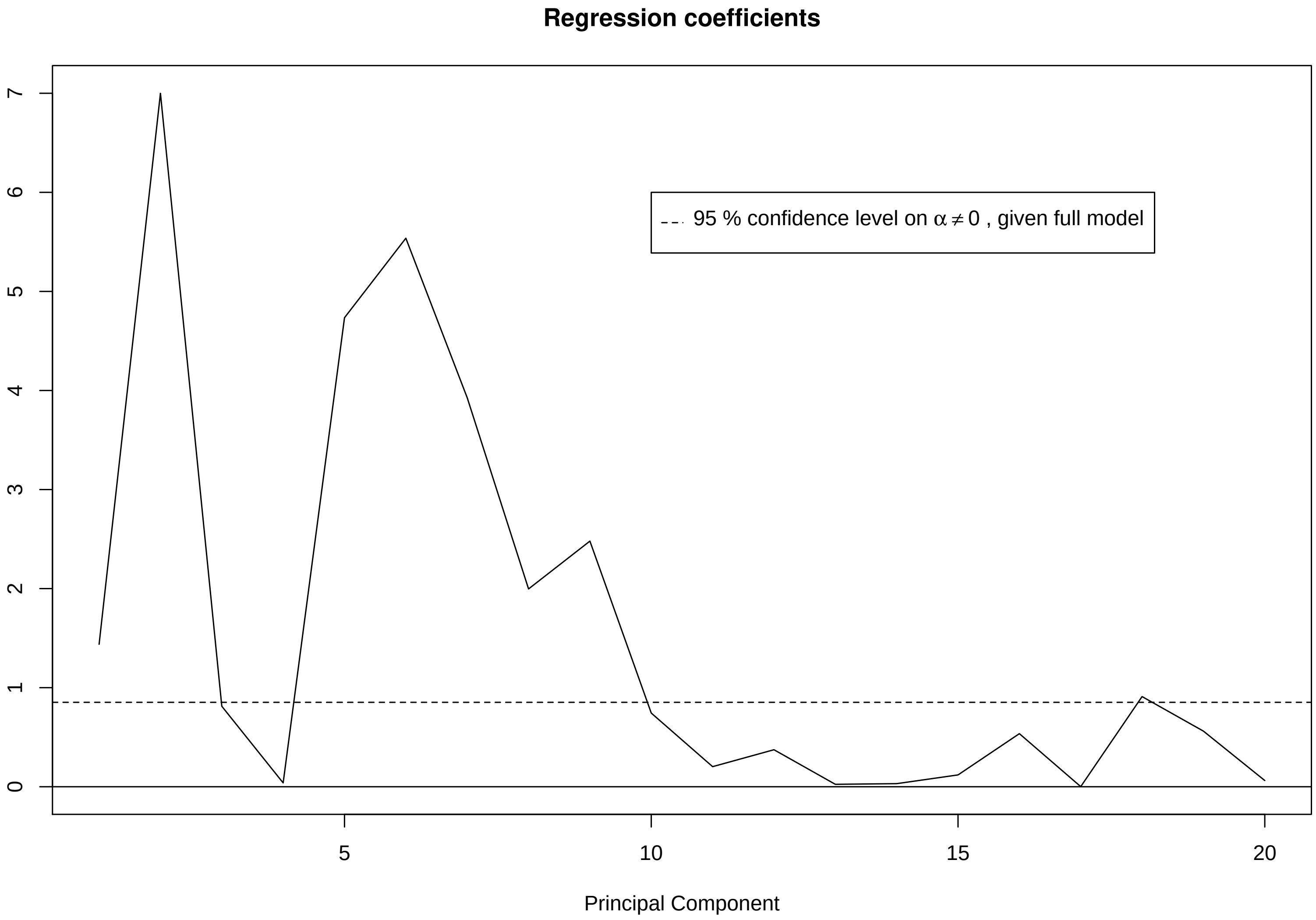
Regression coefficients
95 % confidence level on α ≠ 0 , given full model
|α|
0
1
2
3
4
5
6
7
5
10
15
20
Principal Component

Figure 1

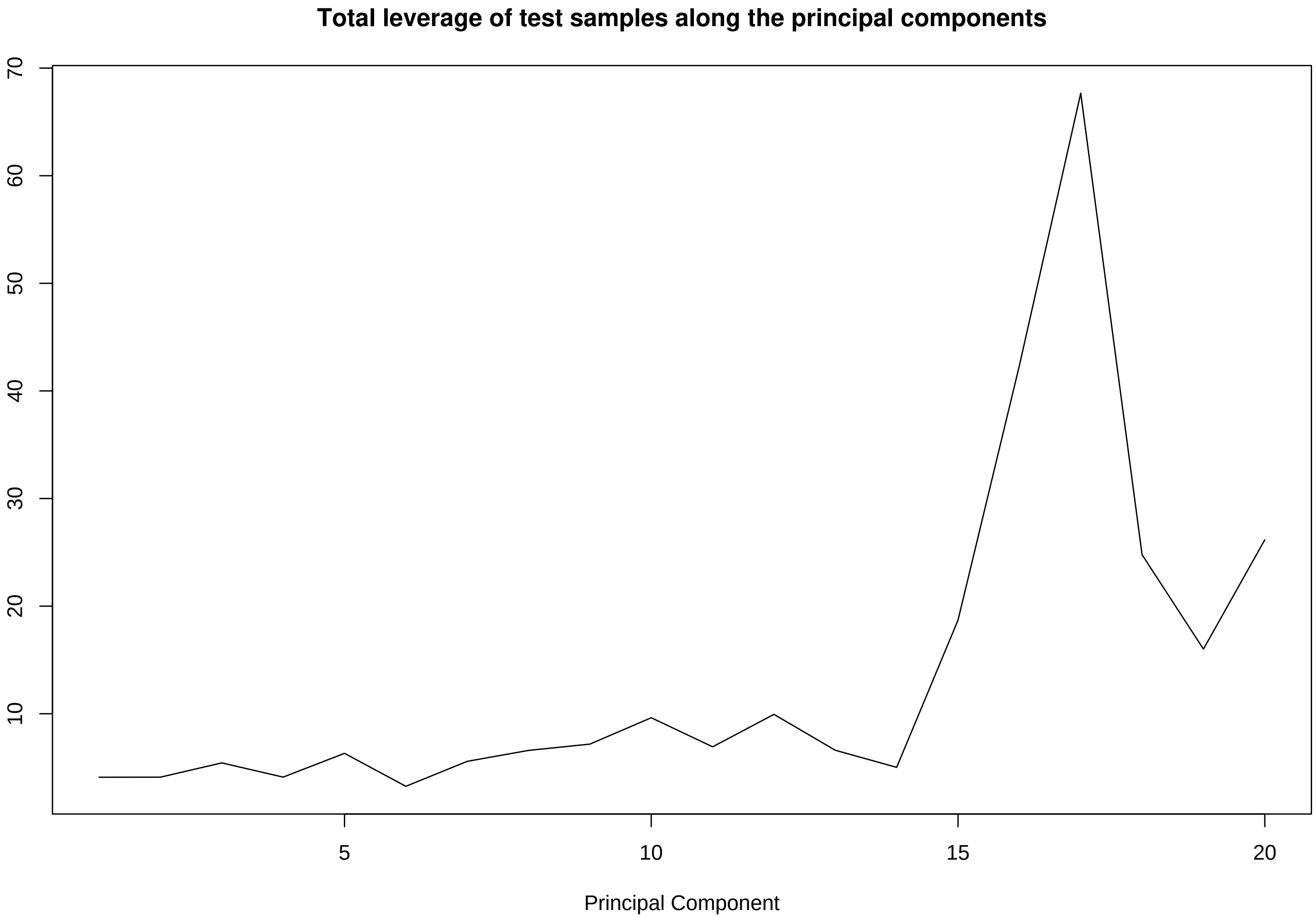
Total leverage of test samples along the principal components
Leverage
Principal Component
10
20
30
40
50
60
70
5
10
15
20

Figure 2

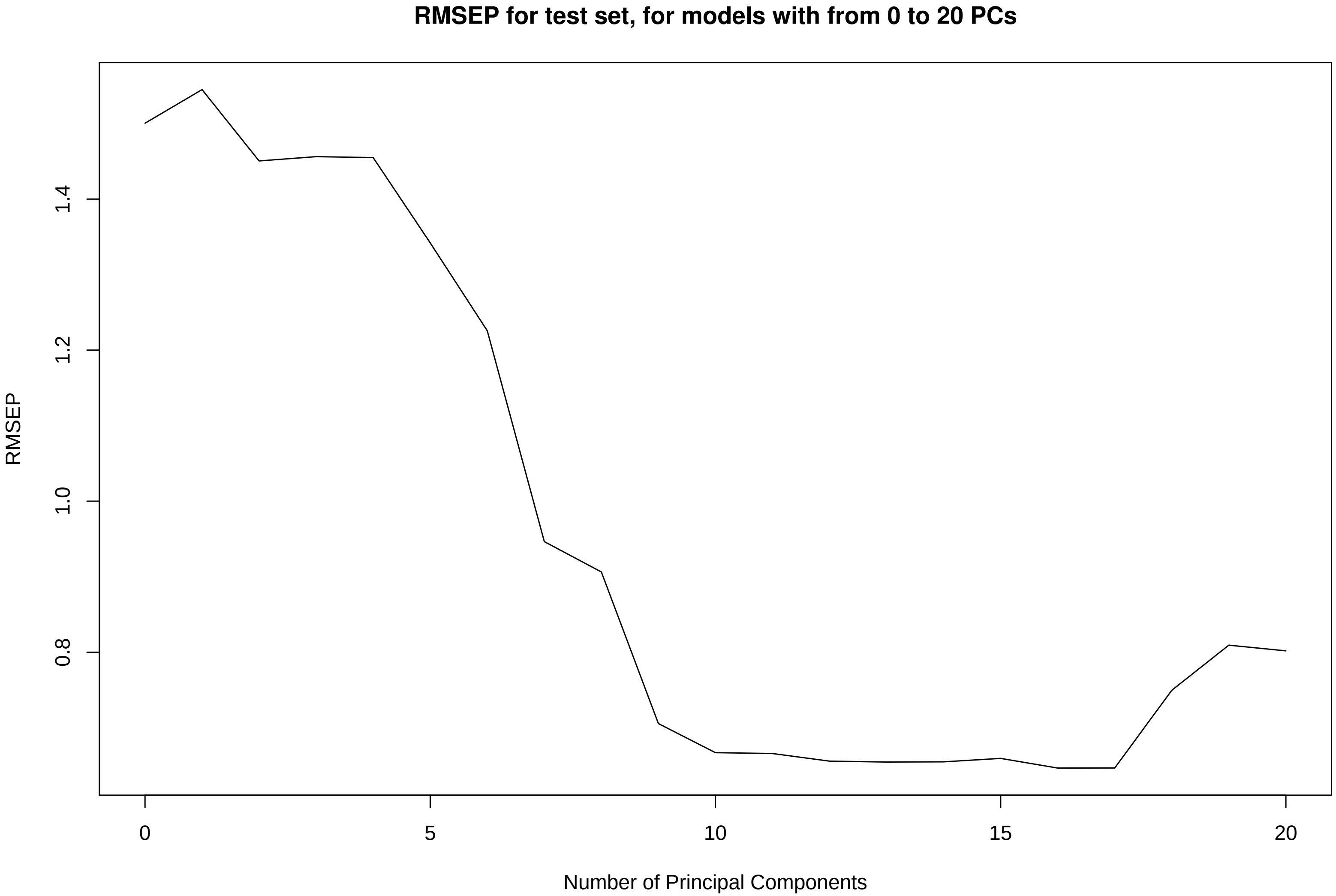
RMSEP for test set, for models with from 0 to 20 PCs
RMSEP
1.4
1.2
1.0
0.8
0
5
10
15
20
Number of Principal Components

Figure 3

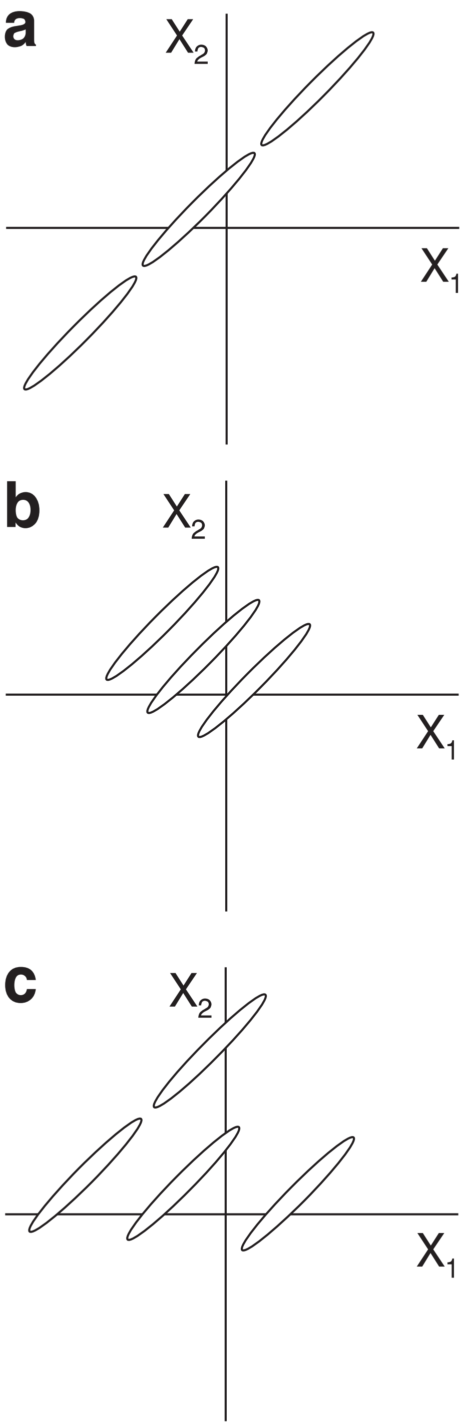

Figure 4

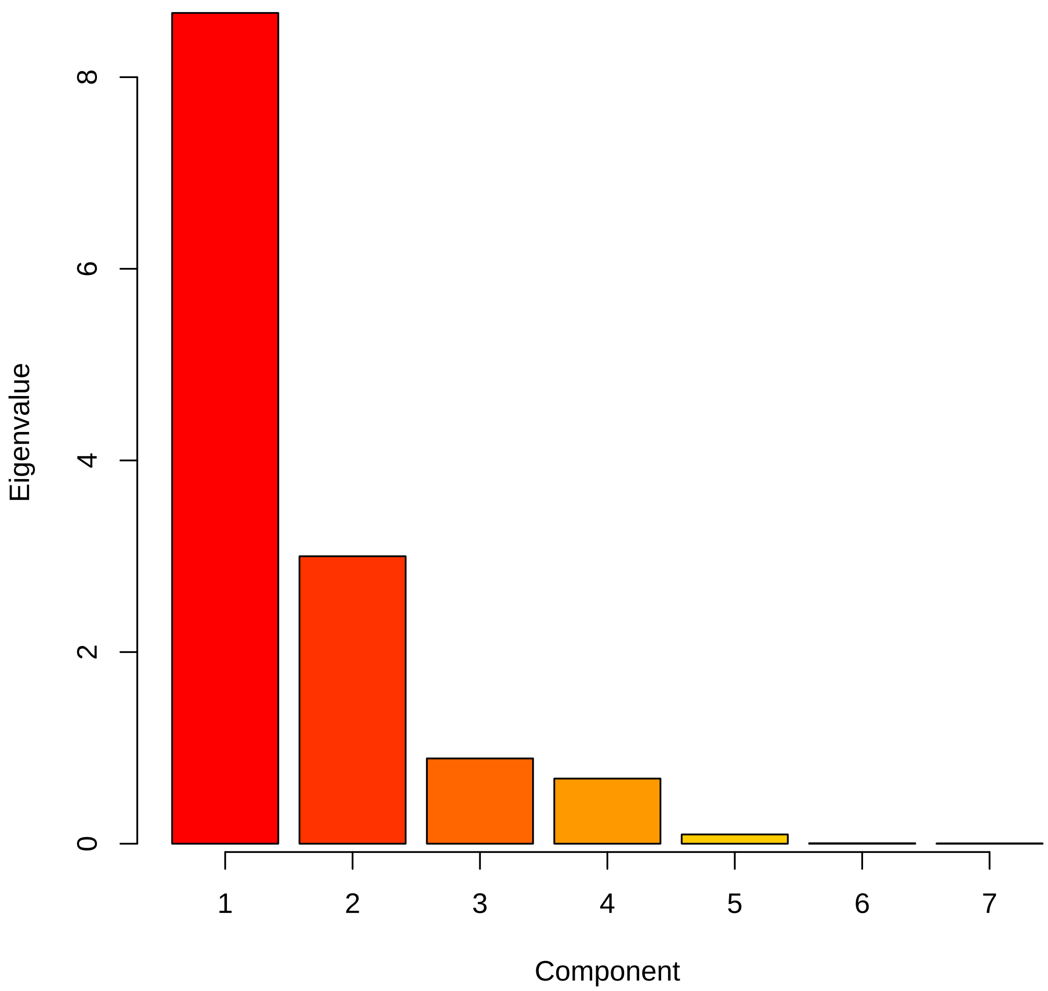

Figure 5