Statistical data processing in clinical proteomics
Berkenbos-Smit, S.

Citation for published version (APA):

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: http://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
Chapter 7

Enhancing classification performance: covariance matters†

An important factor in classifying proteomics samples is the choice of the classification model. Many methods exist and the question of which is best suited is not easily answered. Studies that try to answer this question by comparing a range of methods give different answers because the answer changes with the data. The choice of a certain classification method should be guided by measures such as sample-to-feature ratio, and assumptions about the separability of the classes and covariance structure. In this paper we present an example of a classification problem for which two commonly used methods, Support Vector Machines and Principal Component Discriminant Analysis (PCDA), are unable to create a good classifier. The reason for this is the position of the classes: they are not disjoint (they overlap). Because the within-class covariances are very different, Soft Independent Modelling of Class Analogy (SIMCA) is able to distinguish between the classes, using the residuals from the classes’ PCA models. The difference between PCDA and SIMCA, two seemingly similar methods, can be seen in the metrics they use. Although they can be expressed in a similar fashion, different aspects of the data are stressed, resulting in very different performances. This example shows how choosing a suitable classification method can enhance classification performance.

7.1 Introduction

Proteomics research is an important tool for the discovery of biomarkers. The change in protein composition in blood, urine, tissue or other samples caused by a disease can be measured and used for diagnosis. Multivariate classification methods are used to find differences between patients and healthy subjects that can predict new samples. Many classification methods exist and the question arises which one is most suitable. Several studies have tried to answer this question by comparing the performance of a range of classification methods on one data set. These studies each give a different answer, showing that the best method does not exist. The choice for a certain classification method is guided by assumptions about the data structure: e.g. the sample-to-feature ratio, (linearly) separable classes, and equal class covariances. The success of a method likely depends in part on the experience of the data analyst in preprocessing the data and setting the parameters.

Classification performance depends on the match between data and the classification method. In this paper we give an example of a data set that is poorly classified by two commonly used methods, Principal Component Analysis followed by Linear Discriminant Analysis (PCDA) and Support Vector Machines (SVM), while Soft Independent Modelling of Class Analogy (SIMCA) performs much better. The data are serum protein profiles of recovering and relapsing cervical cancer patients. The detailed SIMCA results show the characteristics of this data set that cause PCDA to fail while SIMCA, using a different metric, is successful.

7.2 Data set

Four groups of patients were included in this study, characterized by the stage of cancer before treatment, the value of the SCC-ag marker, and the disease status 6 to 12 months after treatment, see Table 7.1. Serum samples were taken before and 6 to 12 months after treatment. The serum was depleted of the six most abundant proteins after which the remaining proteins were measured with LC-MS after trypsin digestion. A more detailed description of the samples and LC-MS measurements is given in Chapter 6. The preprocessing was performed for the four groups simultaneously, enabling comparison among the different groups. In the m/z mode the data were reduced. Peak picking
Table 7.1: Overview of patient groups in the recovery and relapse classes.

<table>
<thead>
<tr>
<th>Class</th>
<th>Stage</th>
<th>SCC-ag</th>
<th>group size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recovery</td>
<td>Group I</td>
<td>early</td>
<td>high</td>
</tr>
<tr>
<td>Recovery</td>
<td>Group II</td>
<td>early</td>
<td>low</td>
</tr>
<tr>
<td>Recovery</td>
<td>Group III</td>
<td>advanced</td>
<td>high</td>
</tr>
<tr>
<td>Relapse</td>
<td>Group IV</td>
<td>various</td>
<td>high</td>
</tr>
</tbody>
</table>

was performed using a geometrical algorithm based on the local slope, scanning for local maxima. After determining the peaks and their height, the chromatograms were time aligned (with correlation optimized warping\textsuperscript{153}). The peak matrix, created from the peak lists of the individual samples, was used for multivariate statistical analysis. The samples of the four patient groups were measured on four different days. Our objective is to discriminate between patients that recover (Groups I, II and III) and patients that relapse (Group IV). Therefore, all LC-MS profiles were preprocessed together to allow comparison between the groups. During preprocessing, missing peak locations were filled with zeros. The data were auto scaled to zero mean and unit variance. In cross validation procedures the training data were auto scaled and the test data were scaled with the parameters from the training set. The preprocessing procedure introduced many zeros in the data. This causes problems during auto scaling, especially when autoscaling is performed inside the cross validation loop, necessitating discarding variables with a large percentage of zeros. This is further discussed with the classification results.

7.3 Classification methods: PCDA, SVM, SIMCA

PCDA is a combination of Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA). First, PCA is applied to the data to reduce the dimensionality. The PCA scores are then used in LDA to find a direction that discriminates between the two groups.\textsuperscript{30} The number of components in the PCA step is regulated by internal cross validation to minimize the number of misclassifications as described in Chapter 3. SVM\textsuperscript{39} constructs a hyperplane to separate the classes. Using a kernel function, it is possible to construct a hyperplane in an enlarged feature space, without explicitly transforming the
data. We use the SVM classifier from the Matlab Bioinformatics Toolbox [2.1.1] with commonly used kernels. The basis of the SIMCA\textsuperscript{156} method is PCA. Each class is modelled separately, resulting in a set of principal components per class. The SIMCA classification is made based on two criteria: 1) the object should be close to the model plane (Q statistic), and 2) the object’s projection should be close to the class mean (Mahalanobis distance). A new sample is then assigned to the class which fits best. The PLS Toolbox 3.0 (Eigenvector Research) was used to create the SIMCA model. We set the number of components to minimize the misclassification rate in cross validation, and we use the same number of components in both classes.

7.4 Results

Preliminary analysis

PCA on the complete data gives clear discrimination between the scores on the first two principal components of the ‘recovery’ class and the ‘relapse’ class, see Figure 7.1 A. We cannot distinguish however, whether this result reflects differences between patient classes or that we see an effect of measuring the groups of patients on different days. We can look at the development of the patients instead by subtracting the patient’s profile before treatment from the profile after treatment. Since the profiles of a patient are measured in the same run, they should both contain the same contribution, if present, from the time of measuring. Subtracting one profile from the other removes this contribution if it is linear. Indeed, Figure 7.1 B shows no obvious differences between relapsing and recovering patients in the first two principal components. We therefore use the difference profiles in our further analysis.

Classification

We used PCDA on the difference profiles to classify the recovery and relapse groups and tested the classification performance on new samples in tenfold cross validation. In Chapter 6 we were able to find discriminating patterns with PCDA between profiles before and after treatment of the recovering patients in Group III. However, in the current classification problem PCDA gives poor results, misclassifying 28% (p-value = 0.12) of the profiles. Classification
with SVM with linear, polynomial, quadratic and radial basis function kernels does not result in a better performance. With SIMCA however, we obtain a much better result, with only 11% misclassifications (p-value = 0.001). To avoid problems during auto scaling we removed all variables that were zero for one or more samples. With the removal of all zeros from the data we only look at those variables which are present in all samples. Alternatively, we explored allowing a certain percentage of zero entries for each variable. This did not change the results for PCDA and SVM much, and SIMCA improves slightly when the number of zeros was larger.

Can we explain the difference in performance of PCDA and SVM versus SIMCA? In PCDA and SVM a classification boundary is created and new objects are classified depending on which side of the boundary they are on. This means that it has to be possible to draw a line in (extended) variable space that separates the classes with some degree of success. In other words, for PCDA and SVM to be successful, the classes have to be disjoint. In the next section we take a more detailed look at the SIMCA model to find out why it performs better. After that a mathematical description is given of PCDA and SIMCA, fitting them in one framework.
SIMCA model

In this section we discuss the results for a one component SIMCA model (the classes are described by a one-component model each) using all data to fit the model (no cross validation was used). In Figure 7.2 we see the scores for the models on class ‘recovery’ (A) and ‘relapse’ (B). The scores for the two classes in both models overlap and there can be no discrimination based on these scores. The result is that, based on the $T^2$ statistic, almost all samples fit in both classes. This is in line with the poor performance of PCDA and SVM. In contrast, the residuals plots in Figure 7.3 and Figure 7.4 do show a difference between the classes. Most recovery class objects have smaller residuals than relapse objects in the recovery model and most relapse objects have smaller residuals in the relapse model. This seems enough to discriminate between the two classes. In line with this reasoning we observed that adding more PCs to the SIMCA model increases the misclassification rate: with more of the data explained in the PC model, the residuals become smaller and lose discriminating power. The relative (to their 95% boundaries, given as output of the SIMCA model in the PLS Toolbox) $Q$ and $T^2$ of the predicted objects are given in Figure 7.5. Based on relative $T^2$ (Figure 7.5 A) the samples would almost randomly be assigned to a class, but based on the residuals (Figure 7.5 B), the objects are assigned to the correct classes. Since the relative $Q$ statistic is larger than the relative $T^2$ for all objects, it determines the classification. Originally, Wold used only the $Q$-statistic. He suggested looking also at the scores, but only as an indication of misfit. Applying this to the current data does not change our results. This is in accordance with our finding that for this data set the $Q$-statistic dominates the $T^2$ statistic.

Data structure

In the SIMCA model the loadings of the two classes, each centered on its own class mean, are at a 65° angle. Furthermore, we observed that in the original variable space the class centres lie relatively close together, closer in fact than any of the subjects to the centre of the class they belong to. Figure 7.6 shows schematically how the classes could be positioned, given the scores, residuals, angle of the models and the fact that the class centres lie relatively close to each other. When projected onto the model for the circles class, the scores for the stars do not differ from the scores for the circles. The star residuals, however, are on average larger than the residuals of the circle objects. The same holds,
**Figure 7.2:** Fitted scores for recovered patients (*) and relapsed patients (o) on the SIMCA models A. of the recovery class and B. of the relapse class.

mutatis mutandis, when the circles are projected onto the stars model. The SIMCA results show that the within-class covariances are very different, an attribute of the data that is responsible for the success of SIMCA compared to PCDA and SVM. Because the covariances are different, the residuals of one class are larger when projected onto the loading of the other class. We showed in our discussion of the SIMCA model that it uses mainly the residuals for classification. This distinguishing characteristic of the data is ignored when PCDA and SVM are used for classification.

**SIMCA and PCDA distance metrics**

In some respects PCDA and SIMCA are similar methods. Both use a distance measure in a PCA loading space to classify samples. In this section we show that while PCDA and SIMCA are similar and fit in one framework, the metric they use is different. They stress different aspects of the data, and the loading spaces they operate in are not the same. In the following, the measurements of m variables on N samples are collected in matrix X \((N \times m)\). For clarity the X is assumed to be mean centered, however it does not affect the results if it is not and mean centering of X is done as a first step in PCDA.
Figure 7.3: A. Residuals for recovered patients (⋆) and relapsed patients (o) on Recovery model. Panel B gives a zoomed in view.

Figure 7.4: A. Residuals for recovered patients (⋆) and relapsed patients (o) on Relapse model. Panel B gives a zoomed in view.
Figure 7.5: A. Relative $T^2$ and B. relative $Q$ for predicted objects for the recovery class (*) and the relapse class (o). The first thirty objects show the recovered patients, the last twelve the relapsed patients.

Figure 7.6: Suggestion of data fitting our findings. The models are at a 65° angle.
In SIMCA, a PCA model is constructed on the mean-centered data $X_i$ of each class separately

$$X_i = T_i P_i' + E_i$$

(7.1)

where $T_i (N_i \times q)$ contains the scores, and $P_i (m \times q)$ the loadings for the first q PCs. Subscript $i$ is the class index; in the current problem $i=1,2$. The matrix $E_i$ is the error or residual matrix. A new sample $x_{\text{new}}$ is centered with respect to the class mean, $\bar{x}_i$:

$$\tilde{x}_{\text{new},i} = x_{\text{new}} - \bar{x}_i$$

(7.2)

The score $\tilde{t}_{\text{new},i}$ of the new sample for class $i$ is

$$\tilde{t}_{\text{new},i} = P_i' \tilde{x}_{\text{new},i}$$

(7.3)

and the projection $\hat{x}_{\text{new},i}$ of the new sample on the class model is

$$\hat{x}_{\text{new},i} = P_i P_i' \tilde{x}_{\text{new},i}$$

(7.4)

Classification in SIMCA is based on a combination of two distance measures. The first is the Mahalanobis distance in the PCA loading space of the sample to the class centre ($T_2$ statistic)

$$T_i^2 = (\tilde{t}_{\text{new},i} - \bar{t}_{\text{SIMCA},i})' \left( \frac{1}{N_i - 1} T_i T_i' \right)^{-1} (\tilde{t}_{\text{new},i} - \bar{t}_{\text{SIMCA},i})$$

(7.5)

which, since the class is mean-centered ($\bar{t}_{\text{SIMCA},i} = 0$) equals

$$T_i^2 = \tilde{t}_{\text{new},i}' \left( \frac{1}{N_i - 1} T_i T_i' \right)^{-1} \tilde{t}_{\text{new},i} = \tilde{x}_{\text{new},i}' S_i^{-1} P_i' \tilde{x}_{\text{new},i}$$

(7.6)

where $S_i$ is the sample covariance matrix for class $i$. In the original variable space, the Euclidian distance of the sample to the model plane (the $Q$ statistic) is considered:

$$Q_i = (\tilde{x}_{\text{new},i} - \hat{x}_{\text{new},i})' (\tilde{x}_{\text{new},i} - \hat{x}_{\text{new},i})$$

$$= (\tilde{x}_{\text{new},i} - P_i P_i' \tilde{x}_{\text{new},i})' (\tilde{x}_{\text{new},i} - P_i P_i' \tilde{x}_{\text{new},i})$$

$$= \tilde{x}_{\text{new},i}' (I - P_i P_i') \tilde{x}_{\text{new},i}$$

(7.7)

Both statistics are considered relative to their 95% confidence limits ($Q_i \lim$ and $T_i^2 \lim$) that are established in the training phase and combined in one
‘distance’ measure, \( D_{SIMCA} \):

\[
D_{SIMCA} = \frac{1}{T_i^{2,lim}} \left( \tilde{x}_{new,i} - \bar{x}_i \right)' \left( P_i S_i^{-1} P_i' - I \right) \left( \tilde{x}_{new,i} - \bar{x}_i \right) + \frac{1}{Q_i^{lim}} \left( \bar{x}_{new,i} - \bar{x}_i \right)' \left( I - P_i P_i' \right) \left( \bar{x}_{new,i} - \bar{x}_i \right)
\]

where \( W_{SIMCA,i} \) contains the weights for the squared distance in SIMCA. This is similar to the combined index Yue and Qin introduced for process monitoring\(^{157}\) where process faults are detected and identified.

Classification in PCDA begins with a PCA model of the entire centered data set:

\[
X = TP' + E \tag{7.9}
\]

with the scores \( T \) of size \((N \times q)\) and the loadings \( P \) of size \((m \times q)\). The score and projection of a new sample are given by:

\[
t_{new} = P'x_{new} \tilde{x}_{new} = PP'x_{new} \tag{7.10}
\]

Now classification is performed on the scores, where the sample is assigned to the class for which the \( T^2 \) statistic is lowest:

\[
T^2 = (t_{new} - \bar{t}_i)' S^{-1} (t_{new} - \bar{t}_i) = (P'x_{new} - P'\bar{x}_i)' S^{-1} (P'x_{new} - P'\bar{x}_i)
\]

\[
= (x_{new} - \bar{x}_i)' PS^{-1} P' (x_{new} - \bar{x}_i) = \tilde{x}_{new,i}' W_{PCDA} \tilde{x}_{new,i} \tag{7.11}
\]

where \( S \) is the pooled sample within-class covariance

\[
S = \frac{1}{N-1} T'T \tag{7.12}
\]

and \( W_{PCDA} \) contains the weights for the squared distance in PCDA.

This shows that the SIMCA and PCDA statistics fit in the same format of a weighted squared distance. The weights are different for SIMCA and PCDA, with PCDA looking only at the Mahalanobis distance in the loading space, whereas SIMCA also considers the distance of the sample to the model plane. Note that the model spaces in SIMCA and PCDA are not the same. SIMCA
builds local models for each class, but PCDA builds one PCA model for the entire data, hence the scores and loadings matrices in $W_{SIMCA,i}$ and $W_{PCDA}$ are different. This resembles the observation of Mertens et al. that LDA (without the PCA step) and SIMCA each use a differently weighted Mahalanobis distance. The LDA and SIMCA statistics are both obtained in the original variable space, whereas the PCDA statistics are obtained in the PCA loading space.

7.5 Conclusions

In this chapter we present a classification problem of two classes whose between-class covariance is small compared to the within-class covariance. With PCDA and SVM we are unable to obtain a good classification, because these methods require disjoint classes. SIMCA performs much better, relying on the PCA residuals for classification. The SIMCA results show that the covariance structures of the classes are very different. The difference between PCDA and SIMCA can be seen in the metrics they use. Although they can be expressed in a similar fashion, different aspects of the data are stressed. This results in very different performances.

In this chapter we showed that tuning the choice of a classification method to the structure of the data can generate much better results. Nevertheless, it may not be possible to determine which kind of method is best suitable beforehand. In that case, care should be taken not to move from data mining to ‘method mining’: trying many different classification methods in the hope one will deliver an acceptable performance. This can be avoided by careful validation.