*Metabolomics, Vol. 2, No. 2, June 2006 (© 2006)*
DOI: 10.1007/s11306-006-0022-6


# Assessing the performance of statistical validation tools for megavariate metabolomics data

Carina M. Rubingh,[a,*] Sabina Bijlsma,[a] Eduard P. P. A. Derks,[a] Ivana Bobeldijk,[a] Elwin R. Verheij,[a] Sunil Kochhar,[b] and Age K. Smilde[a]

[a]*Business Unit Analytical Sciences, TNO Quality of Life, P.O. Box 360 3700, AJ Zeist, The Netherlands*
[b]*BioAnalytical Science Department, Nestlé Research Center, P.O. Box 44 CH-1000, Lausanne 26, Switzerland*

Received 13 January 2006; accepted 22 March 2006

Statistical model validation tools such as cross-validation, jack-knifing model parameters and permutation tests are meant to obtain an objective assessment of the performance and stability of a statistical model. However, little is known about the performance of these tools for megavariate data sets, having, for instance, a number of variables larger than 10 times the number of subjects. The performance is assessed for megavariate metabolomics data, but the conclusions also carry over to proteomics, transcriptomics and many other research areas. Partial least squares discriminant analyses models were built for several LC-MS lipidomic training data sets of various numbers of lean and obese subjects. The training data sets were compared on their modelling performance and their predictability using a 10-fold cross-validation, a permutation test, and test data sets. A wide range of cross-validation error rates was found (from 7.5% to 16.3% for the largest trainings set and from 0% to 60% for the smallest training set) and the error rate increased when the number of subjects decreased. The test error rates varied from 5% to 50%. The smaller the number of subjects compared to the number of variables, the less the outcome of validation tools such as cross-validation, jack-knifing model parameters and permutation tests can be trusted. The result depends crucially on the specific sample of subjects that is used for modelling. The validation tools cannot be used as warning mechanism for problems due to sample size or to representativity of the sampling.

**KEY WORDS:** metabolomics; megavariate data; PLS-DA; cross-validation; permutation test; predictability; jack-knife.

## 1. Introduction

Metabolomics studies are performed to investigate responses of biologic systems on environmental influences due to, for instance, toxicological exposure, nutrition or medical treatment. In this field, metabolites in biological samples like plasma or urine are analytically determined using techniques such as nuclear magnetic resonance (NMR; Derome, 1987), liquid or gas chromatography mass spectrometry (LC-MS and GC-MS, respectively; Wilson *et al*., 2005; Lenz *et al*., 2004; Lafaye *et al*., 2003; Plumb *et al*., 2003; Van der Greef *et al*., 2003; Fiehn, 2002). These analytical techniques can generate a large amount of data containing information about a large number of correlated variables, which asks for appropriate statistical tools for data analysis. Multivariate data analysis (MVA) is used to analyze the correlated data. MVA can be used to summarize the data by reducing the dimensions of the data, for regularization purposes, for variable selection, etcetera. One of the applications of MVA is to use correlations and trends in the data in order to discriminate between groups (Massart *et al*., 1997; Vandeginste *et al*., 1998).

* To whom correspondence should be addressed.
E-mail: Rubingh@voeding.tno.nl

Discriminant analysis (DA) is a MVA method that can be used if the interest is focused on differences between groups of objects or on subgroup structures and can serve two slightly different purposes. If it is used to separate distinct sets of objects or observations, discrimination is the main purpose. If it is used to define classification rules to allocate new objects or observations to previously defined groups, it is used for classification (Vandeginste *et al*., 1998).

However, the results found in DA cannot always be trusted as they are sensitive to chance-correlations and/or to the risk of overfitting. Validation tools like cross-validation (Stone, 1974; Martens and Naes, 1989; Hastie *et al*., 2001), permutation tests (Efron and Tibshirani, 1993; Manly, 1997; Good, 2000; Mielke and Berry, 2001), jack-knifing model parameters (Efron, 1982; Martens and Martens, 2000) and test data sets are used to address these problems and provide an objective assessment of the performance and stability of a model. These tools are commonly used to validate the results of multivariate data analyses. When multivariate data become megavariate data, the number of variables is even larger and, due to the curse of dimensionality, the chance of false correlations and the risk of overfit is even higher. In the present study, a data set having a number of variables larger than 10 times the number of subjects is defined to be megavariate. The validity of

cross-validation for small-sample classification was assessed under low dimensionality (Martens and Dardenne, 1998; Braga-Neto and Dougherty, 2004), but still little is known about how validation tools such as cross-validation, jack-knifing and permutation tests will perform for megavariate data.

Cross-validation is used to choose the optimal model parameters as well as to test the predictability of the statistical model. Cross-validation uses the available data minus a particular part (e.g. $1/k$-th part of the total data set) to fit the model and the part that was left out to test the model (Hastie *et al*., 2001). However, the predictability based on a single cross-validation is biased and often too optimistic because the determination of the model meta-parameters (e.g. number of latent variables or any regularization term) is based on the same set as is used to determine the predictability. Hence, still a separate test set is required to determine the predictability for future data. This problem can be addressed by double cross-validation, which makes most efficient use of the available data as all objects are used for model building and validation (Stone, 1974).

The stability of model parameters is assessed by the jack-knife procedure. All available data minus the data of one (or more) objects is used to fit the model and for each perturbed set, the parameters estimates can be obtained. A graphical presentation or an evaluation of (relative) standard deviations of the estimates gives an impression of the stability of the estimates (Efron, 1982; H. Martens and M. Martens, 2000).

A permutation test is used to assess the significance of a classification. The class assignment can be permuted several times and for each permutation, a model between the data and the permuted class-assignment can be built. The discrimination between classes of the model based on the permutated class-assignment is compared to the discrimination of the model based on the original classification (Efron and Tibshirani, 1993; Manly, 1997; Good, 2000; Mielke and Berry, 2001).

The classification of a test data set using the model-parameters based on the training data set, provides information about the generalizability of a model; whether the model is only applicable for the subjects in the training set or whether it can also be used to predict the classification of new subjects. All these tools can be used to prevent that conclusions about the discrimination between classes may be drawn, which cannot be statistically supported.

In order to assess the performance of statistical validation tools for megavariate data sets, several data sets of various sizes, all derived from the same original data set of human LC-MS lipidomic data, are compared on their modelling performance and their predictability. These data were obtained from a co-operative metabolomics study of TNO, Nestlé Research Centre (Lausanne, Switzerland) and the EU NUGENOB project (NUGENOB is the acronym of the project 'Nutrient-Gene interactions in human obesity – implications for dietary guidelines' supported by the European Community (Contract no. QLK1-CT-2000-00618), see the web-site http://www.nugenob.com; Petersen *et al*., 2005; Blaak *et al*., 2006). The main objective of this metabolomics study was to find biomarkers that characterize differences between high and low fat burners in lean and obese subjects. A strategy for data preprocessing, data analysis and validation of statistical models was also developed (Bijlsma *et al*., 2006). The present study was performed in order to investigate the effect of decreasing the number of subjects on the performance of the statistical validation tools. Although metabolomics data were used for the analyses, the issue also carries over to proteomics and transcriptomics data.

## 2. Materials and methods

### 2.1. Data

#### 2.1.1. General

Although real-life data may lead to less distinguishing differences between sets, it was preferred above simulated data because it illustrates the problems researchers have to deal with best. Both biological and analytical variations are present in the data and may be of influence on the results. Data of a co-operative metabolomics study of TNO, Nestlé Research Centre (Lausanne, Switzerland) and the EU NUGENOB project were used. This study involved plasma from 50 lean and 100 obese human subjects, collected at four different time points ($t$ = 0, 1, 2, and 3 h) after a single intake of a fat rich meal. All samples were analysed using four analytical platforms: NMR, GC-MS, LC-MS polar and LC-MS lipid. Details about the study design can be found in Petersen *et al*. (2005), whereas details about the data and data preprocessing can be found in Bijlsma *et al*. (2006). The data set used in the present study was based on the LC-MS lipid data measured at $t = 0$ h, which contained 947 LC-MS peaks (variables).

#### 2.1.2. Base data set

The focus was on differences between lean and obese subjects in the LC-MS lipid. In order to avoid confounding of the results due to an unbalanced number of lean and obese subjects, a random selection of 50 out of the 100 obese subjects was made. The created data set of 50 obese and 50 lean subjects was used as base data set (*data50:50*). Subsets of the *data50:50* were used to study the effect of the decrease of the number of subjects on the analysis and validation results.

#### 2.1.3. Data subsets

A data set was generated containing the data of 40 lean and 40 obese subjects (*data40:40*). The inclusion of a subject into the *data40:40* data set was based on random selection without replacement. The creation of the

*data40:40* set was repeated 10 times, each based on a new random selection of the original *data50:50* set (base data set). This process was repeated for 10 sets of 30 lean and 30 obese subjects (*data30:30*), for 10 sets of 20 lean and 20 obese subjects (*data20:20*), for 10 sets of 10 lean and 10 obese subjects (*data10:10*) and finally, for 10 sets of 5 lean and 5 obese subjects (*data05:05*), all based on random selection out of the *data50:50* base data set. Although additional information about the subjects such as being a high or low fat burner or the center at which the sample was collected (Petersen *et al.*, 2005), was not used in the statistical analysis, equal representation of these factors over the created subsets was secured.

The subset data sets were used for modelling and were used as so called training data sets. The data of the remaining subjects were used as test data sets. The test data set of the *data40:40* set contained the data of the remaining 10 lean and 10 obese subjects, the test data set of the *data30:30* set contained the data of the remaining 20 lean and 20 obese subjects, the test data set of the *data20:20* set contained the data of the remaining 30 lean and 30 obese subjects, the test data set of the *data10:10* set contained the data of the remaining 40 lean and 40 obese subjects, and the test data set of the *data05:05* set contained the data of the remaining 45 lean and 45 obese subjects. As a consquence of this procedure, the size of the test sets differ. To rule out the possible effect of the test data set size, an extra test set, based on a random selection without replacement of the data of 10 obese and 10 lean subjects, was also created for each subset. Summarizing, one base data set was generated as well as 50 training sets (5×10) and 50 test sets (5×10) and 50 extra test sets (5×10). The procedure that was followed to obtain all data sets, is illustrated in figure 1. This procedure was chosen to mimic reality, in which very few samples are available for data analysis. The *data05:05* may be unrealistically small for human studies, but was incorporated for illustrative purposes.

### 2.2. Statistical analysis

Partial least squares discriminant analysis (PLS-DA; Vong *et al.*, 1988; Barker and Rayens, 2003) was used to find a small number of linear combinations of the original variables (called 'latent variables'; LVs), that was predictive for the class membership and that described most of the variability of the LC-MS metabolic profiles. PLS-DA is a linear regression method whereby the multivariate variables (the X-block) corresponding to the observations are related to the class membership (the Y-Block) for each subject. The Y-block contains "1" and "0" only, corresponding to the lean and obese class assignment. It is a classical PLS regression (Geladi and Kowalski, 1986; Martens and Naes, 1989; Massart *et al.*, 1997; Vandeginste *et al.*, 1998) where the response is a categorical one expressing the class membership of a subject. PLS-DA will maximise the covariance between the predicting data set (X block with LC-MS metabolomic profiles) and the data to be predicted (Y-block with class assignments).

Data were mean-centered before analyses. The center-parameters of the training set were used to transform the corresponding test data set. Details about other

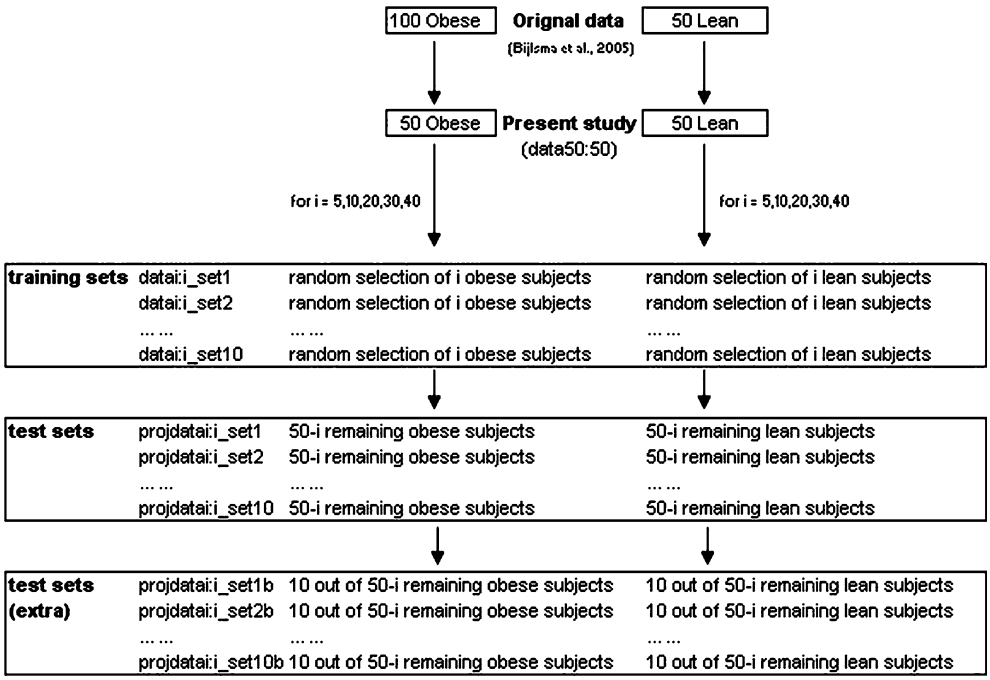


Figure 1. Illustration of the procedure that was followed to obtain the data sets.

aspects of the data pre-processing can be found in Bijlsma *et al*. (2006). All analyses were performed using Matlab Version 7.0.4 R14 (The Mathworks, Inc.) and the PLS Toolbox Version 3.0.4 (Eigenvector Research, Inc.).

### 2.3. Statistical model validation

#### 2.3.1. Cross-validation

The use of a double cross-validation would be preferred (Stone, 1974), because a single cross-validation may lead to bias and overestimation of the true error rate (Hastie *et al*., 2001). However, the issue of bias is in this case of less importance, because only the error rates are compared and it is assumed that the bias in each model is similar. For this reason and for the fact that in case of very small data sets a double cross-validation becomes less appropriate, a single cross-validation was used instead of a double cross-validation. A single 10-fold venetian blind cross-validation based on stratified sampling having the lean and obese class membership as strata, was used to choose the optimal number of LVs as well as to obtain an estimate of the error rate of the PLS-DA model. In the first cross-validation step, 1/10-th of a training data set was left out, under the restriction that the number of lean subjects that was left out was equal to the number of obese subjects that was left out, and data of the remaining subjects were used to build a PLS-DA model. The model was used to predict the class assignment of the ''left out'' subjects. This was repeated until all subjects were left out once. The number of LVs yielding the lowest percentage of misclassifications (error rate) was chosen as the optimal model. Note that by using a 10-fold cross-validation for *data05:05*, only 1 subject is left out at each step of the cross-validation. Hence in this case, the 10-fold cross-validation is equal to a ''leave-one-out'' cross-validation.

#### 2.3.2. Jack-knife

The stability of the regression coefficients of the PLS-DA models was assessed by jack-knifing (Efron, 1982; H. Martens and M. Martens, 2000). In order to be able to use the same data set parts as was used in cross-validation, all available data minus 1/10-th was used to fit the model, instead of leaving-out-one observation per jack-knife step which is a more usual way of jack-knifing. In the first jack-knife step, 1/10-th of a training data set was left out, under the restriction that the number of lean subjects that was left out was equal to the number of obese subjects that was left out, and data of the remaining subjects were used to build a PLS-DA model. This was repeated until all subjects were left out once. The 10 variables having the largest coefficient in the reference model *data50:50* were evaluated graphically using Box-and-Whisker-plots.

#### 2.3.3. Permutation test

Cross-validation can be used to assess the class-predictability of a model. In order to asses the discrimination, an exact or an approximate permutation test can be used (Efron and Tibshirani, 1993; Manly, 1997; Good, 2000; Mielke and Berry, 2001).

The class assignment was permuted in such a way that the ratio between the number of lean (''0'') and obese (''1'') subjects remained equal, and this was done 1000 times with replacement of the class vector. As an exact permutation test would lead to too many combinations, an approximate permutation test was performed on each of the data sets. For the *data05:05* subset, only 100 permutations of the Y-block were performed, because the number of possible permutations is much lower than 1000. For each permutation, a PLS-DA model was built between the X-block and the permuted Y-block using the same optimal number of LVs as determined by cross-validation for the model based on the original class assignment. The ratio of the between sum of squares and the within sum of squares (B/W-ratio) for the class assignment prediction of each model was calculated. If the B/W-ratio of the original class assignment is a part of the distribution based on the permuted class assignments, the contrast between the two classes cannot be considered to be significantly different from a statistical point of view. If, on the other hand, the B/W-ratio based on the original class assignment is much higher compared to the ratios based on the permuted class assignments, the differences between the classes are statistically significant. Because exact accuracy percentages are not important in the scope of this paper, the permutation test is evaluated visually according to figure 2 (Bijlsma *et al*., 2006).

#### 2.3.4. Predictability

Cross-validation, jack-knifing and the permutation test provide information about the validity of the model based on the information in the training data set. The generalizability suggested by the cross-validation error

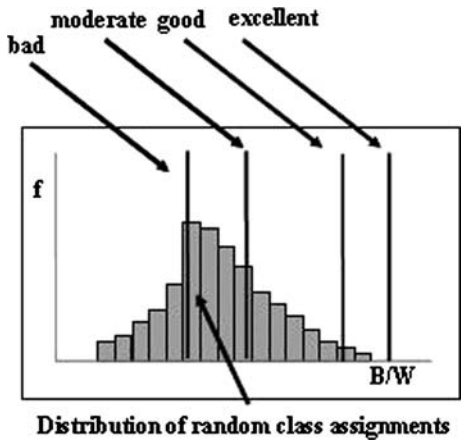


Figure 2. Visual evaluation of the permutation test.

rate was assessed by the prediction of the class assignment of new subjects, which are in this case defined as the subjects in the test data sets. The class assignment prediction of the subjects in these test data sets was determined based on the model parameters of the corresponding training data set. Hence, the prediction was based on the same (number of) LVs as was used for the training set. The error rate of the test data set, being the percentage of misclassified subjects, was calculated and was used as measure for the generalizability of the model. Ideally, the test error rates are comparable to the ones found by 10-fold cross-validation.

## 3. Results and discussion

### 3.1. Training sets

The results of the PLS-DA model for the *data50:50* are presented in figure 3. As this model is based on all lean and obese subjects, this model is considered to be the reference model. For *data50:50*, the cross-validation error rate of the model is 11% (0.11 in figure 3a) based on 11 LVs and is shown in figure 3a. Figure 3b shows the prediction based on the cross-validation for the lean (first 50; marked as 'o') and the obese (second 50; marked as '*') subjects. The overlap between the two classes shown in this sub-figure corresponds to the error rate of 11%. In figure 3c the final fit is shown, which is much more optimistic compared to the prediction based on cross-validation. Finally, in figure 3d the jack-knife results for the 10 largest regression coefficients is given. The results in figure 3 are similar to the results found by Bijlsma *et al.* (2006) in the analyses on the data set based on 100 obese and 50 lean subjects.

A summary of the results of the PLS-DA models based on all training sets is given in table 1. Per data set and per model, the error rate based on the 10-fold cross-validation, the number of used LVs and the evaluation of the permutation test are given. Also the mean and standard deviation of the error rate and the mean number of LVs per data set are presented.

The mean cross-validation error rate and the variance of the error rate both increase if the number of subjects in the data set decreases. The results of the analysis of the *data05:05* sets are the most variable, showing a range of error rates from 0% for the 10th selection to 60% for the 4th selection. The results of the 4th and the 10th selection of *data05:05* are presented in figure 4A and B. The jack-knife results confirm the above described discrepancy between the conclusions based on both sets of *data05:05*. The 10 largest regression coefficients found in the reference model of *data50:50* were considered to be the most important variables for the discrimination between the two groups and therefore, only

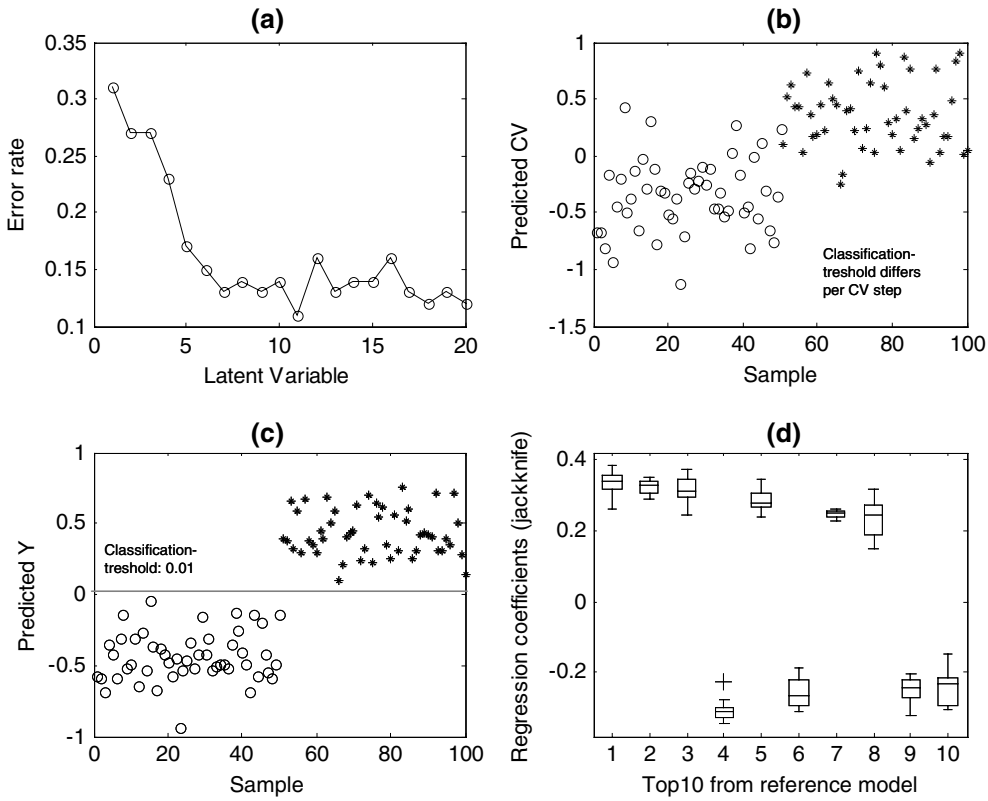


Figure 3. PLS-DA results for data50:50: Cross-validation error rate (a), Prediction based on cross-validation (b; o = lean, * = obese), Prediction based on fit (c; o = lean, * = obese), and Jack-knife (d).

Table 1
Summary of PLS-DA results based on all training sets (ER = cross validation error rate in %, LV = number of latent variables, P = evaluation permutation test with e = excellent, g = good, m = moderate, b = bad)

| Model | Training data set | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 4040 | | | 3030 | | | 2020 | | | 1010 | | | 0505 | | |
| | ER | LV | P | ER | LV | P | ER | LV | P | ER | LV | P | ER | LV | P |
| 1 | 12.5 | 6 | e | 20.0 | 7 | e | 17.5 | 2 | g | 25.0 | 4 | m | 10.0 | 4 | m |
| 2 | 16.3 | 7 | e | 15.0 | 6 | e | 25.0 | 6 | g | 10.0 | 3 | g | 20.0 | 4 | m |
| 3 | 13.8 | 7 | e | 18.3 | 7 | e | 5.0 | 11 | g | 10.0 | 8 | g | 20.0 | 4 | b |
| 4 | 12.5 | 7 | e | 18.3 | 7 | e | 22.5 | 6 | g | 20.0 | 3 | m | 60.0 | 3 | b |
| 5 | 11.3 | 6 | e | 18.3 | 7 | e | 12.5 | 10 | g | 10.0 | 2 | g | 50.0 | 1 | b |
| 6 | 15.0 | 7 | e | 8.3 | 7 | e | 20.0 | 6 | e | 10.0 | 2 | g | 40.0 | 1 | b |
| 7 | 15.0 | 7 | e | 15.0 | 7 | e | 20.0 | 8 | g | 15.0 | 3 | m | 20.0 | 1 | m |
| 8 | 11.3 | 6 | e | 16.7 | 7 | e | 17.5 | 6 | g | 25.0 | 1 | m | 40.0 | 1 | b |
| 9 | 16.3 | 7 | e | 16.7 | 7 | e | 35.0 | 6 | g | 50.0 | 5 | b | 40.0 | 2 | b |
| 10 | 7.5 | 8 | e | 13.3 | 6 | e | 10.0 | 9 | g | 20.0 | 5 | m | 0.0 | 1 | g |
| Mean | 13.1 | 7.0 | | 16.0 | 7.0 | | 18.5 | 7.0 | | 19.5 | 4.0 | | 30.0 | 2.0 | |
| SD | 2.7 | | | 3.4 | | | 8.3 | | | 12.3 | | | 18.9 | | |

these 10 were used to evaluate the jack-knife results. Needless to say, the absolute values presented in figure 4A and B are not comparable to the values presented in figure 3. The coefficients of the 4th selection of *data05:05* show a lot of variation and the coefficients of the 10th selection show only little variation but were almost all equal to zero. This finding confirms that it can be expected that both sets were not representative for the total set of 50 lean and 50 obese subjects.

### 3.2. Test sets

The test data sets were used to determine the generalizability of the models. The number of LVs was based on the number of LVs used for modelling the training data set. The mean and the standard deviation of the test error rate per data set are presented in table 2 and reveal that the predictability of the models based on small training data sets was worse than the predictability of the models based on larger training data sets.

The test error rates varied from 5% to 30% for the test sets corresponding to *data40:40* and from 35% to 50% for the test sets corresponding to *data05:05*. The mean test error rates in table 2 are similar to the cross-validation error rates of the corresponding training data sets in table 1, except for *data10:10*. The standard deviations of the test error rates are less variable compared to the cross-validation error rates of the training data sets presented in table 1.

Although the 10th selection of *data05:05* had a much better cross-validation error rate for the training set (0%) compared to the 4th selection (60%), their test error rate based on the corresponding test set is similar (both 50%). The results of the extra test data sets of 10 obese and 10 lean subjects are also presented in table 2. Although the mean levels of the error rates are similar, the rates are more variable compared to the original test data sets, due to the smaller size of the extra test data sets.

### 3.3. Discussion

The results are predominantly driven by the size of the training data set and the selection of the subjects in that data set, which is especially illustrated by the smaller training data sets. The mean cross-validation error rate increases as the number of subjects in the training data set decreases. In itself this is not a spectacular finding. A model based on a larger training data set can be determined more precisely than a model based on a smaller data set. On the other hand, the larger the test data set, the more precise the mean test error rate can be estimated. Ideally, test error rates are of the same order as cross-validation error rates. The test set error rates and the cross-validation error rates were quite similar at a mean level, except for *data10:10*. However, at individual set level, the cross-validation error rate is in most cases not comparable to the test error rate. This illustrates that the result crucially depends on the specific sample of subjects that was used for modelling.

With only a small selection from a total population it is more likely that the selected subjects are not representative for the studied population, because it is possible that only subjects out of the extremes of the population distribution are selected. This study shows that the selection of subjects is crucial for the conclusions that are drawn about the model.

The effect is best seen in the results of *data05:05*. The 10th selection of *data05:05* had a much better cross-validation error rate for the training set compared to the 4th selection. If the 5 lean and 5 obese subjects of the 10th selection were selected as the representatives of the population under study, the conclusion would be that the 2 groups can be separated based on their LC-MS lipidomic

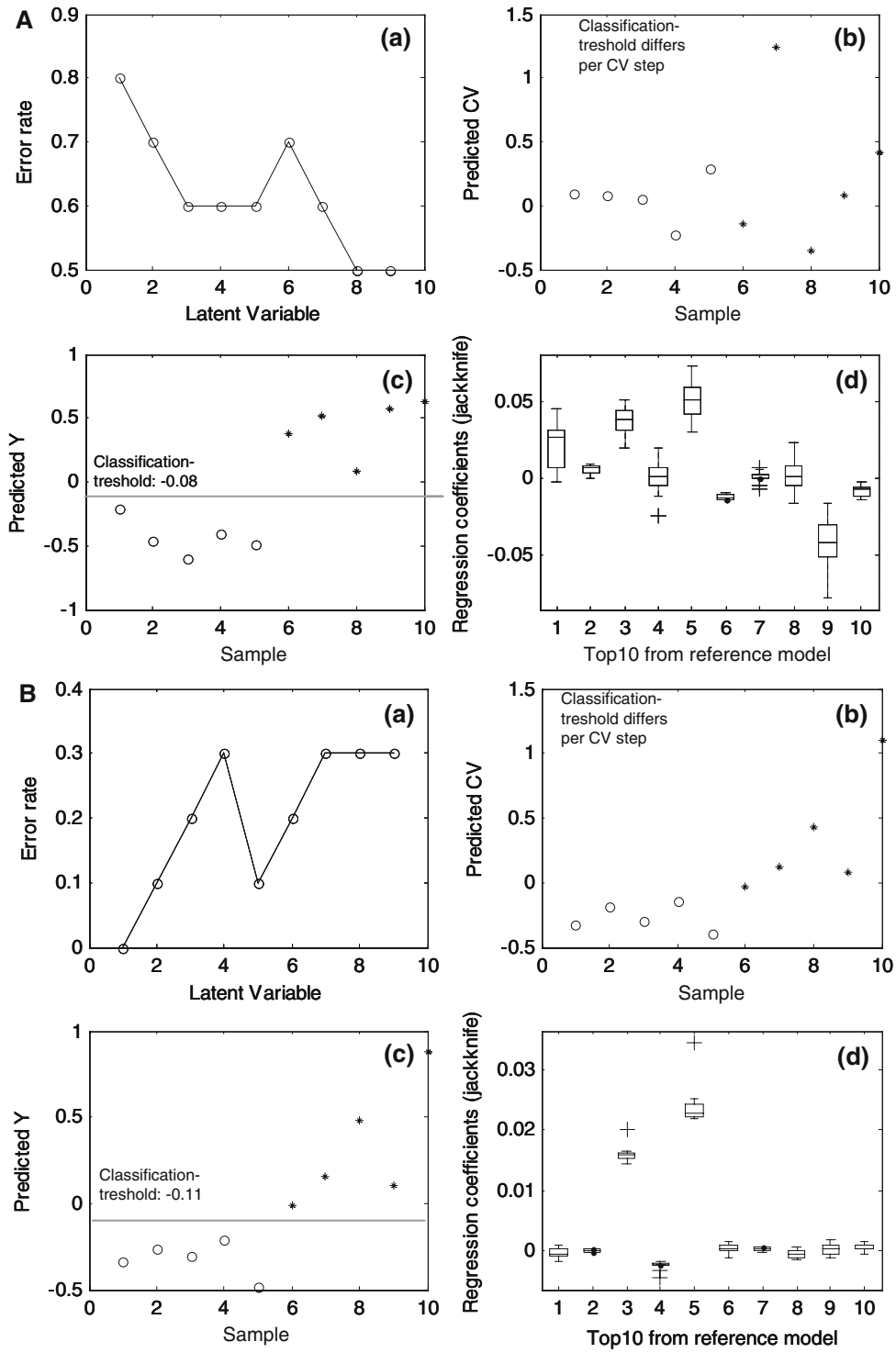


Figure 4. PLS-DA results for the 4th selection (A) and the 10th selection (B) of data05:05: Cross-validation error rate (a), Prediction based on cross-validation (b; o = lean, * = obese), Prediction based on fit (c; o = lean, * = obese), and Jack-knife (d).

profiles, even based on cross-validation results. If the 10 subjects out of set 4 were the subjects selected as the representatives of the population under study, the conclusion would be completely opposite. This means that the conclusions about the model completely depend on the selected 10 subjects. Nevertheless, the error rates based on the corresponding test sets were quite similar. As the predictability of both models was poor, it can be expected that both sets were not representative for the total set of 50 lean and 50 obese subjects. This illustrates how it could go wrong using data sets having considerably less subjects compared to the number of

Table 2
Summary of PLS-DA results based on the projection of all test data sets (number of LVs based on corresponding training data sets).

| Model | Training data set | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 4040 | | 3030 | | 2020 | | 1010 | | 0505 | |
| | Size testset | | Size testset | | Size testset | | Size testset | | Size testset | |
| | 1010 | 1010 | 2020 | 1010 | 3030 | 1010 | 4040 | 1010 | 4545 | 1010 |
| 1 | 15.0 | 15.0 | 35.0 | 45.0 | 30.0 | 35.0 | 40.0 | 45.0 | 50.0 | 50.0 |
| 2 | 10.0 | 10.0 | 30.0 | 30.0 | 41.7 | 35.0 | 30.0 | 10.0 | 38.9 | 40.0 |
| 3 | 30.0 | 30.0 | 27.5 | 25.0 | 50.0 | 50.0 | 46.3 | 55.0 | 50.0 | 50.0 |
| 4 | 30.0 | 30.0 | 42.5 | 40.0 | 18.3 | 10.0 | 48.8 | 45.0 | 46.7 | 50.0 |
| 5 | 30.0 | 30.0 | 17.5 | 10.0 | 13.3 | 15.0 | 50.0 | 50.0 | 50.0 | 50.0 |
| 6 | 25.0 | 25.0 | 22.5 | 30.0 | 48.3 | 50.0 | 43.8 | 45.0 | 50.0 | 50.0 |
| 7 | 15.0 | 15.0 | 17.5 | 10.0 | 33.3 | 40.0 | 38.8 | 55.0 | 50.0 | 50.0 |
| 8 | 20.0 | 20.0 | 20.0 | 25.0 | 50.0 | 50.0 | 50.0 | 50.0 | 50.0 | 50.0 |
| 9 | 5.0 | 5.0 | 35.0 | 30.0 | 35.0 | 35.0 | 50.0 | 50.0 | 37.8 | 35.0 |
| 10 | 25.0 | 25.0 | 35.0 | 40.0 | 30.0 | 20.0 | 50.0 | 50.0 | 50.0 | 50.0 |
| Mean | 20.5 | 20.5 | 28.3 | 28.5 | 35.0 | 34.0 | 44.8 | 45.5 | 47.3 | 47.5 |
| SD | 9.0 | 9.0 | 8.7 | 11.8 | 12.8 | 14.7 | 6.7 | 13.0 | 4.9 | 5.4 |

variables and it also shows the risk of drawing (too) optimistic conclusions about the distinction between the two classes, even based on cross-validation results. The size of the test data set did not seem to be an issue, as the results of the extra test data sets of 10 obese and 10 lean subjects were similar to the results based on the original test data sets.

Because different purposes are served, the conclusions about model validity based on cross-validation are not always comparable to the conclusions drawn based on the permutation test. The variation in performance of the permutation test was lower compared to the variation in error rates. The test only assesses the significance of the classification and does not take the predictability into account, which can explain why a model having a high cross-validation error rate can perform well in the permutation test.

All results indicate that cross-validation, jack-knifing and permutation tests are insufficient validation tools for megavariate data sets with only a few samples. The lower the ratio between the number of subjects and the number of variables, the less the validation results can be trusted. Taking only the results of these validation tools into account can be very misleading and may lead to incorrect conclusions. In order to avoid these problems, the number of samples per group should be large enough. In the present study, the turning point seemed to be between the sets having 10 and 20 subjects per group and based on about 950 variables. Unfortunately, it is impossible to translate this into a "golden rule" for all megavariate data sets.

Due to practical or budgetary limitations, it is often impossible to include the number of subjects that would be necessary to avoid the problems presented above. Another way to deal with megavariate data sets is to make the sets "less megavariate" by reducing the number of variables that are used for the statistical data analysis. This can be done, for instance, based on (i) analytical grounds by using a target approach instead of the total screening approach, (ii) biological grounds by using *a priori* variable selection, (iii) a selection method using statistical tools (Smilde *et al*., 2005), (iv) grey models, in which prior knowledge about (groups of) variables is taken into account (Bijlsma and Smilde, 2000; Gurden *et al*., 2001) or (v) regularization techniques, like using simplified correlation matrices (Schäfer and Strimmer, 2005). The disadvantage of the third and fifth approach is that the variables are selected using MVA methods which use the full data and similar problems as mentioned above can affect the selection. Using this approach, the bias due to selection should be assessed and corrections should be made (Ambroise and McLachlan, 2002). In case of a small number of subjects compared to the number of variables, contradictory results can be expected. Whether more simple statistical methods, e.g, those ignoring correlations like Nearest Shrunken Centroids (Tibshirani *et al*., 2002; Tibshirani *et al*., 2003), can be used to reduce the number of variables, is still under investigation.

The performance is assessed using this specific megavariate metabolomics data, but it is expected that the conclusions will also carry over to many other research areas. It was known that the data represented small differences between obese and lean subjects (Bijlsma *et al*., 2006). It is possible that the findings would be less dramatic if data that represents larger differences between groups is used.

The present study did not take the variable selection into account and only investigated the influence of the number of samples in the data sets. Future research may reveal the impact of the variable selection on the

reliability of the standard statistical validation tools for megavariate data.

## 4. Concluding remarks

The lower the number of subjects compared to the number of variables, the less the outcome of validation tools such as cross-validation, jack-knifing and permutation tests can be trusted. The result depends crucially on the specific sample of subjects that is used for modelling. The validation tools cannot be used as warning mechanism for problems due to sample size or representativity issues.

## References


Ambroise, C. and McLachlan, G.J. (2002). Selection bias in gene extraction on the basis of microarray gene-expression data. *PNAS* **99**, 6562–6566.

Braga-Neto, U.M. and Dougherty, E.R. (2004). Is cross-validation valid for small-sample microarray classification? *Bioinformatics* **20**, 374–380.

Barker, M. and Rayens, W. (2003). Partial least squares for discrimination. *J. Chemometrics* **17**, 166–173.

Bijlsma, S., Bobeldijk, I., Verheij, E.R., Ramaker, R., Kochhar, S., Macdonald, I.A., van Ommen, B. and Smilde, A.K. (2006). Large scale human metabolomics studies. A strategy for data (pre-) processing and validation. *Anal. Chem.* **78**, 567–574.

Bijlsma, S. and Smilde, A.K. (2000). Estimating reaction time constants form a two-step reaction: comparison between two-way and three-way methods. *J. Chemometrics* **14**, 541–560.

Blaak, E.E., Hul, G., Verdich, C., Stich, V., Martinez, A., Petersen, M., Feskens, E.F., Patel, K., Oppert, J.M., Barbe, P., Toubro, S., Anderson, I., Polak, J., Astrup, A., Macdonald, I.A., Langin, D., Holst, C., Sorensen, T.I. and Saris, W. H. (2006). Fat oxidation before and after a high fat load in the obese insulin-resistant state. *J. Clin. Endocrinol. Metabol.* **91**, 1462–1469.

Derome, A.E. (1987). *Modern NMR Techniques for Chemistry Research*. Pergamon Press, Oxford.

Efron, B. (1982). *The Jack-knife, the Bootstrap and Other Resampling Plans*. Society for Industrial and Applied Mathematics, Philadelphia, Pennsylvania, US.

Efron, B. and Tibshirani, R.J. (1993). *An Introduction to the Bootstrap*. Chapman & Hall/CRC. Chapter 15.

Fiehn, O. (2002). Metabolomics – the link between genotypes and phenotypes. *Plant Mol. Biol.* **48**, 155–171.

Geladi, P. and Kowalski, B.R. (1986). Partial least-squares regression: a tutorial. *Anal. Chem. Acta* **185**, 1–17.

Good, P.I. (2000). *Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses*. Springer-Verlag, New York.

Gurden, S.P., Westerhuis, J.A., Bijlsma, S. and Smilde, A.K. (2001). Modelling of spectroscopic batch process data using grey models to incorporate external information. *J. Chemometrics* **15**, 101–121.

Hastie, T., Tibshirani, R. and Friedman, J. (2001). The Elements of statistical learning: Data mining, inference and prediction. Springer Series in Statistics. 533 p; p214–216.

Lafaye, A., Junot, C., Ramounet-Le Gall, B., Fritsch, P., Tabet, J.C. and Ezan, E. (2003). Metabolite profiling in rat uring by liquid chromatography/electrospray ion trap mass spectrometry. Application to the study of heavy metal toxcity. *Rapid Commun. Mass Spectrometry* **17**, 2541–2549.

Lenz, E.M., Bright, J., Knight, R., Wilson, I.D. and Major, H. (2004). Cyclosporin A-induces changes in endogenous metabolites in rat uring: a metabonomic investigation using high field $^{1}$H NMR spectroscopy, HPLC-TOF/MS and chemometrics. *J. Pharmaceut. Biomed. Anal.* **35**, 599–608.

Manly, B.F.J. (1997). *Randomization, Bootstrap and Monte Carlo Methods in Biology*. (second ed.). Chapman and Hall, London.

Martens, H.A. and Dardenne, P. (1998). Chemometrics and intelligent laboratory systems. **44**, 99–121.

Martens, H. and Martens, M. (2000). Modified Jack-knife estimation of parameter uncertainty in bilinear modelling by partial least squares regression (PLSR). *Food Qual. Pref.* **11**, 5–16.

Martens, H. and Naes, T. (1989). *Multivariate Calibration*. Wiley, Chichester.

Massart, D.L., Vandeginste, B.G.M., Buydens, L.M.C., De Jong, S., Lewi, P.J. and Smeyers-Verbeke, J. (1997). *Handbook of Chemometrics and Qualimetrics: Part A*. Elsevier, Amsterdam.

Mielke, P.W. Jr. and Berry, K.J. (2001). *Permutation Methods: A Distance Function Approach. Springer Series in Statistics*, XV, 352 pp.

Petersen, M., Taylor, M.A., Saris, W.H.M., Verdich, C., Toubro, S., Macdonald, I., Rössner, S., Stich, V., Guy-Grand, B., Langin, D., Martinez, A.J., Pedersen, O., Holst, C., Sørense, T.I.A. and Astrup, A., The NUGENOB Consortium (2005). Randomized, multi-center trial of two hypo-energetic diets in obese subjects: high versus low fat content. *Int. J. Obesity* **6**, 1–9.

Plumb, R., Granger, J., Stumpf, C., Wilson, I.D., Evans, J.A. and Lenz, E.M. (2003). Metabonomic analysis of mouse uring by liquid-chromatography-time of flight mass aspectrometry (LC-TOCMS): detection of strain, diurnal and gender differences. *Analyst* **128**, 819–823.

Schäfer, J. and Strimmer, K. (2005). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Stat. Appl. Genet. Mol. Biol.* **4**, 1–29.

Smilde, A.K., van der Werf, M.J., Bijlsma, S., van der Werff-van der Vat, B.J.C and Jellema, R.H (2005). Fusion of mass-spectrometry-based metabolomics data. *Anal. Chem.* **77**, 6729–6736.

Stone, M. (1974). Cross validatory choice and assessment of statistical predictions. *J. R. Stat. Soc. Ser. B (Methodological)* **36**, 111–147.

Thibshirani, R., Hastie, T., Narasimhan, B. and Chu, G. (2003). Class prediction by nearest shrunken centroids, with applications to DNO microarrays. *Stat. Sci.* **18**, 104–117.

Thibshirani, R., Hastie, T., Narasimhan, B. and Chu, G. (2002). Diagnosis of multiple cancer types by shrunken centroids of gene exxpression. *PNAS* **99**, 6567–6572.

Vandeginste, B.G.M., Massart, D.L., Buydens, L.M.C., De Jong, S., Lewi, P.J. and Smeyers-Verbeke, J. (1998). *Handbook of Chemometrics and Qualimetrics: Part B*. Elsevier, Amsterdam.

Van der Greef, J., Davidov, E., Verheij, E., Vogels, J., van der Heijden, R., Adourian, A.S., Oresic, M., Marple, E.W. and Naylor, S. (2003). The role of Metabolomics in Systems Biology. *Metabolic Profiling: Its Role in Biomarker Discovery and Gene Function Analysis*, chapter 10. Harrigan, G.G. and Goodacre, R. (Eds). Springer, New York.

Vong, R., Geladi, P., Wold, S. and Esbensen, K. (1988). Source contributions to ambient aerosol calculated by discriminant partial least squares regression (PLS). *J. Chemometrics* **2**, 281–296.

Wilson, I.D., Plumb, R., Granger, J., Major, H., Williams, R. and Lenz, E.M. (2005). HPLC-MS based methods for the study of metabonomics. *J. Chromatography B* **817**, 67–76.