Real-life metabolomics data analysis: how to deal with complex data?
Rubingh, C.M.

Citation for published version (APA):
Rubingh, C. M. (2010). Real-life metabolomics data analysis: how to deal with complex data?

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.
THE COSTS OF COMPLEX MODEL OPTIMIZATION

Carina M Rubingh¹, Harald Martens², Hilko van der Voet³, Age K Smilde⁴

¹TNO Quality of Life, P.O. Box 360, 3700 AJ Zeist, The Netherlands
²NOFIMA MAT, Osloveien 1, NO-1430 Ås, Norway
³Biometris, Wageningen University and Research Centre, P.O. Box 100, 6700 AC Wageningen, Netherlands
⁴Swammerdam Institute for Life Sciences, University of Amsterdam, Nw Achtergracht 166, 1018 WV Amsterdam, The Netherlands

Submitted for publication
Abstract

Each data-driven action in data modeling consumes degrees of freedom, whether it concerns estimation of parameters, estimation of meta-parameters or selecting variables. By using a double cross validation approach for degrees of freedom calculation the costs for meta-parameter estimation and variable selection can determined explicitly. The only assumptions are independent and identically distributed errors, which makes the approach applicable to many predictive modeling techniques.
Summary

Each data-driven action in data modeling consumes degrees of freedom, whether it concerns estimation of parameters (such as regression coefficients), estimation of meta-parameters (such as optimal model rank) or selecting variables. By using a double cross validation approach for the degrees of freedom calculation, the costs for meta-parameter estimation and variable selection can be determined explicitly. To illustrate the concept, PLS regression was applied to two real-life metabolomics data sets. It is shown that only a small price needs to be paid to determine the meta-parameter (i.e. the optimal model complexity) whereas variable selection is much more ‘expensive’. Concerning a relatively small data set, almost 35% of the total degrees of freedom could be lost due to variable selection. Considering the fact that in functional genomics research the number of objects is mostly even smaller than the smallest subset that was used for the present paper, the impact of ignoring the degrees of freedom consumed by variable selection becomes very alarming.

Introduction

In the field of classical regression models, relating one property (regressand $y$) to a set of explanatory variables (regressors, $X$) in $n$ samples or objects, degrees of freedom is a very common and frequently used expression. But it is often difficult to understand, and even more difficult to keep track of quantitatively. Input data may have fewer degrees of freedom than expected, and data modeling may consume more degrees of freedom than is realised.

Conceptually, a degree of freedom corresponds to an independent way in which something can vary. Physically, an airplane has three degrees of freedom of forward motion through the 3D air space. Independently of that, it can rotate in different ways, which gives additional degrees of freedom for the pilot. In chemistry, a water molecule can likewise move in three directions, and rotate, and in addition it can vibrate in various ways, adding yet more degrees of freedom.

In statistics, the most basic definition of degrees of freedom is that of a parameter of the chi-square distribution, which is used to describe the distribution of quadratic forms under standard statistical assumptions: For a set of standard normal deviates, i.e. variables with independently, identically distributed elements, the distribution of their squared sum is defined as a chi-square distribution, and its degrees of freedom is equal to the number of summed squared standard normal deviates in the set. Another interpretation of degrees of freedom is the costs that are needed to be paid for parameter estimation. For instance, in full-rank multiple linear regression models, each of the $p$ predictor variable as well as the constant are said to use one degree of freedom leaving $n-p-1$ degrees of freedom for the residual variance of a model estimated from $n$ independent observations. In model selection criteria, such as Akaike information criterion (AIC; Akaike, 1973) or Mallows’ $C_p$ (Mallows, 1973), the number of degrees of freedom is used as a model complexity measure.
The more complex the model, the more degrees of freedom are needed to estimate the parameters.

The above mentioned applications of degrees of freedom are straightforward for standard linear models, like full-rank ordinary least squares estimation in multiple regression models. The concept of degrees of freedom is in this case also intuitively understandable. Considering degrees of freedom as independent parts of information in a data set, for each data-driven ‘choice’ that is to be made concerning modeling, a price has to be paid in terms of losing a number of these independent parts. The more choices are made, the less independent information is available, thereby lowering the number of degrees of freedom left over.

Full-rank regression modeling requires independent regressors, and that is unnatural for observational data sets; they call for reduced-rank modeling. Partial Least Squares Regression (PLSR) (Wold et al. 1983, Geladi and Kowalski, 1986; Martens and Naes, 1989) is a reduced-rank regression method commonly used in chemometrics due to its versatility and graphical accessibility. However, for PLSR, no formal definition of degrees of freedom exists. But the intuitive concept as stated above still holds. PLSR is an example of a complex modeling technique, for which the price for modeling in terms of degrees of freedom cannot easily be determined as in linear models. Although PLSR is often referred to as linear model due to its linear form of the final predictive equation, it is a method that depends on non-linear estimators for its coefficients. Hence, better methods for estimating the degrees of freedom in PLSR are needed.

Ye (1998) introduced a concept of generalized degrees of freedom consumed (GDF), that was applicable for evaluation of the final model or fits produced by data mining. The concept of GDF is illustrated in Figure 1a. GDF was defined as the sum of the sensitivity of each fitted value to perturbations in the observed response value. In other words, it measures the flexibility of the modeling procedure by evaluating the fit of the model to small changes in the observed values. If a model is very flexible, the higher the sensitivity of the fitted values to the observed values would be. The fitted values would be close to the observed values, and the GDF would be large. The size of the perturbations is defined by a tuning parameter that must be selected and also a choice needs to be made concerning the number of perturbations. GDF not only depends on the modeling procedure, but also on the underlying true model. Unfortunately, computing GDF suffers from practical difficulties, which makes the approach not very attractive.

A practical suggestion for calculating degrees of freedom for complex modeling was given by Van der Voet (1999). Pseudo degrees of freedom (PDF) were used to assess model complexity and its calculation is based on the predictive performance of a model. The concept of PDF is illustrated in Figure 1b. There are several ways to describe the predictive performance of a model, of which assessing the prediction error based on cross validation is one of the possible manners. Van der Voet (1999) defined PDF using ordinary and cross-
validation residuals as a ratio between a measure of model fit to a measure of predictive performance. This approach is very practical, especially because cross-validation is frequently used to determine the final model fit. So, all elements that are needed to calculate PDF are often already available.

Each choice in modeling will take its degrees of freedom. For complex models, not only model parameters needs to be estimated, but also a choice must be made concerning the estimation of the meta-parameter(s). Another aspect in modeling concerns variable selection: a choice needs to be made about which variables are included in the final model. Again, by making this choice on the basis of the data themselves, the total amount of independent parts of information that are available decreases. In other words, degrees of freedom must be paid in order to be able to make the final model selection. The cost for modeling in terms of meta-parameter selection and variable selection can be determined by using an extension of the PDF calculation as suggested by Van der Voet (1999). To illustrate the concept, PLS regression was applied to real-life metabolomics data sets.

In the PLS regression, the fitted model is \( \hat{y} = Xb_A \), where \( A \) is the number of estimated Latent Variables (LVs) or PLS Components. Here, the meta-parameter is the number of Latent Variables \( A \). It is usually determined based on assessing the mean square error of prediction (MSEP) from the deletion residual error cross-validation (Stone, 1974), in the present context named “Single Cross Validation (SCV)”, in which the \( n \) available samples are split into \( M \) subsets. The MSEP is a measure of how well the model estimated without sample subset \( m \) predicts the true data in subset \( m \). The \( n_m \) samples in each subset are treated as an independent test set for assessing a model developed on the basis of the data in the remaining \( n - n_m \) samples. When all samples, in turn, have been treated as independent test samples, the prediction error MSEP is estimated as a function of the number of LVs, \( a \): 

\[
MSEP_\alpha = \frac{1}{n} \sum_{m=1}^{M} \sum_{i=1}^{n_m} (y_{i(m)} - \hat{y}_{i(m),\alpha})^2, \quad a = 0, 1, 2, ..., A, ..., \text{where } i(m) \text{ is the sample index within subset } m, \text{ and } n = \sum_{m=1}^{M} n_m. \]

In other words, MSEP is a measure for the generalization error and the lack of fit. The optimal number of LVs is found as that which gives minimum MSEP value, or at the lowest number of LVs for which the MSEP is not significantly larger than that minimum in an appropriate significance test (van der Voet 1994).

By using a Double cross-validation approach (DCV; Stone, 1974; also called cross-model-validation or CMV (Anderssen et al., 2006; Smit et al., 2007), the prediction error is determined independently of the meta-parameter estimation. By using this validation approach to calculate the PDF, the costs for meta-parameter estimation can be determined as is shown in the present paper. The concept is also illustrated in Figure 1c.
Figure 1. Background of degrees of freedom calculation by a) Ye (1998): degrees of freedom are calculated from the effect of perturbing each of the observed regressand values (here illustrated for one of the samples only), b) Van der Voet (1999): degrees of freedom are calculated using single cross-validation leaving out one observation per validation step (here illustrated for two of the samples) and c) the current paper: degrees of freedom are calculated using double cross-validation with model estimation in inner loop and estimation of prediction error in outer loop (here illustrated for two of the samples; although no polynomial fitting was applied, a curved line was used to illustrate the differences in model complexity between the different steps of double cross-validation). Legend: ■ = observed value, - = model, ↑↓= perturbation, ▲= left out observation.

The price that needs to be paid for variable selection can also be determined by using a DCV approach for PDF calculation. In functional genomics research (e.g., metabolomics, transcriptomics) or any other field in which the number of available samples is often much lower compared to the number of variables, the impact of selection bias due to variable selection can be significant. Ambroise and McLachlan (2002) described how selection bias can be assessed by a cross-validation or bootstrap external to the selection process. It was demonstrated that it is a necessity to correct for the selection of a subset of variables in order to estimate the true prediction error. The same was demonstrated by Anderssen et al.,
(2006) in the case of variable selection using PLSR. In the present paper it is shown that the impact of variable selection in terms of degrees of freedom, hence in terms of independent parts of information that are lost due to selection, can be made clear.

The costs for meta-parameter estimations and variable selection only makes sense if it is compared to the total degrees of freedom in the data set. Given the fact that predictivity is used to calculate the PDF, the total degrees of freedom in a data set of independent values \( y_i, i=1,2,\ldots,n \) is equal to the number of objects \( n \). So, the cost for meta-parameter estimation or variable selection can be expressed as the number of objects or experiments that needs to be paid for the choice that is made. In other words, there is a relation between the cost for modeling and the number of experiments, hence there is a direct link to the statistical power of statistical tests on differences or relations. Subtle effects can only be found if enough objects are included: the more objects, the more statistical power of tests for finding these effects. In the present paper, also the sample size will be taken into account by using data sets of a varying number of objects to illustrate the impact of the sample size.

Ye (1998) and Van der Voet (1999) suggested two different ways to calculate the degrees of freedom for complex models. However, the methods do not distinguish the various segments of modeling and the cost of each modeling step cannot be determined explicitly. Ye (1998) calculated degrees of freedom for the whole modeling procedure. By defining variable selection as a part of the modeling procedure, the GDF can be used to correct for bias from model selection. Although Van der Voet (1999) is quite general on predictive models, he did not explicitly partition the total PDF among different parts of the model estimation problem. By using an extension of the PDF calculation as suggested by Van der Voet (1999), the cost of meta-parameter estimation and variable selection can be determined explicitly. The common elements of this extended approach and those of Ye (1998) and Van der Voet (1999) is perturbing data and using fitted and predicted values to calculate approximative degrees of freedom. The only assumptions are independent and identically distributed (iid) errors.

**Materials and Methods**

**Data** – Two real-life metabolomics data sets were used. The first data set originates from a quality research to investigate flavours in tomatoes, which are important targets for plant breeders to improve the quality of fresh tomatoes. Metabolic profiles, AFLP marker scores, sensory trait evaluations, plant and fruit morphology measurements and consumer appreciation assessments were obtained on ripe fruits of 94 tomato varieties. Details concerning the data that was obtained for this quality research can be found in Gavai et al. (2009), Ursum et al. (2008), and van Berloo et al. (2008). Due to missing values, 2 varieties were left out of the analysis. A total of 108 GC compounds (volatiles), 25 LC compounds (non-volatiles), 29 agronomical parameters and one of the sensory measurements on 92 tomato varieties were used for the present paper. This data set will be referred to as ‘the
tomato data set’, using the sensory parameter as phenotype and the GC and LC data in combination with the agronomical parameters as ‘metabolic profile’.

The second data set was a microbial metabolomics data set, which originates from a study that was performed in order to identify metabolites that are likely inducers of phenylalanine production (Phe) by E.coli. Sixteen fermentations under eight different environmental conditions were performed according to a 2 times replicated full factorial $2^3$ experimental design, varying Glucose and Succinate as carbon source, high and low phosphate concentration and pH6 and pH7. Samples were drawn at different time intervals during growth. In total, 194 samples were available. Samples were analysed for Phe production and LC-MS and GC-MS were used for metabolite analysis. Details concerning data collection and pre-processing can be found in Rubingh et al. (2009). The Phe data and GC-MS data containing 411 GC components were used for the present paper. In this paper the focus is on the degrees of freedom for the regressand $y$ only, and not on the regressors $X$. The temporal smoothing of Phe for the fermentation data set (Rubingh et al., 2009) reduces the total available degrees of freedom in the input Phe data. However, these initial DF losses are constant and outside the focus of the present paper, and therefore presently ignored. It was known that the Phe production varied a lot between fermentation batches. Two batches had extremely high Phe levels compared to the other batches. Also their variance within the batch was much higher than was seen in the other batches. These two batches could disturb the analysis seriously, hence they were removed before modeling. Finally, 170 samples were included for analysis. This data set will be referred to as ‘the fermentation data set’, using Phe as phenotype and the GC data as metabolic profile.

**Model and model validation** - PLS regression was used to correlate the phenotype to the metabolic data. For each input set, three steps were taken to determine the degrees of freedom used for meta-parameter estimation: (1) a Double Cross Validation, (2) a Single Cross Validation, and (3) a final PLS regression fit to estimate the parameters utilizing all samples. A Double Cross-Validation approach is used to determine the error of modeling independently of the meta-parameter estimation. The meta-parameter of the final model is estimated using the Single Cross-Validation. The final fit of the PLS regression model is based on the meta-parameter estimation of the SCV and the error of the regression model is based on DCV (Smit et al., 2007).

Firstly, a Double Cross Validation (DCV) was performed for each of the two data sets. The analysis of the tomato data was based on a tenfold venetian blind DCV. Ten percent of the samples ($n_{(1)}$) were left out in an outer cross validation loop. The remaining data ($n- n_{(1)}$) were used in an inner loop, in which an 'Inner Single Cross Validation' (ISCV) was used for estimating MSEP and the optimal model rank, $A$. The ISCV consisted here in leaving out, in turn, 10% of the samples ($n_{(2)}$), for testing a model built on the remaining $n- n_{(1)} - n_{(2)}$ samples with respect to MSEP at rank $a_{(2)}=0,1,2,...$. The optimum number of LVs $A_{(2)}$ was determined based on the minimal ISCV prediction error.
The analysis of the fermentation data was based on a 'leave-one-fermentation-out' DCV. For 'leave-one-fermentation-out' Double Cross Validation (DCV), all available time points $n_{m(1)}$ of the one fermentation batch $m(1)$ at a time were left out in an outer cross validation loop. Data of the remaining $M-1$ fermentations were used in an inner loop, in which an 'Inner Single Cross Validation' (ISCV) was used for estimating MSEP and the optimal model rank, $A$. The ISCV consisted here in leaving out, in turn, each of the $M-1$ remaining fermentation batches $m(2)$ with its $n_{m(2)}$ samples, for testing a model built on the remaining $n- n_{m(1)} - n_{m(2)}$ samples with respect to MSEP at rank $a_{m(2)}=0,1,2,...$. The optimum number of LVs $A_{m(2)}$ was determined based on the minimal ISCV prediction error.

To prevent over-fitting, the number of LVs that was used to determine each outer loop prediction was chosen as follows: 1) choose the minimal value of the RMSEP (square root of MSEP) based on the inner loop predictions, 2) compare this minimal value to the RMSEP using one LV less compared to the number that was chosen, 3) for a simple approach, if the reduction in RMSEP between these two values was less than 5%, the model with one LV less was chosen. This number of LVs was used to fit a PLS model on the inner loop samples, which was then used to predict the measurements of the samples that were left out in the outer loop. The outer loop predictions were saved and the procedure was repeated until all samples were left out once in the outer loop. As a consequence of this double cross-validation approach, the outer loop predictions are based on different model complexities, so a slightly different set of samples was used for meta-parameter estimation as was used for error prediction (Smit et al., 2007).

Secondly, a Single Cross Validation (SCV) was used for each of the two data sets. Ten percent of the samples were left out per cross-validation step for modeling the tomato data set (a tenfold venetian blind cross-validation approach) and all available measurements per fermentation were left out per cross-validation step for modeling the fermentation data set (a 'leave-one-fermentation-out' cross validation approach). Based on the minimal prediction error for the measurements that were left out, the optimum number of LVs was determined. The same restriction of 5% change in RMSEP as in the ISCV was used to determine the optimum number of LVs for final modeling. As a consequence of this validation approach, the same set of samples was used for meta-parameter estimation as well as for error prediction.

Finally, the number of LVs that was determined in the previous described SCV procedure was used to build a final PLS model using data of all samples, resulting in a set of final fit predictions. The phenotype levels that were predicted using this model are indicated as 'final fitted values'. As a result, three sets of predictions were available, namely based on DCV, based on SCV and based on the final fit.

Random subsets - To investigate the effect of the number of samples on the used DFs, random subsets of the original data sets were made. Four levels of selection were defined, using about 30%, 50%, 70% and 90% of the data. The random selection for the fermentation
data set was stratified by fermentation. For each level, hundred random selections from the original data sets were made to be able to assess the variation in the estimation of the degrees of freedom, resulting in 4x 100 subsets.

By using 100 different subsets for various sizes of the data set, the variance in modeling can be assessed. However, also bias might be present. The impact of bias can be reduced by leaving out only a few samples per cross-validation step. But if only one sample is left out, which is the case in a leave-one-out cross-validation approach, the variance component cannot be estimated. Therefore, Hastie et al. (2001) recommended the use a tenfold cross-validation approach, leaving out 10% of the data per cross-validation step. The difference between the DCV and the SCV results, might be influenced by bias. However, since the number of samples that were left out in the outer loop of the DCV and each step of the SCV are comparable, the impact of bias will be similar for both approaches. A tenfold cross-validation was used for the tomato data set. By using leave-one-fermentation-out each cross-validation for the fermentation data set, about 7% of the data was left out per cross-validation step, meaning that it was between a leave-one-sample out and a tenfold cross-validation scheme, hence on the safe side.

**Degrees of freedom** - Van der Voet (1999) proposed a definition for pseudo degrees of freedom (PDF) for PLS models based on predictive performance. The PDF are defined as:

\[
PDF = n[1 - \sqrt{(MSEP_{fit} / MSEP_{cv})}],
\]

where \( n \) is the number of objects, \( MSEP_{fit} \) is the mean squared error based on fitted values and \( MSEP_{cv} \) is the mean squared error based on prediction from cross-validation. The PDF are based on an SCV procedure and concern the degrees of freedom that are needed for estimation all model parameters, given the meta-parameter, which is the number of LVs that is needed for modeling. An extension to the calculation by van der Voet was used to investigate the costs for meta-parameter estimation:

\[
PDF_{scv} = n[1 - \sqrt{(MSEP_{fit} / MSEP_{scv})}] \quad (2a)
\]
\[
PDF_{dcv} = n[1 - \sqrt{(MSEP_{fit} / MSEP_{dcv})}] \quad (2b)
\]
\[
PDF_{mp} = PDF_{dcv} - PDF_{scv} \quad (2c)
\]

where \( PDF_{scv} \) is equal to (1), \( PDF_{dcv} \) are the PDF in which the predictions are based on predicted values in the outer loop of the DCV, and \( PDF_{mp} \) are the costs for meta-parameter estimation. \( MSEP_{dcv} \) are the predictions based on DCV, \( MSEP_{scv} \) are the predictions based on SCV, and \( MSEP_{fit} \) are the predictions based on the final fit of the model. (2a)-(2c) were used for the original data set as well as for the 400 random subsets. The degree of variability in the PDF estimates between the 400 subsets was also assessed. The \( PDF_{mp} \) is compared with the total degrees of freedom in the data set, which is considered to be equal to the number of objects in the data set, \( n \). This means for the tomato data set: \( n = 92 \), and when using 30%
of the data, n = 28; using 50%, n = 46; using 70%, n = 65; and using 90%, n = 83. For the fermentation data set this means n = 170, and when using about 30% of the data per fermentation time series, n = 57; using 50%, n = 87; using 70%, n = 127; and using 90%, n = 157.

**Variable selection** – A jackknife approach was used for variable selection, which was performed within both ways of cross-validation. The variable selection in DCV was performed as follows:

1) Define a test set with all observations that are left out in the first outer loop and define a training data set containing all available remaining measurements;

2) Perform an SCV on the training set as described before (ISCV);

3) Determine meta-parameter (# LVs);

4) Calculate the relative standard deviation (RSD) for each variable using the regression coefficients of each step of the inner loop using the number of LVs as determined in 3) and select only those variables with RSD < 50%;

5) Perform a second ISCV on the training set using only the variables which are selected in 4);

6) Determine meta-parameter (# LVs);

7) Use variables from 4) and meta-parameter from 6) to predict the samples which were left out in the outer loop (in 1));

8) Repeat 1) - 7) until all samples are left out once in the outer loop.

The variable selection in SCV was performed is described in 2-6 for DCV. The variables from 5) and the meta-parameter from 6) are used for final modeling to obtain the final model predictions.

The selection procedure was completed for the original data set as well as the 400 subsets. Again, (2a) – (2c) were used to calculate the costs for modeling. Now, $PDF_{mp}$ indicates not only the price for meta-parameter estimation but also for variable selection. $PDF_{scv}$ represent the costs for estimating the parameters of the PLSmodel, given a selection of variables and a number of LVs, whereas the $PDF_{dcv}$ includes the costs of selecting variables and choosing a number of LVs. The difference between $PDF_{scv}$ and $PDF_{dcv}$ is the cost for variable selection and choosing the number of LVs. This measure is compared to the total degrees of freedom in the data set ($n$) in order to estimate the true impact of the modeling choices.

**Model assumptions** – As mentioned before, the proposed approach is not only applicable for PLS regression models, but can be generalized to any predictive model assuming iid errors. For the fermentation data set, multiple measurements were taken for each batch, indicating that both within and between batch variation is modelled. By centering the data
per batch, the contribution of between batch variation could be reduced to almost zero. However, this would imply that the errors are not completely independent anymore since the same data was used for correction as was used for modeling. DCV was used instead of SCV, exactly for that purpose, hence it was decided not to correct for between batch differences in the fermentation data set and to model both within and between batch processes in PLS regression.

**Data pre-processing** - Within each inner loop of the double cross-validation procedure, data were scaled to mean zero and unit variance. The inner loop scaling parameters were used to scale data in the outer loop.

**Software** - All analyses were performed using Matlab Version 7.0.4 R14 (The Mathworks, Inc.) and the n-way toolbox version 2.11.

**Results and Discussion**

The price that needs to be paid for meta-parameter estimation and variable selection were determined by using a DCV approach for PDF. In Table 1, the results for the complete data sets are given, for both the tomato and the fermentation data set and for both modeling with and without variable selection. The degrees of freedom based on the model without variable selection indicate the costs for determining the number of LVs. There is only a small difference in degrees of freedom, which means that only a small price is paid to obtain the number of components.

| Table 1. Costs of meta-parameter estimation and variable selection in PLS-modeling expressed as the absolute number of degrees of freedom (PDF) consumed by the PLS-modeling, calculated using single (SCV) and double (DCV) cross-validation, as well as the difference in degrees of freedom (Δ), which are the cost for modeling. |
|---|---|---|---|---|---|---|
| Data | Variable selection | # var | PDF$^1$ | # var | PDF$^2$ | LVs | Δ$^3$ | n | %Δ$^4$ |
| Tomato | Without | 162 | 25.0 | 162 | 23.0 | 3 | 2 | 92 | 2.2 |
| | With | 109-118 | 26.5 | 114 | 21.1 | 3 | 5.4 | 92 | 5.9 |
| Fermentation | Without | 408 | 78.7 | 408 | 74.5 | 2 | 4.3 | 170 | 2.5 |
| | With | 265-300 | 79.8 | 298 | 63.7 | 2 | 16.1 | 170 | 9.5 |

$^1$PDF$_{dcv} = n[1 - \sqrt{(MSEP_{fit} / MSEP_{dcv})}]$

$^2$PDF$_{scv} = n[1 - \sqrt{(MSEP_{fit} / MSEP_{scv})}]$

$^3$Δ = PDF$_{dcv} - $PDF$_{scv}$

$^4$%Δ = 100*(PDF$_{dcv} - $PDF$_{scv}$)/n

The difference between PDF$_{scv}$ and PDF$_{dcv}$ (Δ in Table 1) that are calculated using variable selection is a measure for the cost of the variable selection and meta-parameter estimation, i.e. how many degrees of freedom are consumed when optimizing the modeling of y from X by testing and removing X-variables based on their apparent predictive contribution for y. In other words, the PDF$_{scv}$ represent the costs for estimating parameters of the PLS model,
given the selected variables and the chosen number of LVs, whereas the \( PDF_{dcv} \) is determined unconditionally of the number of selected variables and chosen number of LVs. The difference between \( PDF_{scv} \) and \( PDF_{dcv} \) after variable selection (\( \Delta \) in Table 1) is larger compared to the difference that was calculated without variable selection, hence variable selection is much more 'expensive' than choosing the number of LVs.

\[ \Delta = \text{median}(PDF_{dcv} - PDF_{scv}) \]

\[ \%\Delta = 100 \times (\text{median}(PDF_{dcv} - PDF_{scv}))/n \]

\[ \text{max}\%\Delta = 100 \times (\text{max}(PDF_{dcv} - PDF_{scv}))/n \]

In Table 2a and 2b, the results are presented for the subset analysis of the tomato data set and the fermentation data set, respectively. For each level of subset analysis, the pseudo-degrees of freedom (PDF) consumed were estimated. The median of these degrees of freedom estimates over the 100 repeated subsets are shown, as well as their minimal and maximal value. The inter-quartile range (IQR), the interval about the median containing 50% of the data, is also calculated as a measure of variation of the degrees of freedom. Index \( n \) gives the number of samples and is considered as the total degrees of freedom in the particular data set, \( DF_{total} \). The relative change in DF (\( \%\Delta \) in Table 2a and 2b) is calculated as percentage of the number of calibration samples used for the four different subset sizes (\( n = 83, 65, 46 \) or 28 for the tomato data, and \( n=156, 127, 87 \) or 57 for the fermentation data). The relative difference clearly shows the price that is paid for meta-parameter estimations and for variable selection. For the tomato data set, the costs for meta-parameter estimation are between 0.7% and 3.2% and the costs for variable selection and meta-parameter estimation are between 6.1% and 11.8%. These are the percentages of the originally

<table>
<thead>
<tr>
<th>Subset</th>
<th>Without variable selection</th>
<th>With variable selection</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>90%</td>
<td>70%</td>
</tr>
<tr>
<td>N</td>
<td>83</td>
<td>65</td>
</tr>
<tr>
<td>( PDF_{dcv} ) Min</td>
<td>13.8</td>
<td>2.7</td>
</tr>
<tr>
<td>Median</td>
<td>21.9</td>
<td>20.6</td>
</tr>
<tr>
<td>Max</td>
<td>29.1</td>
<td>36.5</td>
</tr>
<tr>
<td>IQR</td>
<td>4.4</td>
<td>5.8</td>
</tr>
<tr>
<td>( PDF_{scv} ) Min</td>
<td>15.0</td>
<td>4.4</td>
</tr>
<tr>
<td>Median</td>
<td>21.1</td>
<td>19.9</td>
</tr>
<tr>
<td>Max</td>
<td>27.5</td>
<td>34.0</td>
</tr>
<tr>
<td>IQR</td>
<td>3.0</td>
<td>4.9</td>
</tr>
</tbody>
</table>

\[ \Delta = \text{median}(PDF_{dcv} - PDF_{scv}) \]

\[ \%\Delta = 100 \times (\text{median}(PDF_{dcv} - PDF_{scv}))/n \]

\[ \text{max}\%\Delta = 100 \times (\text{max}(PDF_{dcv} - PDF_{scv}))/n \]
available parts of independent information that must be paid for the choices that are made. For the fermentation data set, these costs for meta-parameter estimation are between 2.7% and 3.5% and for variable selection are between 13.4% and 15.4% (Figure 2).

Table 2b. Costs of meta-parameter estimation and variable selection in PLS-modeling (Δ) expressed as summaries (over 100 repeated resamplings per subset) of the absolute number of degrees of freedom (PDF) consumed by the PLS-modeling, calculated using single (SCV) and double (DCV) cross-validation for several sizes of the fermentation data set.

<table>
<thead>
<tr>
<th>Subset</th>
<th>Without variable selection</th>
<th>With variable selection</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>156</td>
<td>127</td>
</tr>
<tr>
<td>PDF&lt;sub&gt;dcv&lt;/sub&gt; Min</td>
<td>83.2</td>
<td>66.4</td>
</tr>
<tr>
<td>Median</td>
<td>89.5</td>
<td>76.4</td>
</tr>
<tr>
<td>Max</td>
<td>103.8</td>
<td>89.2</td>
</tr>
<tr>
<td>Iqr</td>
<td>3.5</td>
<td>5.3</td>
</tr>
<tr>
<td>PDF&lt;sub&gt;scv&lt;/sub&gt; Min</td>
<td>80.6</td>
<td>60.8</td>
</tr>
<tr>
<td>Median</td>
<td>85.3</td>
<td>71.7</td>
</tr>
<tr>
<td>Max</td>
<td>89.4</td>
<td>78.2</td>
</tr>
<tr>
<td>Iqr</td>
<td>2.5</td>
<td>4.1</td>
</tr>
</tbody>
</table>

Δ = median(PDF<sub>dcv</sub> - PDF<sub>scv</sub>)

%Δ = 100*(median(PDF<sub>dcv</sub> - PDF<sub>scv</sub>))/n

max%Δ = 100*(max(PDF<sub>dcv</sub> - PDF<sub>scv</sub>))/n

The costs for meta-parameter estimation and variable selection are relatively higher for smaller data sets compared to the larger data set, which can be explained in terms of model complexity. In Table 3a and 3b, the frequencies of the number of latent variables are given for the various subsets for the tomato data set and the fermentation data set, respectively. More complex models using more LVs are selected for the larger subsets using 70% and 90% of the data: more data means more possibilities to include details in the model. Much more variation in model complexity is seen in the smaller data sets. Apparently, less effort is needed to determine the complexity of the model for the larger data sets compared to the smaller data sets.
Figure 2. Relative costs in terms of pseudo degrees of freedom (%Δ) for meta-parameter estimation and variable selection for the tomato and the fermentation data set.

Table 3a. Frequency table of model complexity for several sizes of the data set: percentages of chosen number of latent variables in the PLS model applied to the tomato data.

<table>
<thead>
<tr>
<th>#LV</th>
<th>Without variable selection</th>
<th>With variable selection</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>90%</td>
<td>70%</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>27</td>
<td>41</td>
</tr>
<tr>
<td>3</td>
<td>72</td>
<td>42</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>12</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 3b. Frequency table of model complexity for several sizes of the data set: percentages of chosen number of latent variables in the PLS model applied to the fermentation data.

<table>
<thead>
<tr>
<th>#LV</th>
<th>Without variable selection</th>
<th>With variable selection</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>90%</td>
<td>70%</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
The conclusions based on the fermentation data set are comparable to the ones based on the tomato data set: variable selection is more expensive than meta-parameter estimation and the costs for meta-parameter estimation and variable selection are relatively higher for smaller data sets compared to the larger data set. However, the results for the fermentation data set are more extreme than for the tomato data set. This illustrates an important and treacherous phenomenon, called pseudo-replication. As stated before, the experimental design of the fermentation data has been ignored for modeling. By doing so, the hierarchical structure is ignored and a model was applied assuming \( n \) independent observations. Only for independent observations we can state total DF = \( n \). With correlated observations such as arise in hierarchical designs the total DF should be less than \( n \), at least intuitively (remember that there is no formal definition of DF in such a case). As an extreme example consider the case where all observations in a fermentation batch would be exactly equal. Then the total DF would be 14 (= number of batches) rather than 170 (= number of observations). Of course in reality data sets are never that extreme but in general we expect total DF to be less than \( n \) observations in case of dependencies within the larger units. Modeling seems to cost relatively more in terms of DF, but this is due to pseudo-replication, which makes that the initial number of \( n \) DF must be interpreted with great care and must only be used as indicative.

**Figure 3a.** Distribution of estimated pseudo-degrees of freedom (in percentage of number of calibration samples; \( n = 83, 65, 46 \) or 28, respectively) lost by PLSR model optimization for the tomato data set, including rank estimation, variable selection and remaining parameter estimation. For visual comparison, straight lines have been added for \( \text{PDF}_{\text{dev}} = \text{PDF}_{\text{scv}} \).
A lot of variation is seen within a subset, which is illustrated in a density plot of $PDF_{scv}$ versus $PDF_{dcv}$ per subset (Figure 3a and 3b). In general, the smaller the data set, the more variation is seen. The density cloud using 90% of the original data set is less fuzzy than the cloud for the 30% subset. It is also seen that the smaller the data set, the more extreme values of degrees of freedom are found: much more islets are seen for the 50% and 30% data set than for the 70% and 90% subsets. For the smallest subset, using 30% of the data, the costs for meta-parameter estimation and variable selection, differ between 14 and 43 degrees of freedom. At its worst, 30% of the apparent total degrees of freedom are lost. However, the results for the 30% data set must be put in perspective to some extent. Resampling assumes that the sampling distribution reflects the population distribution and that is likely not the case for small sample sizes. As seen before in Rubingh et al. (2006), small data sets show a lot of variation, which is also seen in Figure 3. Nevertheless, even for the 50% data set, which shows less variation than the 30% data set, still 20% of the apparent total degrees of freedom can be lost due to modeling.

**Figure 3b.** Distribution of estimated pseudo-degrees of freedom (in percentage of number of calibration samples; n=156, 127, 87 or 57, respectively) lost by PLSR model optimization for the fermentation data set, including rank estimation, variable selection and remaining parameter estimation. For visual comparison, straight lines have been added for $PDF_{dcv} = PDF_{scv}$. 
Conclusion

Each data-driven action in data modeling consumes degrees of freedom, whether it concerns estimation of parameters (such as regression coefficients), estimation of meta-parameters (such as optimal model rank) or estimation/choice of variable weights (variable selection). Ye (1998) and Van der Voet (1999) introduced the GDF and PDF, respectively, to estimate the degrees of freedom for complex models. However, both approaches do not distinguish the various segments of modeling, like meta-parameter estimation or variable selection. By using a double cross validation approach for PDF calculation, these costs can be determined explicitly. PLS regression was applied to two real-life metabolomics data sets to illustrate the concept. The idea is not restricted to PLS only, but can be generalized to many predictive modeling techniques which assume iid errors. Unfortunately, equality of variance (homoscedasticity) is not always a realistic assumption in chemical analysis, where high measured values tend to have higher uncertainty than low measured values. If necessary, a proper data transformation, like a log transformation, can be applied to obtain homoscedasticity and the method can be applied. There is also no restriction to the method that is used for variable selection.

Considering the fact that in functional genomics research the number of objects is mostly even smaller than the smallest subset that was used for the present paper, the impact of ignoring the costs of variable selection becomes very alarming. Bearing in mind the direct link between the degrees of freedom and the power, it emphasizes the necessity to include enough objects in, for instance, functional genomics studies, in which variable selection is frequently applied such as biomarker discovery studies.

Acknowledgement

We gratefully acknowledge Mariët van der Werf (formerly TNO Quality of Life, Zeist, The Netherlands) and Karin Overkamp (TNO Quality of Life, Zeist, The Netherlands) for making available the microbial metabolomics data. The tomato data that was used as an example in this paper was (co)financed by the Centre for BioSystems Genomics (CBSG) which is part of the Netherlands Genomics Initiative / Netherlands Organisation for Scientific Research.

References


