Received: 23 February 2010, Revised: 31 August 2010, Accepted: 21 September 2010, Published online in Wiley Online Library: 2010

(wileyonlinelibrary.com) DOI: 10.1002/cem.1360

Variable selection in regression—a tutorial

C. M. Andersen[a] and R. Bro[a*]

This paper provides a practical guide to variable selection in chemometrics with a focus on regression-based calibration models. Several approaches, such as genetic algorithms (GAs), jack-knifing, forward selection, etc., are explained; it is also explained how to choose between different kinds of variable selection methods. The emphasis in this paper is on how to use variable selection in practice and avoid the most common pitfalls. Copyright © 2010 John Wiley & Sons, Ltd.

Keywords: variable selection; calibration; chemometrics

1. INTRODUCTION

Multivariate calibration models, such as partial least squares regression (PLS) and principal component regression (PCR), are commonly applied when predicting one or several parameters from a multivariate data set. These methods can handle data sets even when the number of variables is much larger than the number of samples. However, in some situations it can be an advantage to reduce the number of variables in order to, among others, obtain (a) improvement of the model predictions, (b) a better interpretation or (c) lower measurement costs.

Variable selection is used for improving the model performance and give better predictions. With many irrelevant, noisy or unreliable variables, removal of these will typically improve the predictions and/or reduce the model complexity. Improvement of statistical properties can also be a reason for doing variable selection. In some situations, the purpose of variable selection is to obtain a model that is easier to understand, for example, by getting rid of all the variables that do not contribute positively to the model. This may not give better predictions. They may even get a bit worse. However, adequate models using as few variables as possible are sometimes desired. Furthermore, it is sometimes the aim to use measurements obtained from high-resolution instruments to identify the most important variables that can be applied, e.g. in a filter-based instrument. This is relevant for industrial on-line or at-line purposes where high-resolution instruments may be too expensive or scanning a whole spectrum will take too much time. Also, variable selection can be relevant to reduce the risk of overfitting or for computational reasons.

An optimal way to do variable selection is to try all combinations of variables and select the best ones. This sounds simple, but is, in practice, impossible for a number of reasons. With many variables, this may be too cumbersome and take too much time. Even with the most advanced computers, the number of combinations of variables to investigate becomes prohibitive even for, say, 50–100 variables. Furthermore, even if it was possible to test all combinations of variables, the risk of overfitting would be detrimental unless the number of samples was much higher than the number of combinations of variables. Among others, for these reasons, a number of variable selection methods have been developed which try to find a *good* set of variables rather than *the* optimal set of variables. The applicability of each of these methods depends on the data and the purpose of the study. This paper explains some of the most commonly applied methods in chemometric analysis and illustrates when to use which type of method. A discussion on the usability, advantages and disadvantages of each method is included. The various methods are exemplified by an example taken from analysis of beer where a quality parameter, real extract, is predicted from visual and near infrared spectra.

We stress that this paper does not provide an exhaustive review of available methods for variable selection. Rather it highlights generic types of methods by example and illustrates how they can be appropriately used. Also, most of the algorithms and approaches presented are applicable regardless of what regression method is adopted, but in this paper, most variable selection approaches will be illustrated using PLS regression.

2. IMPORTANT ASPECTS OF VARIABLE SELECTION

In this section, some general overall issues are discussed. These issues will hold for all types of variable selection, so they are important to consider generically. An overall statement that is always worth considering is that the actual choice of sensors/measurements is the most basic, fundamental and influential variable selection performed. Choosing the right sensors can hugely affect the outcome of any modeling. For instance, the interest in process analytical technology has led to a number of investigations where near infrared spectroscopy was chosen for modeling certain properties, e.g. in a process stream. Oftentimes, the choice was not based on chemical and physical insight, but simply due to an incorrect understanding that near infrared spectroscopy is the method of choice in process analytical technology.

* *Correspondence to: R. Bro, Quality and Technology, Department of Food Science, Faculty of Life Sciences, University of Copenhagen, Rolighedsvej 30, DK-1958 Frederiksberg C, Denmark.*

a *C. M. Andersen, R. Bro*
Quality and Technology, Department of Food Science, Faculty of Life Sciences, University of Copenhagen, Rolighedsvej 30, DK-1958 Frederiksberg C, Denmark

2.1. Spectral and similar data

For most spectral, measurements at one wavelength are correlated to the measurements at the neighboring wavelength. For such data, it is wasteful and overly complex to work on choosing single wavelengths independently, since the information from one component in the sample will influence more than one wavelength. It is better to choose windows of wavelengths [1]. This reduces the complexity of the problem enormously and at the same time reduces the risk of spurious results and speeds up the analysis. Several of the methods described below can be applied on intervals rather than on individual variables. For interval PLS (*i*PLS) and variations of this method, it is obvious but also for other methods, such as genetic algorithms (GAs), it is a possibility.

2.2. Overfitting

If there are many more variables than samples, it is possible by chance and/or overfitting to find a certain number of variables that correlate to the property to be predicted. If such variables are chosen and the model is applied on new samples, the predictions may be very poor or there may be no relationship at all. Therefore, validation is fundamental. This is especially so

- the larger the ratio of variables to samples is,
- the higher the rank of the data is,
- the more possibilities (combinations) the variable selection evaluates,
- the less is known about the relevance of the data for predicting.

The problem with overfitting can be illustrated using a small simulated data set. In the following, a data set of four samples is used. A response vector, **y**, is generated randomly (normally distributed random numbers) and then a set of dependent variables, held in the matrix **X**, is generated using 1–5 variables (also random and normally distributed). As all data are random, there should be no relation between **X** and **y**. For each number of variables, we look for the subset of variables that produces the lowest mean squared calibration error. Thus, for, e.g., an **X** with three columns (and four rows), we test using column one, column two, column three, column one *and* two, etc., until all combinations have been tested. This is repeated 900 times with different random numbers and the resulting models are shown in terms of the correlation between **y** and the predicted **y**. In Figure 1 (top row), it is clear that with only four samples, a perfect fit is always obtained using four or more variables. Clearly, the perfect correlations are merely a result of overfitting and as such are meaningless. This is, in fact, trivial but it is still important to remember in practical applications of variable selection with more variables than samples. As soon as the number of variables exceeds the number of samples, the fit will be trivially perfect. It is also interesting that with just three variables, it is seen that quite often a reasonable or good model is obtained even though it is known that there is no real relation between **X** and **y**.

Rather than selecting variables based on fit, the prediction error obtained from cross-validation can be used. In Figure 1 (bottom row), the results corresponding to the upper row of plots are shown but now using cross-validation rather than fit for selecting variables. As can be seen, the cross-validation helps somewhat in minimizing the overfit, but it is also apparent that, e.g. with five variables, most of the models obtained will seem to be predicting well even during cross-validation regardless that there is no real relation between the variables. Evidently, cross-validation is of limited use when the number of variables is in the neighborhood of the number of samples.

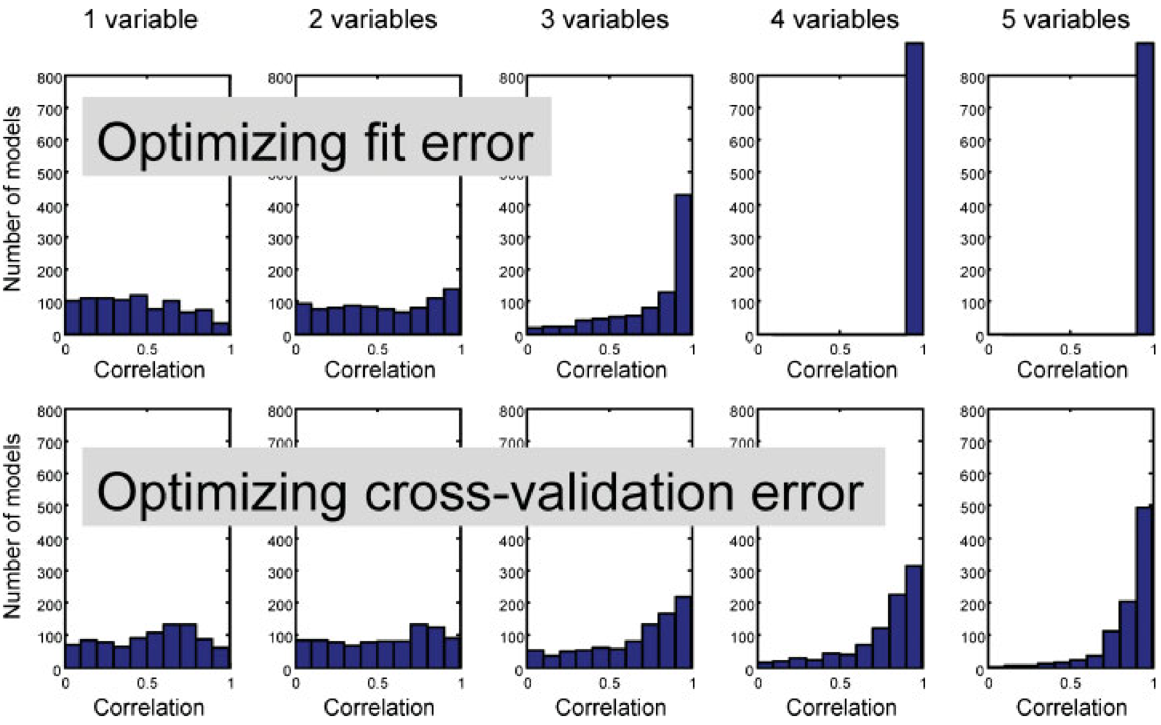

Figure 1. Histogram showing the correlations obtained from 900 repetitions of best variable selection result between a random **y** and one to five random **X** variables. Top shows correlation of predictions and response when selecting variables based on fit and bottom shows correlation off predictions and response when selection variables based on cross-validation. The data set has four observations

With even more variables, the problem becomes even more pronounced.

Other ways to reduce the risk of overfitting are to use a real test set, cross model validation or permutation tests. Cross model validation is a resampling technique where the whole model development step *including* variable selection is subjected to cross-validation [2]. By permutation test, one randomly permutes the response variable and looks at how the prediction performs. If it is as good as for the estimated calibration model, the model is considered as being overfitted [3].

2.3. Outliers

It is well known that proper handling of outliers is essential in data analysis. This is even more so in variable selection. Many variable selection methods are based on assessing minor differences in model quality or even in assessing statistics such as significance calculated from model parameters. Therefore, the result of variable selection is even more sensitive to outliers than the actual model fitting. Ideally, every result during the process of variable selection should therefore be complemented by careful outlier detection. This may be difficult in practice, but at the very least, the resulting model obtained after variable selection should be carefully assessed and the variable selection possibly re-run upon handling new outliers in order to verify the result.

It may also sometimes be advised to remove even minor outliers prior to variable selection. That way it is ensured that the selected variables are not merely a spurious effect of some moderate outliers. If such an approach of 'generous' outlier removal is adopted, such outliers should be reintroduced subsequently and the overall model assessed by more conventional outlier handling.

When using a variable selection model for prediction, the predictions are based on the selected variables. It is sometimes conjectured that outlier detection can suffer if only the selected variables are used for outlier detection. It may, therefore, be beneficial to base the outlier detection during prediction on the original data set. That way robustness and sensitivity are gained from using the whole 'fingerprint'. Note, though, that while this is true for some applications, the opposite may also very well be true if the removed variables are simply irrelevant in the context. In such a situation, outliers in the original data set may be completely satisfactory and normal samples. This is actually the case for the data used as an example here (see later), where the original data set has three outlying beer samples because of a difference in color of the beer. These samples, though, are not outliers in the context of the prediction problem investigated here and outlier detection should preferably be done on reduced data. In essence, the analyst should use insight on the data and problem to derive a suitable approach for outlier detection.

2.4. Redundancy

Two or more variables may be approximately similar. In a loading plot of a sound model, such variables will be located close to each other. If the purpose of the variable selection is to get a simple model, it is possible to remove the redundant variables. The predictions will probably not improve but the model will be based on fewer variables, each with a unique appearance, and it may therefore be easier to understand and interpret.

However, removing redundant variables can be risky when interpretation is the issue. Assume that two variables are correlated in a given model. One of the variables may be causally related to the response. For example, various types of foods exposed to light and oxygen will develop compounds that can react with thiobarbituric acid (TBA). The measurement of these compounds is directly related to off-flavor formation. Oxygen exposure is one of the factors leading to oxidation. It may happen that exposure to oxygen is more accurately determined and hence has a higher correlation to off-flavor. Therefore, variable selection will select oxygen rather than TBA and the causal explanation will not be visible in the final model.

2.5. Blind trust in variable selection

Sometimes, variable selection is used in an automated fashion in a black-box approach. This is not the optimal approach. In order to build proper models, it is necessary to realize that variable selection methods work under certain assumptions and will seldom provide *the* solution. However, when used appropriately, they can provide useful insight on what variables seem important, what variables seem unimportant and what variables are of intermediate importance. With such insight combined with a thorough understanding of the data, the data analyst can make qualified decisions on how to use the results of the variable selection. Quite often, for example, only the very least significant variables are removed and the variable selection is re-run (after re-checking for possible new outliers). Upon removing some of the variables, the role of the remaining ones may change, leading to new discoveries. Hence, proper variable selection is often an iterative process, where the analyst is working his or her way towards a good solution.

2.6. Pre-processing

It is important to stress that the application of pre-processing techniques such as scaling, multiplicative scatter correction, derivatives among others may affect the result of a variable selection method. A natural flow when using pre-processing and variable selection is: (1) A data set is pre-processed by, e.g. a first-order Savitzky–Golay derivative function, (2) the pre-processed data set is used as the basis for a variable selection/elimination method such as forward selection of single variables and (3) the selected variables are used in PLS model for predicting a quality parameter. In order to apply this procedure on new samples it requires access to the full spectrum in order to apply the same pre-processing as during model building. This might not be attractive if the purpose was to decrease measurement time.

An erroneous approach that is sometimes seen is that the derivatives are calculated on the reduced data set; this means that the derivative might be calculated on non-neighboring variables which, in the worst case, does not make sense. Repeated use of variable selection techniques, such as the jack-knifing or variable importance for projection (VIP) method, will exaggerate this problem. Furthermore, note that the pre-processed data *after* variable selection are different from the data set that the variables were selected from because the derivatives are calculated differently. This is not necessarily bad if predictions improve, but when it is found that the predictions do not improve upon variable selection, the user will have to check whether it was due to (1) bad variables selected or (2) the change in the pre-processing.

3. THE DATA SET

The various variable selection methods are exemplified using a data set consisting of visual and near infrared spectra of 60 beer samples (Figure 2 top) [4]. There is a large variation in the visual part of the spectra going from 400 nm to approximately 700 nm. This is due to variation in the visual appearance of the beers, which vary from very light beers to very dark beers (see also correlations in Figure 2 bottom). The area has a high variance, but it has little or no relationship to the chemical property to be predicted. In the high spectral range, dominated by absorbance of water, high absorbances lead to noisy measurements, which may also have a certain influence on the regression models but are not related to the parameter to be predicted. The remaining part of the spectrum is dominated by C–H and N–H stretching overtones except for the O–H second overtone of water at approximately 970 nm. This data set is interesting because it contains the two features that mostly lead to suboptimal models when using non-relevant variables. The left (low wavelength) part of the spectrum contains *highly systematic but non-relevant variation*. The right (high wavelength) part is mainly *unsystematic noise* and also irrelevant for predictions. Note that different measurement conditions could have provided more sound data in this region, which, however, is outside the scope of this paper.

The noisy part is typically not too difficult to handle but it leads to spurious correlations. In this case, such spuriously selected variables are easily detected visually because of the spectral nature of the data; if one variable is selected and the neighboring variables are not, then it is an indication that the result is not to be trusted. Such visual aids are not feasible for all types of data, but we can use them here to see how different approaches handle noisy data. This is an important point for the following examples. In several cases, it will be easily detected from the visual appearance of the model that certain selected areas are not really useful based on spectral insight. While such spectral insight should definitely be used when applicable, there are also many other types of data where such insight is not available. Transcriptomic data may be such an example. Still, such data will also have both types of irrelevant variations: systematic and random. We stress that the example here is chosen because of these problematic regions and the simplicity of visualizing these with spectral data, but the variable selection techniques are equally applicable on data of more discrete nature.

The systematic but irrelevant left part of the spectrum is more challenging and will typically lead to problems. The information in that area has no physical relation to the quality but because of the high variance and the limited number of samples, variables from that region are typically included by variable selection methods due to the high variance and accidental moderate correlation.

The purpose of acquiring the data is to predict the real extract concentration, which is a measure of the ability of the yeast to ferment alcohol. It is used as quality parameter in the beer production. It is measured by the brewery and mostly found in the range 4.23–18.76% plato. The uncertainty is estimated to be 0.02–0.04% plato. Predicting the real extract from spectroscopic measurements can provide a fast quality measurement in the beer production.

The data set is divided into two parts. One first part containing 40 samples is a calibration set, which is used to build the models. All decisions on variables to include as well as the number of components to use are based on the calibration set only. These are validated using a test set consisting of the remaining 20 samples. As the 60 samples were independent beer samples and as there are many samples with intermediate real extract values, the split was based on taking every third sample in the test set upon sorting by real extract value. Full (leave-one-out) cross-validation is applied in the development of the calibration model. Using all variables, this results in a four-component PLS model with a root mean squared error of cross-validation (RMSECV) of 0.88% plato and a correlation between the measured and predicted real extract of 0.88. When evaluated on the test set, an

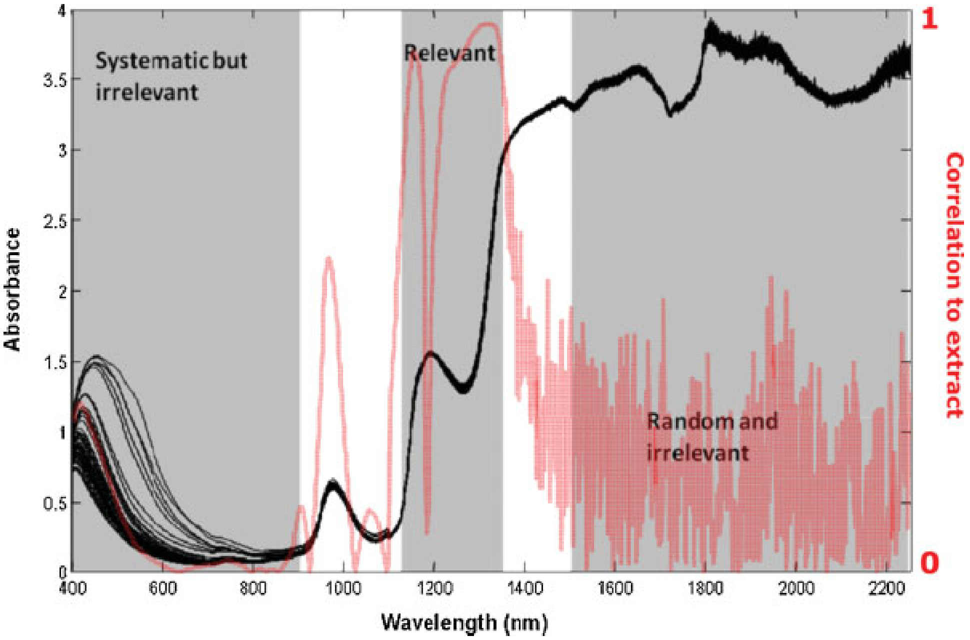

Figure 2. Visual and near infrared spectra of the 60 beer samples. In red a correlation spectrum showing correlation between each wavelength and real extract is given (scale is given on the right y-axis).

Table I. Results obtained by the various variable selection methods. The wavelengths chosen are given in the parenthesis. #-comp denotes the number of components, *r* denotes the correlation between the measured and predicted values, and RMSECV and RMSEP denote the prediction errors for the calibration model and when using the test set

Model	#-comp	r^a	RMSECV/P
No variable selection	4	0.93/0.97	0.88/0.65
Selectivity ratio (1144–1166, 1224–1350 nm)	3	0.99/0.99	0.20/0.20
*i*PLS—five intervals (1142–1510 nm)	3	0.99/0.99	0.22/0.17
*i*PLS—20 intervals (1240–1332 nm)	7	1.00/1.00	0.14/0.15
*i*PLS—40 intervals (1194–1238 nm)	4	0.99/0.99	0.22/0.26
*i*PLS—60 intervals (1200–1230 nm)	4	0.99/0.99	0.23/0.28
Genetic algorithm (vary)	3–10	-/-	0.13−0.17/-
Forward selection (1152–1244 nm)	4	0.99/0,99	0.14/0.14
Backward selection (588–774, 870–1054, 1148–1230 nm)	4	0.99/0.99	0.29/0.25
Best subset selection (588–680, 964–1054, 1148–1230 nm)	7	0.99/0.99	0.11/0.18
Jack-knifing (1184, 1310–1320, 1324–1326, 1340 nm)	2	1.00/1.00	0.23/0.16

[a] The correlation of the calibration model is given before the slash and the value obtained when using the modeling-independent test set is given after the slash.

RMSEP of 0.65% plato is obtained (Table I). For a sample set of 40 samples, many would consider a slightly more conservative segmentation than full cross-validation appropriate. In this particular case, though, results did not change when using six segments cross-validation.

4. DIFFERENT TYPES OF VARIABLE SELECTION METHODS

4.1. Using model parameters and diagnostics

In many situations, a reasonable and statistically valid model can be made using all variables. The model is perhaps not perfect, so improvements are desired. However, if a reasonable model can be made, it can help in improving the situation. This is so because a model that works is reflecting real phenomena in its parameters. Hence, variable selection methods based on appearance of model parameters can be applied to identify the variables giving the best model. Before such techniques are applied, it is extremely important to verify that the model is indeed valid, which could be done by appropriate cross- or test set validation. If the model is not valid, model parameters, such as loadings and regression coefficients, are *not* meaningful and hence wrong conclusions would be reached by using these for selecting variables.

Methods for assessing variable relevance are well known. For example, variables that have relatively low loadings in all components are often feasible to remove from the model. Likewise, variables which have low regression coefficients are feasible to remove. Figure 3 shows the loadings and regression coefficients of a PLS regression made on the beer data. The high spectral area is noisy and should be removed, since this behavior is not expected in smooth spectroscopic data where neighbor variables are highly correlated. At low wavelengths, the loadings are large, indicating that these variables should not be removed. However, the regression coefficients in the same area are small, illustrating that overall it may be an advantage to remove these variables.

It is also possible to dig a little further into a model using more derived diagnostics. For example, if the explained calibrated variance for a variable is high but the explained validated variance is low, it shows that even though this variable is reasonably modeled, it is not important when predicting new samples. It is likely that the model will improve when such variables are removed. Also, it is possible to remove variables that are highly correlated. It may not improve the predictions but the model will get simpler.

4.2. Model-based variable importance

VIP is a combined measure of how much a variable contributes to describe the two sets of data; the dependent (**y**/**Y**) and the independent variables (**X**). The VIP value for the *j*th variable is given as

$$\mathrm{VIP}_j = \sqrt{\frac{\sum_{f=1}^{F} w_{jf}^2 \cdot \mathrm{SSY}_f \cdot J}{\mathrm{SSY}_{\mathrm{total}} \cdot F}}$$

where w_{jf} is the weight value for variable *j* component *f*, SSY_f is the sum of squares of explained variance for the *f*th component and *J* the number of variables. $\mathrm{SSY}_{\mathrm{total}}$ is the total sum of squares explained of the dependent variable, and *F* is the total number of components. The weights in a PLS model reflect the covariance between the independent and dependent variables and the inclusion of the weights is what allows VIP to reflect not only how well the dependent variable is described but also how important that information is for the model of the independent variables.

A VIP smaller than one indicates a non-important variable, which could probably be removed. However, it is important also to look at the model appearance before a variable is included in the model or thrown away [5]. It is not advisable to simply remove *everything* below one. Instead, a few of the variables with the very lowest VIP values should be removed. If the model improves, the method can be repeated on the reduced data set until no more improvements are found.

The selectivity ratio (SR) is another similar measure. It is defined by the ratio between the explained variance of each variable and the residual variance. A high value denotes a variable with good predictive performance [6].

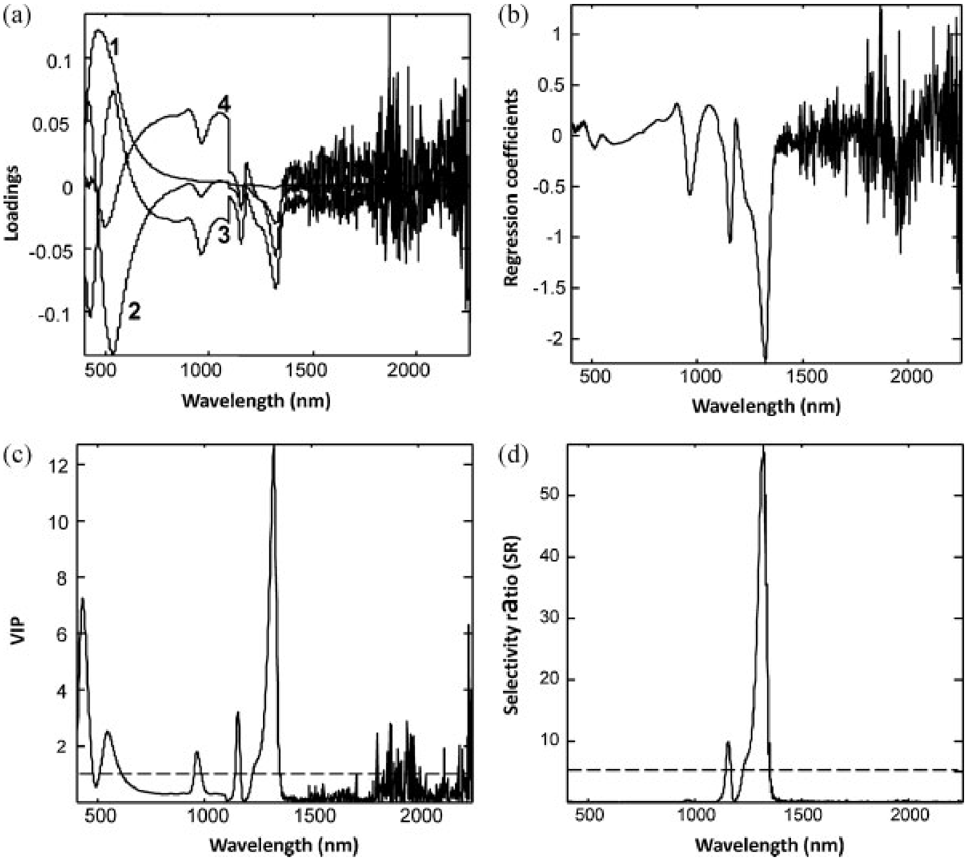

Figure 3. (a) Loadings of a four-component PLS model containing the NIR spectra of the calibration set as X-variables and the real extract of the same samples as the y-variables. Numbers indicate component number. (b) Regression coefficients of the same model. (c) The variables important for projection (VIP) calculated from the PLS model as described by Chong and Jun [7]. (d) The selectivity ratios calculated from the PLS model as explained by Rajalahti [6]. The dashed lines indicate the thresholds of one and five for the VIP and SR, respectively.

Both VIP and SR are calculated on individual variables and can therefore be displayed as the spectral data, as illustrated in Figure 3c and d. The wavelength area above 1400 nm has large but noisy regression coefficients. When calculating the VIP, these wavelengths show lower importance. The visual but irrelevant area shows some importance. Even though the visual area is not really useful for prediction, this is in accordance with what VIP is aiming at. It finds variables that are important not only for prediction but also for describing **X**. In this case, the lower visual part of the spectrum has such a high variance compared to the remaining part of **X** that it is only natural that VIP also highlights this area. Only variables in the chemical and relevant wavelength area are chosen by SR, which illustrates the advantage of this method for these data. For other types of data, SR may not necessarily provide an equally clear picture. Especially when the relevant signal is severely overlapped by other signals, SR may be more difficult to interpret than VIP. A PLS model is made on the variables with SR above five and the results are shown in Table I. The thresholds of one and five are chosen rather arbitrarily. Other values could have been tried giving similar, worse or even better predictions.

4.3. Interval partial least squares regression (iPLS)

If data are highly correlated, such as spectral data, windows of variables should be used instead of doing variable selection on each variable individually. There are several ways to do this. *i*PLS is one of the more commonly used methods. It provides an overall picture of the data set including parts that contain relevant information, interferences, noise, etc. The spectra are divided into a number of intervals of equal length and PLS models are made on each of these intervals [4]. The purpose is to find one or a few intervals, which give better predictions than the predictions obtained when using the full spectrum, and to provide an overview to help in spectroscopic interpretation. The comparison of interval performance is mainly based on the RMSECV but also on the number of components and the correlation between measured and predicted values. The number of components is chosen automatically based on the minimum RMSECV value.

It is also possible to manually make intervals instead of using intervals of equal size. This may be relevant when one wants to have a whole peak in the same interval. Furthermore, the number of intervals has to be considered. A low number of intervals may have the risk of being too broad, hindering the effects of small peaks to be seen. However, a large number of intervals may not fully take the correlations between the variables into account. Thus, when doing *i*PLS in practice, it can be an advantage to try different number of intervals and inspect and interpret any significant differences in the result.

An *i*PLS model with 20 intervals is made using the calibration samples. Figure 4 gives a graphical illustration of the predictions obtained by each of the intervals, and can, besides the use in variable selection, be of help in the chemical interpretation. As expected, the RMSECVs show high values in the visual area where

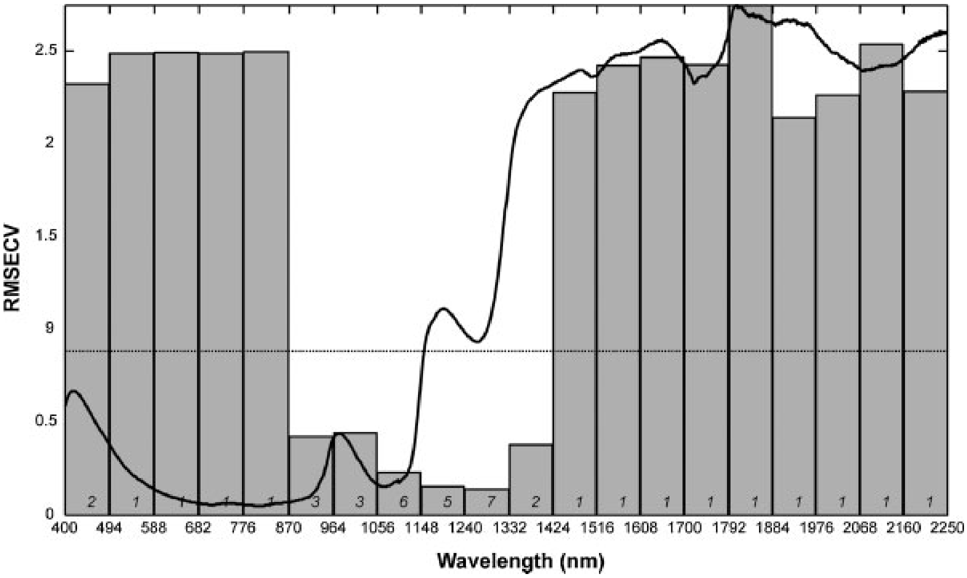

Figure 4. Illustration of the results from iPLS. The columns denote the RMSECV obtained from each of the intervals. The number in italics indicates the optimal number of PLS components in these models. The horizontal line gives the RMSECV of the full-spectrum model with four PLS components. It is plotted together with the normalized mean spectrum.

there is no chemical information. The error estimate is also high in the high wavelength area with noisy variables. Note how the use of intervals makes it more unlikely to find spurious correlations with single wavelengths in the noisy area (compare with the noisy variables with high VIP values in Figure 3c).

The use of wavelengths between approximately 900 and 1400 nm give the lowest RMSECVs also when compared with the full spectrum model. The interval with the lowest RMSECV is the one going from 1240 to 1332 nm where the correlation between the measured and predicted value is 1.00 and the RMSECV is 0.14 for a model with seven PLS components (Table I). When validated using a test set, the correlation is 1.00 and the RMSEP is 0.15. It is seen from the figure that the local models with the lowest RMSECV are the models with the highest number of components. Four components were used in the full spectrum model.

The *i*PLS approach should really be considered as an exploratory visualization of a data set rather than actual variable selection. For example, there is no way to see whether the interval around 1300 will lead to even better predictions when combined with, for example, an area in the visual part of the spectrum. Later, we will describe how the *i*PLS approach can be modified into a more dedicated variable selection approach. What remains, though, is that *i*PLS is really an effective tool for screening the importance of different parts of a data set consisting of smooth curves such as spectra.

4.4. Genetic algorithms (GA)

A GA is a technique somewhat inspired by the theory of evolution. It mimics selection in nature by evaluating models consisting of certain combinations of variables in a number of generations. Other often used methods based on the evolution theory and resembling GA are genetic programming, evolution strategies and evolutionary algorithms [8,9].

The GA is made up by a number of steps [10,11]. First, a vector consisting of zeros and ones is made with the size corresponding to the number of variables. It is denoted a chromosome. The randomly defined zeros and ones indicate the variables that should be included. Each zero or one is a gene and a PLS model made with the chosen genes is defined as an individual.

A start population is made by making a number of different individuals. The amount of individuals is defined as the population size. It is typically in the range between 20 and 500 and stays constant throughout the calculations. Each individual defines a PLS model with a subset of the original variables. For each of these models, the quality can be given, typically in terms of the RMSECV. The individuals in this first generation start producing offspring, which is a recombination of the initial chromosomes. A new population is made by randomly copying the chromosomes of the first generation where the ones that give the best predictions have a larger chance of being copied, resulting in the likely disappearance of the chromosomes that do not give as good predictions. Furthermore, cross-over and mutations of the best chromosomes are made. In cross-over, randomly selected parts of the genes in a pair of chromosomes are interchanged with a predefined frequency whereas a mutation is a change in a single gene, which takes place with a very low probability. The whole process is repeated until a stop criterion has been met, which can be a predefined number of generations, predefined time of elaboration, attainment of a predefined response value, etc. There are many variants of GAs, but this is the essential idea.

Figure 5 shows results from a GA made on the beer data. The figure shows how many times each variable has been selected after 100 generations. The variables that give the best predictions are chosen most times. Ideally, such results could be used for selecting or eliminating variables. However, the result is not convincing for the beer data where variables that have no relationship to the real extract have been chosen a high number of times in the visual part of the spectrum. Reasons may be that there are too many variables and they are correlated. The performance of the algorithm tends to decrease when there are many (e.g. more than 200) variables [12]. This is due to the risk of overfitting when the data set gets too large in relation to the number of samples. Often, this problem can be circumvented using windows of variables or by backward interval PLS (biPLS) as

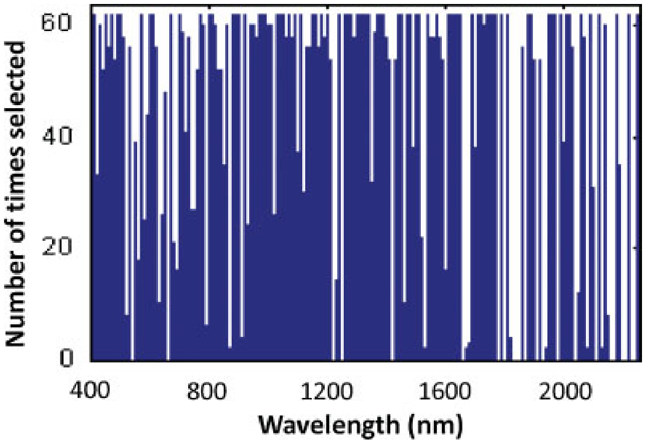

Figure 5. Number of times each variable has been selected as a function of the variable number. The GA is run with a population size of 100 and 100 generations. The mutation and cross-over rates are 0.005 and 2, respectively.

suggested by Leardi and Nørgaard [13]. Leardi *et al*. [12] made independent GAs using a large window size and removed the windows of variables that were chosen the least. This was repeated a number of times until there were less than 200 variables remaining and the use of windows of variables was no longer required.

Running a number of GAs will give slightly different results. GAs with similar parameter settings have been run six times with the use of windows to circumvent the above-mentioned problem. Figure 6 shows the number of times each window has been selected for each of the six runs. The chosen windows vary somewhat between the six runs highlighting that a single result should not be taken as the absolute truth. However, an overall tendency can be identified. The wavelength area containing chemical information has been chosen most often. Also, windows in the visual area have been selected in most of the runs. However, the high wavelengths containing mainly noise are not selected in any of the models. All runs are made with a population size of 100 and the division into 20 windows as for *i*PLS. The predictions and the number of components vary somewhat between the six runs and between the individuals in one run. For example, the average RMSECV goes from 0.13 to 0.17% using between three and 10 PLS components. The fact that one does not get one best model illustrates that GAs have a large random component in them, both in that the initial population is chosen randomly and that the following populations are chosen by guided chance. This just helps in highlighting that this variable selection method provides a helpful tool for deciding on what variables to include but should best be used as an initial screening of what parts of the data to definitely exclude. Then final decision can be made based on the evaluation of an actual model in detail. Note also that even when used blindly, all the shown models provide a significant improvement compared to a model on the raw data even though the actual variables included differ a lot.

4.5. Classical statistical approaches

There are a number of more classical statistically based approaches that have also found use in chemometrics. Forward selection is based on choosing the variables that give the best predictions one by one. Initially, the variable that results in the lowest prediction error is chosen. Alternatively, cross-validation or test set validation can also be used. Then the variable, which combined with the first variable give the best model of all two-variable models, is selected. This continues until the predictions no longer improve by adding new variables [14]. Advantages of this approach are that it is fast and it handles data sets with many irrelevant variables because it is not based on the global model. However, it disregards combined effects of several

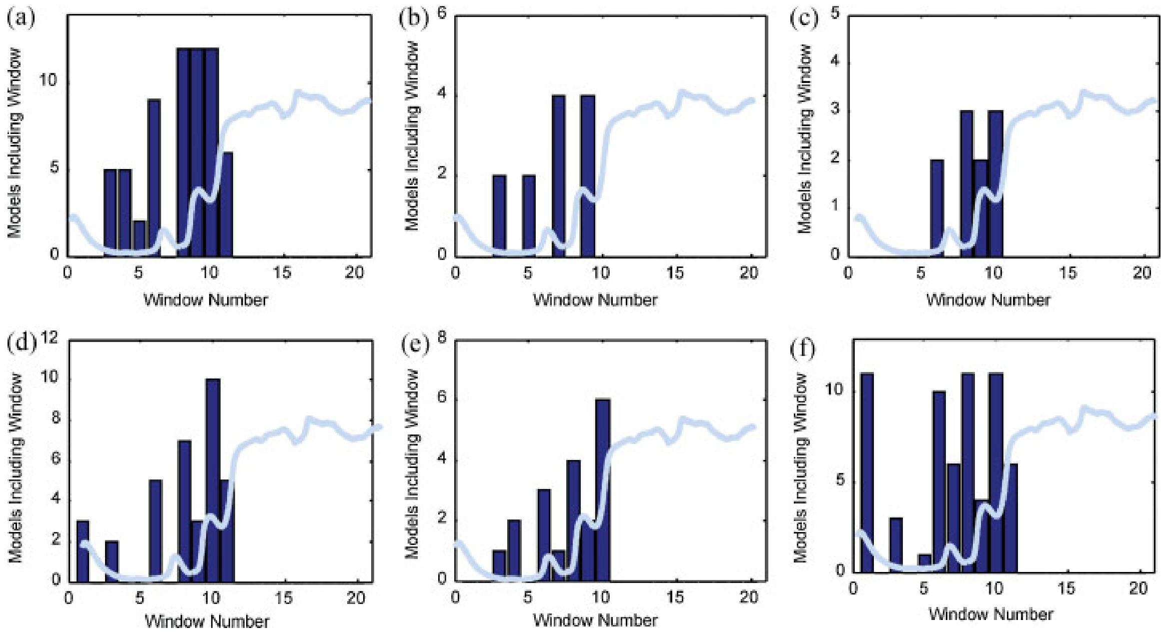

Figure 6. Results from genetic algorithms made on the beer data six times. The figure shows the number of times each window number has been chosen in the final generation. Number of generations: (a) 13, (b) 13, (c) 12, (d) 11, (e) 10 and (f) 13. On each figure, the average spectrum is overlaid to enable comparison with original data.

variables to some extent because it selects the variables one by one.

Backward selection resembles forward selection but uses the full model from which variables that contribute least to the predictions are removed one by one. This is continued until the best possible result is obtained [14]. Backward selection is reasonably fast and has the advantage that it takes combined effects of variables into account. In practice, however, it is usually only used when there are fairly few variables because otherwise, the de-selection of variables becomes cumbersome and also random to some extent. This could be so because with very few good variables and many more bad ones, it is rather arbitrary which variables are excluded initially.

A combination of forward and backward selection is named stepwise multiple regression [14]. It starts by selecting variables as in forward selection but with the condition that a variable can be removed if subsequently added variables make it less important. This is done in a way resembling backward selection.

A third method is best subset selection, which typically analyses MLR (multiple linear regression) models on all combinations of variables and chooses the combination that gives the best fit. It is exhaustive as all combinations of variables are tested, and is only possible when there are few variables [15]. Furthermore, overfitting is a serious problem because smaller sample sizes cannot support the large number of degrees of freedom that is often required [16].

4.6. Chemometric techniques based on classical approaches

There are a number of chemometric approaches resembling the classical techniques. An example is statistical significance testing, which tries to eliminate variables with non-significant regression coefficients [17]. It is fast but works only if there are not too many variables. Furthermore, it requires that the model is valid because otherwise the estimated significances will not be meaningful. Hence, the model should be able to predict reasonably well before such an approach can be pursued.

Jack-knifing is often applied for significance testing in multivariate calibration [18,19]. Jack-knifing can be conveniently done using the models generated during cross-validation. In cross-validation, a number of models on subsets of the data are estimated. From these models, the uncertainty of the parameter estimates can be calculated. Hence, it can be determined whether a regression coefficient of a variable is significantly different from zero. The jack-knife uncertainty estimate can be calculated on each of the model parameters, but in relation to variable selection, loadings or mostly regression coefficients are the natural choice. The uncertainty is estimated from the equation:

$$\mathrm{se} = \sqrt{\frac{n-1}{n}\sum\left(p_{(i)}-\bar{p}\right)^2}$$

where n is the number of jack-knife models, $p_{(i)}$ is the estimated parameter and $\bar{p}$ is the mean of the estimated parameter from all the calculated models [20]. The least significant variables can be removed and the jack-knife procedure can be continued until there is no longer an improvement in the model performance by removing insignificant variables or until all variables included are significant.

Westad and Martens [21] have previously applied jack-knifing on the beer data set. They obtained RMSECVs between 0.16 and 0.18% depending on how cross-validation was performed, the number of times jack-knife was run and how many variables were removed. It is comparable to the results given in Table I where the wavelengths at 1184, 1310–1320, 1324–1326 and 1340 nm are identified to be significant using the criterion two times the standard deviation as significance limit.

The classical statistical techniques, forward selection, backward selection and best subset selection can be applied using windows of variables rather than individual variables. When forward selection is performed with intervals, it is named forward interval PLS [22]; backward selection results in backward interval PLS (biPLS) [13] and best subset selection gives synergy interval PLS (siPLS) [23]. Furthermore, there is a moving window approach, which is named moving window PLS [24]. An example of the applicability of biPLS is as input in GA where it can be an advantage to remove non-relevant spectral parts prior to the data analysis [13]. To do this, a version named dynamic biPLS has been developed where biPLS is done several times using different number of intervals and different composition of the deletion groups in the cross-validation. This results in a vector showing how many times each variable has been included. GA will be performed on the variables included most often.

The chemometric variants of the three classical methods are applied on the beer data using 20 intervals. The results are shown in Table I. The best predictions are obtained by forward interval PLS, which gives a prediction error of 0.14 when allowing up to four intervals to be used. However, only one interval is chosen. siPLS is allowed to consist of up to four intervals and is found to give reasonable predictions for a model where a combination of three intervals is used. The poorest predictions are obtained with biPLS, which also used the highest number of intervals.

4.7. Other approaches

There are many methods that have not been mentioned in this tutorial, but we have covered a large number of typical approaches. Currently, there is a fair bit of attention on methods such as least absolute shrinkage and selection operator (LASSO) in the applied mathematics area and such methods may also be useful in chemometrics. LASSO aims to minimize the sum of squared errors as for least squares regression but with the criterion that the 1-norm of the regression coefficient vector should be below a predefined threshold. The value of this threshold determines the degree of variable selection. A small value will make many coefficients zero and this essentially provide a variable selection [25,26].

5. DISCUSSION

Table I shows that all variable selection methods improve the model performance compared with the full spectrum model for the data used here. What is of much more importance, though, is that all the methods can be used to visualize what parts of the data are assessed as important and what parts are not. In that respect, it is possible to use the results of variable selection as a starting point for producing a feasible model.

It is important that the resulting model is carefully analyzed and interpretations are performed thoroughly in order not to reach incorrect conclusions. For a certain data set, there may be several variable selection methods with similar applicability and it is always instructive to compare the results from several types of variable selection. Furthermore, some of the variable selection

Table II. Characteristics of various variable selection methods. The list of methods is non-exclusive. The following comments apply to the characteristics. The ability to work in windows rather than individual variables is extremely important for data where each underlying variable is expressed in many measured variables (e.g. many types of spectral data). The characteristic 'works with many irrelevant variables' really refers to whether the methods need a sound model initially. For example, VIP values are only useful when the overall model is valid and reasonable. The last characteristic 'can be applied repeatedly' implies that the method can be used for highlighting bad variables and remove a few of those before re-running the method

	Allows windows	Requires a valid initial model	Works with many irrelevant variables	High risk of overfitting	Can be applied repeatedly
Model parameters and diagnostics		X			X
Variable imp. projection (VIP)		X			X
Selectivity ratio (SR)		X			X
Interval PLS (*i*PLS)	X	X			
Genetic algorithms (GA)	X		X	X	X
Jack-knife		X			X
Forward interval PLS	X		X		
Backward interval PLS (biPLS)	X	X			
Synergy interval PLS (siPLS)	X	X	X		
LASSO type methods		X			X

methods apply better to the data than others. Table II shows the characteristics of the methods described here. It may be of help in the decision of which methods to use. For example, jack-knifing is as mentioned not optimal for the highly correlated beer data because it looks at individual variables rather than at windows of variables. There may also be different purposes of the analysis. GA is relevant when a fast overview of the data is the aim, but other methods are more relevant when the aim is to refine an already good model. Both GAs and *i*PLS give a visual illustration of the parts of the spectra that are important and that parts that are mainly noise or contain other irrelevant information. In essence, variable selection should rather be considered as variable *elimination* where the clearly irrelevant parts are removed and the remaining parts containing potentially useful information are kept for further data analysis.

REFERENCES

1. Höskuldsson A. Variable and subset selection in PLS regression. *Chemometr. Intell. Lab. Syst.* 2001; **55**: 23–38.
2. Anderssen E, Dyrstad K, Westad F, Martens H. Reducing over-optimism in variable selection by cross-model validation. *Chemometr. Intell. Lab. Syst.* 2006; **84**: 69–74.
3. Baumann K. Cross-validation as the objective function for variable-selection techniques. *Trends Analyt. Chem.* 2003; **22**: 395–406.
4. Nørgaard L, Saudland A, Wagner J, Nielsen JP, Munck L, Engelsen SB. Interval partial least-squares regression (*i*PLS): a comparative chemometric study with an example from near-infrared spectroscopy. *Appl. Spec.* 2000; **54**: 413–419.
5. Wold S, Johansson E, Cocchi M. *3D QSAR in Drug Design; Theory, Methods and applications*. ESCOM: Leiden, Holland, 1993; 523–550.
6. Rajalahti T, Arneberg R, Berven FS, Myhr K-M, Ulvik RJ, Kvalheim O. Biomarker discovery in mass spectral profiles by means of selectivity ratio plot. *Chemometr. Intell. Lab. Syst.* 2009; **95**: 35–48.
7. Chong IG, Jun CH. Performance of som variable selection methods when multicollinearity is present. *Chemometr. Intell. Lab. Syst* 2005; **78**: 103–112.
8. Pena-Reyes AC, Sipper M. Evolutionary computation in medicine: an overview. *Artif. Intell. Med.* 2000; **19**: 1–23.
9. Whitley D. An overview of evolutionary algorithms: practical issues and common pitfalls. *Inf. Soft. Tech.* 2001; **43**: 817–831.
10. Leardi R. Genetic algorithms in chemistry. *J. Chromatogr.* 2007; **1158**: 226–233.
11. Leardi R, Boggia R, Terrile M. Genetic algorithms as a strategy for feature selection. *J. Chemom.* 1992; **6**: 267–281.
12. Leardi R, Seasholtz MB, Pell RJ. Variable selection for multivariate calibration using a genetic algorithm: prediction of additive concentrations in polymer films from Fourier transform-infrared spectral data. *Anal. Chim. Acta* 2002; **461**: 189–200.
13. Leardi R, Nørgaard L. Sequential application of backward interval partial lest squares and genetic algorithms for the selection of relevant spectral regions. *J. Chemom.* 2004; **18**: 486–497.
14. Broadhurst D, Goodacre R, Jones A, Rowland JJ, Kell DB. Genetic algorithms as a method for variable selection in multiple liner regression and partial least squares regression with applications to pyrolysis mass spectrometry. *Anal. Chim. Acta* 1997; **348**: 71–86.
15. André CDS, Narula SC, Elian SN, Tavares RA. An overview of the variables selection methods for the minimum sum of absolute errors regression. *Stat. Med.* 2003; **22**: 2101–2111.
16. Babyak MA. What you see may not be what you get: a brief, nontechnical introduction to overfitting in regression-type models. *Psychosom. Med.* 2004; **66**: 411–421.
17. Brown PJ, Spiegelman CH, Denham MC. Chemometrics and spectral frequency selection. *Philos. Trans. Phys. Sci. Eng.* 1991; **337**: 311–322.
18. Faber NM. Uncertainty estimation for multivariate regression coefficients. *Chemometr. Intell. Lab. Syst.* 2002; **64**: 169–179.
19. Martens H, Martens M. Modified Jack-knife estimation of parameter uncertainty in bilinear modeling by partial least squares regression (PLSR). *Food Qual. Prefer.* 2000; **11**: 5–16.
20. Wehrens R, Putter H, Buydens LMC. The bootstrap: a tutorial. *Chemometr. Intell. Lab. Syst.* 2000; **54**: 35–52.
21. Westad F, Martens H. Variable selection in near infrared spectroscopy based on significance testing in partial least squares regression. *J. Near Infrared Spec.* 2000; **8**: 117–124.
22. Xiaobo Z, Jiewen Z, Xingyi H, Yanxiao L. Use of FT-NIR spectrometry in non-invasive measurements of soluble solid contents (SSC) of Fuji apple based on different PLS models. *Chemometr. Intell. Lab. Syst.* 2007; **87**: 43–51.
23. Munck L, Nielsen JP, Møller B, Jacobsen S, Søndergaard I, Engelsen SB, Nørgaard L, Bro R. Exploring the phenotypic expression of a regulatory proteome-altering gene by spectroscopy and chemometrics. *Anal. Chim. Acta* 2001; **446**: 169–184.
24. Jiang J-H, Berry J, Siesler HW, Ozaki Y. Wavelength interval selection in multicomponent spectral analysis by moving window partial least squares regression with applications to mid-infrared and near-infrared spectroscopic data. *Anal. Chem.* 2002; **74**: 3555–3565.
25. Tibshirani R. (1996) Shrinkage and selection via the LASSO. *J. R. Stat. Soc. Series B* 1996; **58**: 267–288.
26. Osborne MR, Presnell B, Turlach BA. On the LASSO and its dual. *J. Comput. Graph. Stat.* 2000; **9**: 319–337.