MODEL VALIDATION BY PERMUTATION TESTS: APPLICATIONS TO VARIABLE SELECTION

FREDRIK LINDGREN, BJÖRN HANSEN AND WALTER KARCHER
Commission of the European Communities, European Chemical Bureau, JRC Ispra Establishment, I-210 20 Ispra, Italy

MICHAEL SJÖSTRÖM
Research Group for Chemometrics, Department of Organic Chemistry, Umeå University, S-901 87 Umeå, Sweden

AND

LENNART ERIKSSON
Umetri AB, PO Box 7960, S-907 19 Umeå, Sweden

SUMMARY

Regression model validation by permutation tests was explored. Especially in cases where the model significance is doubtful, a permutation test adds crucial information which can often can be decisive for the existence of the model. The background and applicability of the test procedure are described. As an example, the use of permutation tests was extended to validation and investigation of four predictor variable selection techniques, namely MUSEUM, GOLPE, VIP and IVS-PLS. The selection methods are briefly reviewed and compared. The permutation tests were applied before, during and after variable selection. Some similarities and differences in the behaviour of the variable selection techniques were found and are commented upon. © 1996 by John Wiley & Sons, Ltd.

KEY WORDS Permutation; Variable selection; IVS-PLS; GOLPE; MUSEUM; VIP

1. INTRODUCTION

In a sound strategy for regression model development, many sequential steps are considered, such as training set selection, data pretreatment, selection of data analytical tools, variable selection or not,etc. The list of steps can be made long. An often neglected but important step is model validation. Before interpreting and using a model in prediction, it is mandatory to thoroughly investigate its predictive power.

There exist a number of ways to express the performance of a regression model. The most common parameters are the 'explained variance' for the response variable **y**, denoted R^2, and the residual standard deviation (RSD). The term R^2 is often referred to as 'model fitness' and should preferably be as close to unity as possible, while the RSD should be kept small. For judging a

* Author to whom correspondence should be addressed.

CCC 0886–9383/96/060521–12 $17.50

model's predictive power, meaning how well it performs in forecasting, techniques such as cross-validation[1-3] (CV) and bootstrapping[4] are frequently used. Both approaches give an internal assessment of the predictive ability, often denoted Q^2 or R^2-cross-validated. Compared with R^2, the measure Q^2 can at maximum be unity (same as R^2) but with no lower boundary (minus infinity). The extension of CV and bootstrapping is of course external validation, performed by keeping a test set of compounds outside all model development steps. When the final model has been established, predictions are made for the test set, creating an external Q^2. A comparison of internal and external Q^2s gives an idea of the validity and robustness of the model.

However, knowing R^2, RSD and Q^2 might sometimes not be sufficient for judging the validity of a model. Prior to modelling, one seldom knows the relevance of a predictor variable matrix **X** for modelling a response variable **y**. Thus each model has a possibility of being developed by pure chance, a so-called coincidental or chance correlation. Further, every regression data set contains structures which cause so-called 'background correlation'. This can easily be verified by modelling data sets based on randomly generated numbers.[5] Especially in cases where the number of samples is small ($N < 15$) and the number of variables is large ($K > 100$), the background correlation increases drastically. Also, the latent structure in the predictor variable matrix **X** and the variable distribution in both **X** and the response variable **y** are influential factors for the background correlation.

An easy and efficient way to check how far a model is from being a coincidence correlation is by performing a permutation test. Such tests have been known for a long time,[6-10] but their use has been limited in the area of chemometrics. Especially in cases where the significance of a model is doubtful, such a test provides crucial information on whether to keep or reject a model. The explanation and use of permutation tests are more thoroughly described in a separate section of this paper.

One critical step in a model development scheme which requires extensive validation is predictor variable selection. Today, variable selection has become the hottest fashion in chemometrics and several approaches have recently been suggested in the literature.[11-18] The reason for variable selection is usually to improve some statistical parameter (R^2, Q^2 or RSD) or simply to achieve a model simplification. The improvements should of course be of a global nature and they are supposed to hold in a broader perspective. A 'locally' improved model may have limited or no consequence in a wider application. The only way to verify such enhancements is by the use of an external validation set. Unfortunately, this is seldom seen in reality, mostly owing to a limited number of samples. In many cases one still likes to perform predictor variable selection even though external validation is not feasible. In these cases a permutation test is a valuable addition to the internal validation scheme.

In this paper, permutation tests were applied to investigate variable selection. Four different variable selection techniques, namely GOLPE[11,12] MUSEUM[13,14] IVS-PLS[15,16] and VIP,[17] were used in a small comparative study using the Selwood data set.[19] The four methods and the QSAR-type data set will be briefly reviewed. The permutation test was applied prior to, during and after the variable selection procedures. Some differences and similarities between techniques were recorded.

2. PERMUTATION TEST IN REGRESSION MODELLING

2.1. Background

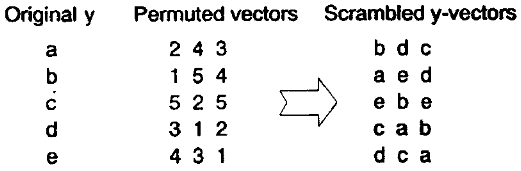
Regression modelling sometimes produce models of questionable significance. Even though a model is determined significant, the efficient rank can often be a problem using latent variable techniques such as PCR[20] or PLS.[21-23] As mentioned earlier, all regression data sets may have a so-called background correlation. Usually, chance models are detected by bad statistical performance (R^2, RSD and Q^2), but this may not always be the case.

In a previous investigation, tables of significance levels for cross-validation in connection with PLS were proposed based on modelling of data sets of random numbers.[24] It was discovered that several factors influenced the so-called 'background of predictive power' (Q^2 of models based on random numbers). The number of samples (N), the number of variables (K) and the number of model dimensions (A) were found to be the most important ones. However, it was also seen that the latent structure in the predictor variable matrix **X** and the variable distribution in both **X** and the response variable **y** are influential factors. In addition, the influence of various numbers of cross-validation groups was never investigated, which might be one of the more crucial factors. Consequently, for a correct use of these published tables, one needs specific knowledge and experience about the particular data structure. A simple way to solve this problem is by performing a permutation test, which was also indicated by the authors of Reference 24.

2.2. Test procedure

An easy and efficient way to check for background correlations is described here. The method is based on a repetitive reordering of the N entries in the response variable **y**. First a random number generator is applied to produce permuted vectors containing integers between 1 and N. In this investigation, 25 vectors were used as default. In the second step the elements in the original **y**-variable are reordered according to the integer vectors, creating 25 scrambled response variables just by switching their internal positions, as illustrated in Figure 1. These new **y**-variables should have no or very limited association with the predictor variables in **X**. As a third step the scrambled **y**-vectors are modelled, one by one, using the intact **X**-data. In every run, R^2 and Q^2 are calculated and recorded. Finally, when all scrambled ys have been modelled, their R^2-values and Q^2-values are presented as distributions (background distributions) and compared with the 'real' ones calculated from the original data; see Figure 2. The result can also be presented as a percentage overlap between the real R^2 and Q^2 and their respective background distributions. A 0% overlap is of course the optimal condition.

When creating the permuted vectors of integers, two slightly different approaches can be used. The difference lies in the association with the original **X**-matrix. The first approach is simply to generate the permuted vectors freely without positional restrictions. This implies that for each permuted vector, each vector element might end up in its original position (original **y**-element no. 3 is placed at the permuted position no. 3, meaning no change). The extreme case is of course when a scrambled **y**-vector is identical with the real one. The statistical implication of this is that the real R^2 and Q^2 are both inside their respective background distributions if all possible permutation combinations are considered (faculty of N). This might complicate the formulation of a statistically correct significance criterion.



Original y	Permuted vectors	Scrambled y-vectors
a	2 4 3	b d c
b	1 5 4	a e d
c	5 2 5	e b e
d	3 1 2	c a b
e	4 3 1	d c a

Figure 1. A y-variable consisting of letters a–e is rearranged into three new scrambled y-vectors by following the integer combinations of the three permuted vectors

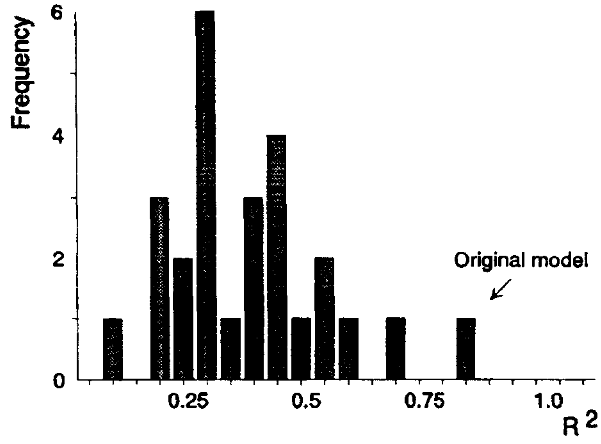


Figure 2. Results of a permutation test presented as a histogram. The R^2 from the original model is compared with the distribution of 25 R^2s from the modelling of scrambled y-vectors. In this case the original R^2 is significantly higher than the others

A second way is to control the permutation and check that the association between X and y is completely eliminated. One ascertains that no vector element is placed on its right location. The number of possible combinations is heavily reduced by this constraint; however, the real R^2 and Q^2 are in this case non-members of their corresponding background distribution This facilitates a simple statistical evaluation of the result, since both the real R^2 and Q^2 are assumed to be outside the background distributions for a significant model.

We note that the difference between these two approaches will only be seen when performing permutation test on data sets with a small number of samples ($N < 15$). The larger N gets, the more the two procedures will resemble one another.

In a third way the two approaches described above are merged into an intermediate and more transparent procedure. The permutation is made as outlined in the first approach (no constraints), but additionally the correlation between each scrambled y-vector and the real y is registered.[25] With this extension the result, usually presented as a distribution, can be visualized as a two-dimensional scatter plot; see Figure 3. On the horizontal axis the R^2 between the real y and the scrambled y-vectors is given and on the vertical axis their respective explained variances (R^2-model) when modelled by X. This means that the background R^2 distribution can be recreated (as in Figure 2) simply by projecting all points down on the vertical axis.

The advantages of this plot are several. In this way one can keep track of the correlation between the real y and the scrambled ones. For a scrambled y-vector with high correlation to the real y a high R^2-model is also expected. The calculated regression line and equation also give some additional information. The intercept with the vertical axis gives an estimation of the background R^2-model in the case where no correlation between the real y and the scrambled ones exists (background $R^2 = 0.14$, Figure 3). A near-zero slope of the regression line and a high intercept indicate cases where extra caution should be taken in using and interpreting the 'real' model.

Compared with using predefined tables of significance levels based on random numbers, a permutation test allows the user to test the specific behaviour of a particular data set. The unique structure of each data set is exclusively investigated, which makes this technique preferable.

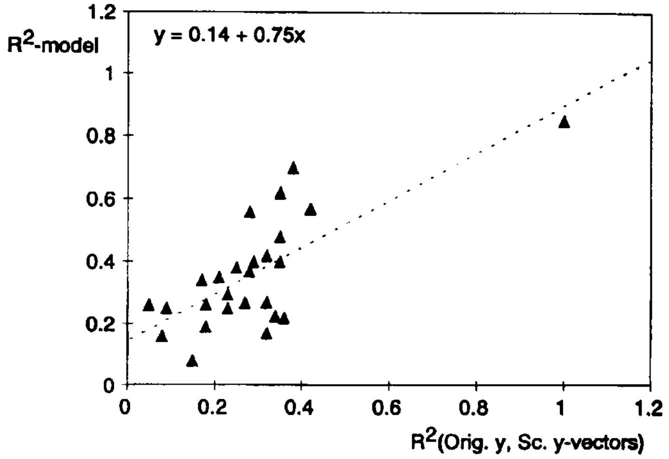


Figure 3. Scatter plot of results from a permutation test. On the horizontal axis the R^2 for the intercorrelation between the original y and the scrambled y-vectors is given. The vertical axis shows the respective R^2 when modelled by **X**

2.3. Validation of variable selection by permutation tests

In this paper the use of permutation tests was extended to predictor variable selection. The test was used in two different ways. In the first study it was applied to models based on all predictor variables and on selected subsets. The background R^2 and Q^2 before and after variable selection were recorded and used as the key measurement. In a second study the permutation test was applied to the variable selection technique itself. Predictor variable selection was performed using permuted y-variables as well as the parent one. The improved predictive power when modelling the 'real' y was compared with the overall improvement for the permuted ones. This test is crucial for variable selection techniques and some differences between the investigated methods were noticed.

3. VARIABLE SELECTION TECHNIQUES

3.1. Review

In the study, four variable selection techniques were used. Emphasis was placed on comparing their behaviour in a permutation test, not necessarily as a competition between them. For ease of understanding, they are briefly reviewed.

MUSEUM[13,14] (mutation and selection uncover models) is an evolutionary strategy that includes mutation and selection but avoids crossover of regression models. The algorithm starts from any number of randomly chosen variables. Random mutation, first by addition or elimination of a very limited number of variables and later by larger fractions, leads to new variable-reduced models which are all evaluated by an appropriate fitness function. Compared with a common genetic algorithm, only the 'fittest' subset is kept and used for further mutation. As a final step of mutation, all variables inside the model are eliminated and all variables outside the model are added, one by one, to detect new possible mutations. The described procedure is repeated several times, always re-entering the best variable subset as the starting point in the mutation scheme. In the very last step, variables not significant at a 95% level are eliminated, starting with the least significant one.

GOLPE[11,12] (generating optimal linear PLS estimators) is a technique for variable selection developed for PLS, but the strategy is possible to apply to any regression method. In an early version a preselection was performed removing redundant variables by a D-optimal selection in the PLS weights. This step has been deleted in later versions. The method is based on evaluation of the predictive power of a number of PLS models built by different combinations of predictor variables selected according to a factorial design. The models are made to systematically evaluate each variable's contribution to a model. The variable subset is concluded after all combinations have been investigated through an evaluation of the design matrix. In GOLPE, considerable cross-validation is applied, together with a dummy variable check to justify the validity of the result. This elevates the reliability but makes the computation more tedious.

IVS-PLS[15,16] (interactive variable selection for PLS) is a recent development aiming at improving PLS prediction through selective reweighting of single elements in the PLS weight vector **w** (indirect X-variables). Different to other techniques, the selection is made dimensional-wise. A cut-off limit regulating the size of the rejected elements is introduced by two approaches, inside-out (removing small **w**-values) and outside-in (removing large ones). A rejected variable simply receives a **w**-value equal to zero. However, the cut-off limit is not fixed. It is gradually changed and for each new threshold a CV-value is calculated. Finally the CV-values are plotted versus the cut-off limit, producing a curve (CV-plot). A minimum in this curve indicates a preferable cut-off limit for reducing the CV-value (equal to improving Q^2). At the starting point of each new dimension, all variables are included and because of only using two approaches for selection, the number of possible variable combinations is very limited.

VIP[17] (variable importance in the projection) is a value indicating the contribution of each predictor variable to a model. VIP is the sum over all model dimensions of the variable influence (VIN). For a given PLS dimension a, $(\mathrm{VIN}_{ak})^2$ is equal to the squared PLS weight $(a_{kq})^2$ multiplied by the percentage explained SS by that PLS dimension. The VIP-value is accumulated (over all PLS dimensions) and divided by the total explained SS by the model and multiplied by the number of **X**-variables in the model. This scaling makes the squared sum of all VIP-values equal to the number of **X**-variables in the model. Finally, one can compare the VIP-value for one **X**-variable with another and the ones above unity (used in this study) are generally considered as the most relevant to explain **y**. It is noted that for a one-dimensional PLS model, VIP is proportional to the squared values of **w**:

$$\mathrm{VIP}_k = \left(\sum_a (\mathrm{VIN})^2_k/\mathrm{SS}_a\right)K \tag{1}$$

As seen, VIP is actually not an independent selection technique but the measurement can still be used for variable subset selection. Using VIP, simply the explanation of **y** is considered and not the predictive capability.

3.2. A brief comparison of selection techniques

A clear difference between the four methods is noticed in simplicity, speed of calculation and the number of possible variable combinations, VIP being the simplest, fastest but maybe also the most limited one. The VIP approach is very similar to a variable selection by interpretation of PLS weight plots. The IVS method is close to VIP, but two clear differences remain. In IVS the selection is made according to predictive power instead of covariance with **y**, and by the outside-in approach, removal of variables with a large PLS weight is possible. This was found to be beneficial in cases where an inside-out approach was used in a previous dimension. However, common to both are that the starting point is a PLS model with all variables included and the required computation time is short (seconds for VIP and a few minutes for IVS-PLS when $K = 500$).

In both GOLPE and MUSEUM the number of variable combinations investigated is substantially increased. Nevertheless, a distinct difference exists between the two in the selection scheme. In GOLPE a predefined number of runs is set up by the number of rows in the factorial design. The selected subset is not the variables from the best-performing model but rather the ones considered important in evaluating the complete factorial design matrix.

In MUSEUM the best variable subset is constantly updated through an evolutionary algorithm. As also seen in the closely related genetic algorithms, the final subset is often dependent on the randomly selected starting set and the mutation itself. This implies that several 'final' models are generated with similar statistical performance, but based on rather different variable subsets. When the total number of available variables increases, this becomes a serious problem.

4. DATA SET AND EXPERIMENTAL DETAILS

The Selwood data set[19] is frequently used in studies of variable selection. The data describe the antifilarial activity for a set of 31 antimycin A_1 analogues. The compounds were characterized by 53 physico–chemical descriptors. The response variable is their *in vitro* biological activity ($\log[1/EC_{50}(\mu M)]$. For this data set, several variable subsets found by the MUSEUM method have been published. This makes it possible to include MUSEUM in the comparison even though the algorithm was not accessible to us.

The SIMCA-P programme[26] was used for PLS modelling of the full data set as well as for reinvestigation of the subsets selected by MUSEUM and VIP calculations. For GOLPE the PC-version 1.0 was applied and for IVS-PLS an in-house version running in MATLAB[27] was utilized. The criteria to reach the best-performing subset were very different between techniques. However, when the variable subsets were eventually determined, the comparison was made through a reinvestigation by a two-dimensional PLS model (SIMCA-P) using seven CV-groups for estimating Q^2. Because of the specific dimensional building of an IVS-PLS model, recalculation in SIMCA-P could not be made. However, the CV-procedure in IVS-PLS is identical with SIMCA-P, which justifies the comparison. Further, all negative Q^2s were set to zero. This implies that for the second study, improvements solely in negative Q^2-values ($-0{\cdot}3$ to $-0{\cdot}1$) were not recorded.

5. RESULTS AND DISCUSSION

5.1. Permutation test before and after predictor variable selection

In the first study a permutation test was applied to the full data ($K = 53$) prior to variable selection and to the subsets suggested by the four techniques. The subsets are given in Table 1. The number of used variables ranged from 10 (MUSEUM) to 23 (IVS-PLS). The four subsets contained some overlaps and three predictor variables were found by all techniques (nos. 4, 50 and 52).

The result of the permutation test (R^2 and Q^2) is presented in Table 2. The R^2-values of the full data set models are presented in a two-dimensional scatter plot, Figure 4. A wide range of R^2-values ($0{\cdot}308$–$0{\cdot}711$) is found for the 25 scrambled **y**-vectors. The intercept also confirms that even when no correlation between the real **y** and the scrambled ones exists, a background correlation of $0{\cdot}50$ is present. R^2 for the original model was still found to be unique compared with

Table 1. Best predictor variable subsets judged by four different techniques

Method	Predictor variable subset (column no.)	*K*
MUSEUM	4, 11, 16, 18, 37, 38, 40, 50, 52, 53	10
GOLPE	1, 3, 4, 6, 7, 9, 12, 14, 17, 22, 32, 50, 52	13
VIP	1, 2, 3, 4, 5, 6, 7, 12, 14, 17, 25, 31, 37, 42, 43, 50, 51, 52	18
IVS-PLS (Dim. 1)	1, 2, 3, 5, 6, 7, 17, 25, 26, 31, 32, 33, 34, 37, 42, 43, 50, 51, 52, 53	
(Dim. 2)	4, 12, 14, 50	23

Table 2. Permutation test results for first study

	Full model		MUSEUM		GOLPE		VIP		IVS-PLS		Int. Corr.
Number of	($K = 53$, $A = 3$)		($K = 10$, $A = 2$)		($K = 13$, $A = 2$)		($K = 18$, $A = 2$)		($K = 23$, $A = 2$)		[y, y-sc][b]
y-variable	R^2	Q^2	R^2	Q^2	R^2	Q^2	R^2	Q^2	R^2	Q^2	R^2
Orig.	0·781	0·449	0·795	0·494	0·714	0·609	0·730	0·542	0·743	0·671	1
1	0·690	0·102	0·263	0·091	0·288	0·038	0·233	0	0·248	0·108	0·008154
2	0·560	0	0·248	0·004	0·186	0	0·300	0	0·217	0	0·027093
3	0·387	0	0·175	0	0·133	0	0·170	0	0·131	0	0·05598
4	0·654	0	0·307	0	0·427	0	0·375	0	0·330	0	0·020851
5	0·438	0	0·145	0	0·107	0	0·134	0	0·138	0	0·003318
6	0·574	0·069	0·331	0·147	0·177	0	0·264	0	0·157	0	0·069064
7	0·411	0	0·239	0	0·319	0	0·252	0	0·262	0	0·000025
8	0·501	0	0·067	0	0·171	0	0·203	0	0·236	0	0·01004
9	0·476	0	0·217	0	0·257	0	0·285	0·004	0·168	0·021	0·007208
10	0·381	0	0·160	0	0·163	0	0·232	0	0·090	0	0·003125
11	0·542	0	0·100	0	0·145	0	0·201	0	0·105	0	0·015326
12	0·711	0·066	0·234	0	0·272	0	0·268	0	0·359	0	0·00093
13	0·571	0	0·262	0	0·276	0	0·253	0	0·165	0	0·004396
14	0·665	0·330	0·525	0·327	0·554	0·354	0·530	0·393	0·466	0·380	0·136235
15	0·426	0	0·300	0·012	0·187	0	0·288	0	0·266	0	0·018824
16	0·308	0	0·126	0	0·058	0	0·081	0	0·059	0	0·2·6 × 10^6
17	0·501	0	0·237	0	0·177	0	0·176	0	0·224	0	0·009351
18	0·557	0	0·103	0	0·234	0	0·313	0	0·165	0	0·013363
19	0·398	0·009	0·273	0·044	0·201	0	0·195	0	0·179	0	0·002694
20	0·631	0	0·152	0	0·238	0	0·295	0	0·215	0	9·2 × 10^5
21	0·576	0·040	0·275	0	0·176	0·029	0·242	0	0·178	0·048	0·02873
22	0·353	0	0·169	0	0·230	0	0·190	0	0·255	0	0·033782
23	0·498	0	0·283	0	0·138	0	0·209	0	0·170	0·019	0·001436
24	0·483	0·004	0·308	0	0·191	0·030	0·206	0	0·201	0	0·102848
25	0·469	0	0·170	0	0·145	0	0·129	0	0·146	0	0·003856
Avg·[a]	0·510	0·024	0·226	0·025	0·218	0·018	0·240	0·015	0·205	0·023	0·0230
SD[a]	0·109	0·069	0·096	0·071	0·103	0·070	0·088	0·078	0·088	0·078	0·0342

[a] The original model results were not included in the calculation of the average or standard deviation.
[b] The internal correlation between the scrambled ys and the original y-variable.

the permutation test models, although the significance level was not very high. This indicates that R^2 alone is not a sufficient parameter for judging the model validity for the present data set.

The related comparison of Q^2s gave a clearer indication that the original model performs significantly better than the permutation test models. The background Q^2 was estimated to be 0·01, compared with 0·45 for the real model.

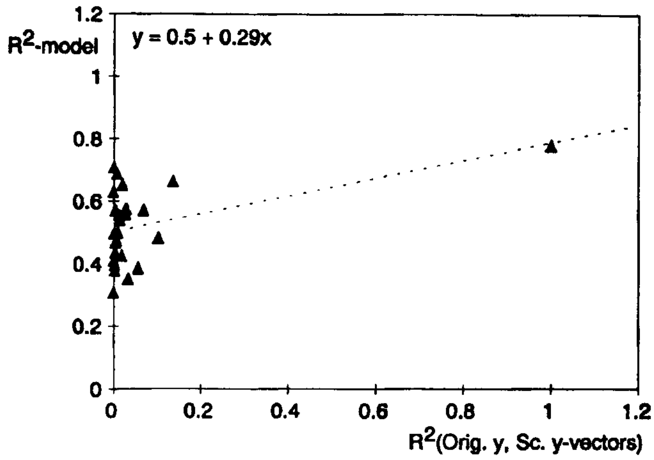


Figure 4. Scatter plot of permutation test results from first study. The models are based on all 53 predictor variables. The axes are the same as in Figure 3

In the bottom rows of Table 2 the mean values of each column are given (the original model is not included). Comparing the average R^2 from models based on the four subsets with the average R^2 for the full data set models, a clear drop (>50%) in background correlation was observed. Interestingly, the four subsets performed rather similarly (average R^2 between 0·20 and 0·24). The distinct reduction in background R^2 has several explanations. Firstly, with a small number of variables (K) the theoretical chance of having high background R^2 is lower. Secondly, the variable subsets are optimized for explaining the real y, which implies that their descriptive power for other y-vectors is limited. Also, the lower dimensionality of the variable-reduced models has a positive effect for the background R^2. No difference was found in comparing average Q^2s, mainly because the average Q^2 for the full data set models was already very low (0·024).

5.2. Variable selection applied to permuted response vectors

In a second study a permutation test was directly applied to the variable selection itself. VIP, IVS-PLS and two versions of GOLPE (selected variables and selected plus uncertain variables) were tested. Because of no access to the algorithm, MUSEUM was left out.

For each of the 25 scrambled y-vectors 'improved' models were generated and the results (R^2, Q^2 and the improvement in Q^2 compared with the full data set model) are given in Table 3. The improvement in Q^2 for this type of permutation test should preferably be as small as possible. This measurement indicates the robustness of a variable selection technique used for particular data. The Q^2 improvements can be seen as the background improvement that each technique produces for the specific data structure.

The results in Table 3 are summarized in four bar charts (Figures 5(a)–5(d)). GOLPE (selected plus uncertain) and IVS-PLS gave the smallest average improvements in Q^2 (Figure 5(a). GOLPE (selected) and VIP gave overall higher improvements. The reverse was found in Figure 5(b), displaying the average R^2 for the same models. In this case, GOLPE (selected) gave models with significantly lower R^2s than the other methods. This implies that for GOLPE (selected) the explained variance in y was clearly sacrificed for improvement in predictive power.

Table 3. Permutation test results from second study

Number	VIP			IVS-PLS			GOLPE (selected)			GOLPE (selected plus uncertain)		
of y-variable	R^2	Q^2	Imp.	R^2	Q^2	Imp.	R^2	Q^2	Imp.	R^2	Q^2	Imp.
Orig.	0·730	0·542	0·098	0·743	0·671	0·222	0·741	0·609	0·160	0·673	0·473	0·024
1	0·537	0·428	0·326	0·590	0·334	0·232	0·615	0·457	0·355	0·557	0·257	0·155
2	0·404	0·026	0·026	0·410	0·083	0·083	0·316	0·168	0·168	0·403	0·095	0·095
3	0·303	0·002	0·002	0·306	0	0	0·194	0·019	0·019	0·209	0·019	0·019
4	0·518	0·220	0·220	0·546	0	0	0·448	0·202	0·202	0·554	0	0
5	0·407	0·065	0·065	0·364	0	0	0·357	0·068	0·068	0·346	0	0
6	0·513	0·312	0·243	0·300	0·199	0·130	0·443	0·233	0·164	0·474	0·214	0·145
7	0·320	0	0	0·310	0	0	0·315	0	0	0·315	0	0
8	0·324	0	0	0·404	0	0	0·076	0	0	0·182	0	0
9	0·382	0·021	0·021	0·375	0	0	0·236	0·083	0·083	0·356	0·004	0·004
10	0·262	0·089	0·089	0·244	0·010	0·010	0·171	0·070	0·070	0·238	0	0
11	0·512	0·008	0·008	0·422	0	0	0·169	0·031	0·031	0·390	0	0
12	0·621	0·421	0·355	0·584	0·351	0·285	0·618	0·446	0·380	0·638	0·077	0·011
13	0·480	0	0	0·500	0	0	0·360	0·007	0·007	0·496	0	0
14	0·581	0·461	0·131	0·620	0·406	0·076	0·607	0·464	0·134	0·611	0·433	0·103
15	0·404	0·037	0·037	0·357	0	0	0·394	0·070	0·070	0·402	0·020	0·020
16	0·206	0	0	0·218	0	0	0·159	0·021	0·021	0·159	0·021	0·021
17	0·429	0	0	0·416	0	0	0·391	0·112	0·112	0·453	0	0
18	0·399	0	0	0·436	0	0	0·298	0·095	0·095	0·365	0	0
19	0·325	0·144	0·135	0·306	0·131	0·122	0·279	0·168	0·159	0·351	0·131	0·122
20	0·524	0·052	0·052	0·532	0	0	0·432	0	0	0·462	0	0
21	0·443	0·193	0·153	0·416	0·076	0·036	0·475	0·184	0·144	0·471	0·184	0·144
22	0·220	0	0	0·238	0	0	0·215	0	0	0·239	0	0
23	0·385	0·039	0·039	0·361	0	0	0·225	0·090	0·090	0·272	0·082	0·082
24	0·404	0·068	0·064	0·317	0·136	0·132	0·312	0·226	0·222	0·398	0·123	0·119
25	0·313	0	0	0·404	0	0	0·131	0	0	0·168	0	0
Avg.[a]	0·409	0·103	0·079	0·399	0·069	0·044	0·329	0·129	0·104	0·380	0·066	0·042
SD[a]	0·109	0·149	0·105	0·112	0·124	0·079	0·151	0·144	0·105	0·136	0·108	0·057

[a] The original model results were not included in the calculation of the average or standard deviation.

The number of improved models also varied substantially (Figure 5(c), with IVS-PLS and GOLPE (selected plus uncertain) being the most moderate ones (nine and 13 respectively). Finally, for comparative reasons, the improvements in Q^2 for the original data set are given in Figure 5(d). GOLPE (selected plus uncertain) was very restrictive for the original data, while IVS-PLS and GOLPE (selected) gave significant improvements (0·22 and 0·16, respectively).

Concluding the bar charts (Figures 5(a)–5(d)), it may be stated that the improvement made for the model based on the original data is not significantly different from the improvements for the 25 permutation test models. No variable selection technique treated the original data in a unique way. As mentioned earlier, the outcome of a permutation test depends on several factors: number of objects, number of variables, etc. Thus the results of this study (Table 3) are specific for this particular data set. Consequently, the individual values of background R^2 and improvements in Q^2 for different variable selection techniques will most likely change when other data sets are investigated. However, the overall differences in behaviour between techniques might not change very much.

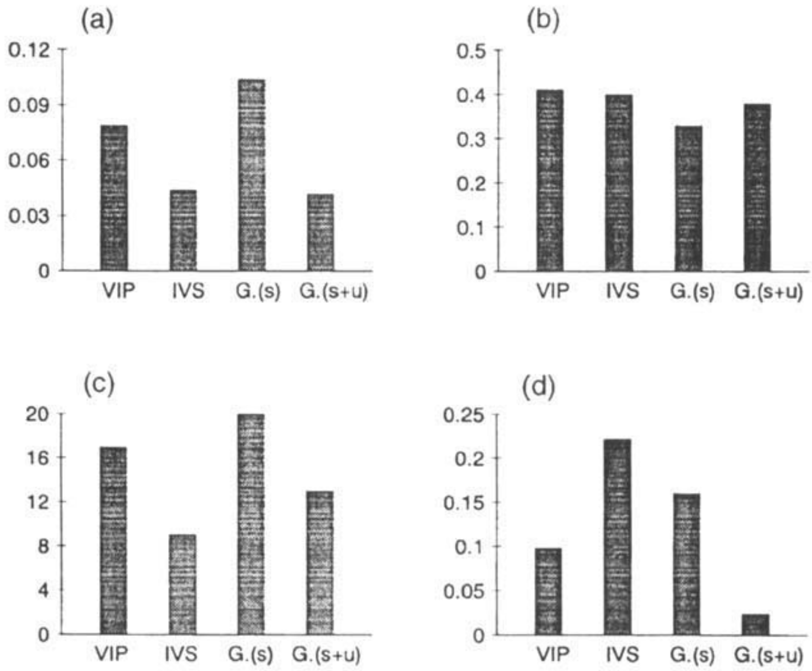


Figure 5. Four bar charts summarizing the permutation test results in second study: (a) average improvement in Q^2 for four different variable selection techniques modelling 25 scrambled **y**-vectors; (b) average R^2 for 25 models; (c) total number of improved models considering Q^2; (d) improvement in Q^2 for original data set

6. CONCLUSIONS

A permutation test is a useful and complementary tool for assessing the validity of a regression model. In this paper the procedure is described and used in two small examples. In the first study a three-dimensional PLS model on the Selwood data set was determined significant with respect to both R^2 and Q^2, based on the outcome of the permutation test. It was also seen that the background R^2 is considerably lowered using a predictor variable subset instead of the full **X**-matrix.

In the second study the reliability of variable selection techniques was investigated by permutation tests. It was seen that all four methods generated model improvements for data sets with scrambled **y**-vectors. The results indicate that the improvements found by any of the used variable selection techniques are partly non-relevant and can be considered as an overfitting of the measure Q^2. The average improvements can he considered as a threshold for each individual technique. The study also indicates that the use of external validation is strongly recommended when variable selection has been applied.

In future investigations, several data sets with varying numbers of objects and variables will be considered in a similar approach. Hopefully, this will give more information on the possibilities, limitations and pitfalls of feature selection.

ACKNOWLEDGEMENTS

The financial support from the Directorate General XII of the Commission of the European Communities is gratefully acknowledged. We thank Professor Sergio Clementi and Dr. Gabriele

Cruciani, University of Perugia, for introduction and guidance to using GOLPE and for inspiring collaboration. Finally, Professor Hugo Kubinyi, BASF, Ludwigshafen, and Professor Svante Wold, Umeå University, are thanked for stimulating discussions on various approaches concerning predictor variable selection.

REFERENCES

1. M. Stone, *J. R. Stat. Soc. B*, **36**, 111–133 (1974).
2. S. Geisser, *Biometrika*, **61**, 101–107 (1974).
3. S. Wold, *Technometrics*, **20**, 379–406 (1978).
4. C. Leger, D. N. Politis and J. P. Romano, *Technometrics*, **34**, 378–398 (1992).
5. M. Clark and R. D. Cramer III, *Quant. Struct.-Act. Relat.* **12**, 137–145 (1993).
6. A. Walf and J. Wolfowitz, *Ann. Math. Stat.* **15**, 358–372 (1944).
7. W. Hoeffding, *Ann. Math. Stat.* **23**, 169–192 (1952).
8. E. S. Edgington, *Statistics: Textbooks and Monographs*, No. 77, *Randomization Tests*, 2nd edn, Marcel Dekker, New York (1987).
9. R. L. Schmoyer, *J. Am. Stat. Assoc.* **89**, 1507–1516 (1994).
10. I. N. Wakeling, M. M. Raats and H. J. H. MacFie, *J. Sens. Stud.* **7**, 91–96 (1992).
11. M. Baroni, S. Clementi, G. Cruciani, G. Costantino, D. Riganelli and E. Oberrauch, *J. Chemometrics*, **6**, 347–356 (1992).
12. M. Baroni, G. Costantino, G. Cruciani, D. Riganelli, R. Valigi and S. Clementi, *Quant. Struct.-Act. Relat.* **12**, 9–20 (1993).
13. H. Kubinyi, *Quant. Struct.-Act. Relat.* **13**, 985–294 (1994).
14. H. Kubinyi, *Quant. Struct.-Act. Relat.* **13**, 393–401 (1994).
15. F. Lindgren, P. Geladi, S. Rannar and S. Wold, *J. Chemometrics*, **8**, 349–363 (1994).
16. F. Lindgren, P. Geladi, A. Berglund, M. Sjöström and S. Wold, *J. Chemometrics*, **9**, 331–342 (1995).
17. S. Wold, in *Methods and Principles in Medicinal Chemistry*, Vol. 2, *Chemometric Methods in Molecular Design*, ed. by H. van de Waterbeemd, pp. 195–218, Chemie, Weinheim (1994).
18. R. Leardi, *J. Med. Chem.* **33**, 136–142 (1994).
19. D. L. Selwood, D. J. Livingstone, J. C. W. Comely, A. B. O'Dowd, A. T. Hudson, P. Jackson, K. S. Jandu, V. S. Rose and J. N. Stables, *J. Med. Chem.* **33**, 136–142 (1990).
20. H. Martens and T. Naes, *Multivariate Calibration*, Wiley, Chichester (1989).
21. S. Wold, A. Ruhe, H. Wold and W. J. Dunn III, *SIAM J. Sci. Stat. Comput.* **5**, 735–743 (1984).
22. S. Wold, C. Albano, W. J. Dunn III, U. Edlund, K. Esbensen, P. Geladi, S. Hellberg, E. Johansson, W. Lindberg and M. Sjöström, in *Chemometrics, Mathematics and Statistics in Chemistry*, ed. B. R. Kowalski, pp. 17–19, Reidel, Dordrecht (1984).
23. A. Höskuldsson, *J. Chemometrics*, **2**, 211–228 (1988).
24. I. N. Wakeling and J. J. Morris, *J. Chemometrics*, **7**, 291–304 (1993).
25. L. Eriksson and S. Wold, in *Methods and Principles in Medicinal Chemistry*, Vol. 2, *Chemometric Methods in Molecular Design*, ed. by H. van de Waterbeemd, pp. 309–318, Verlag Chemie, Weinheim (1995).
26. *SIMCA-P Software Manual*, Umetri AB (1994).
27. *MATLAB Software User's Guide*, The MathWorks Inc., Natick, MA (1992).