Improvements in Sample and Variable Selection in Multivariate Calibration

J.P.M. Andries

Verbeteringen aan monster- en variabelenselectie bij multivariate kalibratie

Proefschrift

ter verkrijging van de graad van doctor
aan de Radboud Universiteit Nijmegen
op gezag van de rector magnificus prof. dr. Th.L.M. Engelen,
volgens besluit van het college van decanen
en van de graad van doctor in de Farmaceutische Wetenschappen
aan de Vrije Universiteit Brussel
in het openbaar te verdedigen op
vrijdag 16 januari 2015 om 12:30 uur precies

door

Johannes Petrus Maria Andries

geboren op 20 november 1945
te Tilburg

Promotoren:

Prof. dr. L.M.C. Buydens
Prof. dr. Y. Vander Heyden (*Vrije Universiteit Brussel*)

Manuscriptcommissie:

Prof. dr. A.P.M. Kentgens
Prof. dr. D. Coomans (*Vrije Universiteit Brussel*)
Dr. B.G.M. Vandeginste (*Unilever*)

ISBN: 978-94-6295-033-7

Cover design: Proefschriftmaken.nl || Uitgeverij BOXPress
Printed & Lay Out by: Proefschriftmaken.nl || Uitgeverij BOXPress
Published by: Uitgeverij BOXPress, ’s-Hertogenbosch

Improvements in Sample and Variable Selection in Multivariate Calibration

Doctoral thesis

to obtain the degree of doctor
from Radboud University Nijmegen
on the authority of the Rector Magnificus prof. dr. Th.L.M. Engelen,
according to the decision of the Council of Deans
and the degree of doctor in Pharmaceutical Sciences
from the Vrije Universiteit Brussel
to be defended in public on
Friday, January 16, 2015 at 12:30 hours

by

Johannes Petrus Maria Andries

born on November 20, 1945
in Tilburg, the Netherlands

Supervisors:

Prof. dr. L.M.C. Buydens
Prof. dr. Y. Vander Heyden (*Vrije Universiteit Brussel*)

Doctoral Thesis Comittee:

Prof. dr. A.P.M. Kentgens
Prof. dr. D. Coomans (*Vrije Universiteit Brussel*)
Dr. B.G.M. Vandeginste (*Unilever*)

ISBN: 978-94-6295-033-7

Cover design:	Proefschriftmaken.nl || Uitgeverij BOXPress
Printed & Lay Out by:	Proefschriftmaken.nl || Uitgeverij BOXPress
Published by:	Uitgeverij BOXPress, ’s-Hertogenbosch

Dankwoord

Hierbij wil ik iedereen bedanken die op enigerlei wijze betrokken is geweest bij de totstandkoming van dit proefschrift.

Het onderzoek dat beschreven is in dit proefschrift is in deeltijd uitgevoerd bij het lectoraat Analysetechnieken in de Life Sciences (ALS) van de Avans Hogeschool in Breda. Het betreft een samengevoegd promotieonderzoek van het Departement Analytische Chemie van de Radboud Universiteit in Nijmegen onder supervisie van Professor Lutgarde Buydens, en het Departement Analytische Scheikunde en Farmaceutische Technologie van de Vrije Universiteit Brussel in België onder supervisie van Professor Yvan Vander Heyden.

Op de eerste plaats bedank ik mijn promotoren Lutgarde en Yvan hartelijk voor de bereidheid om mij als externe promovendus te willen begeleiden bij de uitvoering van dit onderzoek. Ik heb hun begeleiding steeds als zeer prettig en stimulerend ervaren.

Ook bedank ik mijn lectoren bij het lectoraat ALS, Henk Claessens, Govert Somsen, Ad de Jong en Theo Noij, voor de support die zij gegeven hebben aan dit onderzoek.

Mijn collega's in het lectoraat ALS, Edward Knaven, Ben de Rooij, Martie Verschuren, bedank ik hartelijk voor de prettige samenwerking gedurende de uitvoering van mijn onderzoek.

Tot slotte bedank ik mijn vrouw Corrie als onmisbare steun en toeverlaat aan het thuisfront.

Contents

Dankwoord

1 General Introduction ... 1

1.1 Thesis project ... 2

1.2 Outline of this thesis ... 2

2 Introduction to sample selection in High Performance Liquid Chromatography ... 5

2.1 Introduction ... 5

2.2 Characterization of stationary phases ... 5

2.3 Retention models in high performance liquid chromatography ... 6

2.4 Strategy for sample selection for the development of QSRR models ... 7

3 Strategy for reduced calibration sets to develop quantitative structure–retention relationships in high-performance liquid chromatography ... 11

3.1 Abstract ... 11

3.2 Introduction ... 12

3.3 Theory ... 13

3.4 Strategy for selection of test analytes for a reduced calibration set ... 14

3.5 Quality criteria for the resulting QSRR models ... 15

3.6 Data and methodology ... 17

3.7 Results and Discussion ... 19

3.8 Conclusions ... 29

4 Introduction to variable selection ... 33

4.1 PLS model ... 34

4.2 PLS modelling ... 35

4.3 'Large *K* - small *N*' problem ... 38

4.4 Classification of variable selection methods ... 38

4.5 Scope of variable selection ... 40

4.6 Predictor-variable property based methods ... 41

4.7 Uninformative Variable Elimination ... 44

4.8 Penalised methods ... 45

4.9 Genetic algorithms ... 45

4.10 Interval PLS ... 46

4.11 Variable selection for PLS2 ... 48

4.12 Variable selection for classification ... 48

4.13 Variable selection for QSAR and QSRR modelling ... 48

4.14 Summary ... 50

4.15 Variable selection in this thesis ... 51

5 Improved Variable Reduction in partial least squares modelling based on Predictive-Property-Ranked Variables and adaptation of partial least squares complexity 57

5.1 Abstract 57

5.2 Introduction 58

5.3 Theory 60

5.4 Reference methods 64

5.5 Data and methodology 66

5.6 Results and discussion 68

5.7 Conclusions 82

6 Predictive-Property-Ranked Variable Reduction in Partial Least Squares Modelling with Final Complexity Adapted Models: Comparison of Properties for Ranking 87

6.1 Abstract 87

6.2 Introduction 88

6.3 Theory 89

6.4 FCAM method 94

6.5 Wilcoxon signed rank test 96

6.6 Data and methodology 98

6.7 Results and discussion 99

6.8 Conclusions 108

7 Predictive-Property-Ranked Variable Reduction with Final Complexity Adapted Models in Partial Least Squares Modelling for Multiple Responses 111

7.1 Abstract 111

7.2 Introduction 112

7.3 Theory 113

7.4 Data and methodology 116

7.5 Results and discussion 119

7.6 Conclusions 127

8 Summary, conclusions, discussion and future perspectives 131

8.1 Summary and conclusions 131

8.2 Discussion and future perspectives 133

Appendix 135

A. List of abbreviations 135

B. List of publications 137

Samenvatting 138

Curriculum Vitae 141

1 General Introduction

Chemometrics has now been used for some 40 years [1,2]. The combination of modern information-rich analytical techniques with efficient multivariate regression tools for quantitative and qualitative analysis makes that chemometric applications for prediction and classification of samples are nowadays widespread in analytical chemistry [2-4].

The regression problem, i.e., how to model one or several dependent variables (responses) **y** or **Y**, by means of a set of predictor variables, **X**, is one of the most common data-analytical problems. Examples in chemistry include relating (*i*) concentrations of different components in chemical samples to their mixture spectra, (*ii*) chemical properties, reactivity or biological activity of a set of molecules to their chemical structure, and (*iii*) the origin or activity of samples to their chromatographic or spectral profiles [5].
Traditionally, the relationship between **y** or **Y** and **X** is modelled using linear regression (LR) or multiple linear regression (MLR) [6,7]. This works well if there are few fairly uncorrelated independent **X**-variables and more samples than **X**-variables. However, with modern analytical instruments, including spectrometers and chromatographs, many **X**-variables are measured, which are usually correlated, and many are uninformative and noisy. Partial least squares (PLS) regression is a modern multivariate regression method, which is able to model the relationship between **y** or **Y** and a large number of noisy and correlated **X**-variables, for a data set with small numbers of samples [5,7,8].

In the last decades, highly sophisticated instrumental analysis techniques like Nuclear Magnetic Resonance (NMR) Spectroscopy, Fourier Transform Infrared (FT-IR) Spectroscopy, and hyphenated techniques such as Gas Chromatography-Mass Spectrometry (GC-MS), Liquid Chromatography-Mass Spectrometry (LC-MS), and Capillary Electrophoresis-Mass Spectrometry (CE-MS) are introduced in routine analysis and generate huge data sets. Additionally, the trend to investigate very complex problems, for example in life sciences, makes that chemometricians are now faced with an enormous flood of data, a real data tsunami according to Buydens [9]. Chemometric tools are now increasingly applied in bio-informatics, especially in the strongly developing field of metabolomics [10-14] which increases the data flood. In metabolomics all metabolites of a biological system are identified and quantified [11]. Consequently, the analysis of these data will be more and more demanding.

To master the data flood, new or improved chemometric methods should be developed. One strategy can be to upgrade the applied multivariate methods such as PLS. These methods can be improved by the development of new or modified methods (*i*) to select the most informative samples, and/or (*ii*) to reveal the informative signals in the data while removing noise and uninformative variables. This may not only reduce the signals in the data flood to be investigated, but also improve the extracted information.

1.1 Thesis project

The goal for the research presented in this thesis is to contribute in coping with the data flood and to develop new or improved chemometric methods both for sample and variable selection.

Sample selection is focussed on Quantitative Structure-Retention Relationships (QSRRs) in Reversed-Phase Liquid Chromatography (RPLC). The QSRR models were (multiple) linear regression models and the goal of the work was the selection of reduced calibration sets. QSRRs are mathematical relationships between a chromatographic retention parameter and variables (descriptors) related to the molecular structure of the analytes [15,16]. RPLC, combined with mass spectrometric detection, now plays a key role in the life sciences applications [17]. However, the wide variety of commercially available RPLC stationary phases makes the effective selection of an appropriate stationary phase for a particular separation a challenging task [18]. QSRRs are used to characterise RPLC stationary phases [19-21] and can help appropriately selecting a suitable starting point (i.e., the initially selected chromatographic system formed by the stationary and mobile phase) for further method development [22].

Variable selection in the presented work is focussed on PLS modelling because this technique now dominates the practice of multivariate modelling. Reasons are for the latter the quality of the obtained models and the ease of their implementation due to the availability of appropriate PLS software [4].
The aim of this thesis work is to develop new or improved variable selection methods for PLS modelling, which can be applied both for continuous and non-continuous data, and which must be widely applicable in chemometrics and in new emerging fields, such as metabolomics.

1.2 Outline of this thesis

This thesis is divided into eight chapters, which besides this General Introduction (**Chapter 1**) is organised as follows. Chapters 2 and 3 form a first part of the presented research and deal with the sample selection to build QSRR models in HPLC, while Chapters 4 till 7 concern the variable selection prior to PLS modelling.

In **Chapter 2**, an introduction is given on sample selection for Quantitative Structure-Retention Relationships in High-Performance Reversed-Phase Liquid Chromatography.

In **Chapter 3**, a study is presented about a strategy for the construction of reduced calibration sets to be used for the development of Quantitative Structure-Retention Relationships in High-Performance Reversed-Phase Liquid Chromatography. The application of the proposed strategy provides small calibration sets suitable for future QSRR model building to describe and predict retentions on new RPLC systems.

In **Chapter 4**, an introduction is given in variable selection for PLS. The characteristics of the most widely used methods and their advantages and drawbacks are described. These methods are compared in order to select the most promising variable-selection method as a starting point for the development of new or improved methods that may help mastering the data tsunami in chemometrics and bioinformatics. A strategy for the development of these methods is formulated.

In **Chapter 5**, a study is presented about the development of three new stepwise variable selection methods for PLS modelling with one response (PLS1). The Final Complexity Adapted Models method, denoted as FCAM, is proposed as preferred. The results of this study form the basis for the studies presented in Chapters 6 and 7.

In **Chapter 6**, the utility and effectiveness of different predictor-variable properties in variable selection are investigated and compared, when using the FCAM method from the study in Chapter 5, and the best properties are identified.

In **Chapter 7**, the development and testing is presented of a new variable-selection method for multiple-response partial-least-squares (PLS2) modelling, using an adapted FCAM method, FCAM-PLS2.

Finally, in **Chapter 8**, the findings of the research in this thesis project are summarized with reference to the research objectives, along with conclusions and recommendations for future research.

References

[1] K. Esbensen, P. Geladi, J. Chemometr. 4 (1990) 389.
[2] P. Geladi, Spectrochim. Acta, Part B, 58 (2003) 767.
[3] B. Lavine, J. Workman, Anal. Chem. 82 (2010) 4699.
[4] B. Lavine, J. Workman , Anal. Chem. 85 (2013) 705.
[5] S. Wold, M. Sjöström, L. Eriksson, Chemom. Intell. Lab. Syst. 58 (2001) 109.
[6] N.R. Draper, H. Smith, Applied Regression Analysis, Second edition, John Wiley and Sons, New York, 1981.
[7] P. Geladi, B.R. Kowalski, Anal. Chim. Acta 185 (1986) 1.
[8] H. Martens, T. Næs, Multivariate Calibration, (2nd edn), Wiley, NewYork, 1993.
[9] L. Buydens, The Analytical Scientist 1 (2013) 24.
[10] J. van der Greef, A.K. Smilde, J. Chemometr. 19 (2005) 376.
[11] J.C. Lindon, J.K. Nicholson, Annu. Rev. Anal. Chem. 1 (2008) 45.
[12] S.L. Robinette, J.C. Lindon, J.K. Nicholson, Anal. Chem. 85 (2013) 5297.
[13] R. Madsen, T. Lundstedt, J. Trygg, Anal. Chim. Acta 659 (2010) 23.
[14] J. Boccard, S. Rudaz, J. Chemometr. 28 (2014) 1.
[15] R. Kaliszan, Structure and Retention in Chromatography. A Chemometric Approach, Harwood Academic Publishers, Amsterdam, 1997.
[16] R. Kaliszan, Chem. Rev. 107 (2007) 3212.
[17] K.K. Unger, R. Ditz, E. Machtejevas, R. Skudas, Angewandte Chemie International Edition 49 (2010) 2300.
[18] H.A. Claessens, Trends Anal. Chem. 20 (2001) 563.
[19] M.H. Abraham, M. Rozés, C.F. Poole, S.K. Poole, J. Phys. Org. Chem. 10 (1997) 358.
[20] R. Kaliszan, M. A. van Straten, M. Markuszewski, C.A. Cramers, H.A. Claessens, J. Chromatogr. A 855 (1999) 455.
[21] K. Héberger, J. Chromatogr. A 1158 (2007) 273.
[22] T. Hancock, R. Put, D. Coomans, Y. Vander Heyden, Y. Everingham, Chemom. Intell. Lab. Syst. 76 (2005) 185.

2 Introduction to sample selection in High Performance Liquid Chromatography

2.1 Introduction

Efficient sampling methods can help reducing the number of experiments and therefore also reducing the amount of generated data, especially when they are applied for a widely used analysis technique such as High Performance Liquid Chromatography (HPLC). HPLC is probably one of the most powerful separation techniques in analytical chemistry and biochemistry. Using HPLC, mixtures of compounds can be separated and individual components identified and quantified [1]. At present, approximately 90% of all HPLC separations are carried out by Reversed-Phase Liquid Chromatography (RPLC) because of its broad application range. Except for the high molecular weight range, nearly all substances can be separated by RPLC [2]. Additionally, liquid chromatography, coupled with a mass spectrometer detector (LC-MS), is in the field of quantitative bioanalysis the preferred technique for quantitating small molecules, because of its specificity, sensitivity, and speed [3,4]. Because of its wide use and the huge sets of data generated, HPLC now contributes to the data tsunami. Chemometric tools are needed to extract information from these ever-increasing amount of data [5,6].

Solving two kinds of problems in HPLC can help reducing the number of experiments. First, the efficient selection of an appropriate stationary phase, and second, the a priori prediction of the retention of analytes for a specific chromatographic system, i.e. the combination of the mobile and stationary phase. Efficient and cost effective sample selection for HPLC can reduce the number of experiments. In this introduction is described how this is realised in this thesis project.

2.2 Characterization of stationary phases

In RPLC, the selection of a suitable stationary phase is an important starting condition prior to the development of a robust separation method. However, the wide variety of commercially available RPLC stationary phases [2] makes the effective selection of an appropriate column for a particular separation a challenging task.

Columns can be selected using chromatographic characterization methods [2]. These methods can be subdivided into two groups:

- *Empirically based* characterization methods [2] or *test set* methods [7]. The chromatographic information is obtained using sets of rather arbitrarily selected test compounds, which are supposed to reflect a specific column property, e.g. silanol activity or hydrophobicity, see Refs. [2,7-11].
- *Model-based* characterization methods. The chromatographic information is obtained using mathematical models describing the relationship between chromatographic parameters and structure related properties of test compounds, i.e. molecular descriptors, see Refs. [12-17].

The empirically based methods use a relatively low number of test compounds representing the column properties. For example in Ref. [7] eight methods, using 1 to 8 test analytes, and in Ref. [9] five methods, using 1 to 9 test analytes, are described and investigated. These chromatographic tests often produce conflicting results [9]. Until now none of these has been widely accepted [7,10]. Therefore an urgent need exists to select RPLC columns based on more objective criteria.

A promising approach of model-based chromatographic retention prediction is the use of Quantitative Structure-Retention Relationships (QSRRs). QSRRs are mathematically derived relationships between chromatographic parameters and descriptors related to the molecular structure of the analytes. In QSRRs these descriptors are used to model the molecular interaction of the analytes with a given chromatographic system, formed by the combination of a mobile and stationary phase. QSRRs are used to characterise RPLC systems, and to describe and predict retentions of analytes on these RPLC systems, see Refs. [12-23]. Therefore, they can help selecting an appropriate chromatographic system for a particular RPLC separation.

Contrary to the test set methods, a substantial number of test analytes are used to obtain proper QSRR models. For example in Ref. [15] 25 test analytes are used, in Ref. [21] 87 analytes, and in Ref. [23] 67 analytes. This makes the application of QSRRs laborious and time consuming. The use of QSRRs will be more attractive if QSRRs could be built with a small number of test analytes, comparable to or only slightly higher than those for the test set methods. This requires the development of a new methodology for the construction of small calibration sets for QSRRs.

2.3 Retention models in high performance liquid chromatography

Properties of chemical compounds depend on their structure and on the physicochemical environment. In QSRRs the relation between a retention-related property *y* in a specific chromatographic system and structure-related variables (descriptors) x_1, x_2, …, x_m is described by a model, generally a multiple linear regression model [13]. The general QSRR model has the form:

$$y = \beta_0 + \beta_1 \cdot x_1 + \cdots \beta_m \cdot x_m \qquad \textbf{(1)}$$

The model parameter β_0 is a constant, while β_1, β_2, …, β_m are coefficients which describe the dependence of the chromatographic property *y* on the independent variables x_1, x_2, …, x_m, respectively. The set of β_i coefficients [$\beta_{0,}\beta_1$, β_2, …, β_m] is characteristic for the chromatographic system and for the calibration set of molecules used to build the model.

The estimated set of coefficients will be denoted as [$b_{0,}$ b_1, b_2, …, b_m]. Different analyte properties are reflected by different values of the analyte-dependent variables x_i and the chromatographic property *y* on a specific chromatographic system. The differences between chromatographic systems will be reflected by differences in the set of coefficients. Therefore, QSRRs can be used to characterise chromatographic columns by the set of estimated regression coefficients [$b_{0,}$ b_1, b_2, …, b_m]. This can help appropriately selecting a suitable chromatographic system for further method development [24].

Meaningful and statistical significant QSRR models are also used to predict the retention of new analytes under the same chromatographic conditions, from the estimated regression coefficients and the molecular properties included in the model, without additional experiments [12]. This can also help selecting an appropriate chromatographic system for a particular RPLC separation. Additionally, this may reduce the number of laboratory experiments and hence also the amount of generated data. It requires that the descriptor variables x_i are known for the analytes. Descriptor values of analytes can be determined experimentally, found in the literature or calculated [16,25].

The classical QSRR models contain small numbers (1-5) of descriptors [12-17] for which linear regression (LR) or multiple linear regression (MLR) is used for model building [12]. With the introduction of theoretical molecular descriptors, generated by calculation chemistry, much larger sets of descriptors were introduced in QSRR modelling. As an example, the Dragon software (http://www.talete.mi.it/) allows the computation of more than 4000 molecular descriptors, see [26]. Then, either MLR, or more advanced modelling techniques such as partial least squares regression, both combined with feature selection, are needed [12].

For the development of QSRRs, the retentions of a representative set of analytes, called the calibration set, are measured on a specific column under well-defined chromatographic conditions, and the regression coefficients estimated. The analytes must be selected such that both the retention and the descriptors span a range relevant for the intended use of the QSRR. The regression coefficients in the classical QSRR models are estimated by MLR [27,28]. For MLR the correlations between the descriptors should be as low as possible [29], and the number of analytes should be larger than the number of coefficients. Traditionally, as a rule of thumb, a minimum of 4 to 6 analytes per descriptor are applied to account for the uncertainties in the calculation of the descriptors and the experimental error in determining the retention [13,19,29]. Examples of calibration sets which meet these requirements can be found in Refs. [15,19,20,30-33].

One of the goals of this thesis project is to propose a strategy for the construction of reliable reduced calibration sets for classical QSRRs, with a smaller number than 4 to 6 analytes per descriptor. This will reduce the number of experiments and therefore can help reducing the workload. Additionally it will make QSRR methods more attractive and useful in laboratory practice.

2.4 Strategy for sample selection for the development of QSRR models

The number of experiments can be reduced if the selection of calibration samples is based on the descriptor set only. Then, the calibration samples (analytes) can be selected without experiments. Thereafter, the retentions have only to be measured for the selected samples. This makes sample selection cost efficient. Preferably the calibration samples are selected with a uniform distribution [34].
The Kennard-Stone algorithm (KS) [35] is a well-known selection method which is suitable for the selection of a representative uniformly distributed subset from a larger pool of samples along the independent *x*-variables space [36,37].

Using the information given above, a strategy for the development of a new sample selection method for the construction of small reliable calibration sets for classical Quantitative Structure–Retention Relationships in High-Performance Reversed-Phase Liquid Chromatography, containing 1-5 descriptors, for which LR or MLR is used for model building, is presented in the first part of this thesis project in chapter 3. Later, after this thesis project, this strategy will also be applied for QSRRs, containing many more descriptors generated by calculation chemistry, for which PLS models are built, after the application of one of the variable reduction methods presented in the second part of this thesis project.

References

[1] L.R. Snyder, J.J. Kirkland, J.L. Glajch, Practical HPLC Method Development, 2nd edition, Wiley, New York, 2001.

[2] H.A. Claessens, Trends Anal. Chem. 20 (2001) 563.

[3] R.N. Xu, L. Fan, M.J. Rieser, T.A. El-Shourbagy, J. Pharm. Biomed. Anal. 44 (2007) 342.

[4] I.D. Wilson, R. Plumb, J. Granger, H. Major, R. Williams, E.M. Lenz, J. Chromatogr. B, 817 (2005) 67.

[5] J. Boccard, S. Rudaz, J. Chemometr. 28 (2014) 1.

[6] J. van der Greef, A.K. Smilde, J. Chemometr. 19 (2005) 376.

[7] D. Visky, Y. Vander Heyden, T. Iványi, P. Baten, J. De Beer, Z. Kovács, B. Noszál, E. Roets, D.L. Massart, J. Hoogmartens, J. Chromatogr. A 977 (2002) 39.

[8] D. Visky, Y. Vander Heyden, T. Iványi, P. Baten, J. De Beer, Z. Kovács, B. Noszál, P. Dehouck, E. Roets, D.L. Massart, J. Hoogmartens, J. Chromatogr. A 1012 (2003) 11.

[9] H.A. Claessens, M.A. van Straten, C.A. Cramers, M. Jezierska, B. Buszewski, J. Chromatogr. A 826 (1998) 135.

[10] C. Stella, S. Rudaz, J.-L. Veuthey, A. Tchapla, Chromatographia 53 (2001) S-132.

[11] S.D. Rogers, J.G. Dorsey, J. Chromatogr. A 892 (2000) 57.

[12] R. Put, Y. Vander Heyden, Anal. Chim. Acta 602 (2007) 164.

[13] R. Kaliszan, Structure and Retention in Chromatography. A Chemometric Approach, Harwood Academic Publishers, Amsterdam, 1997.

[14] M.H. Abraham, M. Rozés, C.F. Poole, S.K. Poole, J. Phys. Org. Chem. 10 (1997) 358.

[15] R. Kaliszan, M. A. van Straten, M. Markuszewski, C.A. Cramers, H.A. Claessens, J. Chromatogr. A 855 (1999) 455.

[16] R. Kaliszan, Chem. Rev. 107 (2007) 3212.

[17] K. Héberger, J. Chromatogr. A 1158 (2007) 273.

[18] T. Baczek, R. Kaliszan, J. Chromatogr. A 962 (2002) 41.

[19] M.A. Al-Haj, R. Kaliszan, A. Nasal, Anal. Chem. 71 (1999) 2976.

[20] M.A. Al-Haj, R. Kaliszan, B. Buszewski, J. Chromatogr. Sci. 39 (2001) 29.

[21] L.C. Tan, P.W. Carr, M.H. Abraham, J. Chromatogr. A 752 (1996) 1.

[22] J. Zhao, P.W. Carr, Anal. Chem. 70 (1998) 3619.

[23] N.S. Wilson, M.D. Nelson, J.W. Dolan, L.R. Snyder, R.G. Wolcott, P.W. Carr, J. Chromatogr. A 961 (2002) 171.

[24] T. Hancock, R. Put, D. Coomans, Y. Vander Heyden, Y. Everingham, Chemom. Intell. Lab. Syst. 76 (2005) 185.

[25] C. Wang, M.J. Skibic, R.E. Higgs, I.A.Watson, H. Bui, J. Wang, J.M. Cintron, J. Chromatogr. A, 1216 (2009) 5030.

[26] R. Todeschini, V. Consonni, Handbook of Molecular Descriptors, Wiley, Weinheim, 2000.

[27] N.R. Draper, H. Smith, Applied Regression Analysis, Second edition, John Wiley and Sons, New York, 1981.

[28] D.L. Massart, B.G.M. Vandeginste, L.M.C. Buydens, S. de Jong, P.J. Lewi, J. Smeyers-Verbeke, Handbook of Chemometrics and Qualimetrics, Part A, Elsevier, Amsterdam, 1997.

[29] M. Vitha, P.W. Carr, J. Chromatogr. A 1126 (2006) 143.

[30] J. Jiskra, H.A. Claessens, C.A. Cramers, R. Kaliszan, J. Chromatogr. A 977 (2002) 193.

[31] T. Baczek, R. Kaliszan, J. Chromatogr. A 987 (2003) 29.

[32] M.A. Al-Haj, P. Haber, R. Kaliszan, B. Buszewski, M. Jezierska, Z. Chilmonzyk, J. Pharm. Biomed. Anal. 18 (1998) 721.

[33] R. Kaliszan, T. Baczek, A. Bucinski, B. Buszewski, M. Sztupecka, J. Sep. Sci. 26 (2003) 271.

[34] M. Forina, S. Lanteri, M. Casale, J. Chromatogr. A, 1158 (2007) 61.

[35] R.W. Kennard, L.A. Stone, Technometrics 11 (1969) 137.

[36] J. Yoon, B. Lee, C. Han, Chemom. Intell. Lab. Syst. 64 (2002) 1.

[37] E. Bouveresse, D.L. Massart, Chemom. Intell. Lab. Syst. 32 (1996) 201.

3 Strategy for reduced calibration sets to develop quantitative structure–retention relationships in high-performance liquid chromatography[1]

3.1 Abstract

In high-performance liquid chromatography, quantitative structure–retention relationships (QSRRs) are applied to model the relation between chromatographic retention and quantities derived from molecular structure of analytes. Classically a substantial number of test analytes is used to build QSRR models. This makes their application laborious and time consuming. In this work a strategy is presented to build QSRR models based on selected reduced calibration sets. The analytes in the reduced calibration sets are selected from larger sets of analytes by applying the algorithm of Kennard and Stone on the molecular descriptors used in the QSRR concerned. The strategy was applied on three QSRR models of different complexity, relating log k_w or log k with either: (*i*) log P, the *n*-octanol–water partition coefficient, (*ii*) calculated quantum chemical indices (QCI), or (*iii*) descriptors from the linear solvation energy relationship (LSER). Models were developed and validated for 76 reversed-phase high-performance liquid chromatography systems.

From the results we can conclude that it is possible to develop log P models suitable for the future prediction of retentions with as few as seven analytes. For the QCI and LSER models we derived the rule that three selected analytes per descriptor are sufficient. Both the dependent variable space, formed by the retention values, and the independent variable space, formed by the descriptors, are covered well by the reduced calibration sets. Finally guidelines to construct small calibration sets are formulated.

Keywords: Liquid chromatography; Quantitative structure–retention relationships; Samples selection; Reduced calibration sets; Retention modelling; Retention prediction

[1] **Jan P.M. Andries**, Henk A. Claessens, Yvan Vander Heyden, Lutgarde M.C. Buydens
Published in Analytica Chimica Acta 652 (2009) 180-188

3.2 Introduction

In high-performance liquid chromatography, the selection of a suitable stationary phase is an important starting condition prior to the development of a robust separation method. This is particularly true for reversed-phase liquid chromatography (RPLC), a technique applied in 80-90% of all HPLC separations [1]. Presently an estimated number of more than 600 different RPLC columns are available on the market and this number is still increasing [2]. This wide variety of commercially available RPLC stationary phases makes the selection of a suitable column for a particular separation a challenging task. The selection of a stationary phase is often based on the results of a number of chromatographic tests [2-4] or on the empirical knowledge of the analyst [5]. However, the majority of the chromatographic tests, has an empirical basis and often produce conflicting results [6]. Therefore an urgent need exists to select RPLC columns based on more objective criteria.
In addition to that chromatographic retention prediction methodologies can be valuable starting points for RPLC method development [7]. A promising approach is the use of quantitative structure-retention relationships (QSRRs) [7,8]. QSRRs are statistically derived relationships between chromatographic parameters and descriptors related to the molecular structure of the analytes. In QSRRs these descriptors are used to model the molecular interaction of the analytes with a given stationary phase and eluent [9].

In chromatography, QSRRs have been applied to: (*i*) gain a better understanding of the molecular mechanism of the chromatographic separation process; (*ii*) identify the most informative structure-related properties of analytes; (*iii*) characterise stationary phases, and (*iv*) predict retention for new analytes [8]. Abraham et al. [10] and Kaliszan et al. [11] have used different types and numbers of molecular descriptors in different models to model the retention of a representative set of test analytes on a specific chromatographic system: (*i*) the logarithm of the *n*-octanol–water partition coefficient, (*ii*) three calculated quantum chemical indices, and (*iii*) a set of five solvation parameters (see further). With these models, the retention of new solutes under the same chromatographic conditions is predicted [7]. However many other models can be found in the literature. Recently, reviews on QSRR applications in column liquid chromatography were written by Put and Vander Heyden [7], Kaliszan [12] and Héberger [13].

Until now a substantial number of test analytes has been used to obtain proper QSRR models [11,14-18]. This makes the application of QSRRs laborious, time consuming and therefore less attractive from a practical point of view.

In [14,15] QSRR models with the above mentioned three descriptor sets were developed with a reduced set of 18 test analytes, selected from a starting set of 58. In [17] QSRR models with solvation parameters were developed with a reduced set of 22 structurally diverse analytes, chosen from a starting set of 87 analytes, described in [16]. In [19] five reduced sets of five or seven analytes were selected from a set of 67 analytes described in [18] to allow quantitative prediction of retention and selectivity.

The goal of this work is to propose a strategy for the development of reliable reduced calibration sets for QSSR models, based on molecular structure properties. This to make QSRR methods more attractive and useful in laboratory practice.

3.3 Theory

3.3.1 QSRR models

Properties of chemical compounds depend on their structure and on the physicochemical environment. In QSRRs the relation between a retention related property *y* in a specific chromatographic system and structure-related variables (descriptors) x_1, x_2, …, x_m is described by a model, generally a multiple linear regression model [8]. The general QSRR model has the form:

$$y = \beta_0 + \beta_1 \cdot x_1 + \cdots \beta_m \cdot x_m \qquad \textbf{(1)}$$

The model parameter β_0 is a constant, while β_1, β_2, …, β_m are coefficients which describe the dependence of the chromatographic property *y* on the independent variables x_1, x_2, …, x_m, respectively. The set of β_i coefficients [$\beta_{0,}\beta_1$, β_2, …, β_m] is characteristic for the chromatographic system, i.e. the combination of the mobile and stationary phase, and for the calibration set of molecules used to build the model. The estimated set of coefficients will be denoted as [$b_{0,}$ b_1, b_2, …, b_m].
Different analyte properties are reflected by different values of the analyte-dependent variables x_i and the chromatographic property *y* on a specific chromatographic system. The differences between chromatographic systems will be reflected by differences in the set of coefficients.

QSRRs can therefore be used to characterise chromatographic columns by the set of estimated regression coefficients [$b_{0,}$ b_1, b_2, …, b_m]. QSRRs can also be used to predict the retention of a new analyte on a specific chromatographic system if the set of regression coefficients is known for that system. This requires that the descriptor variables x_i are known for the analyte. Descriptor values of test analytes can be found in the literature or are calculated [12].

In this study, log k_w or log *k* is used as the chromatographic property *y*. Log k_w is the logarithm of the retention factor *k* of the analyte extrapolated to a virtual mobile phase of pure water or pure buffer. It is the intercept of the (linear) relationship between the isocratic log *k* values and the corresponding organic modifier fraction in the eluent [20]. It is known that apart from the analyte, log k_w also depends on both the nature of the organic modifier [21,22] and on the kind of relationship (linear or polynomial) used for extrapolation [22]. Therefore, log $\mathrm{k_w}$ cannot be considered as a pure solute property.

In comparative studies of retention properties of RPLC stationary phases three main types of QSRR, containing different descriptors, have been investigated. They are discussed below. The first QSRR model relates log k_w to the logarithm of the calculated *n*-octanol–water partition coefficient, log *P* [11,23]:

$$\log k_w = \beta_0 + \beta_1 \log P \qquad \textbf{(2)}$$

Log *P* accounts for the hydrophobic properties of the analyte. Applications of this QSRR type can be found in Refs. [11,14,15,23-25].

The second QSRR model relates log k_w of an analyte to three calculated quantum chemical indices (QCI): (*i*) electron excess charge of the most negatively charged atom, δ_{min}; (*ii*) square of total dipole moment, μ^2; (*iii*) water-accessible molecular surface area, A_{WAS} [11,23]:

$$\log k_w = \beta_0 + \beta_1\delta_{\min} + \beta_2\mu^2 + \beta_3 A_{WAS} \qquad \textbf{(3)}$$

δ_{min} accounts for the ability of the analyte to participate in polar interactions; μ accounts for the dipole–dipole and dipole-induced dipole attractive interactions of the analyte; A_{WAS} accounts for the strength of London-type interactions of the analyte [24]. Applications of this QSRR type can be found in Refs. [11,14,15,23-26].

The third QSRR model is formed by the linear solvation energy relationship (LSER). LSERs relate log k_w of an analyte to five solvation parameters. The LSER model is given by [27-30]:

$$\log k_w = \gamma + \varepsilon E + \sigma S + \alpha A + \beta B + \nu V \qquad \textbf{(4)}$$

Each of the descriptors *E*, *S*, *A*, *B* and *V* accounts for a specific molecular interaction. *E* is the excess molar refraction, *S* is the dipolarity/polarizability, *A* the overall hydrogen bond acidity, *B* the overall hydrogen bond basicity and *V* the McGowan volume. LSER models are very general. They provide an understanding of the importance of various chemical interactions in a chromatographic system. LSER values for a large number of analytes are available [27]. Reviews concerning this model can be found in [29,30]. Applications of this QSRR type can be found in Refs. [11,14,15,23,25,27].
The three models are indicated further as log *P*, QCI and LSER models, respectively.

3.3.2 Calibration set

To apply QSRRs for retention prediction, the retention factors of a representative set of test analytes, called the calibration set, are measured on a specific column under well-defined chromatographic conditions, and log k_w is predicted . The analytes must be selected such that both the chromatographic property and the descriptors span a range relevant for the intended use of the QSRR. The coefficients in equations (2-4) are estimated by multiple linear regression (MLR).
For MLR the correlations between the descriptors should be as low as possible [30], and the number of analytes should be larger than the number of coefficients. Traditionally, as a rule of thumb, a minimum of 4 to 6 analytes per descriptor are applied to account for the uncertainties in the calculation of the descriptors and the experimental error in determining log k_w [8,14,30]. Examples of calibration sets which meet these requirements can be found in Refs. [11,14,15,23-26].

3.4 Strategy for selection of test analytes for a reduced calibration set

In this study a reduction of the number of test analytes for QSRR modelling, below a minimum of 4 to 6 analytes per descriptor, is investigated. A selection of test analytes from a larger set is made based on auto scaled descriptor values, applied in the QSRR, using the algorithm of Kennard and Stone [31].

The Kennard and Stone algorithm is a sequential method that makes a selection covering the variable space uniformly. Preferably the procedure starts with an analyte closest to the mean of the variable space to prevent that the analytes that are initially selected are all situated at the boundaries of the variable space. This is the case for a reduced data set where only few analytes are selected. As second analyte, that has the largest distance to the first one is selected. The third analyte then is the one furthest from the already selected, and so on.

3.5 Quality criteria for the resulting QSRR models

3.5.1 Calibration error

A carefully designed selection procedure should result in a reduced calibration set that is a representative subset of the full calibration set. The residual variance of the QSRR model developed with the reduced calibration set, s_{red}^2, should not be significantly larger than the residual variance of the equivalent QSRR model developed with the full calibration set, s_{full}^2. The residual variances s_{full}^2 and s_{red}^2 are calculated with equations (5) and (6), respectively,

$$s_{full}^2 = \frac{1}{n_{full} - p} \sum_{i=1}^{n_{full}} (y_i - \hat{y}_i)^2 \qquad \textbf{(5)}$$

$$s_{red}^2 = \frac{1}{n_{red} - p} \sum_{i=1}^{n_{red}} (y_i - \hat{y}_i)^2 \qquad \textbf{(6)}$$

where y_i and $\hat{y}_i$ are the experimental and predicted properties, respectively, of the i^{th} analyte in the calibration set, n_{full} and n_{red} are the number of analytes in the full and the reduced calibration sets, respectively, and p is the number of estimated parameters in the model.

For models based on the reduced calibration sets, for which holds $s_{red}^2 \le s_{full}^2$, the reduced calibration sets have a modelling power which is equal to or better than that of the full calibration set. For reduced models, for which holds $s_{red}^2 > s_{full}^2$, it is tested whether this difference is significant by means of a one-tailed F-test [32],

$$F_1 = \frac{s_{red}^2}{s_{full}^2} \qquad s_{red}^2 > s_{full}^2 \qquad F_{1,crit} = F_{\left(1-\alpha/2, n_{red}-p, n_{full}-p\right)} \qquad \textbf{(7)}$$

where α is the level of significance and $F_{(1-\alpha/2,nred-p,nfull-p)}$ is the F value at a confidence level of $1-\alpha/2$ and $n_{red}-p$ degrees of freedom of the numerator and $n_{full}-p$ degrees of freedom of the denominator.

The residual variance of the reduced calibration set, s_{red}^2, is not significantly larger than that of the full calibration set, s_{full}^2, if the calculated F_1-value is smaller than a one-tailed critical F-value, $F_{1,crit}$ at a confidence level of $1-\alpha/2$.

Thus, for the QSRR-model, developed with a reduced calibration set, for which $s_{red}^2 > s_{full}^2$ and

$$\frac{s^2_{red}}{s^2_{full}} < F_{1,crit}$$

it follows that

$$s^2_{red} < F_{1,crit} \cdot s^2_{full} \qquad \textbf{(8)}$$

The term $F_{1,crit} \cdot s_{full}^2$ determines a critical upper limit for the variance of the residuals of the reduced calibration set. This critical upper limit is called the critical calibration variance $s_{1,crit}^2$.

$$s^2_{1,crit} = F_{1,crit} \cdot s^2_{full} \qquad \textbf{(9)}$$

The residual variance s_{red}^2 of the QSRR model developed with the reduced calibration set is, at a given significance level α, not significantly larger than the residual variance s_{full}^2 of the QSRR model which is developed with the full calibration set, if $s_{red}^2 \leq s_{full}^2$ or if

$$\Delta s^2_1 = s^2_{1,crit} - s^2_{red} > 0 \qquad \textbf{(10)}$$

3.5.2 Prediction error

The resulting models are validated with a test set formed by the analytes, belonging to the full calibration set, but not selected for the reduced set.

For a test set the variance of the residuals of the predicted chromatographic property, s_{test}^2, is calculated with

$$s^2_{test} = \frac{1}{n_{test}} \sum_{i=1}^{n_{test}} (y_i - \hat{y}_i)^2 \qquad \textbf{(11)}$$

n_{test} is the number of analytes in the test set; n_{test}=n_{full}-n_{red}.

The QSRR model, developed with the reduced calibration set, is valid if s_{test}^2 is not significantly larger than the variance of the residuals of the chromatographic property calculated from the full calibration set, s_{full}^2. This is again tested by a one-tailed F-test:

$$F_2 = \frac{s^2_{test}}{s^2_{full}} \qquad s^2_{test} > s^2_{full} \qquad F_{2,crit} = F_{(1-\alpha/2, n_{test}, n_{full}-p)} \qquad \textbf{(12)}$$

This F-test is only applied if $s_{test}^2 > s_{full}^2$. In fact it is evaluated whether the prediction error of the QSRR model, developed with the reduced calibration set, is not worse than the calibration error of the model developed with the full calibration set.

Analogously to (8)-(10), equations (13)-(15) can be derived.

$$s_{test}^2 < F_{2,crit} \cdot s_{full}^2 \qquad \textbf{(13)}$$

$$s_{2,crit}^2 = F_{2,crit} \cdot s_{full}^2 \qquad \textbf{(14)}$$

$s_{2,crit}^2$ is called the critical validation variance. The term $F_{2,crit}{\cdot}s_{full}^2$ again determines a critical upper limit, now for the variance of the residuals of the test set.

The QSRR model, developed with the reduced calibration set, is valid if $s_{test}^2 \leq s_{full}^2$ or if

$$\Delta s_2^2 = s_{2,crit}^2 - s_{test}^2 > 0 \qquad \textbf{(15)}$$

In model validation it is usual to validate models by comparing calibration error (s_{red}^2) with the prediction error (s_{test}^2) with

$$F_3 = \frac{s_{test}^2}{s_{red}^2} \qquad s_{test}^2 > s_{red}^2 \qquad F_{3,crit} = F_{(1-\alpha/2, n_{test}, n_{red}-p)} \qquad \textbf{(16)}$$

We are however interested in how well the test set is predicted by our proposed models as compared to the models based on the full calibration set. Therefore we use the stricter condition of equation (12).

3.6 Data and methodology

3.6.1 Datasets

Five data sets, called *Kaliszan*, *Wilson*, *Al-Haj I*, *Al-Haj II* and *Tan,* were used to test the strategy.

Kaliszan-data set [11]*:* The chromatographic data concern log k_w values of 25 structurally diverse test analytes on 12 C_{18} and 6 C_8 columns in combination with one to four mobile phases: methanol-water, acetonitrile-water, methanol-buffer and acetonitrile-buffer, resulting in 42 chromatographic systems. Phosphate buffers were used with a concentration of 20 mM and pH of 3.0. The 25 test analytes were selected by Abraham et al. [33]. Table 3-1shows the columns and the numbering of the 42 systems. Table 3-2 shows the analytes and the values of the descriptors for the log *P*, LSER and QCI models. Analytes with missing values for either log k_w or a descriptor are not entered into the calibration or test sets.

Wilson-data set [18]*:* The data concern log *k* values of 45 neutral test analytes on 10 columns with the mobile phase acetonitrile-water 50% (v/v), forming 10 chromatographic systems. The values of the descriptors for the log *P* and LSER models were calculated with ADME Boxes [34] and are shown in Table 3-3.

Al-Haj I-data set [14]*:* The data set contains log k_w values of 58 test analytes on 3 columns using the mobile phases methanol-water, acetonitrile-water or acetonitrile-phosphate buffer (0.1 M, pH 7.0), forming 5 chromatographic systems. Descriptor values of 48 analytes are given for the log *P* model, of all analytes for the QCI model and of 40 analytes for the LSER model. In the QCI set values are given for the total dipole moment μ instead of μ^2 used in equation (3).

Table 3-1 Chromatographic columns, manufacturers, dimensions, abbreviations, and numbering of the chromatographic systems (1-42) for the *Kaliszan* data set. Extracted from Reference [11].

Stationary phases				Mobile phases			
C18 columns	**Manufacturer**	**Dimensions L x i.d. (mm x mm)**	**Abbreviations**	**MeOH-water**	**ACN-water**	**MeOH-buffer**	**ACN -buffer**
Zorbax RX-C18	Hewlett-Packard, Newport, DE, USA	150 x 4.6	RX	1	2	3	4
HypersilODS	Shandon HPLC, Runcom, UK	125 x 4.6	Hyper	5	6	7	8
Polygosil-60-5-C18	Macherey-Nagel, Diiren, Germany	125 x 4.6	Poly	9	10	11	12
Alltima C18 5U	Alltech, Deerfield, IL, USA	150 x 4.6	All	13	14	15	16
TSKgel OD-2PW	TosoHaas, Stuttgart, Germany	150 x 4.6	TPW	17	18	19	20
Eclipse X DB-CI8	Hewlett-Packard, Newport, DE, USA	150 x 4.6	XC18	21	22	23	24
Hypersil HyPURlTY C18	Shandon HPLC, Runcom, UK	150 x 4.6	HyPUR	25			
Kromasil KRl00-5C18	Eka Nobel, Bohus, Sweden	150 x 4.6	Krom	26			
Nucleosil 100-5 CI8 HD	Macherey-Nagel, Düren, Germany	150 x 4	NuC18	27			
Purospher RP-18 e	Merck, Darmstadt, Germany	125 x 4	Puro	28			
Symmetry C18	Waters, Milford, MA, USA	150 x 4.6	Sym18	29			
TSKgel ODS-80TS	TosoHaas, Stuttgart, Germany	150 x 4.6	TTS	30			
C8 columns							
LiChrospher RP-Select B	Merck, Darmstadt, Germany	125 x 4	SelB	31	32	33	34
Aluspher RP-Select B	Merck, Darmstadt, Germany	125 x 4	Alu	35	36	37	38
Nova-Pak C8	Waters, Milford, MA, USA	150 x 3.9	Nova	39			
Nucleosil 100-5 C8	Macherey-Nagel, Düren, Germany	150 x 4	NuC8	40			
SymmetryShield RP8	Waters, Milford, MA, USA	150 x 4.6	Sym8	41			
Eclipse XDB-C8	Hewlett-Packard, Newport, DE, USA	150 x 4.6	XC8	42			

Al-Haj II-data set [15]*:* The data set consists of log k_w values of 27 solutes on 7 columns, each with the mobile phases methanol-water and acetonitrile-water, forming 14 chromatographic systems. Descriptor values of 23 analytes are given for the log *P* model, of all analytes for the QCI model and of 25 analytes for the LSER model. In the QCI set values are given for the total dipole moment μ. The analytes in data set *Al-Haj II* form a subset of those in data set *Al-Haj I,* but the tested chromatographic systems are different.
Tan-data set [16]*:* The data concern log *k* values of 87 solutes on 5 columns with the mobile phase acetonitrile-water 50% (v/v), forming 5 chromatographic systems. For all analytes the values of the *S*, *A*, *B* and *V* descriptors are given. They are used in an adapted LSER model without *E*.

Together, these data sets contain retention values of 76 chromatographic systems, while 208 models (log *P*, QCI, LSER and adapted LSER) were built with the available data.

3.6.2 Software

All calculations are made with in-house made programs developed in Matlab (V. 5.3) (The Math Works, Natick, MA, USA) [35]. The Kennard and Stone algorithm from the ChemoAC Standard Functions Toolbox for MATLAB [36] is used for the selection of analytes. Log *P* and LSER descriptors for the analytes in the *Wilson* data set are calculated with ADME Boxes [34].

3.7 Results and Discussion

3.7.1 Determination of the minimal number of analytes for reduced calibration sets

The three QSRR models from equations (2), (3) and (4) require three descriptor sets. In this study we will try to reduce substantially the size of the calibration sets. The selection of a reduced set of analytes from a full calibration set is performed with the Kennard and Stone algorithm applied on the sets of descriptors. The descriptors for these reduced calibration sets are then used to develop the relevant QSRR models.

For each chromatographic system tested with a given data set, the QSRR models are developed with the full calibration set and s_{full}^2 is calculated. Thereafter, for each system, a series of models is developed with reduced calibration sets. The analytes for the reduced calibration sets are selected from the full set in the sequence as proposed by the Kennard and Stone procedure after auto scaling the variables. Each series of QSRR models with reduced calibration sets starts with the minimal number of analytes needed for MLR, being the number of coefficients in the model plus one. For instance, to develop the log *P* model (equation (2)), the 3 first analytes, selected by Kennard and Stone from their log *P* values, are used. Analogously, for the QCI model (equation (3)), modelling starts with the 5 first selected analytes, and for the LSER model (equation (4)) with the 7 first selected analytes.

Table 3-2 Structural descriptors of the test analytes that were employed in the QSRR equations for the Kaliszan data set [11]. For the meaning of the descriptors see text.

No.	Analyte	Log *P*	LSER descriptors					QCI descriptors		
			E	S	A	B	V	δ_{min}	μ^2	A_{WAS}
1	n-Hexylbenzene	5.52	0.591	0.50	0.00	0.15	1.562	-0.2104	0.03880	415.40
2	1,3,5-Triisopropylbenzene	-	0.627	0.40	0.00	0.22	1.985	-0.2057	0.00624	478.27
3	1,4-Dinitrobenzene	1.47	1.130	1.63	0.00	0.41	1.065	-0.3418	0.00012	312.07
4	3-Trifluoromethylphenol	2.95	0.425	0.87	0.72	0.09	0.969	-0.2454	4.39321	302.54
5	3,5-Dichlorophenol	3.62	1.020	1.10	0.83	0.00	1.020	-0.2434	1.98246	306.77
6	4-Cyanophenol	1.60	0.940	1.63	0.79	0.29	0.930	-0.2440	10.9693	290.61
7	4-Iodophenol	2.91	1.380	1.22	0.68	0.20	1.033	-0.3021	2.51856	301.47
8	Methylphenylether	2.11	0.708	0.75	0.00	0.29	0.916	-0.2116	1.56000	288.13
9	Benzamide	0.64	0.990	1.50	0.49	0.67	0.973	-0.4334	12.8450	293.30
10	Benzene	2.13	0.610	0.52	0.00	0.14	0.716	-0.1301	0.00000	244.95
11	Chlorobenzene	2.89	0.718	0.65	0.00	0.07	0.839	-0.1295	1.70824	269.49
12	Cyclohexanone	0.81	0.403	0.86	0.00	0.56	0.861	-0.2944	8.83278	269.31
13	Dibenzothiophene	4.38	1.959	1.31	0.00	0.18	1.379	-0.2709	0.27457	364.54
14	Phenol	1.47	0.805	0.89	0.60	0.30	0.775	-0.2526	1.52028	256.72
15	Hexachlorobutadiene	4.78	1.019	0.85	0.00	0.00	1.321	-0.0750	0.06708	352.14
16	Indazole	1.77	1.180	1.25	0.54	0.34	0.905	-0.2034	2.39011	285.46
17	Caffeine	-0.07	1.500	1.60	0.00	1.35	1.363	-0.3620	13.3298	367.02
18	4-Nitrobenzoic acid	1.89	0.990	1.07	0.62	0.54	1.106	-0.3495	11.7786	321.77
19	N-Methyl-2-pyrrolidinone	-0.54	0.491	1.50	0.00	0.95	0.820	-0.3532	12.9168	270.53
20	Naphthalene	3.30	1.340	0.92	0.00	0.20	1.085	-0.1277	0.00000	313.25
21	4-Chlorophenol	2.39	0.915	1.08	0.67	0.20	0.898	-0.2482	2.18448	280.38
22	Toluene	2.73	0.601	0.52	0.00	0.14	0.716	-0.1792	0.06916	274.50
23	Benzonitrile	1.56	0.742	1.11	0.00	0.33	0.871	-0.1349	11.1222	277.91
24	Benzoic acid	1.87	0.730	0.90	0.59	0.40	0.932	-0.3651	5.85156	288.00
25	1,3-Diisopropylbenzene	-	0.605	0.46	0.00	0.20	1.562	-0.2055	0.08820	399.79

All QSRR models developed with the reduced sets are validated with a test set consisting of the remaining analytes of the full calibration set, measured in the considered chromatographic system. For all QSRR-models the quality criteria as described in section 3.5 are calculated. For each combination of selected analytes in the reduced sets, n_{red}, and the corresponding number of analytes in the test set, n_{test}, a combined test for the calibration and validation is performed to evaluate whether the calibration and validation criteria hold at the one-sided confidence level of 97.5%.
Starting from the minimal number of analytes in the reduced sets, the selection was ended if the combined test for calibration and validation passes for all systems, three consecutive times, in order to avoid a pass by chance. For each set of descriptors, the minimal number of required analytes in the reduced calibration set was then the number at which the combined test passed for the first time.

As an example, for one chromatographic system of the *Kaliszan* data set in Fig. 3-1A (top window), the calibration variances s_{red}^2 of the log *P* models, the corresponding critical calibration variances $s_{1,crit}^2$ and s_{full}^2 are depicted against the number of analytes in the reduced calibration set. In the bottom window of Fig. 3-1A the validation variances s_{test}^2, the corresponding critical validation variances $s_{2,crit}^2$ and s_{full}^2 are shown against the number of analytes in the test set. Similar graphs for the QCI and LSER models are given in Fig. 3-1B and C. For each model on a chromatographic system the residual variance of the full calibration set, s_{full}^2, is constant. If $s_{red}^2 \leq s_{full}^2$, s_{red}^2 lies below the horizontal line of s_{full}^2, and $s_{1,crit}^2$ is not calculated. Analogously, when $s_{test}^2 \leq s_{full}^2$, $s_{2,crit}^2$ is not calculated.

Table 3-3 Neutral test analytes of the Wilson data set taken from [18]. Structural descriptors were calculated with ADME boxes (see text). For the meaning of the descriptors see text.

nr	component	Log *P*	*E*	*S*	*A*	*B*	*V*
1	Benzene	2.124	0.63	0.57	0.00	0.13	0.716
2	Toluene	2.598	0.64	0.57	0.00	0.15	0.857
3	Ethylbenzene	3.284	0.58	0.64	0.00	0.12	0.998
4	p-Xylene	3.077	0.66	0.57	0.00	0.18	0.998
5	Propylbenzene	3.579	0.64	0.56	0.00	0.18	1.139
6	Butylbenzene	4.117	0.64	0.56	0.00	0.18	1.280
7	Naphthalene	3.376	1.38	0.92	0.00	0.19	1.085
8	p-Chlorotoluene	3.142	0.78	0.67	0.01	0.12	0.980
9	Dichlorobenzene	3.246	0.91	0.77	0.00	0.07	0.961
10	Benzotrichloride	4.198	0.88	0.90	0.00	0.10	1.225
11	Bromobenzene	2.909	0.95	0.76	0.00	0.09	0.891
12	1-Nitropropane	1.239	0.22	0.72	0.00	0.25	0.706
13	Nitrobenzene	2.040	0.87	1.08	0.00	0.23	0.891
14	p-Nitrololuene	2.513	0.88	1.08	0.00	0.25	1.032
15	p-Nitrobenzyl chloride	2.933	1.01	1.26	0.03	0.24	1.154
16	N-Benzylformamide	0.679	0.91	1.56	0.26	0.66	1.114
17	Anisole	1.911	0.62	0.79	0.00	0.33	0.916
18	Benzyl alcohol	1.083	0.80	0.84	0.39	0.61	0.916
19	3-Phenyl propanol	2.191	0.80	0.84	0.37	0.58	1.198
20	5-Phenyl pentanol	3.275	0.79	0.86	0.37	0.58	1.480
21	Phenol	1.014	0.78	0.90	0.50	0.39	0.775
22	p-Chlorophenol	1.531	0.94	1.01	0.67	0.38	0.898
23	2,3-Dihydroxynaphthalene	1.868	1.65	1.40	0.77	0.59	1.203
24	1,3-Dihydroxynaphthalene	1.574	1.71	1.42	1.00	0.66	1.203
25	Eugenol	2.860	0.91	0.97	0.27	0.53	1.354
26	Danthron	2.220	2.19	2.18	0.49	0.82	1.646
27	n-Propyl formate	0.815	0.12	0.73	0.00	0.42	0.747
28	Melhyl benzoate	2.030	0.71	0.94	0.00	0.45	1.073
29	Benzonitrile	1.748	0.81	1.09	0.00	0.27	0.871
30	Coumarin	1.432	1.13	1.30	0.03	0.56	1.062
31	Acetophenone	1.671	0.70	1.03	0.00	0.46	1.014
32	Benzophenone	3.379	1.35	1.38	0.00	0.40	1.481
33	*cis*-Chalcone	3.731	1.52	1.70	0.00	0.58	1.720
34	*trans*-Chalcone	3.731	1.52	1.70	0.00	0.58	1.720
35	*cis*-4-Nitrochalcone	3.580	1.79	2.27	0.00	0.68	1.894
36	*trans*-4-Nitrochalcone	3.580	1.79	2.27	0.00	0.68	1.894
37	*cis*-4-Methoxychalcone	3.609	1.58	1.80	0.00	0.79	1.919
38	*trans*-4-Methoxychalcone	3.609	1.58	1.80	0.00	0.79	1.919
39	Prednisone	0.729	2.19	3.25	0.41	1.97	2.712
40	Hydrocortisone	1.651	2.04	2.92	0.73	1.90	2.798
41	Mephenytoin	1.380	1.38	1.59	0.16	1.11	1.684
42	0xazepam	2.177	2.40	1.83	0.60	1.43	1.992
43	Flunitrazepam	2.938	2.14	2.15	0.00	1.15	2.143
44	5,5-Diphenylhydantoin	1.831	1.94	2.04	0.44	1.14	1.869
45	N,N-Dimethyl acetamide	-0.664	0.26	1.01	0.00	0.83	0.788

In Fig. 3-1A, the combined test for calibration and validation is passed for all analyte numbers 3-18 in the reduced calibration sets and this is passed for the first time for 3 analytes. Therefore, for the log *P* model, for the chromatographic system concerned, the minimal number of analytes in the reduced calibration set is considered to be three.
In Fig. 3-1B, the combined test is passed for all analyte numbers 5-18 and this for the first time for 5 analytes. Therefore, for the QCI model, for the system concerned, the minimal number of analytes in the reduced calibration set is considered to be five.
Fig. 3-1C shows that for 7-8 analytes in the calibration sets and in the corresponding test sets, $s_{test}^2 > s_{2,crit}^2$ holds. The combined test passed for 9-18 analytes. For the LSER model, for the system concerned, the minimal number of analytes in the reduced calibration set is 9.
The minimal number of analytes for the reduced calibration sets for the three models, are determined similarly on all chromatographic systems of the five data sets.

In Fig. 3-2, for the *Kaliszan* data set, the numbers of chromatographic systems for which the combined calibration and validation test passed for a given model are shown against the number of the reduced-calibration set analytes.
For the log *P* model, a calibration set with 3 analytes allows passing the test for all systems. On all systems for the QCI and LSER models, calibration sets with 6 and 13 analytes, respectively, pass the combined calibration and validation test.

Therefore, for the *Kaliszan* data set, log *P* models developed with reduced calibration sets with 3 properly selected analytes are suitable for QSRR modelling and prediction. This is also true for QCI models with 6 and LSER models with 13 analytes.

The minimal numbers of analytes in reduced calibration sets are also determined for the *Wilson*, *Al-Haj I*, *Al-Haj II* and *Tan* data sets according to the above procedure, see Table 3-4.

For each model the maximal number of analytes in the reduced calibration set, $max(n_{red})$, is determined, and was found to be 7 for the log *P* models, 8 for the QCI models, 15 for the LSER models, and 9 for the adapted LSER model with 4 descriptors (for the *Tan* data set).

The analytes in the reduced calibration sets are selected by the Kennard and Stone procedure, starting from the center. Except for the center point the remaining analytes of the reduced sets will be located at the extremes of the variable space formed by the descriptors. Using uncorrelated descriptors, which one expects in good QSRR models, the number of analytes to describe the experimental domain will be equal to two times the number of descriptors. Therefore, this number of analytes plus one, the center point analyte, could be considered the minimal number of analytes to build proper QSRR models. From the case studies (Table 3-4) a somewhat higher number of analytes seems to be required, in practice 7, 8 and 15 versus 3, 7 and 11 respectively.

In Table 3-4, at the one-sided confidence level of 97.5%, useful QSRR models are developed by using three times the number of descriptors, *m*, in the model as the number of analytes in the reduced calibration set for the QCI-, LSER- and adapted LSER-models.
The finding that 3 analytes per descriptor, selected as described above, are sufficient to develop useful QSRR models, allows smaller calibration sets than applying the traditional rule of thumb suggesting 4 to 6 analytes per descriptor.
For the log *P* models, 7 analytes are required in the worst case. Thus even for a simple model with one descriptor a minimum of analytes seems to be required. Here, it will depend on the correlation between retention and log *P*, and on the linearity of this combination.

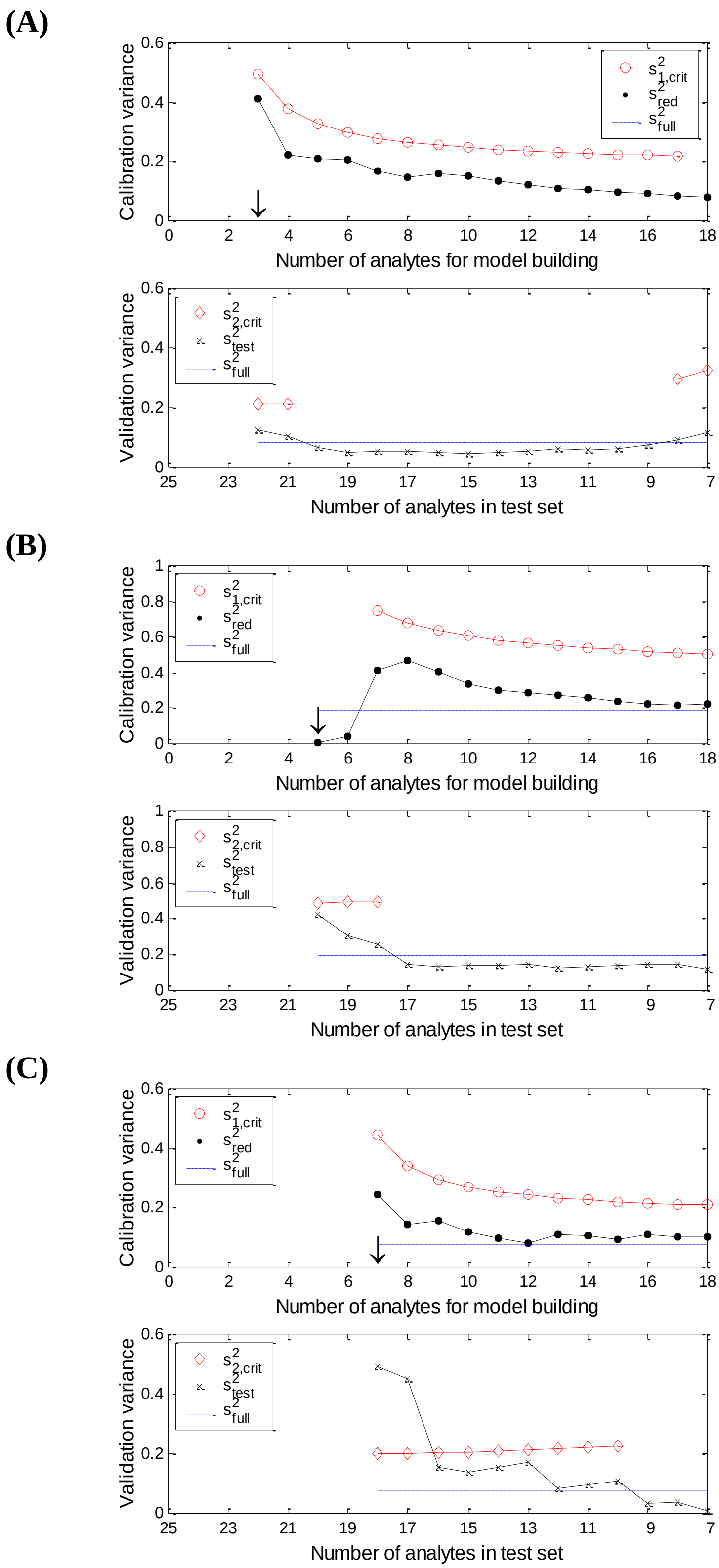


Fig. 3-1 Calibration and validation variances as a function of the numbers of analytes used for model building and for testing, respectively, for chromatographic system 1 of the *Kaliszan* data set; (A) for the log *P* models, (B) for the QCI models, (C) for the LSER models; Top window: O $s_{1,crit}^2$, critical calibration variance (one-sided, 97.5%), • s_{red}^2, variance of the residuals of the reduced calibration sets, – s_{full}^2, variance of the residuals of the full calibration set, and ↓ minimal number of analytes required; Bottom window: ◊ $s_{2,crit}^2$, critical validation variance (one-sided, 97.5%), x s_{test}^2, variance of the residuals of the test sets, and – s_{full}^2, variance of the residuals of the full calibration set.

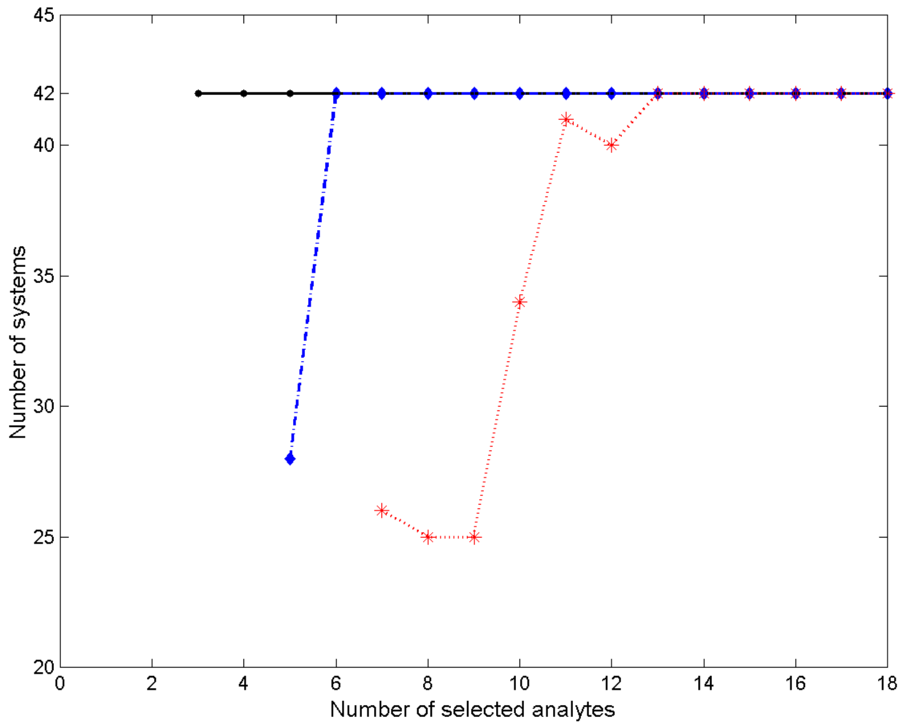


Fig. 3-2 Results for the *Kaliszan* data set; Number of chromatographic systems (out of 42) for which the combined calibration and validation test at a one-sided confidence level of 97.5% passed for a given number of calibration set analytes; (●) for the log *P* models, (♦) for the QCI models, (✳) for the LSER models

Table 3-4 The numbers of analytes in the reduced calibration sets for the different models (at one-sided confidence level of 97.5%)

No	Data set	No of systems	Log *P*		QCI		LSER		adapted LSER	
			n_{full}	n_{red}	n_{full}	n_{red}	n_{full}	n_{red}	n_{full}	n_{red}
1	Kaliszan	42	23	3	25	6	25	13		
2	Wilson	10	45	6			45	**15**		
3	Al-Haj I	5	48	4	58	**8**[*]	40	12		
4	Al-Haj II	14	23	**7**	27	6	25	11		
5	Tan	5							87	**9**
max(n_{red})				7		8		15		9
number of descriptors *m*				1		3		5		4
3**m*				3		9		15		12

[*] after the deletion of one outlying chromatographic system

In Table 3-5 and Table 3-6 the correlation matrices are given for the full and reduced calibration sets between the QCI and LSER descriptors respectively of the *Kaliszan* data set. It is seen that (*i*) correlations between descriptors are low, and (*ii*) similarity between both correlation matrices is high.

Table 3-5 Correlation matrix of the QCI descriptors for the Kaliszan data set, (a) full calibration set, (b) reduced calibration set with 6 analytes

(a)

	δ_{min}	μ^2	A_{WAS}
δ_{min}	1	-0.56	0.03
μ^2	-0.56	1	-0.25
A_{WAS}	0.03	-0.25	1

(b)

	δ_{min}	μ^2	A_{WAS}
δ_{min}	1	-0.63	0.04
μ^2	-0.63	1	-0.42
A_{WAS}	0.04	-0.42	1

Selected analytes in selection order: 3-Trifluoromethylphenol; 1,3,5-Triisopropylbenzene; Benzamide; Hexachlorobutadiene; Benzonitrile; Benzene

Table 3-6 Correlation matrix of the LSER descriptors for the Kaliszan data set, (a) full calibration set, (b) reduced calibration set with 13 analytes

(a)

	E	*S*	*A*	*B*	*V*
E	1	0.54	0.09	0.13	0.19
S	0.54	1	0.34	0.57	-0.23
A	0.09	0.34	1	-0.19	-0.29
B	0.13	0.57	-0.19	1	-0.02
V	0.19	-0.23	-0.29	-0.02	1

(b)

	E	*S*	*A*	*B*	*V*
E	1	0.46	-0.20	0.17	0.32
S	0.46	1	0.09	0.59	-0.25
A	-0.20	0.09	1	-0.30	-0.35
B	0.17	0.59	-0.30	1	-0.01
V	0.32	-0.25	-0.35	-0.01	1

Selected analytes in selection order: Benzonitrile; Caffeine; 1,3,5-Triisopropylbenzene; Dibenzothiophene; 3,5-Dichlorophenol; N-Methyl-2-pyrrolidinone; Benzamide; Hexachlorobutadiene; 1,4-Dinitrobenzene; 3-Trifluoromethylphenol; Toluene; 4-Cyanophenol; Phenol

3.7.2 Covering of variable spaces by reduced calibration sets

In Fig. 3-3 the experimental and predicted log k_w values of the analytes for one chromatographic system of the *Kaliszan* data set are depicted for the log *P*, QCI and LSER models based on reduced calibration sets with 7 analytes for the log *P* model and three analytes per descriptor for the QCI and LSER models. Coefficients of multiple determination were estimated by linear regression. Fig. 3-3 shows that the retentions of the analytes in the reduced calibration sets are distributed over the entire range of log k_w values. This observation was seen for all other combinations of models and systems in the five data sets. This indicates that the extremes of the descriptor space (which were selected by the Kennard and Stone algorithm) also include the analytes with the extreme log k_w values.

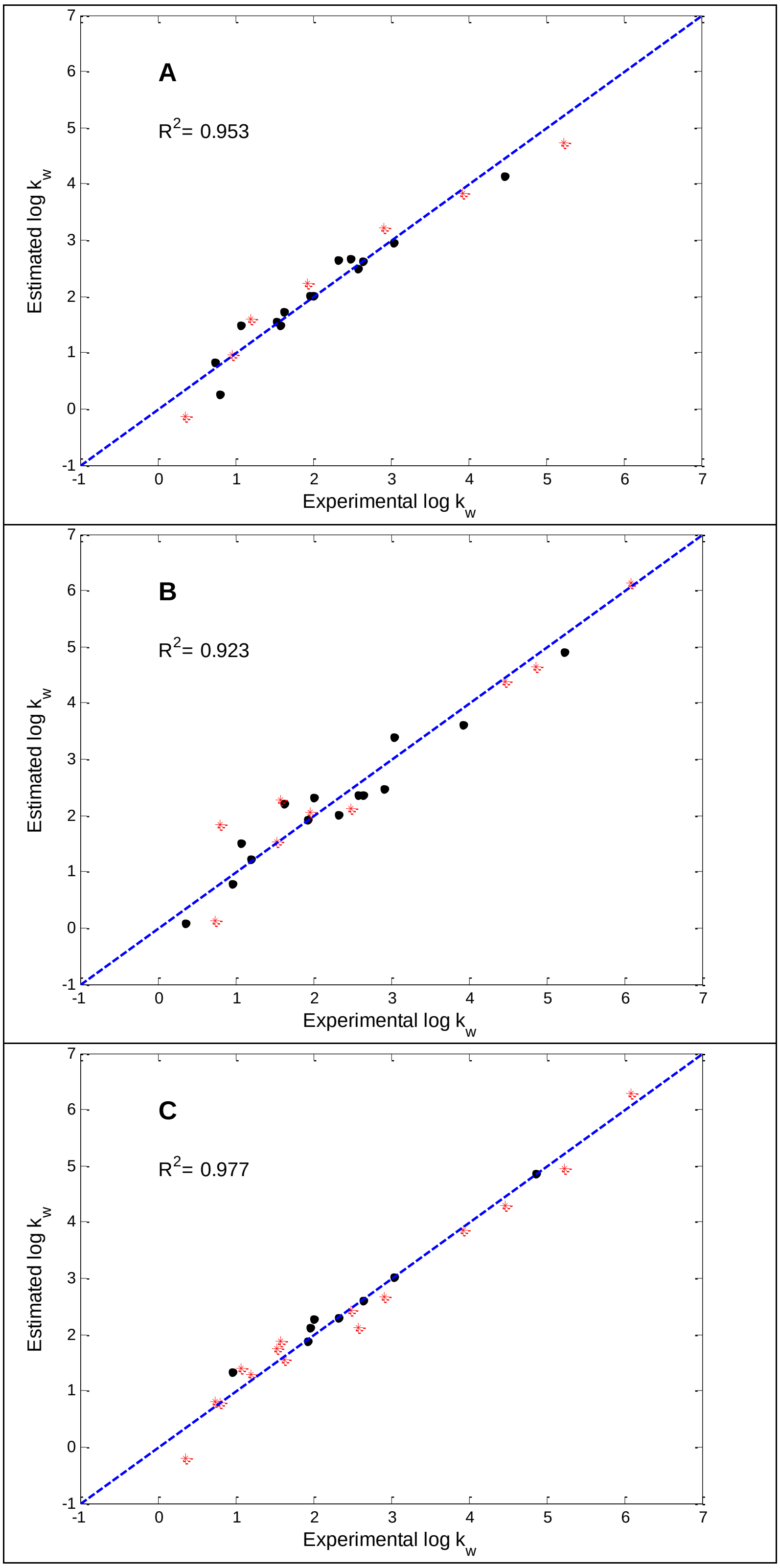


Fig. 3-3 Kaliszan data set; Estimated log kw vs. experimental log kw for chromatographic system 1, (A) for the log P model, (B) for the QCI model, (C) for the LSER model, (✳) analytes of reduced calibration set, (•) test analytes

Fig. 3-4 shows that the log *P* values of the 7 selected analytes in the five data sets are well distributed over the entire range. In Fig. 3-5, for the *Kaliszan* data set, the values of the auto scaled QCI descriptors are shown of all analytes in the full calibration set and of the nine selected analytes in the reduced set. The latter are well distributed over the whole range of descriptor values. This observation is seen for all QCI sets in the five data sets.
In Fig. 3-6, for the *Kaliszan* data set, a similar graph is given for the LSER descriptors. The values of the fifteen selected analytes are again well distributed over the whole range of descriptor values, an observation valid for all LSER sets in the five data sets.

Both the dependent and independent variable spaces are thus covered well by the reduced calibration sets. The retention values of the analytes in the reduced calibration sets are distributed over the entire log k_w range and the descriptor values in the log *P*-, QCI- and LSER-models are distributed over the descriptor ranges.

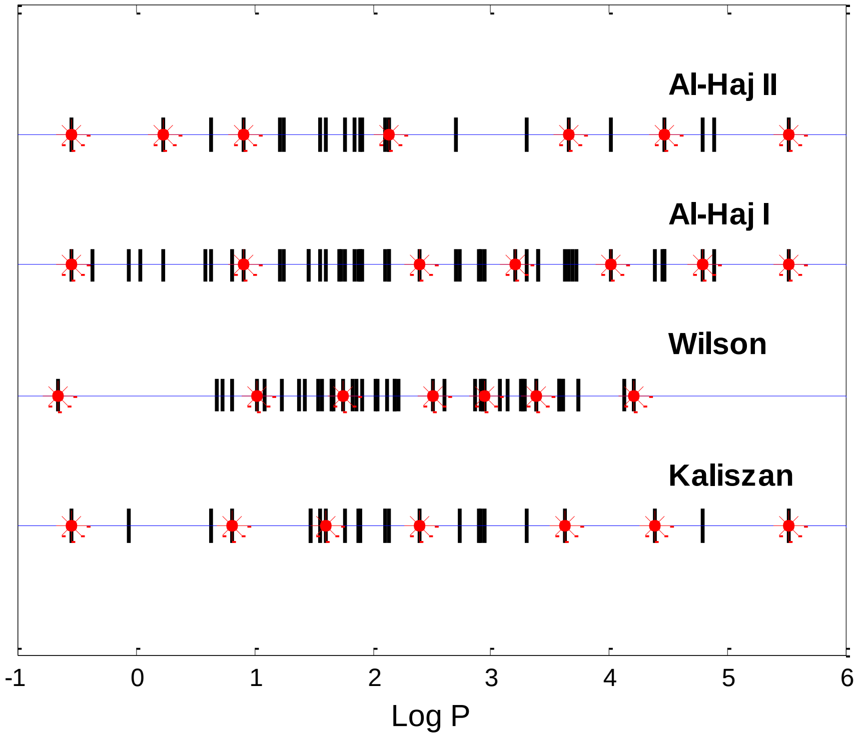


Fig. 3-4 Log *P* values of all analytes (|) and of the 7 selected analytes (✳) in the different data sets

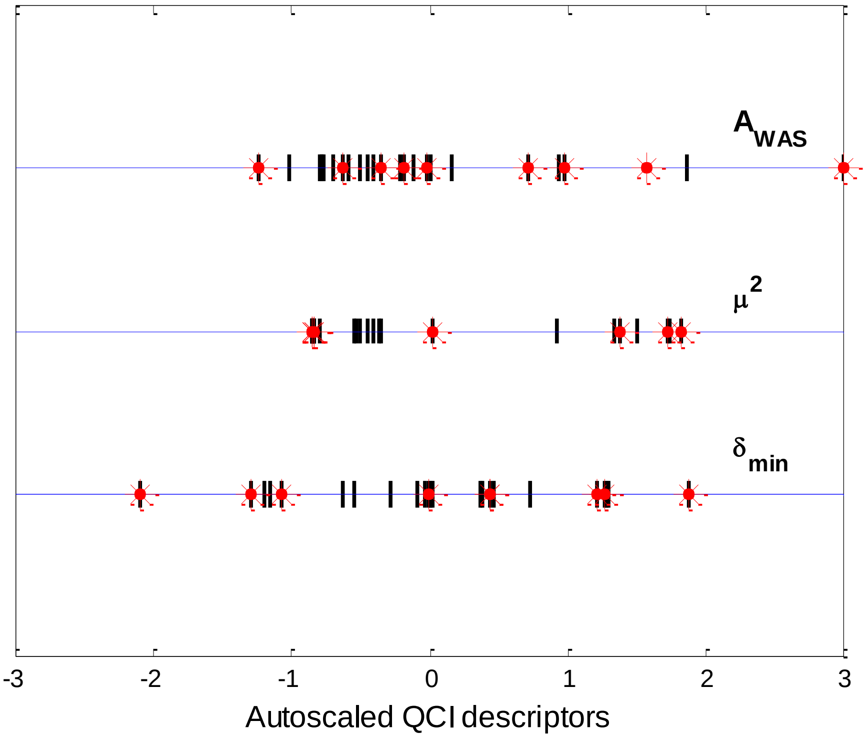


Fig. 3-5 ***Kaliszan*** **data set; Values of the auto scaled QCI descriptors of all analytes (|) and 9 selected analytes (✳)**

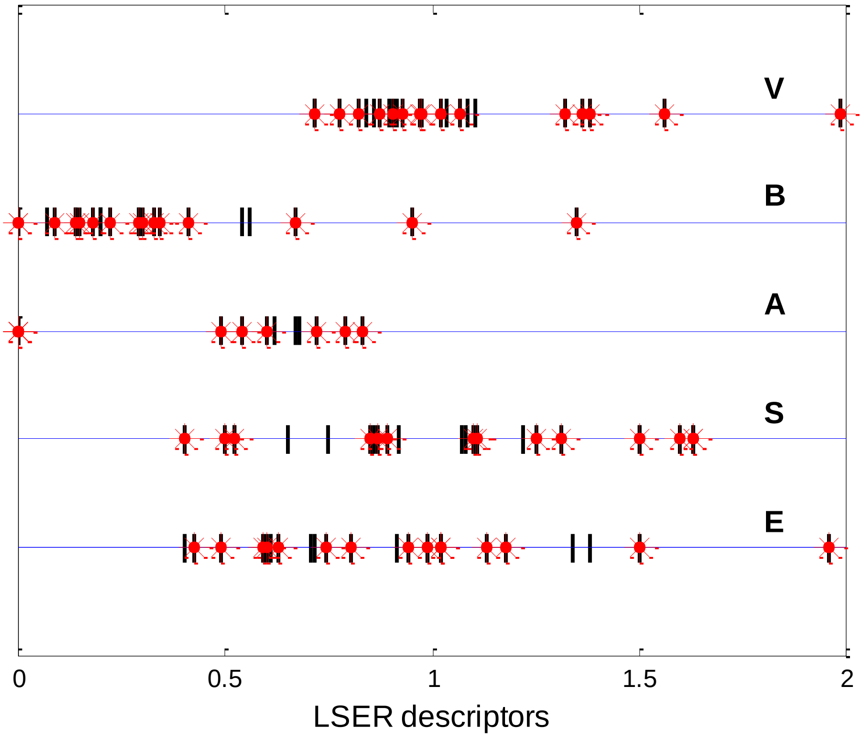


Fig. 3-6 ***Kaliszan*** **data set; Values of the LSER descriptors of all analytes (|) and 15 selected analytes (✳)**

3.7.3 Guidelines to construct small calibration sets

In QSRR-studies, for instance using another data set or another equation, the experimental work can be reduced substantially by using small calibration sets. Time is saved when these small calibration sets are constructed before any experiment is carried out. The results in this study allow defining some guidelines to construct small calibration sets.

1. Determine the descriptors to be included in the model.
2. Select a large set of candidate analytes which are considered representative in relation to the application involved.
3. Calculate the descriptors of the analytes.
4. Select analytes by the method of Kennard and Stone, until the number of analytes is equal to 3*m*.
5. Carry out the experiments with the small calibration set.
6. Check whether the range of retention values is sufficiently large for the application at hand and develop the QSRR model.

3.8 Conclusions

The aim of this work was to develop a strategy for the construction of reliable reduced calibration sets for QSSR models, based on molecular structure properties.

It has been demonstrated, using 76 reversed-phase high-performance liquid chromatography systems, that it is possible to develop useful QSRR models based on selected reduced calibration sets. The analytes in the reduced calibration sets were selected based on their distribution in the molecular-descriptor space. Selection was carried out by the algorithm of Kennard and Stone on the auto scaled descriptors. The calibration and prediction errors of the reduced calibration sets are not significantly larger than the calibration errors of the corresponding full calibration sets. Both the dependent variable space, formed by the retention values log k_w or log k, and the independent variable space, formed by the in the model considered descriptor values, are covered well by the reduced calibration sets.

The results show that application of the proposed strategy provides log *P* models with seven, and QCI and LSER models with three selected analytes per descriptor, which are suitable for the future prediction of retentions. Substantial reductions of calibration sets for log *P*, QCI and LSER models can thus be realised. Guidelines to construct small calibration sets are formulated.

The use of these reduced calibration sets will reduce the experimental workload for the development of solid QSRR models substantially. This may encourage the use of QSRRs in laboratory practice.

Acknowledgements

The authors thank Michael L.M. Kromdijk for technical assistance

References

[1] U.D. Neue, HPLC Columns: Theory, Technology and Practice, Wiley, New York, 1997.
[2] H.A. Claessens, Trends Anal. Chem. 20 (2001) 563.
[3] D. Visky, Y. Vander Heyden, T. Iványi, P. Baten, J. De Beer, Z. Kovács, B. Noszál, E. Roets, D.L. Massart, J. Hoogmartens, J. Chromatogr. A 977 (2002) 39.
[4] D. Visky, Y. Vander Heyden, T. Iványi, P. Baten, J. De Beer, Z. Kovács, B. Noszál, P. Dehouck, E. Roets, D.L. Massart, J. Hoogmartens, J. Chromatogr. A 1012 (2003) 11.
[5] T. Hancock, R. Put, D. Coomans, Y. Vander Heyden, Y. Everingham Chemometr. Intell. Lab. Syst. 76 (2005) 185.
[6] H.A. Claessens, M.A. van Straten, C.A. Cramers, M. Jezierska, B. Buszewski, J. Chromatogr. A 826 (1998) 135.
[7] R. Put, Y. Vander Heyden, Anal. Chim. Acta 602 (2007) 164.
[8] R. Kaliszan, Structure and Retention in Chromatography. A Chemometric Approach, Harwood Academic Publishers, Amsterdam, 1997.
[9] T. Baczek, R. Kaliszan, J. Chromatogr. A 962 (2002) 41.
[10] M.H. Abraham, M. Rozés, C.F. Poole, S.K. Poole, J. Phys. Org. Chem. 10 (1997) 358.
[11] R. Kaliszan, M. A. van Straten, M. Markuszewski, C.A. Cramers, H.A. Claessens, J. Chromatogr. A 855 (1999) 455.
[12] R. Kaliszan, Chem. Rev. 107 (2007) 3212.
[13] K. Héberger, J. Chromatogr. A 1158 (2007) 273.
[14] M.A. Al-Haj, R. Kaliszan, A. Nasal, Anal. Chem. 71 (1999) 2976.
[15] M.A. Al-Haj, R. Kaliszan, B. Buszewski, J. Chromatogr. Sci. 39 (2001) 29.
[16] L.C. Tan, P.W. Carr, M.H. Abraham, J. Chromatogr. A 752 (1996) 1.
[17] J. Zhao, P.W. Carr, Anal. Chem. 70 (1998) 3619.
[18] N.S. Wilson, M.D. Nelson, J.W. Dolan, L.R. Snyder, R.G. Wolcott, P.W. Carr, J. Chromatogr. A 961 (2002) 171.
[19] L.A. Lopez, S.C. Rutan, Journal of Chromatography A 965 (2002) 301
[20] E. Soczewinski, C.A.J. Wachtmeister, J. Chromatogr. 7 (1962) 311.
[21] T. Dzido, H. Engelhardt, Chromatographia 39 (1994) 51.
[22] C. F. Poole, S. K. Poole, A. D. Gunatilleka, Adv. Chromatogr. 40 (2000) 159.
[23] J. Jiskra, H.A. Claessens, C.A. Cramers, R. Kaliszan, J. Chromatogr. A 977 (2002) 193.
[24] T. Baczek, R. Kaliszan, J. Chromatogr. A 987 (2003) 29.
[25] M.A. Al-Haj, P. Haber, R. Kaliszan, B. Buszewski, M. Jezierska, Z. Chilmonzyk, J. Pharm. Biomed. Anal. 18 (1998) 721.
[26] R. Kaliszan, T. Baczek, A. Bucinski, B. Buszewski, M. Sztupecka, J. Sep. Sci. 26 (2003) 271.
[27] M.H. Abraham, Chem. Soc. Rev. 22 (1993) 73.
[28] M.H. Abraham, A. Ibrahim, A.M. Zissimos, J. Chromatogr. A 1037 (2004) 29.
[29] C.F. Poole, S.K. Poole, J. Chromatogr. A 965 (2002) 263.
[30] M. Vitha, P.W. Carr, J. Chromatogr. A 1126 (2006) 143.
[31] R.W. Kennard, L.A. Stone, Technometrics 11 (1969) 137.

[32] D.L. Massart, B.G.M. Vandeginste, L.M.C. Buydens, S. de Jong, P.J. Lewi, J. Smeyers-Verbeke, Handbook of Chemometrics and Qualimetrics, Part A, Elsevier, Amsterdam, 1997.

[33] M.H. Abraham, H.S. Chadha, R.A.E. Leitao, R.C. Mitchell, W.J. Lambert, R. Kaliszan, A. Nasal, P. Haber, J. Chromatogr. A 766 (1997) 35.

[34] http://pharma-algorithms.com/ (accessed on Feb 2, 2009)

[35] http://www.mathworks.com/ (accessed on Feb 2, 2009)

[36] CHEMOAC Standard Function Toolbox (http://www.vub.ac.be/fabi/publiek/index.html) (accessed on Feb 2, 2009)

4 Introduction to variable selection

Efficient variable selection methods can help reducing the data flood in chemometrics, especially when they are applied on widely used multivariate regression techniques in analytical chemistry. These techniques are used for the extraction of relevant chemical information about analytes, products or processes [1-3]. With multivariate regression models chemical quantities can frequently be estimated with reasonable accuracy and with minimal data treatment [3]. Partial least squares (PLS) regression is a commonly used multivariate technique. PLS models the relationship between the variables in a data matrix **X** and a response matrix **Y** by defining a set of latent variables which maximizes the explained covariance [1,2,4]. PLS is considered able to deal with a large number of noisy and correlated variables, and with small numbers of samples. It is a versatile technique, used for both qualitative and quantitative analysis, in many different application fields, such as food chemistry, pharmaceutical analysis, agriculture, environment, and industrial and clinical chemistry [5].

PLS regression has foremost been used for quantitative tasks in multivariate calibration [1,2], but has also been applied for qualitative classification tasks in the form of partial least squares discrimination analysis (PLS-DA) [6,7]. The PLS-DA method is especially useful for high-dimensional data, where classical discrimination methods such as linear discriminant analysis (LDA) have numerical difficulties because of singularity issues [8]. PLS-DA is one of the most widely used classification methods, not only in chemometrics but also in bioinformatics [8,9].

Modern analytical techniques produce huge amounts of data. However, most of it is noisy or uninformative data. With variable selection, noisy and uninformative variables can be eliminated, and subsets containing informative variables retained. Using only informative subsets of variables, simple, robust and interpretable PLS models can be obtained, both in chemometrics and bioinformatics [10-18].

In bioinformatics, especially in metabolomics, variable selection is used for biomarker discovery [9,19-23]. Biomarkers are measurable biological characteristics which can be used as indicators of a biological state or condition [22]. For biomarker discovery, it is important to find the simplest combination of metabolites that can produce a suitably effective predictive result [22]. Hence there is a need for highly selective variable selection methods. Existing methods should be modified or new methods developed to meet this challenge. Additionally, variable selection will also help to master the data tsunami in chemometrics and bioinformatics [24].

In this introduction, an overview is given of variable selection methods for PLS1 and PLS2, both for quantitative and qualitative tasks, because PLS now dominates multivariate modelling in chemometrics [25]. Also, an overview is given of variable selection for Quantitative Structure-Activity Relationships (QSAR) modelling and the related Quantitative Structure-Retention Relationships (QSRRs) for Reversed-Phase Liquid Chromatography (RPLC). The PLS modelling, including data pre-processing and validation, and the scope of the variable selection process are also described. The characteristics of the most widely used types of variable selection methods and their advantages and drawbacks are highlighted.

Methods which are mostly applied for multiple linear regression (MLR), such as variable selection in a stepwise mode [26] and successive projections algorithms (SPA) [27], or methods in which variables are selected independently of PLS modelling are not included.

Finally conclusions are formulated for the development of new variable selection methods in this PhD project.

4.1 PLS model

A PLS model for multiple responses (PLS2) is developed from a calibration set of N objects or observations with M responses or dependent variables in the **Y** matrix and K independent predictor variables in the **X** matrix. The $\mathbf{Y}(N \times M)$ matrix consist of M column vectors of dependent response variables denoted by $\mathbf{y_m}$ (m=1, …, M). The $\mathbf{X}(N \times K)$ matrix consist of K column vectors of independent predictor variables denoted by $\mathbf{x}_\mathrm{k}$ (k=1, …, K). The objective of PLS is to select the optimal number A ($A \leq K$) of latent variables or PLS2 factors, which are linear combinations of the original variables $\mathbf{x}_\mathrm{k}$. The PLS2 model is given by Eqs. (**1**) and (**2**).

$$\mathbf{X} = \mathbf{TP^T} + \mathbf{E_A} \qquad \textbf{(1)}$$

$$\mathbf{Y} = \mathbf{TQ^T} + \mathbf{F_A} \qquad (2)$$

where $\mathbf{T}(N \times A)$ is a score matrix, $\mathbf{P}(K \times A)$ a matrix with the x-loading vectors $\mathbf{p}_a$ (a=1, 2, …, A) as columns, $\mathbf{Q}(M \times A)$ a matrix with the y-loading vectors $\mathbf{q}_a$ (a=1, 2, …, A) as columns, $\mathbf{E_A}(N \times K)$ and $\mathbf{F_A}(N \times M)$ the residual matrices for **X** and **Y**, respectively, after the extraction of A factors. The optimal number of PLS factors, A, can be determined using cross-validation (CV).

The matrix $\mathbf{B}(K \times M)$, with PLS2 regression coefficients b_{km}, can be estimated after calibration, with,

$$\mathbf{B} = \mathbf{W}\left(\mathbf{P}^T\mathbf{W}\right)^{-1}\mathbf{Q} \qquad \textbf{(3)}$$

where $\mathbf{W}(K \times A)$ is the **X** weight matrix [2].

The responses of the samples in the test set can be predicted with,

$$\hat{\mathbf{Y}}_{Test} = \mathbf{X}_{Test}\mathbf{B} \qquad \textbf{(4)}$$

where $\hat{\mathbf{Y}}_{Test}$ $(N_{Test} \times M)$ is the predicted response matrix of the test set samples, $\mathbf{X}_{Test}$ $(N_{Test} \times K)$ is the data matrix of the test set, and N_{Test} is the number of test-set samples.

For a PLS model with one response (PLS1), similar equations can be used with M=1. Further details on PLS can be obtained in Refs. [1,2,4].

4.2 PLS modelling

PLS model building encompasses the following steps: (*i*) data pre-processing, (*ii*) modelling, and (*iii*) validation. Each step in this process has an effect on the following steps. These steps are described below.

4.2.1 Data pre-processing

There are many experimental and instrumental effects causing additional variations and non-linearities in the data which are not related to the composition of the samples. Examples of these effects are, sample collection, sample preparation and instrumental artefacts [28,29]. PLS has a high modelling power and these additional variations and non-linearities can be modelled in conjunction with the target information, at the expense of higher model complexities. Proper data pre-processing can eliminate these unwanted variations beforehand and concentrate the relevant information in the first PLS factors, which results in more parsimonious models [28].

The results of PLS modelling depend on the pre-processing of the data [12]. The influence of more informative **X**-variables can be increased by appropriate pre-processing [1]. Important pre-processing techniques are centering, scaling, normalisation, standard normal variate transformation, multiplicative scatter correction, Savitzky-Golay smoothing, differentiation, and orthogonal signal correction. However, pre-processing affects the data analysis depending on the analytical technique used and there is no single recipe that can be used for all data [29]. The pre-processed data set is used as the basis for variable selection [12].

Variables are *centered* by subtracting their averages. Centering removes the offset from the data. It may (*i*) reduce the rank of the model, (*ii*) increase the fit to the data, or (*iii*) avoid numerical problems. Centering will not remove scale differences between variables [30].

Scaling is used to adjust scale differences or to accommodate for heteroscedasticity. It changes the weights of the variables [30]. Variables which ranges are different more than one magnitude of 10 are often *logarithmically scaled*. This make their distributions fairly symmetrical. If the relative importance of variables is unknown [1], or when variables have different scales [3], variables are first centred by subtracting their averages, followed by division by their standard deviations. This so-called *auto-scaling* gives each variable the same prior importance in the analysis [1].

Normalisation is applied if size effects of samples, such as those of concentration, should be removed. Chromatographic or spectral profiles of samples can be normalised by division of each value by the sum [31] or norm [32] of the profile.

The *standard normal variate* (SNV) *transformation* reduces multiplicative effects of scattering, particle size and multi-colinearity changes over spectra. In SNV each spectrum is first centred and then scaled by its standard deviation [33,34].

Multiplicative scatter correction (MSC) eliminates the effect of light scattering of particles of different sizes and shapes in solutions. It corrects for both multiplicative and additive scatter effects [34]. MSC improves the linearity of the **X**-**y** relation. A linear regression is performed between a sample spectrum $\mathbf{x}_i$ and a reference spectrum $\mathbf{x}_{ref}$, most frequently the mean spectrum: $\mathbf{x}_i$=b_0+$b_1\mathbf{x}_{ref}$. Thereafter, the sample spectrum is corrected by subtraction of the intercept b_0 and division by the slope b_1: $\mathbf{x}_{i,MSC}$= $(\mathbf{x}_i$-$b_0)/b_1$, [2,33-35].

The noise in each sample profile point consists of random changes in the amplitude of the signal. *Smoothing* reduces the signal-to-noise ratio of these profiles. The best-known algorithm used for smoothing is that of Savitzky and Golay (SG) [36]. In SG smoothing, the noise fluctuations in the data are reduced by the application of a $2m+1$ (m=1, 2, …) wide moving window. A polynomial of a chosen degree n ($n < 2m+1$) is fitted to equally spaced data in the window by least squares regression analysis, and the central point of the window is interchanged with the corresponding fitted value of the polynomial. Thereafter, the window is moved one point, a new polynomial calculated, and a new fitted value interchanged with the new central point, etc.. The least squares regression procedure is accelerated by the use of pre-calculated arrays of convoluting integers and array norms for each order n of the polynomial (n=2, 3, …) and each window size ($2m+1$=5, 6, …). The convolution arrays with the corresponding norms are also called SG filters. The method is introduced by Savitzky and Golay in [36]. Corrections are published in [37,38]. Equations for the calculation of SG filters are given in [38].

Differentiation is widely applied to eliminate background or baseline effects and to enhance differences between profiles [34,39]. The first derivative removes constant baseline or background effects and the second derivative eliminates linear baseline shifts [34]. The Savitzky-Golay procedure is the recommended method for the calculation of derivatives [39]. It combines smoothing and differentiation into one single step. Signals can be differentiated by SG filters and derivatives of smoothed signals are obtained. The central point of the moving window is interchanged with the corresponding derivative of a chosen degree of the fitted least squares polynomial [36].

In *Orthogonal signal correction* (OSC), information is removed that is orthogonal to the response **y** [33,40-43].

4.2.2 Modelling and validation

In PLS modelling, an optimal model is developed, based on a representative set of calibration samples and using a suitable PLS algorithm. The main purpose is to estimate predictor parameters from the PLS model in such a way that predictions of the response **y** with measured **X** values of future unknown samples have as low prediction errors as possible. The optimal PLS model complexity must be determined, and predictor parameters must be estimated. Finally, the obtained model must be validated [44]. These steps make the development of an operational PLS model a complicated process.

Representative samples can be selected using accurate and reproducible sampling procedures. They must also be representative for future unknown samples of the same kind in relation to the problem at hand. See for more details Ref. [45].

Two PLS algorithms are widely used: the Nonlinear Iterative Partial Least Squares (NIPALS) algorithm with orthogonal scores, and SIMPLS. The NIPALS algorithm is introduced by Wold et al. in [46], see also [1,2,4]. It can be considered as the standard PLS algorithm [47,48]. SIMPLS is introduced by de Jong in [48]. Faber and Ferré showed that NIPALS is the most stable and SIMPLS the fastest algorithm [47].

The PLS model is developed with two independent sets of samples, a calibration or *training* set and a *test* set. The model is built with the calibration samples in the training set. The model must be validated before it is used for prediction of response values **y** of new samples. Therefore, during model building, the predictive ability of the PLS model is assessed by internal validation with the training set. Finally, the PLS model is assessed by external validation with a test set. The samples in the test set are independent from the samples in the training set. Mostly, the training and test sets are obtained by partitioning the original data set, for instance using the Kennard-Stone [49] or the Duplex [50] algorithm, or by random selection [44].

First, during model building, the optimal model complexity A_{Opt} must be established. A compromise must be found between under-fitting and over-fitting. In under-fitting, the model complexity is too low ($A<A_{\mathrm{Opt}}$), leaving a part of the structure in the data unexplained. In over-fitting, the model complexity is too high ($A>A_{\mathrm{Opt}}$), including a part of the measurement noise in the model. Both under- and over-fitting may result in poor future model performance. With numerous and correlated **X**-variables there is a substantial risk for over-fitting, i.e., getting a well-fitting model with little or no predictive power. Cross-validation (CV) is a practical and reliable way to test this predictive ability in the training set. It has become the standard in PLS modelling [1].

Cross-validation is a resampling method for internal validation with the calibration set. This set is split into *M* subsets, often five to ten. Repeatedly, sub-models are developed with the reduced calibration set with one of the subsets left out, until each subset has been kept out once. This produces *M* sub-models. With each sub-model, predictions of the responses of the samples in the left out subset are estimated, and differences between the experimental and predicted responses calculated. When all sub-sets (m=1,2, …, M) have been left out in cross-validation, the root mean squared errors of cross-validation (RMSECV) is determined. Cross-validation is repeatedly conducted with increasing model complexities A (A= 1, 2, …), and a graph of RMSECV against the model complexities is made. The complexity corresponding to the minimum in this graph is considered as the optimal model complexity A_{Opt} [1,51].

Although the minimum RMSECV is a reasonable choice, it is based on a finite number of samples, and therefore, it is subject to error. Thus, using the number of factors corresponding to the minimum can lead to some over-fitting. Therefore, one can choose for a less complex and probably more robust model than that corresponding to the absolute minimum [3,52,53]. Using a model with less parameters, may result in less propagation of errors from the data into the parameter estimates, and so over-fitting will be minimized [54]. In case of a steady decrease in RMSECV, without a minimum, the complexity is chosen for which the decrease in predictive ability is below a given threshold [55,56].

The commonly applied leave-one-out cross validation has a strong tendency to over-fitting [1,57]. A segmented cross-validation procedure with more than one sample in the left-out segment (*n*-fold or leave-more-out cross-validation) is therefore preferred.

In PLS modelling for classification, the optimal model complexity should not be determined by CV with respect to RMSECV, because this is most often not optimal for classification purposes [58]. In this case, the optimal model complexity can be determined based on the percentage of correctly classified samples that have been left out [59,60].

4.3 ‘Large *K* - small *N*’ problem

Modern analytical methods produce a large number of variables *K*, while the number of samples *N* is often limited [61]. However, often most of the variables are uninformative because they are noisy, originate from the analytical background or from factors that are irrelevant to the problem at hand [10,62]. Additionally, when the number of variables is much larger than the number of samples (*K*>>*N*) it is possible that variables by chance correlate to the dependent property and over-fitting occurs. Predictions will be worsened by uninformative variables. Therefore, they should be removed. The ‘large *K* - small *N*’ problem can be solved by a search for a small set of informative variables to model the dependent property [63].

Both theoretical [61,63-66] and experimental evidence [3,11,67-71] exist that elimination of uninformative variables from the original data set improves the performance of PLS models. By elimination of uninformative variables, the risk of over-fitting is reduced and better predictions may be obtained. This may result in simpler models, which can help in the interpretation of the multivariate models. Elimination of uninformative variables can also be important for cost reduction in process control by reducing the number of sensors in filter-based instruments for industrial on-line or at-line purposes [3,10,12]. Finally, variable reduction can also be relevant for computational reasons [12]. It is now widely accepted that a well-performed variable selection can improve PLS models [13].

In practice, it is impossible to investigate all models based on all possible combinations of variables. For *K* variables, 2^K-1 models should be evaluated. For example, for 50 variables, $1.13\cdot 10^{15}$ combinations and models are possible. If it took 1 second per model, this would take $3.57\cdot 10^{7}$ years. Even the investigation of all models based on a specific number of variables may be impossible. For instance for the selection of a subset of *J* from *K* variables, $K!/\{J!(K-J)!\}$ models should be investigated [72]. For example, if one wants to build a model based on a selection of 10 out of 50 variables, $50!/\{10!\cdot(50-10)!\}=1.03\cdot 10^{10}$ possible combinations should be investigated, and this would take 326 years.

In multivariate data analysis in analytical chemistry, the number of variables is often much larger. Spectroscopic data may contain several hundreds to some ten thousands of variables [73-75]. Hence, mostly, it is impossible to test all combinations of variables. In variable selection methods the number of combinations is restricted by an appropriate algorithm. Usually a small subset is obtained from the original variables.

4.4 Classification of variable selection methods

Variable selection methods can be classified based on the use of individual variables or intervals, on the initial selection and on the kind of algorithm used for the selection. All variable selection methods start with an initial selection of variables followed by a further optimisation of the selection by an appropriate algorithm. The variable selection method can be based on selecting either individual variables or intervals of variables. For methods using individual variables, the selection can start with variables selected either randomly or based on given variable properties, such as PLS regression coefficients (see section 4.6).
The methods based on predictive properties can further be optimised either by deleting variables below a specified threshold, or after ranking on a given property, followed by an

iterative process consisting of variable elimination, remodelling and re-ranking of variables [16,76]. In penalised predictive property based methods, simultaneously a PLS model is built and variables are selected using a constraint for the regression coefficients [8,77].

In the selection methods starting with randomly selected variables, the selection is further optimised either iteratively in Iterative PLS [78], or by a genetic algorithm for PLS in GA-PLS [79-81].

Iterative PLS (IPLS) starts with a small number of randomly selected variables. Thereafter, iteratively, new variables are added to or already selected variables removed from the selection if that improves the model [78]. Iterative PLS can also be applied with intervals. In genetic algorithms, a start population is created consisting of a set of vectors each with randomly selected variables. The following optimisation of the variable selection is conducted by an algorithm that mimics the natural selection in biologic evolution.

In the methods based on intervals, spectra are subdivided into intervals of equal width and separate PLS models developed for each interval. Variable selection is optimised either by adjusting the interval width or combining intervals, the latter occasionally also combined with adjusting interval widths.

Table 4-1 provides an overview of the classification of variable selection methods, including the most important methods within the different classes and their references. The most widely used techniques are shown in bold italics.

Table 4-1 Classification of variable selection methods, with the most important methods and their references

Selection type	**Initial selection**	**Type of further optimisation of the selection**	**Examples of (types of) methods**	**References**
Individual variables	Predictive property based	Non-iterative optimisation	***Threshold PPRV-methods***	[3,16,69,87,88,90,93]
		Iterative optimisation	***Iterative PPRV-methods*** ***UVE*** MCUVE CARS	[3,11,16,76,88,97] [3,14,90,103-109] [43,71,82,99,103,110,111] [82,98-100]
		Penalised		[8,87,112,113]
	Random	Iterative optimisation	Iterative PLS ***GA-PLS***	[78,89,164] [16,68,79,80,81,114-117]
Intervals of variables	Individual intervals	Adjusting interval width	***iPLS***	[125,127-130]
		Combining intervals and/or adjusting interval width	Iterative PLS SiPLS FiPLS BiPLS GA-iPLS MWPLS MCSMWPLS, CSMWPLS, SCMWPLS	[89,164] [125,127,128] [129,130] [126,127,129,130] [126,130] [131-136] [132,134,137]

Most widely used types of methods shown in italics (see text); Abbreviations see text.

The characteristics of the most widely used types of methods, as well their advantages and drawbacks are described below. These types include (*i*) methods based on predictor-variable properties using a threshold, e.g. Threshold PPRV methods, (*ii*) iterative methods based on predictor-variable properties, e.g. Iterative PPRV methods, (*iii*) Uninformative Variable Elimination (UVE) methods, (*iv*) Penalised methods, (*v*) Genetic Algorithms for PLS (GA-PLS), and (*vi*) Interval PLS (iPLS). They are shown in Table 4-1 in bold italics.

At the end of this introducing chapter, these most widely used types of methods are compared for their advantages and drawbacks. The comparison is made to select the most promising type of variable selection method as a starting point for the development of new or improved methods to help mastering the data tsunami in chemometrics and bioinformatics.

4.5 Scope of variable selection

According to Andersen and Bro [12], variable selection should be considered as variable elimination where the clearly irrelevant variables are removed and the remaining variables containing potentially useful information are kept for further data analysis. Variable-selection methods are developed to find a good set of variables rather than the optimal set.

Essentially, a variable-selection procedure consists of two parts. First, a variable selection part in which variables are selected based on their influence on the model. This requires the choice of a search algorithm and an influence measure for the variables. Secondly, a model evaluation part to evaluate the performance of the PLS models built with the selected variables [12,68,82].

Mostly, a reasonable and statistically valid model can be made using all variables. The model validity is tested by appropriate cross- and test-set validation [12]. A reasonable and valid model gives a satisfying description of the relation between independent and dependent variables and has acceptable predictive properties. This model is not perfect, so it can be improved. Therefore, it is a good reference point for models built after variable selection. It is also reasonable to assume that model parameters, such as PLS regression coefficients or their significance, can be applied to find a reduced set with informative variables giving finally the best model. It must be stressed that the application of pre-processing techniques may affect the result of a variable-reduction method [3,12].

The initial model is improved during the variable-selection procedure, either in a forward or backward mode. During variable selection, properties of the models built with the remaining sets, such as predictive ability, will change. Therefore, these properties will often be evaluated in an iterative process.

Selection of the most correlated variables with the response *y* may not always result in the best performing models because variables that correct for interferents may be eliminated. Combinations of variables, which have low individual correlation coefficients with the response, may be better correlated when combined [82]. Variables that individually are rather useless, may provide well-performing models in combination with others [10].

In spectral data analysis, analytical chemists are not interested in the most correlated variables, but in combinations of variables found in chemically meaningful absorption bands or combinations of bands [82]. Additionally, information from a set of variables, combined in a multivariate model, makes it possible to determine the concentration of an analyte in the presence of interferents, provided that the signals of the interferents are not completely identical to that of the analyte [58,83].

4.5.1 Forward and backward modes

During the variable-selection procedure the model is improved, either in a forward or backward mode. The forward mode or forward selection starts building a model with the variable that results in the best prediction. Then, the variable is added which gives the best prediction in combination with the first. Thereafter, variables that give the best and improved predictions in combination with the already selected are added one by one. The forward mode may disregard combined effects of variables because it selects the variables sequentially [12]. The backward mode or backward elimination starts with the full original variable set, followed by eliminating one by one the variables that contribute least to the prediction. The backward mode is reasonably fast and has the advantage that it takes combined effects of variables into account [12]. In each mode, the variable-selection process is repeated until a stopping criterion, such as an optimal predictivity, is met.

4.5.2 Model evaluation

The predictive ability of a PLS model is assessed by internal and external validation. The model is built with the samples in the training set. During variable selection, PLS models are assessed by internal validation in the training set, using cross validation. The model developed with the finally selected variable set is assessed by external validation with a test set. The samples in the test set are independent from the samples in the training set. Mostly, the training and test sets are obtained by partitioning the original data set, for instance using the Kennard-Stone [49] or Duplex [50] algorithms, or by random selection [44]. The prediction error is evaluated either on the objects in the training set, on those in the left-out segments of cross validation, or those on the objects in the test set, as the root mean squared errors of calibration (RMSEC), cross validation (RMSECV) or prediction (RMSEP), respectively, or as the squared values of the correlation coefficient between estimated and experimental properties, for cross-validation and test objects, R_{CV}^2 and R_{Test}^2, respectively.

4.5.3 Chemical relevance of variable selection

The goal of variable reduction is to obtain models with small sets of variables showing improved or similar predictability. Variable selection can provide useful insight in which variables are informative and which are not. Therefore, variable reduction may help in the chemical interpretation of the PLS model. As an example, in near infrared (NIR) spectroscopy, organic molecules have specific absorption bands. Therefore, NIR spectra of samples, containing organic molecules, are influenced by these absorptions. In fact, the functional group effect is by far the most dominant of all effects in NIR [10]. It may be expected that informative variables are located in these absorption bands.

4.6 Predictor-variable property based methods

For PLS1, with one response variable y, many methods are based on so-called predictor-variable properties. Mostly, the properties are related to model parameters or model performance. They can indicate the influence of the variables on the PLS1 model. The higher the magnitude of the predictor-variable property, the more important the variable.

The predictor-variable properties can be divided into four groups.

1. Model dependent parameters:
   - magnitude of PLS regression coefficients [11,16,69,71,76,84-87].
   - significance of PLS regression coefficients assessed by the student t value, calculated from the ratio of the PLS regression coefficient and its standard deviation, and estimated by a resampling technique [3,16,71,87-90].
2. Combined model dependent parameters:
   - variable importance in the projection (VIP) score of a variable [11,76,86,87,91,92].
   - norm of the loading weights [86].
3. Parameters related to the predictive ability of the model,
   - selectivity ratio (SR) [19,20].
4. Model independent parameters,
   - correlation coefficient between predictor variables and the dependent variable [3,11,93].

As an example, predictor-variable properties are calculated for a full-spectrum PLS model, of a data set consisting of near infrared spectra of corn samples, with as response their moisture content, provided by Eigenvector Research (http://www.eigenvector.com/, accessed on March 21, 2014). All properties are calculated for centred data, with the exclusion of the correlation coefficient. In Fig. 4-1, the above-mentioned six predictor-variable properties are shown, as well as the water absorption bands, for the full-spectrum PLS model of this data set. In the NIR region, water has strong absorption bands between 1400 and 1450 nm and between 1900 and 1940 nm [94]. In the graphs of the predictor-variable properties related to the PLS model (Fig. 4-1A-E), positive peaks are seen inside both water bands. The peaks in the second water band (from 1900 to 1940 nm) are always higher than those in the first water band (from 1400 to 1450 nm). In Ref. [95] is described that the NIR absorbance in the second water band is often used for the quantitative analysis of water contents in dry food samples, such as corn. These positive property peaks in the water bands indicate that important variables result in high values of predictor-variable properties. Therefore, predictor-variable properties, which are related to the PLS model parameters or model performance, can be applied to select informative variables and/or to eliminate uninformative.

In Fig. 4-1F, correlation coefficients between the original x-variables or absorbances at a wavelength and the original moisture contents are given. These correlation coefficients are independent of the PLS model. All correlation coefficients are high and a minor peak is observed in the second water band. The correlation coefficient between predictor variables and the dependent variable is often used as predictor-variable property [3,11,93].

The methods based on predictor-variable properties start with building a model, mostly the PLS model developed for the original data set, for which one of the above-mentioned properties is calculated. Variables then are ranked in descending order of the considered property. This ranking reflects their importance for the PLS model. We call this Predictive-Property-Ranked Variables based methods, denoted as PPRV methods. These PPRV methods can be split into two sub-categories: non-iterative PPRV methods using a threshold and iterative PPRV methods. The characteristics of these methods are described below.

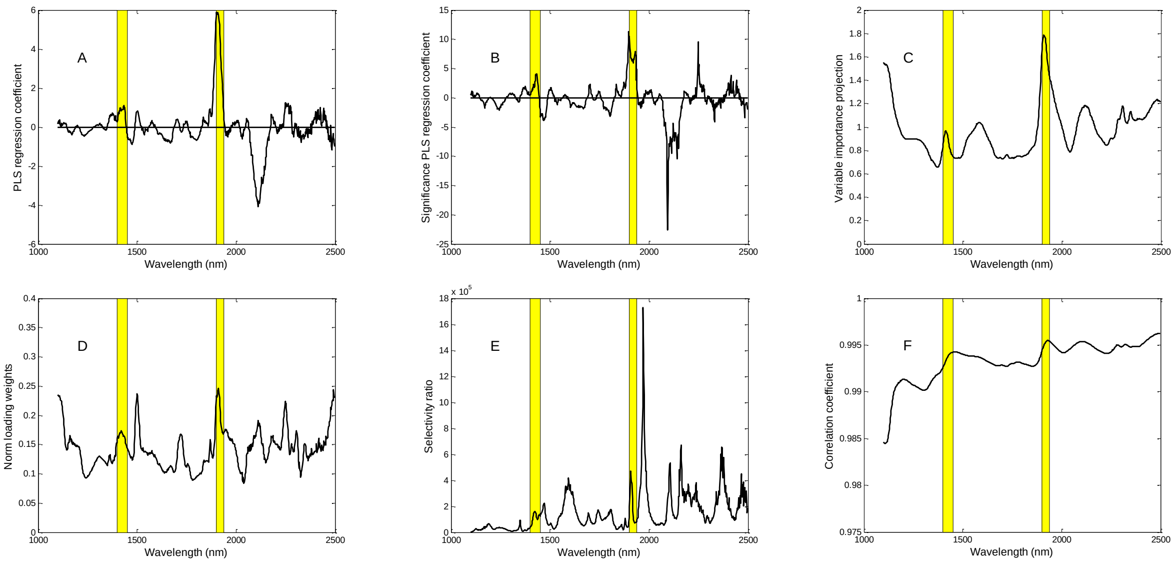


Fig. 4-1 Predictor-variable properties for data set corn, response moisture; (A) PLS regression coefficient; (B) Significance of PLS regression coefficients; (C) Variable importance in the projection; (D) Norm loading weights; (E) Selectivity ratio; (F) Correlation coefficient; the yellow columns represent the water absorption bands

4.6.1 Non-iterative PPRV methods

In the non-iterative PPRV methods using a threshold, variables with property values below a pre-defined threshold are considered uninformative and removed. Thereafter, for the reduced variable set, the final PLS model is calculated. These methods are fast and easy to compute [15]. However, a common disadvantage is that they neglect both the interactions of variables with the response and the interactions among variables [18].

The selection is highly affected by the chosen threshold and choosing a good threshold level may be a challenge [15]. The threshold is either determined arbitrarily [93], or through statistical assessment of the significance of the properties using bootstrap [91], jack knife [16,88,90,96] or Monte Carlo re-sampling methods [71,82]. The performance depend on the applied property [16,88]. These methods have been widely applied in analytical chemistry [16,20,87,88,90,93]. They are also the most widely used methods for biomarker discovery in metabolomics [22].

4.6.2 Iterative PPRV methods

In iterative PPRV methods, iteratively, the variable with the smallest value is eliminated and a new PLS model calculated. In the stepwise removal of variables, the predictive abilities of the PLS models are assessed, mostly by the RMSECV. The set of variables, resulting in the optimal model, is then selected [3,11].

These iterative PPRV methods are time consuming [15]. They are effective because their selective and predictive abilities are good, especially when using the PLS regression coefficients. They are robust, and avoid over-fitting and chance correlations. They are useful for different types of data sets. Their performance depend on the applied property [11].

Contrary to the methods using thresholds, they account both for interactions of variables with the response and for interactions among variables [18]. Methods based on predictor-variable properties have been widely applied in analytical chemistry [3,11,16,76,88,97]. They are also often used for biomarker selection [22].

Other iterative PPRV-methods are Uninformative Variable Elimination for PLS (UVE-PLS), including Monte-Carlo UVE (MCUVE), Covariance Procedures (CovProc) [11,65], Competitive Adaptive Reweighted Sampling (CARS) [82,98-100] and Covariance Selection (CovSel) [101]. They all use some predictor-variable property, but the algorithms for variable selection are different from that described above. UVE-PLS is a widely used method in chemometrics. The characteristics of UVE-PLS are described in section 4.7.

4.7 Uninformative Variable Elimination

Uninformative Variable Elimination for PLS (UVE-PLS) is based on the significance (or fitness) of PLS regression coefficients as predictor-variable property. UVE-PLS is introduced in Ref. [90]. It determines the fitness of each predictor variable k in the $\mathbf{X}$ matrix against those of L artificial random variables added to the data set. These added random variables have very small absolute values, of the order of about 10^{-10}, so that their influence on the regression coefficients of the predictors is negligible. For the optimal complexity A, the $K+L$ mean PLS regression coefficients $\bar{b}_k$ and their standard deviations $s(b_k)$ are calculated from vectors of regression coefficients, obtained by a resampling method, such as jack-knifing. The fitness c_k of each variable k is determined by the ratio of the mean regression coefficient and its standard deviation: $c_k=\bar{b}_k/s(b_k)$. A suitable cut-off value $|c_k|_{cut\text{-}off}$ is calculated from the L artificial variables, taking the maximum of their absolute c_k values. Predictor variables with $|c_k|$ below the cut-off value are classified as uninformative and eliminated. A new PLS model is built with the reduced set and cross-validated. The algorithm is repeated for complexities A-1, A-2, … until the predictive ability is not improved anymore.

An advantage of UVE-PLS is that it is user independent and therefore does not present any configuration problems [89]. A drawback is that in a replicated UVE-PLS, the number of eliminated variables is variable because of the variability in the added artificial random noise. Additionally, the number of retained variables by UVE-PLS is rather large [10,102]. It is better not to use UVE-PLS in Quantitative Structure–Activity Relationship (QSAR) modelling, because bad models are obtained [17]. In QSAR the $\mathbf{X}$ matrix does not contain uninformative noise variables, which are to be removed. However, the method has been widely applied in analytical chemistry [3,14,43,90,103-109].

Modifications of the UVE method are obtained by the use of other resampling techniques than jack-knifing. In Monte-Carlo UVE (MCUVE), a large number (typically 100) of subsets of training samples are selected randomly from the training set, and PLS sub-models generated. The fitness c_k of each variable is calculated from the corresponding regression coefficients of the sub-models. No random noise variables are added to the original data matrix. The method is introduced in [71]. Applications are described in [43,71,82,99,103,110,111].

4.8 Penalised methods

Penalised (or sparse) methods are based on PLS regression coefficients as predictor-variable property. They simultaneously build a regression model and perform wavelength selection by setting regression coefficients of uninformative variables to zero.
They are increasingly applied in chemometrics [77]. An early example of a penalised method is the Least Absolute Shrinkage and Selection Operator (LASSO) method. The method is introduced in [112]. In the LASSO method the sum of squared errors for least squares regression is minimized with the constraint that the sum of the absolute value of the regression coefficients, i.e. the L_1 norm, should be below a predefined threshold. Because of this constraint, also called the L_1 penalty, coefficients will be made zero. This can be regarded as a variable selection technique. The value of the threshold determines the degree of variable selection. A low threshold will make many coefficients zero and a lower number of variables will be retained [8,12,77,112]. Applications can be found in [8,87,113]. In [87] was found that the performance of LASSO was worse than that of a method using a threshold for VIP values. In [77] is stated that the LASSO does not perform as well as classical multivariate calibration methods in combination with other variable selection approaches.

Other penalised methods include Ridge Regression, Elastic Net, Sparse PLS, and Sparse Partial Least-Squares Discriminant Analysis (see section 4.12), Support Vector Regression, see Ref. [8,77]. Sparseness, with estimated parameter vectors containing many zero's, can lead to an improved prediction or classification performance compared to non-penalised methods. However, it depends on the data structure and on the sample size whether penalised methods give better results [8]. Penalised methods are still not as fast and efficient as classical multivariate methods [77].

4.9 Genetic algorithms

Genetic algorithms (GAs) are methods based on the principles of natural selection in biologic evolution. Species adapt over a high number of generations, because the fittest survive and spread their genetic material to following generations [79]. There are many variants of GAs. However, all have four fundamental steps in common:

1. creation of the original population
2. evaluation of the models
3. reproduction
4. mutations

These steps are discussed below. In different GAs, these steps are carried out in various ways.

1. Creation of the original population
A start population is created consisting of a number of vectors with randomly generated zeros and ones. The size of each vector is equal to the number of variables. The vector is called a *chromosome*. Each zero or one is a *gene*. A one indicates that the corresponding variable should be included in the model. The included variables form a subset of the original variables. For each chromosome, a PLS model is developed, called an *individual*. The number of chromosomes in the start population is the *population size*. It is mostly chosen in the range between 20 and 500 and remains constant during the calculations.

2. Evaluation of the models
The predictive ability of each individual is evaluated by the RMSECV, called the *fitness* of the individual.

3. Reproduction
A new generation of chromosomes is created in two sub steps. First, chromosomes from the former generation are copied with a probability related to its fitness. The best chromosomes have a higher probability to be copied than the worst. Secondly, the copied chromosomes are randomly paired and the pairs undergo a *crossover*. In a crossover, *offspring* is formed by interchanging randomly selected parts of the genes in pairs of chromosomes. Crossover is conducted with a high probability, so that almost all pairs undergo this operation.

4. Mutation
In this step some randomly selected genes are changed from a 1 to a 0 or vice versa with a very low probability, typically about 1%.

The steps 2 to 4 are repeated until a stop criterion is met, such as a predefined number of iterations, the attainment of a predefined response value, or after some percentage of the individuals in the population are using identical variable subsets.

More details about GAs can be found in [79-81]. Genetic algorithms for PLS (GA-PLS) have successfully been used for variable selection in analytical chemistry [16,68,79-81,114-117] and in bioinformatics [18]. They explore the space of all possible subsets fairly well in a rather long time [10]. However, GAs do have significant drawbacks. First, they tend to be slow. Secondly, they require a considerable level of expertise because numerous adjustable factors have to be set for the algorithm [10]. Thirdly, there is a large variability of solutions [16]. Fourthly, preferably, the number of variables should be kept below 200, to avoid a decrease in the performance of the algorithm [117,126]. For data sets with more than 200 variables the number of variables should be reduced before the application of a GA [80,126].

4.10 Interval PLS

Variable selection methods can be based on either individual variables or on intervals of variables. They make use of simple metrics and are readily available in commercial software [91]. Individual variable selection methods are widely used, both for continuous data in spectroscopy [10,11,15,80,82,84,90,91,118,119], and for non-continuous data, like those for Quantitative Structure-Activity Relationships (QSARs) [70,120-122], biomarker identification in GC-MS and LC-MS [59,123], and gene selection [119].

However, regarding the use of individual variable selection methods for spectral data, it is argued that the results are more difficult to interpret since the selected wavelengths are often distributed across the complete spectra instead of within a few confined intervals [91,124]. Therefore, it is recommended to select intervals of consecutive variables, instead of individual spectral variables [124]. In spectral data, adjacent variables may be highly correlated. With interval methods, the most informative wavelength bands are identified, which makes model interpretation easier [91].

Interval PLS (iPLS), introduced by Nørgaard et al. [125], is one of the more commonly used interval methods. In iPLS, the spectra are subdivided into intervals of equal width. Separate PLS models are developed for each interval, usually with a different number of PLS factors.

The prediction performance of these interval models and the full-spectrum model are compared, mostly based on the RMSECV, to determine the interval with the best predictive ability [10,125]. iPLS provides an overall picture of the data set and primarily locates the most relevant spectral regions. However, it does not take into account possible synergism between different spectral regions [126]. Applications can be found in [127,128].

The probability is very low to find the optimal set of variables with the best predictive ability in iPLS by the selection of only one interval. Therefore several extensions of iPLS are developed to further optimize variable selection by intervals.

In Synergy interval PLS (SiPLS) [125] the combination of intervals with the best predictive ability is searched for. First iPLS is conducted, and thereafter PLS models are developed for all possible combinations of two, three or four intervals. The combination of intervals with the lowest RMSECV is selected [10,125]. The computation time can be very long depending on the number of intervals and the selected number of intervals to combine [127]. Applications can be found in [127,128].

In Forward interval PLS (FiPLS), first iPLS is conducted, and the interval with the lowest RMSECV selected. Thereafter, forward selection is performed with intervals. Finally, the combination of intervals with the minimal RMSECV is selected [10,129,130].

In Backward interval PLS (BiPLS), first the data set is split into a given number of intervals, similar to iPLS, and PLS models are calculated with each interval left out in a sequence. The left out interval, resulting in the highest RMSECV for the included intervals, is deleted. Thereafter, backward selection is performed with intervals. Finally, the combination of included intervals with the minimal RMSECV is selected [10,126,127,129]. BiPLS can also pre-select variables which can be used as an input for GA-PLS [126].

In GA-iPLS a genetic algorithm is applied using intervals of variables instead of pure variables [10,130].

In Moving Window PLS (MWPLS) an *H* variables wide spectral window is constructed forming a $N \times H$ sub matrix of the calibration set. The spectral window is moved through the entire spectrum. For each window position, PLS models with varying complexities are developed for the corresponding sub matrices and the sums of squared residues (SSR's) or the RMSECV´s calculated. The SSR's or RMSECV´s are plotted as a function of the window position and the spectral regions with a minimal SSR [131-134] or RMSECV [135] over all windows are determined. MWPLS is introduced in Ref. [131]. Applications are found in [131-136].

In Changeable Size Moving Window Partial Least Squares (CSMWPLS), an optimized sub-region is searched in a selected informative region. In [132] a Modified Changeable Size Moving Window Partial Least Squares (MCSMWPLS) is proposed. In Searching Combination Moving Window Partial Least Squares (SCMWPLS), an optimized combination of informative regions based on CSMWPLS is searched. CSMWPLS and SCMWPLS are introduced in [137]. Applications can be found in [132,134,137].

In [14] is concluded that the effectivity of interval PLS methods for variable selection in near-infrared spectroscopy is low. In [127] is concluded that UVE performs better than iPLS, SiPLS and BiPLS.

4.11 Variable selection for PLS2

Variable selection for PLS models for multiple responses (PLS2) is complicated by the fact that each variable may have a different influence on the different responses. This can, at least partly, explain why for PLS1 numerous procedures for variable selection have been developed, see the reviews in [3,10,12,14-17] and the references therein, while only a few address those for PLS2 [101,138-140].

Like for PLS1, variable selection for PLS2 is often based on PLS model parameters. In [139] variables with the minimum PLS2 regression coefficient in the corresponding rows of the PLS2 regression coefficient matrix **B** are stepwise eliminated. In [138], variables with a cumulative absolute PLS2 regression coefficient in the corresponding rows of the **B** matrix are selected when above a threshold, which is set to the mean of these cumulative values for all variables. In [140] variable selection is based on the magnitude of absolute weights in the PLS2 weight vectors. In [101] variables are stepwise selected based on their global covariance with all responses, which are independent of the PLS2 model.

4.12 Variable selection for classification

Partial Least Squares Discriminant Analysis (PLS-DA) is the application of PLS for classification problems in which the response vector **y** codifies the class of each sample [141]. In the two-class case, usually the values of the dependent variable y are given 1 for one class and 0 or -1 for the other. In the case of more than two classes, for each class, dummy response variables are created, and a PLS2 algorithm applied [142]. The class label of an unknown sample is determined on the basis of the *y* value predicted by the PLS model. Ideally, the predicted *y* should be close to the coded class values. In practice, it is a real number and different approaches can be used to convert the predicted *y* into a class label [141].
PLS-DA is especially useful for high-dimensional data, where classical discrimination Methods, such as linear discriminant analysis (LDA) have numerical difficulties because of singularity issues [8]. Therefore, PLS-DA is not only used in chemometrics [8,100,140,141,143] but also in bioinformatics [7,19,20,98,111,140,142,144,145]. PLS-DA is one of the most frequently applied methods for classification problems in metabolomics [146].

Classification by PLS-DA can be improved by variable selection. Variable selection using predictor-variable properties, based on PLS regression coefficients is used in [21,139,143-145], on VIP in [100,111], on the selectivity ratio in [19,20,141], and on the largest absolute values of PLS weights in [140]. UVE is used in [98,111], CARS in [98], and genetic algorithms in [139,147,148].

4.13 Variable selection for QSAR and QSRR modelling

Quantitative Structure-Activity Relationships (QSAR) are mathematical models for a series of chemical compounds relating structural, physical, and/or chemical properties (descriptors) to one of their biological activities. A statistically validated QSAR model is capable of predicting the biological activity of a new compound within the same series, as an alternative to time-consuming and labour-intensive processes of chemical synthesis and biological evaluation.

QSAR models can help in the design of new compounds. Therefore, they have become useful tools in the pharmaceutical industry [17].

Similar to QSARs, QSRRs are statistically derived relationships between chromatographic parameters and descriptors related to the molecular structure of the analytes. In QSRRs these descriptors are used to model the molecular interaction of analytes with a given stationary phase and eluent of a chromatographic system. Using a validated QSRR model, the retention of new analytes can be predicted for the chromatographic system considered, see Refs. [149-153].

Generally, QSAR and QSRR models with a large number of variables or descriptors are not desirable for the following reasons. First, only a few descriptors have an important influence on a biological activity or chromatographic property, respectively. Second, the interpretation of a model containing a large number of descriptors is difficult. Thus to build simple QSAR or QSRR models, a variable selection technique is needed [17,149,154].

Variable selection for QSAR is based on PLS regression coefficients in [11,155], on VIP in [156], on GA-PLS in [157-159], and on the correlation between predictor-variables and the response in [160]. In [159] the Replacement Method (RM) and Forward Stepwise Regression Method [26] are used. In the RM a chosen variable is replaced by another one to minimize the total standard deviation [159].
In [161] an evolutionary Museum algorithm has been used. This algorithm starts from a random model containing any combination of variables of the data set. In the next steps one or a very few variables are added to or eliminated from this model. Any model with increased fitness defined by a certain criterion, e.g. the standard deviation *s* or the Fischer significance value F of the regression equation, is taken as a new breeding organism which is further mutated by variable additions or eliminations.

In the review of Goodarzi et al. [17] about variable selection for QSAR, is concluded that often models are obtained with a quality similar to that with all variables. Only the RM systematically selected few variables. GAs and Backward Elimination PLS selected much larger numbers of variables than RM. CARS, CovSel, UVE and predictive property based methods using VIP generally led to bad QSAR models and should therefore not be used in QSAR modelling.

For QSRR, often classical models are used with small numbers (1-5) of descriptors [149-151], for which multiple linear regression is used for model building [149]. PLS is used for later introduced descriptor sets containing large numbers of theoretical molecular descriptors generated by calculation chemistry. For these sets, variable selection is needed.
Variable selection for QSRRs is conducted with UVE-PLS in [105,108,162], and with GA-PLS in [162,163].

4.14 Summary

The development of new variable selection methods may help to reveal the informative signals in the huge data sets generated by modern sophisticated instrumental analysis methods. It can help the chemometricians to master the data tsunami.

One of the goals of this research is to develop new or improved variable selection methods for PLS modelling, with a high specificity and which must be widely applicable both in chemometrics and in new emerging fields such as metabolomics. Therefore, they must be suited both for continuous and non-continuous data. Additionally, they must be applicable for either PLS1 or PLS2.

In this introduction, an overview of variable reduction methods for PLS is given. The characteristics of six widely applied types of methods are described. Their advantages and drawbacks are summarized in Table 4-2.

Table 4-2 Comparison of widely applied types of variable selection methods for PLS

Method	Advantages	Drawbacks
Threshold PPRV methods	fast; easy to compute	ignore interactions of variables with the response and interactions between variables; selection is highly affected by the chosen threshold; performance depend on the applied property
Iterative PPRV methods	good selective and predictive ability; robust; avoid over-fitting and chance correlations; account for interactions of variables with the response and for interactions between variables; useful for different types of data sets	time consuming; performance depend on the applied property
UVE-PLS	user independent; no configuration problems	large variability of solutions; large number of retained variables
Penalised methods	simultaneously build a regression model and perform wavelength selection	not as fast and efficient as traditional multivariate methods
GA-PLS	explores the variable space fairly well	tend to be slow; require a considerable level of expertise; large variability of solutions; number of variables < 200
Interval PLS methods	only suited for continuous data in spectroscopy; most informative wavelength bands in spectral data are identified;	not suited for non-continuous data; low effectivity

Interval methods are not suited for our purposes because they are only applicable for continuous data and not for non-continuous data. GA-PLS is not suited, because these algorithms work only well with less than 200 variables. Therefore, for large data sets, a pre-selection of variables will be needed for GA-PLS. UVE-PLS is not suited because of its low selectivity. Often large numbers of variables will be retained. Additionally, both GA-PLS and UVE-PLS have a large variability of solutions, which make them also not suited for biomarker discovery because it requires the selection of simple and stable combinations of metabolites [22]. Penalised methods are not suited because they are still not as fast and efficient as traditional multivariate methods.

The threshold-PPRV methods have the disadvantage that they ignore both the interactions of variables with the response and interactions among variables. That is not the case in iterative PPRV methods. Given the advantages for iterative PPRV methods mentioned in Table 4-2, this type of methods seem most promising as starting point for the development of new or improved variable selection methods.

4.15 Variable selection in this thesis

Using the information in the preceding sections, the following requirements for the development of new variable selection methods for PLS are defined.

1. The new methods must have the characteristics of iterative PPRV methods.
2. They must work in the backward mode because of the advantage that it accounts for combined effects of variables.
3. The new methods must first be developed based on one predictor-variable property and tested for PLS1, and the best selected.
4. For the best new method for PLS1, the selective and predictive performance of different kinds of predictor-variable properties must be investigated, and the best property selected.
5. The best method for PLS1 will be adapted to PLS2.
6. The new methods will be developed and tested with spectral and simulated data, because for these data no alignment procedures have to be applied.

Following this strategy, the results of the research done in this PhD project for the development of new variable selection methods for PLS is presented in the following chapters.

In chapter 5, a study is presented about the development of three new stepwise variable selection methods for PLS modelling with one response (PLS1), with a possibility to decrease the PLS model complexity during the variable reduction process. These methods are based on variables ranked on the absolute values of the PLS1 regression coefficients as predictor-variable property. The selective and predictive performances of these methods are compared with two existing methods as reference. The results of this study form the basis for the studies presented in chapters 6 and 7.

In chapter 6, the utility and effectiveness of six individual and nine combined predictor-variable properties are investigated and compared, when using the FCAM method resulting from the study in chapter 5. The selective and predictive performances of the models resulting from the use of these properties are statistically compared using the one-tailed Wilcoxon signed rank test.

In chapter 7, a study is presented about the development of a new variable selection method for multiple-response partial-least-squares (PLS2) modelling, using an adapted FCAM method for PLS2, FCAM-PLS2. The utility and effectiveness of four new predictor-variable properties, derived from the multiple response PLS2 regression coefficients, are investigated.

References

[1] S. Wold, M. Sjöström, L. Eriksson, Chemom. Intell. Lab. Syst. 58 (2001) 109.

[2] H. Martens, T. Næs, Multivariate Calibration, (2nd edn), Wiley, NewYork, 1993.

[3] M. Forina, S. Lanteri, M.C. Cerrato Oliveros, C. Pizarro Millan, Anal. Bioanal. Chem. 380 (2004) 397.

[4] P. Geladi, B.R. Kowalski, Anal. Chim. Acta 185 (1986) 1.

[5] M. Forina, S. Lanteri, M. Casale, J. Chromatogr. A. 1158 (2007) 61.

[6] M. Barker, W. Rayens, J. Chemometr. 17 (2003) 166.

[7] M. Bylesjö, M. Rantalainen, O. Cloarec, J.K. Nicholson, E. Holmes, J. Trygg, J. Chemometr. 20 (2006) 341.

[8] P. Filzmoser, M. Gschwandtner, V. Todorov, J. Chemometr. 26 (2012) 42.

[9] A. Smolinska, L. Blanchet, L.M.C. Buydens, S.S. Wijmenga, Anal. Chim. Acta 750 (2012) 82.

[10] Z. Xiaobo, Z. Jiewen, M.J.W. Povey, M. Holmes, M. Hanpin, Anal. Chim. Acta 667 (2010) 14.

[11] R.F. Teófilo, J.P.A. Martins, M.M.C. Ferreira, J. Chemometr. 23 (2009) 32.

[12] C. M. Andersen, R. Bro, J. Chemometr. 24 (2010) 728.

[13] N. Boaz, R.C. Ronald, J. Chemometr. 19 (2005) 107.

[14] R.M. Balabin, S.V. Smirnov, Anal. Chim. Acta 692 (2011) 63.

[15] T. Mehmood, K.H. Liland, L. Snipen, S. Sæbø, Chemom. Intell. Lab. Syst. 118 (2012) 62.

[16] J.P. Gauchi, P. Chagnon, Chemom. Intell. Lab. Syst. 58 (2001) 171.

[17] M. Goodarzi, S. Funar-Timofei, Y. Vander Heyden, Trends Anal. Chem. 42 (2013) 49.

[18] Y. Saeys, I. Inza, P. Larrañaga, Bioinformatics 23 (2007) 2507.

[19] T. Rajalahti, R. Arneberg, A.C. Kroksveen, M. Berle, K.M. Myhr, O.M. Kvalheim, Anal. Chem. 81 (2009) 2581.

[20] T. Rajalahti, R. Arneberg, F.S. Berven, K.M. Myhr, R.J. Ulvik, O.M. Kvalheim, Chemom. Intell. Lab. Syst. 95 (2009) 35.

[21] S. Bijlsma, I. Bobeldijk, E.R. Verheij, R. Ramaker, S. Kochhar, I.A. Macdonald, B. van Ommen, A.K. Smilde, Anal. Chem. 78 (2006) 567.

[22] J. Xia, D.I. Broadhurst, M. Wilson, D.S. Wishart, Metabolomics 9 (2013) 280.

[23] J. Boccard, S. Rudaz, J. Chemometr. 28 (2014) 1.

[24] L. Buydens, The Analytical Scientist, 1 (2013) 24.

[25] B. Lavine, J. Workman, Anal. Chem. 85 (2013) 705.

[26] N.R. Draper, H. Smith, Applied Regression Analysis, Second edition, John Wiley and Sons, New York, 1981.

[27] M.C.U. Araújo, T.C.B. Saldanha, R.K.H. Galvão, T. Yoneyama, H.C. Chame, V. Visani, Chemom. Intell. Lab. Syst. 57 (2001) 65.

[28] O.E. de Noord, Chemom. Intell. Lab. Syst. 23 (1994) 65.

[29] T. Rajalahti, O.M. Kvalheim, Int. J. Pharm. 417 (2011) 280.

[30] R. Bro, A.K. Smilde, J. Chemometr. 17 (2003) 16.

[31] S. Dunkerley, J. Crosby, R.G. Brereton, K.D. Zissis, R.E.A. Escott, Analyst, 123 (1998) 2021.

[32] L. Stordrange, F.O. Libnau, D. Malthe-Sørenssen, O.M. Kvalheim, J. Chemometr. 16 (2002) 529.

[33] P. Chalus, Y. Roggo, S. Walter, M. Ulmschneider, Talanta 66 (2005) 1294.

[34] M. Zeaiter, J.M. Roger, V. Bellon-Maurel, Trends Anal. Chem. 24 (2005) 437.

[35] Å. Rinnan, F. van den Berg, S. Balling Engelsen, Trends Anal. Chem. 28 (2009) 1201.

[36] A. Savitzky, M.J. E. Golay, Anal. Chem. 36 (1964) 1627.

[37] J. Steinier, Y. Termonia, J. Deltour, Anal. Chem. 44 (1972) 1906.
[38] H.H. Madden, Anal. Chem. 50 (1978) 1383.
[39] B.G.M. Vandeginste, D.L. Massart, L.M.C. Buydens, S. de Jong, P.J. Lewi, J. Smeyers-Verbeke, Handbook of Chemometrics and Qualimetrics, Part B, Elsevier, Amsterdam, 1998.
[40] S. Wold, H. Antti, F. Lindgren, J. Öhman, Chemom. Intell. Lab. Syst. 44 (1998) 175.
[41] T. Fearn, Chemom. Intell. Lab. Syst. 50 (2000) 47.
[42] J.A. Westerhuis, S. de Jong, A.K. Smilde, Chemom. Intell. Lab. Syst. 56 (2001) 13.
[43] M. Daszykowski, M.S. Wrobel, H. Czarnik-Matusewicz, B. Walczak, Analyst, 133 (2008) 1523.
[44] H.A. Martens, P. Dardenne, Chemom. Intell. Lab. Syst. 44 (1998) 99.
[45] L. Petersen, P. Minkkinen, K. H. Esbensen, Chemom. Intell. Lab. Syst. 77 (2005) 261.
[46] S. Wold, A. Ruhe, H. Wold, W.J. Dunn, SIAM J. Sci. Stat. Comput. 5 (1984) 735.
[47] N.M. Faber, J. Ferré, J. Chemometr. 22 (2008) 101.
[48] S. de Jong, Chemom. Intell. Lab. Syst. 18 (1993) 251.
[49] R.W. Kennard, L.A. Stone, Technometrics 11 (1969) 137.
[50] R.D. Snee, Technometrics 19 (1977) 415.
[51] E. Anderssen, K. Dyrstad, F. Westad, H. Martens, Chemom. Intell. Lab. Syst. 84 (2006) 69.
[52] L. Kooistra, R. Wehrens, R.S.E.W. Leuven, L.M.C. Buydens, Anal. Chim. Acta 446 (2001) 97.
[53] D.M. Haaland, E.V. Thomas, Anal. Chem. 60 (1988) 1193.
[54] M.B Seasholtz, B. Kowalski, Anal. Chim. Acta, 277 (1993) 165.
[55] B. Li, J. Morris, E.B. Martin, Chemom. Intell. Lab. Syst. 64 (2002) 79.
[56] S. Wold, Technometrics 24 (1978) 397.
[57] K. Baumann, Trends Anal. Chem. 22 (2003) 395.
[58] K. Kjeldahl, R, Bro, J. Chemometr. 24 (2010) 558.
[59] K. Wongravee, N. Heinrich, M. Holmboe, M. L. Schaefer, R.R. Reed, J. Trevejo, R.G. Brereton, Anal. Chem. 81 (2009) 5204.
[60] R.G. Brereton, Trends Anal. Chem. 25 (2006) 1103.
[61] A. Höskuldsson, J. Chemometr. 22 (2008) 150.
[62] M. Shariati-Rad, M. Hasani, J. Chemometr. 24 (2010) 45.
[63] B. Nadler, R.R. Coifman, J. Chemometr. 19 (2005) 107.
[64] C.H. Spiegelman, M.J. McShane, M.J. Goetz, M. Motamedi, Q.L. Yue, G.L. Coté, Anal. Chem. 70 (1998) 35.
[65] S.P. Reinikainen, A. Höskuldsson, J. Chemometr. 17 (2003) 130.
[66] L. Xu, I. Schechter, Anal. Chem. 68 (1996) 2392.
[67] J.A. Hageman, M. Streppel, R. Wehrens, L.M.C. Buydens, J. Chemometr. 17 (2003) 427.
[68] A.S. Bangalore, R.E. Shaffer, G.W. Small, M.A. Amold, Anal. Chem. 68 (1996) 4200.
[69] A. Garrido Frenich, D. Jouan-Rimbaud, D.L. Massart, S. Kuttatharmmakul, M. Martinez Galera, J.L. Martinez Vidal, Analyst 120 (1995) 2787.
[70] H.J. Kubinyi, J. Chemometr. 10 (1996) 119.
[71] W. Cai, Y. Li, X. Shao, Chemom. Intell. Lab. Syst. 90 (2008) 188.
[72] W.J. Krzanowski, Principles of Multivariate Analysis – A User's Perspective, Oxford University Press, 1988.
[73] K. Janné, J. Pettersen, N.O. Lindberg, T. Lundstedt, J. Chemometr. 14 (2001) 203.
[74] A. Smolinska, L. Blanchet, L.M.C. Buydens, S.S. Wijmenga, Anal. Chim. Acta 750 (2012) 82.

[75] H. Idborg, L. Zamani, P.O Edlund, I. Schuppe-Koistinen, S.P. Jacobsson, Journal of Chromatography B, 828 (2005) 14–20.
[76] A. Lazraq, R. Cléroux, J.P. Gauchi, Chemom. Intell. Lab. Syst. 66 (2003) 117.
[77] E. Andries, J. Chemometr. 27 (2013) 50.
[78] S.D. Osborne, R.B. Jordan, R. Künnemeyer, Analyst 122 (1997) 1531.
[79] R. Leardi, J. Chromatogr. A 1158 (2007) 226.
[80] R. Leardi, A.L. Gonzalez, Chemom. Intell. Lab. Syst. 41 (1998) 195.
[81] R. Leardi, R. Boggia, M. Terrile, J. Chemometr. 6 (1992) 267.
[82] H.D. Li, Y. Liang, Q. Xu, D. Cao, Anal. Chim. Acta 648 (2009) 77.
[83] R. Bro, Anal. Chim. Acta 500 (2003) 185.
[84] H. Xu, Z. Liu, W. Cai, X. Shao, Chemom. Intell. Lab. Syst. 97 (2009) 189.
[85] S.A. Dodds, W.P. Heath, Chemom. Intell. Lab. Syst. 76 (2005) 37.
[86] M.J. Anzanello, S.L. Albin, W.A. Chaovalitwongse, Chemom. Intell. Lab. Syst. 97 (2009) 111.
[87] I.G. Chong, C.H. Jun, Chemom. Intell. Lab. Syst. 78 (2005) 103.
[88] F. Westad, H. Martens, J. Near Infrared Spectrosc. 8 (2000) 117.
[89] C. Abrahamsson, J. Johansson, A. Sparén, F. Lindgren, Chemom. Intell. Lab. Syst. 69 (2003) 3.
[90] V. Centner, D.L. Massart, O.E. de Noord, S. de Jong, B.G.M. Vandeginste, C. Sterna, Anal. Chem. 68 (1996) 3851.
[91] R. Gosselin, D. Rodrigue, C. Duchesne, Chemom. Intell. Lab. Syst. 100 (2010) 12.
[92] S. Wold, E. Johansson, M. Cocchi, 3D QSAR in Drug Design; Theory, Methods, and Applications, ESCOM, Leiden, Holland, 1993.
[93] A. Höskuldsson, Chemom. Intell. Lab. Syst. 55 (2001) 23.
[94] J. Luypaert, D.L. Massart, Y. Vander Heyden, Talanta 72 (2007) 865.
[95] H. Büning-Pfaue, Food Chemistry 82 (2003) 107.
[96] J. Moros, J. Kuligowski, G. Quintás, S. Garrigues, M. de la Guardia, Anal. Chim. Acta 630 (2008) 150.
[97] M. Forina, C. Casolino, C.P. Millán, J. Chemometr. 13 (1999) 165.
[98] H.D. Li, Y.Z. Liang, Q.S. Xu, D.S. Cao, J. Chemometr. 24 (2010) 418.
[99] K. Zheng, Q. Li, J. Wang, J. Geng, P. Cao, T. Sui, X. Wang, Y. Du, Chemom. Intell. Lab. Syst. 112 (2012) 48.
[100] W. Fan, H.D. Li, Y. Shan, H. Lv, H. Zhang, Y. Liang, Anal. Methods, 3 (2011) 1872.
[101] J.M. Roger, B. Palagos, D. Bertrand, E. Fernandez-Ahumada, Chemometr. Intell. Lab. Syst. 106 (2011) 216.
[102] S. Ye, D.Wang, S. Min, Chemom. Intell. Lab. Syst. 91 (2008) 194.
[103] W. Cai, Y. Li, X. Shao, Chemom. Intell. Lab. Syst. 90 (2008) 188.
[104] C. Abrahamsson, J. Johansson, A. Sparén, F. Lindgren, Chemom. Intell. Lab. Syst. 69 (2003) 3.
[105] R. Put, M. Daszykowski, Baczek, Y. Vander Heyden, J. Proteome Res. 5 (2006) 1618.
[106] R. Put, Y. Vander Heyden, Proteomics 7 (2007) 1664.
[107] A.M. van Nederkassel, M. Daszykowski, D.L. Massart, Y. Vander Heyden, J. Chromatogr. A 1096 (2005) 177.
[108] T. Hancock, R. Put, D. Coomans, Y. Vander Heyden, Y. Everingham, Chemom. Intell. Lab. Syst. 76 (2005) 185.
[109] H. Swierenga, F. Wülfert, O.E. de Noord, A.P. de Weijer, A.K. Smilde, L.M.C. Buydens, Anal. Chim. Acta 411 (2000) 121.
[110] Q.J. Han, H.L. Wu, C.B. Cai, L. Xu, R.Q. Yu, Anal. Chim. Acta 612 (2008) 121.
[111] X.M. Sun, X.P. Yu, Y. Liu, L. Xu, D.L. Di, Chemom. Intell. Lab. Syst. 115 (2012) 37.
[112] R. Tibshirani, J. R. Stat. Soc. Series B, 58 (1996) 267.

[113] H. Öjelund, H. Madsen, P. Thyregod, J. Chemometr. 15 (2001) 497.
[114] C.B. Lucasius, G. Kateman, Chemom. Intell. Lab. Syst. 19 (1993) 1.
[115] C.B. Lucasius, G. Kateman, Chemom. Intell. Lab. Syst. 25 (1994) 99.
[116] R. Wehrens, L.M.C. Buydens, Trends Anal. Chem. 17 (1998) 193.
[117] R. Leardi, M.B. Seasholtz, R.J. Pell, Anal. Chim. Acta 461 (2002) 189.
[118] J. Ghasemi, A. Niazi, R. Leardi, Talanta 59 (2003) 311.
[119] K.H. Liland, M. Høy, H. Martens, S. Sæbø, Chemom. Intell. Lab. Syst. 122 (2013) 103.
[120] A. Yasri, D. Hartsough, J. Chem. Inf. Comput. Sci. 41 (2001) 1218.
[121] A.M. Helguera, P.R. Duchowicz, M.A.C. Pérez, E.A. Castro, M.N.D.S. Cordeiro, M.P. González, Chemometr. Intell. Lab. 81 (2006) 180.
[122] P.P. Roy, K. Roy, QSAR Comb. Sci. 27 (2008) 302.
[123] M. Daszykowski, W. Wu, A.W. Nicholls, R.J. Ball, T. Czekaj, B. Walczak, J. Chemometr. 21 (2007) 292.
[124] F. Rossi, D. Francois, V. Wertz, M. Meurens, M. Verleysen, Chemom. Intell. Lab. Syst. 86 (2007) 208.
[125] A.S.L. Nørgaard, J. Wagner, J.P. Nielsen, L. Munck, S.B. Engelsen, Applied Spectroscopy, 54 (2000) 413.
[126] R. Leardi, L. Nørgaard, J. Chemometr. 18 (2004) 486.
[127] D. Wu, Y. He, P. Nie, F. Cao, Y. Bao, Anal. Chim. Acta, 659 (2010) 229.
[128] Q. Chen, J. Zhao, M. Liu, J. Cai, J. Liu, J. Pharm. Biomed. Anal. 46 (2008) 568.
[129] Z. Xiaobo, Z. Jiewen, L. Yanxiao, Vibrational Spectroscopy 44 (2007) 220.
[130] Z. Xiaobo, Z. Jiewen, H. Xingyi, L. Yanxiao, Chemom. Intell. Lab. Syst. 87 (2007) 43.
[131] J.H. Jiang, R.J. Berry, H.W. Siesler, Y. Ozaki, Anal. Chem. 74 (2002) 3555.
[132] S. Kasemsumran, Y.P. Du, K. Maruo, Y. Ozaki, Chemom. Intell. Lab. Syst. 82 (2006) 97.
[133] S. Kasemsumran, Y.P. Du, K. Murayama, M. Huehne, Y. Ozaki, Analyst 128 (2003) 1471.
[134] S. Kasemsumran, Y.P. Du, K. Murayama, M. Huehne, Y. Ozaki, Anal. Chim. Acta 512 (2004) 223.
[135] H. Chen, T. Pan, J. Chen, Q. Lu, Chemom. Intell. Lab. Syst. 107 (2011) 139.
[136] B. Hemmateenejad, M. Akhond, F. Samari, Spectrochimica Acta Part A 67 (2007) 958.
[137] Y.P. Du, Y.Z. Liang, J.H. Jiang, R.J. Berry, Y. Ozaki, Anal. Chim. Acta 501 (2004) 183.
[138] M. De Luca, F. Oliverio, G. Ioele, G. Ragno, Chemometr. Intell. Lab. Syst. 96 (2009) 14.
[139] Z. Ramadan, X.H. Song, P.K. Hopke, M.J. Johnson, K.M. Scow, Anal. Chim. Acta, 446 (2001) 233.
[140] B.K. Alsberg, D.B. Kell, R. Goodacre, Anal. Chem. 70 (1998) 4126.
[141] M.P. Gómez-Carracedo, J. Ferré, J. M. Andrade, R. Fernández-Varela, R. Boqué, Anal. Bioanal. Chem. 403 (2012) 2027.
[142] J.A. Westerhuis, H.C.J. Hoefsloot, S. Smit, D.J. Vis, A.K. Smilde, E.J.J. van Velzen, J.P.M. van Duijnhoven, F.A. van Dorsten, Metabolomics 4 (2008) 81.
[143] F. Liu, Y. He, L. Wang, Anal. Chim. Acta 615 (2008) 10.
[144] R. Rousseau, B. Govaerts, M. Verleysen, B. Boulanger, Chemom. Intell. Lab. Syst. 91 (2008) 54.
[145] K. Wongravee, N. Heinrich, M. Holmboe, M.L. Schaefer, R.R. Reed, J. Trevejo, R.G. Brereton, Anal. Chem. 81 (2009), 5204.
[146] J. van der Greef, A.K. Smilde, J. Chemometr. 19 (2005) 376.
[147] Z. Ramadan, D. Jacobs, M. Grigorov, S. Kochhar, Talanta 68 (2006) 1683.

[148] S. Duraipandian, W. Zheng, J. Ng, J.J. H. Low, A. Ilancheran, Z. Huang, Analyst, 136 (2011) 4328.
[149] R. Put, Y. Vander Heyden, Anal. Chim. Acta 602 (2007) 164.
[150] R. Kaliszan, Chem. Rev. 107 (2007) 3212.
[151] K. Héberger, J. Chromatogr. A 1158 (2007) 273.
[152] C.F. Poole, S.K. Poole, J. Chromatogr. A 965 (2002) 263.
[153] M. Vitha, P.W. Carr, J. Chromatogr. A 1126 (2006) 143.
[154] M. Goodarzi, R. Jensen, Y. Vander Heyden, J. Chromatogr. B, 910 (2012) 84.
[155] P.P. Roy, K. Roy, QSAR Comb. Sci. 27 (2008) 302.
[156] S. Wold, E. Johansson and M. Cocchi, in 30 QSAR in Drug Design. Theory, Methods and Applications, ed. by H. Kubinyi, pp. 523-550, ESCOM, Leiden (1993).
[157] K. Hasegawa, Y. Miyashita, K. Funatsu, J. Chem. Inf. Comput. Sci. 37 (1997) 306.
[158] B. Hemmateenejad, R. Miri, M. Akhond, M. Shamsipur, Chemom. Intell. Lab. Syst. 64 (2002) 91.
[159] A.H. Morales, P.R. Duchowicz, M.Á.C. Pérez, E.A. Castro, M.N.D.S. Cordeiro, M.P. González, Chemom. Intell. Lab. Syst. 81 (2006) 180.
[160] M. Shamsipur, B. Hemmateenejad, M. Akhond, H. Sharghi, Talanta 54 (2001) 1113.
[161] H. Kubinyi, J. Chemometr. 10 (1996) 119.
[162] K. Bodzioch, A. Durand, R. Kaliszan, T.Baczek, Y. Vander Heyden, Talanta 81 (2010) 1711.
[163] F. Tian, L. Yang, F. Lv, P. Zhou, J. Sep. Sci. 2009, 32, 2159.
[164] S.D. Osborne, R.B. Jordan, R. Künnemeyer, Analyst, 122 (1997) 1531.

5 Improved Variable Reduction in partial least squares modelling based on Predictive-Property-Ranked Variables and adaptation of partial least squares complexity[2]

5.1 Abstract

The calibration performance of partial least squares for one response variable (PLS1) can be improved by elimination of uninformative variables. Many methods are based on so-called predictive variable properties, which are functions of various PLS-model parameters, and which may change during the variable-reduction process. In these methods variable reduction is made on the variables ranked in descending order for a given variable property. The methods start with full spectrum modelling. Iteratively, until a specified number of remaining variables is reached, the variable with the smallest property value is eliminated; a new PLS model is calculated, followed by a renewed ranking of the variables. The Stepwise Variable Reduction methods using Predictive-Property-Ranked Variables are denoted as SVR-PPRV. In the existing SVR-PPRV methods the PLS model complexity is kept constant during the variable-reduction process. In this study, three new SVR-PPRV methods are proposed, in which a possibility for decreasing the PLS model complexity during the variable-reduction process is built in.
Therefore we denote our methods as PPRVR-CAM methods (Predictive-Property-Ranked Variable Reduction with Complexity Adapted Models). The selective and predictive abilities of the new methods are investigated and tested, using the absolute PLS regression coefficients as predictive property. They were compared with two modifications of existing SVR-PPRV methods (with constant PLS model complexity) and with two reference methods: uninformative variable elimination followed by either a genetic algorithm for PLS (UVE-GA-PLS) or an interval PLS (UVE-iPLS). The performance of the methods is investigated in conjunction with two data sets from near-infrared sources (NIR) and one simulated set. The selective and predictive performances of the variable reduction methods are compared statistically using the Wilcoxon signed rank test.

The three newly developed PPRVR-CAM methods were able to retain significantly smaller numbers of informative variables than the existing SVR-PPRV, UVE-GA-PLS and UVE-iPLS methods without loss of prediction ability. Contrary to UVE-GA-PLS and UVE-iPLS, there is no variability in the number of retained variables in each PRV(R) method. Renewed variable ranking, after deletion of a variable, followed by remodelling, combined with the possibility to decrease the PLS model complexity, is beneficial. A preferred PPRVR-CAM method is proposed.

Keywords: Variable reduction, PLS1, PPRVR-CAM, UVE-GA-PLS, UVE-iPLS, Wilcoxon signed rank test

[2] **Jan P.M. Andries**, Yvan Vander Heyden, Lutgarde M.C. Buydens
Published in Analytica Chimica Acta 705 (2011) 292-305

5.2 Introduction

Multivariate regression techniques are widely used in analytical chemistry for the extraction of chemical information about analytes [1,2,3]. Using multivariate regression models chemical quantities can frequently be estimated with reasonable accuracy and with minimum data treatment [3]. Partial least squares (PLS) regression is a commonly used multivariate technique, which is considered able to deal with a large number of noisy and correlated variables, and with small numbers of samples. It is a versatile method, used for both qualitative and quantitative analysis, in many different application fields, such as food chemistry, pharmaceutical analysis, agriculture, environment, and industrial and clinical chemistry [4].

Both theoretical [5-9] and experimental evidence [3,10-15] exist that elimination of noisy and uninformative variables from the original data set can improve the performance of PLS calibration. In addition, elimination of uninformative variables can be important for cost reduction in process control by reducing the number of sensors, and can help in the interpretation of multivariate models [0].

Several methods have been developed for the selection of informative subsets of variables, such as uninformative variable elimination (UVE) [15-20], genetic algorithms (GA) [12,21,22], interval PLS (iPLS) [23,24], methods based on predictive-variable properties [3,10,13,15,16,25-38], tabu search [11], simulated annealing [39], mutual information (together with support vector machines) [40] and Monte Carlo variable selection [15,41].

For PLS1, with one response variable y, many methods are based on so-called predictive-variable properties, which are functions of various PLS1-model parameters, such as weights, loadings, PLS regression coefficients, or combinations of these parameters. Common examples of predictive-variable properties used are: (*i*) magnitude of PLS regression coefficients [10,13,15,25-29], (*ii*) magnitude of PLS regression coefficients multiplied [3,30] or divided [13] by the standard deviation of the predictor variable, (*iii*) correlation coefficients between predictor variables and the dependent variable [3,10,31], (*iv*) variable importance in the projection (VIP) score of a variable [10,26,29,32-34], (*v*) reliability, uncertainty or significance of PLS regression coefficients assessed by the student t value, calculated from the ratio of the PLS regression coefficient and its standard deviation, estimated by jack knifing [3,15,16,35,36], (*vi*) selectivity ratio (SR) [37,38]. The ranking of the variables on the predictive properties reflects their importance for the PLS model. The higher the magnitude of the property, the more important the variable.

The methods based on predictive-variable properties can be grouped into two categories, either using a threshold or a ranking of the property values. In the first category, after the development of a PLS model with the original data set, variables with property values below a defined threshold are considered uninformative and removed. The final PLS model is calculated after the removal of uninformative variables. The threshold is either determined arbitrarily [31], or through statistical assessment of the significance of the properties using bootstrap [34], jack knife [16,20,25,35] or Monte Carlo re-sampling methods [15,42].

In the second category a PLS model is built with the original data set and the variables are ranked in descending order of the considered property. Iteratively, the variable with the smallest value is eliminated and a new PLS model calculated. We call this Stepwise Variable

Reduction methods using Predictive-Property-Ranked Variables, denoted as SVR-PPRV methods.
In the stepwise removal of variables, the predictive abilities of the PLS models are assessed by the root mean squared error of cross validation (RMSECV) or the squared correlation coefficient for prediction Q^2. The set of variables, resulting in the optimal model, is then selected [10,25,26,30,31]. The goal is thus to obtain small sets of variables with improved or similar predictability, for a test set estimated as the root mean squared error of prediction (RMSEP), as the original data set.

The predictive property values of the variables may change during the variable reduction process, because they are functions of the parameters of the PLS algorithm which also can change in this process. In the stepwise variable reduction process the data matrix is changing continuously and the optimal number of PLS1 factors, i.e. the best PLS1 model complexity, can change as well. If the same PLS model complexity is used during the variable reduction procedure, RMSECV values may become overoptimistic [43], since it is possible that the best model complexity decreases due to the elimination of uninformative variables [16]. Therefore, SVR-PPRV methods should account for these changing variable property values and best PLS model complexity. Three steps thus need to be considered. First, after the removal of a variable, a new PLS model has to be calculated generating new PLS parameters and hence also new property values. Secondly, after remodelling, variables have to be re-ranked. Thirdly, a decrease in PLS model complexity must be considered.

In the existing SVR-PPRV methods [10,25,26,30,31] the PLS model complexity is fixed during the variable reduction process, and only the first or the second of the above steps is performed. In this study three new SVR-PPRV methods are proposed with a possibility to decrease the PLS1 complexity, and with the three steps integrated. They are different in the way the model complexity is decreased. They are called Predictive-Property-Ranked Variable Reduction with Complexity Adapted Models methods, denoted as PPRVR-CAM methods.

In this study, the performances, i.e. the selective and predictive abilities of the new PPRVR-CAM methods are investigated and compared with two related SVR-PPRV methods and two non-stepwise reference variable reduction methods, by built PLS1 models. The absolute value of the PLS regression coefficients is used as predictive-variable property because of the good performance reported for this property [10,25,26]. In a following study the effectiveness of other predictive properties will be investigated in combination with the preferred PPRVR-CAM method resulting from this study.

The two existing SVR-PPRV methods have a constant PLS complexity during variable reduction. They are modifications of methods described by Gauchi and Chagnon [25] and by Teófilo et al. [10]. The reference methods are hybrid methods: uninformative variable elimination (UVE) followed by either a genetic algorithm for PLS (GA-PLS) or an interval PLS, denoted as UVE-GA-PLS and UVE-iPLS, respectively. In the UVE step uninformative variables are eliminated to reduce computing time in the following GA or iPLS step, and to improve the performance of the GA-step [12,25].

The utility and effectiveness of the methods are investigated in conjunction with near-infrared (NIR) spectra and simulated data. NIR spectroscopy is chosen as application field because PLS is extensively used in analysis of these spectra [44,45]. Two NIR data sets and one simulated set were investigated. The latter is used to test the general applicability of the methods.

The data sets contain a total of 16 responses (see Table 5-1-Table 5-3). With this high number of responses, more reliable results were obtained for the statistical tests, carried out for the performance comparison of the variable reduction methods.

5.3 Theory

5.3.1 PLS1 regression coefficients

The variable reduction is based on the PLS1 regression coefficients b_k, which are elements of the regression vector $\mathbf{b}(K \times 1)$, calculated with,

$$\mathbf{b} = \mathbf{W}\left(\mathbf{P}^T\mathbf{W}\right)^{-1}\mathbf{q} \qquad \textbf{(1)}$$

where $\mathbf{W}(K \times A)$ is the $\mathbf{X}$ weight matrix, $\mathbf{P}(K \times A)$ is a x-loading matrix and $\mathbf{q}(1 \times A)$ is the y-loading vector [2]. The PLS1 regression coefficients b_k are dependent from each other unless A equals K [1]. K is the number of predictor variables in the $\mathbf{X}(N \times K)$ matrix, A is the number of PLS1 factors and N is the number of objects. Further details of PLS1 can be obtained in Refs. [1,2,46]. Influential variables have large positive or negative regression coefficients. The absolute value of the PLS1 regression coefficient of variable k, denoted as REG_k, is used in this study as a variable property for variable reduction.

$$REG_k = \left|b_k\right| \qquad \textbf{(2)}$$

5.3.2 Stepwise Variable Reduction methods using Predictive-Property-Ranked Variables

Two SVR-PPRV and three PPRVR-CAM methods, are investigated. The methods start building a PLS1 model from the original data set, followed by ranking the variables in descending order of magnitude of the considered property REG_k. The selective and predictive abilities of the methods are compared. Until a specified number of remaining variables is reached, iteratively, the variable with the smallest REG_k is eliminated and a new PLS1 model calculated.

Three new PPRVR-CAM methods, in which it is accounted for the fact that the properties may change during the variable reduction process, are introduced. Properties such as weights, loadings and PLS regression coefficients are functions of the parameters of the PLS algorithm, which are dependent on each other because they are calculated in a sequence of programming steps [2]. During the stepwise variable reduction process, the composition of the data matrix is changing continuously and parameters of the PLS algorithm can change simultaneously. As a result, variable properties can also change. Therefore, after each variable removal, a new PLS1 model is developed, generating new PLS parameters and hence also new property values. After remodelling, variables are reranked. During variable reduction, uninformative variables are eliminated. Therefore, the best PLS model complexity A may decrease. In the PPRVR-CAM methods a possibility for decreasing the model complexity is built in.

In summary, the PPRVR-CAM methods have the following characteristics: (*i*) remodelling after removal of a variable, (*ii*) renewed ranking of variables and (*iii*) best PLS1 model complexity evolution during the variable reduction process. The PPRVR-CAM methods have the first two characteristics in common but are different in decreasing model complexity.

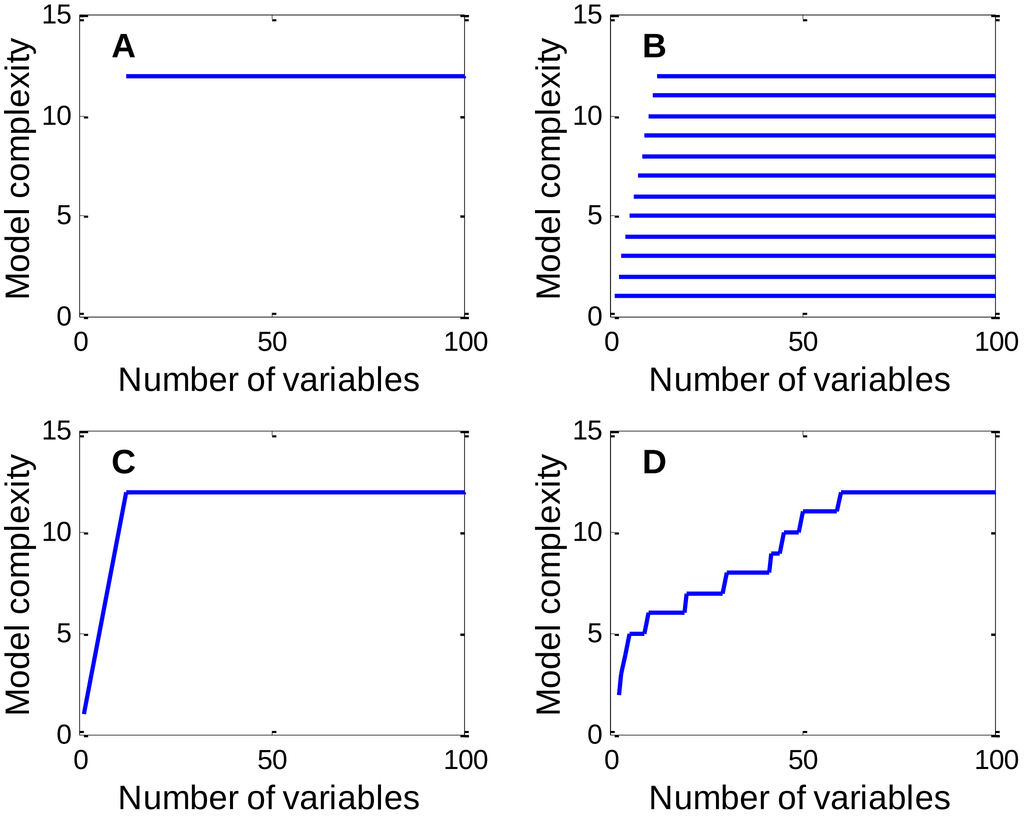


Fig. 5-1 PLS model complexity vs number of remaining variables for the five PPRV(R) methods, (A) SVR-1 and 2, (B) RCAM, (C) FCAM, (D) ICAM

The SVR-PPRV methods are modifications of existing methods [10,25]. They have a related methodology, but keep a constant PLS complexity during variable reduction, while of the first two characteristics one or both are considered.

For the five PPRV(R) methods, the differences in PLS model complexity during variable reduction are described below. As an example, model complexity changes are shown in Fig. 5-1 for a data set with 100 variables and a full spectrum PLS model complexity of 12.

The first SVR-PPRV method, denoted as SVR-1, is a modification of that described by Gauchi and Chagnon [25]. The *RMSECV* is used as criterion to select the best variable set, see section 5.3.4. Variable reduction is conducted at constant model complexity *A*, determined for the full spectrum, until *A* remaining variables (Fig. 5-1A).

The second SVR-PPRV method, denoted as SVR-2, is a modification of that recently described by Teófilo et al. [10]. Contrary to [10] variable reduction is conducted at constant model complexity *A*. Variable reduction stops at *A* remaining variables (Fig. 5-1A). In the SVR--2 method, variables are ranked only once (difference with SVR-1), at the start of the variable reduction process, based on the full spectrum PLS model result.

The first PPRVR-CAM method is an extended version of SVR-1. The variable reduction procedure starts with model complexity *A*, and is repeated with stepwise descending complexities *A*-1, *A*-2, …, 1 (Fig. 5-1B). At each model complexity, variable reduction stops when the number of remaining variables equals the model complexity. This method is called Predictive-Property-Ranked Variable Reduction with Repetitive Complexity Adapted Models, denoted as PPRVR-RCAM and abbreviated to RCAM.

A limitation of SVR-1 is that the minimal number of remaining variables equals the complexity *A* of the full spectrum PLS model. The second PPRVR-CAM method consist of a first variable reduction part, identical to SVR-1, with constant PLS model complexity *A* until the selection of *A* variables, and a second part with stepwise decreasing PLS model complexity *A*-1, *A*-2, …,1 after each variable removal. Variable reduction stops at one retained variable (Fig. 5-1C). This method is called Predictive-Property-Ranked Variable Reduction with Final Complexity Adapted Models, denoted as PPRVR-FCAM and abbreviated to FCAM.

In the third PPRVR-CAM method the procedure starts with model complexity *A*, while the possibility of decreasing the PLS model complexity is built in from the beginning. Two RMSECV values are calculated after each removal of a variable, one for the model complexity *A*, $RMSECV_A$, and one for a complexity *A*-1, $RMSECV_{A-1}$. The model complexity *A* is decreased by one if $RMSECV_{A-1} < RMSECV_A$ holds twice in a row (Fig. 5-1D). Because the minimal value for *A*-1=1, the complexity *A* is not decreased below 2. Variable reduction stops at two retained variables. This method is called Predictive-Property-Ranked Variable Reduction with Integral Complexity Adapted Models, denoted as PPRVR-ICAM and abbreviated to ICAM.

5.3.3 Model validation

The predictive ability of the models is assessed by internal validation in the training set, using segmented (*n*-fold) cross validation, resulting in the root mean squared error of cross validation (RMSECV),

$$RMSECV = \sqrt{\frac{1}{N_{cal}} \sum_{i=1}^{N_{cal}} \left(y_i - \hat{y}_i\right)^2} \qquad \textbf{(3)}$$

where y_i and $\hat{y}_i$ are the experimental and predicted properties, respectively, of the i^{th} calibration sample when situated in a left out segment, N_{cal} is the number of calibration samples in the training set.

The predictive ability of the models is also assessed by external validation with a test set, resulting in the root mean squared error of prediction (RMSEP),

$$RMSEP = \sqrt{\frac{1}{N_{test}} \sum_{i=1}^{N_{test}} \left(y_i - \hat{y}_i\right)^2} \qquad \textbf{(4)}$$

where y_i and $\hat{y}_i$ are the experimental and predicted properties, respectively, of the i^{th} sample in the test set, N_{test} is the number of samples in the test set. After variable reduction, using the reduced variable set, the best PLS model complexity is redetermined by segmented cross validation (SCV), which is then used for the external validation.

R^2_{Cal}, R^2_{Test}: Squared values of the correlation coefficient R between estimated and experimental properties are calculated with the reduced variable sets, for calibration (R^2_{Cal}) in the training set, and prediction (R^2_{Test}) with a test set, using the model complexity determined for internal and external validation, respectively.

The best complexity of a PLS model is determined by SCV. In order to avoid overfitting an adjusted Wold's R criterion, R_{adj}, is applied [47,48]. Initially, the minimum in the RMSECV vs model complexity curve is determined. Thereafter, going from the minimum to a lower number of PLS factors, the following equation is determined:

$$R_{adj} = \frac{RMSECV_{A+1}}{RMSECV_A} \tag{5}$$

where $RMSECV_{A+1}$ and $RMSECV_A$ are the error values of PLS models with A+1 and A factors, respectively. When R_{adj} <0.98 then the A factor model is considered as the best complexity [49].

5.3.4 Selection criterion for the preferred variable set

In the five PPRV(R) methods, *RMSECV* values are plotted as a function of the number of remaining variables. The model with the global minimal value, $RMSECV_{Min}$, corresponds to the variable set with optimal predictive capability. However, a smaller variable set with *RMSECV* not significantly higher than that corresponding to $RMSECV_{Min}$ is preferred. Its maximal value, $RMSECV_{Crit}$ is defined as the *RMSECV* not significantly larger than $RMSECV_{Min}$, by means of a one-tailed F-test [50],

$$RMSECV^2_{Crit} = F_{(\alpha,N_{cal},N_{cal})} RMSECV^2_{Min} \tag{6}$$

with $F_{(\alpha,N_{cal},N_{cal})}$ at the significance level α=0.05 and N_{cal} degrees of freedom of both the numerator and denominator, being the number of calibration samples in the training set.

Small variable sets with improved or at least equivalent predictability compared to the original data set can only be obtained if $RMSECV_{Crit}$ is smaller than or equal to the *RMSECV* of the full spectrum model, $RMSECV_{FS}$. Therefore, if $RMSECV_{Crit} > RMSECV_{FS}$, then $RMSECV_{Crit}$ is set to $RMSECV_{FS}$.

Thus, the smallest variable set with K_{Best} variables and a $RMSECV_{Best}$ smaller than or equal to $RMSECV_{Crit}$ is considered the best set for a given PPRV(R) method. A low number of variables can be beneficial with regard to (*i*) a better understanding of the model, and (*ii*) selection of a viable set of sensors in process control.

5.4 Reference methods

In stepwise backwards variable selection methods it is possible that variables are excluded which could be important when added to the finally selected set [51]. Therefore, two reference methods are chosen based on completely different selection mechanisms, i.e. the hybrid methods UVE-GA-PLS and UVE-iPLS. In a first step, the search range is reduced by the elimination of uninformative variables from the original data set by UVE-PLS. In the following step, further variable reduction is carried out by either a genetic algorithm or interval PLS, resulting in a number of remaining variables comparable to that of the PPRV(R) methods.
UVE-GA-PLS is a fully non-stepwise method. In UVE-iPLS, variables are selected stepwise in the iPLS part, but the selection is conducted in the forward mode, i.e. in a direction opposite to that of the PPRV(R) methods. However, the selected number of variables in both hybrid methods will vary because of the variability in the UVE-step, and for UVE-GA-PLS, also in the GA step.

Uninformative variable elimination for PLS (UVE-PLS) [16] determines the fitness of each predictor variable k in the **X** matrix against those of L artificial random variables added to the data set. These added random variables have very small absolute values, of the order of about 10^{-10}, so that their influence on the regression coefficients of the predictors is negligible. The $K+L$ mean PLS regression coefficients $\overline{b}_k$ and their standard deviations $s(b_k)$ are calculated from i vectors of regression coefficients, obtained by leave-one-out jack-knifing (i=1, …, N_{Cal}). The fitness c_k of each variable k is determined by the ratio of the mean regression coefficient and its standard deviation: $c_k=\overline{b}_k/s(b_k)$. A suitable cut-off value $|c_k|_{cut\text{-}off}$ is calculated from the L artificial variables, taking the maximum of their absolute c_k values. Predictor variables with $|c_k|$ below the cut-off value are classified as uninformative and eliminated. In UVE-PLS, the number of eliminated variables is variable because of the variability in the added artificial random variables.

Genetic algorithms (GAs) are variable selection methods based on the principles of natural selection in biologic evolution. Species adapt over a high number of generations, because the fittest survive and spread their genetic material to following generations [52]. GAs have successfully been used for variable selection [12,21,52-55]. Details about the method can be found in [21,52].
According to Leardi et al. [56] the performance of GAs improves when the number of variables is kept below 200. In Ref. [12] it was concluded that better results with GA in wavelength selection for NIR can be obtained by using a subset of relevant spectral points instead of the full spectrum. One of the disadvantages of GAs is the large variability of solutions [25].

In interval PLS, a subset of variables is selected by a sequential search for the best variables. Spectra are split into small equidistant intervals which can be either a single variable or a window of adjacent variables. iPLS can operate in the forward or backward mode by successively including or excluding intervals, respectively. Details about the method, are described in [23].

Table 5-1 Results of variable reduction methods for the Diesel data set

Model	Response	Method characteristics	Full spectrum	UVE-GA-PLS	UVE-iPLS	Method SVR-1	Method SVR-2	Method RCAM	Method FCAM	Method ICAM
1	**Viscosity**	PLS complexity selecting best set	11	7	15	11	11	7	11	5
		Number of variables, K_{Best}	401	24	17	13	34	7	13	7
		$RMSECV_{Best}$	0.121	0.107	0.104	0.105	0.119	0.118	0.105	0.116
		PLS complexity best set	11	7	15	7	9	5	7	4
		RMSEP	0.102	0.104	0.099	0.099	0.116	0.131	0.099	0.124
		R^2_{Test}	0.934	0.931	0.938	0.938	0.914	0.891	0.938	0.905
2	**BP50**	PLS complexity selecting best set	11	5	15	11	11	10	11	6
		Number of variables, K_{Best}	401	30	20	11	108	10	11	10
		$RMSECV_{Best}$	3.47	3.08	3.15	3.06	3.47	3.31	3.06	3.46
		PLS complexity best set	11	5	15	7	9	5	7	5
		RMSEP	3.60	3.62	3.11	3.49	3.50	3.92	3.49	3.96
		R^2_{Test}	0.956	0.955	0.968	0.959	0.958	0.948	0.959	0.946
3	**CN**	PLS complexity selecting best set	5	4	5	5	5	4	4	4
		Number of variables, K_{Best}	401	22	11	5	10	4	4	4
		$RMSECV_{Best}$	1.99	1.89	1.88	1.91	1.99	1.99	1.91	1.92
		PLS complexity best set	5	4	5	4	6	4	4	4
		RMSEP	2.11	2.07	2.05	2.08	2.16	2.15	2.08	2.08
		R^2_{Test}	0.654	0.661	0.664	0.660	0.630	0.638	0.661	0.657
4	**D4052**	PLS complexity selecting best set	15	6	15	15	15	15	15	8
		Number of variables, K_{Best}	401	28	24	17	67	17	17	16
		$RMSECV_{Best}$	$1.05\cdot10^{-3}$	$1.06\cdot10^{-3}$	$9.43\cdot10^{-4}$	$9.54\cdot10^{-4}$	$9.37\cdot10^{-4}$	$9.54\cdot10^{-4}$	$9.54\cdot10^{-4}$	$1.05\cdot10^{-3}$
		PLS complexity best set	15	6	15	10	15	10	10	6
		RMSEP	$9.20\cdot10^{-4}$	$1.09\cdot10^{-3}$	$1.04\cdot10^{-3}$	$1.07\cdot10^{-3}$	$9.13\cdot10^{-4}$	$1.07\cdot10^{-3}$	$1.07\cdot10^{-3}$	$1.09\cdot10^{-3}$
		R^2_{Test}	0.991	0.988	0.989	0.989	0.992	0.989	0.989	0.988
5	**Freeze**	PLS complexity selecting best set	9	9	10	9	9	8	7	5
		Number of variables, K_{Best}	401	16	12	9	25	8	7	11
		$RMSECV_{Best}$	2.57	2.30	2.26	2.35	2.54	2.45	2.43	2.53
		PLS complexity best set	9	9	10	6	12	5	7	7
		RMSEP	2.49	2.93	2.72	2.76	2.92	2.68	2.69	2.56
		R^2_{Test}	0.624	0.482	0.564	0.545	0.483	0.571	0.570	0.609
6	**Total**	PLS complexity selecting best set	14	9	15	14	14	11	14	6
		Number of variables, K_{Best}	401	29	22	15	25	11	15	20
		$RMSECV_{Best}$	0.600	0.526	0.503	0.583	0.593	0.577	0.583	0.594
		PLS complexity best set	14	9	15	9	11	8	9	14
		RMSEP	0.592	0.605	0.710	0.617	0.703	0.692	0.617	0.620
		R^2_{Test}	0.991	0.990	0.986	0.990	0.986	0.987	0.990	0.990

5.5 Data and methodology

5.5.1 Diesel data set

The first data set was composed of 252 diesel samples with first derivative spectral NIR data at 401 wavelengths. The spectral data were provided without wavelengths. The data set was downloaded from the Eigenvector Research homepage (http://www.eigenvector.com). It was split as provided with 20 high leverage and 116 low leverage samples in the training set and 116 low leverage samples in the test set. The physical properties viscosity (Visc), boiling point (BP50), cetane number (CN), density (D4052), freezing temperature (Freeze) and total aromatics (Total) are used as responses. These 6 responses were each modelled as a function of the NIR data (Table 5-1). To determine the PLS model complexity, the *RMSECV* values were obtained from 10-fold cross validation.

5.5.2 Corn data set

The second data set consists of NIR spectra of 80 corn samples with a wavelength range of 1100–2498 nm at 2 nm intervals, resulting in 700 predictor variables. This data set is part of a data set labelled corn, provided by Eigenvector Research. The spectra used in this study were obtained from the spectrometer denoted as “m5”. The moisture, oil, protein and starch contents of the samples are used as response variables (Table 5-2). The data set is split into a training set of 60 and a test set of 20 samples using the duplex method [57]. Eight fold cross validation is conducted during model building.

5.5.3 Simulated data set

The third data set is simulated. It represents the spectra or chromatograms of mixtures containing one to four compounds, indicated by A, B, C and D. Six sample types of mixtures, A, AB, AD, ABC, ABD and ABCD, are created (Table 5-3). The pure spectral/ chromatographic profiles of the analytes were formed by Gaussian peaks, measured within the first 100 variables of the global profile (Fig. 5-2). The concentrations of the analytes were randomly generated between 0 and 1. To study the selective abilities of the variable reduction methods, the response vector **y** was formed by the concentrations of compound A.

Each **X**-**y** combination consists of one of 120 samples of simulated spectra with 200 predictor variables. The first 100 variables are informative, with *x* values used for the calculation of the analyte profiles in the mixtures. The last 100 variables are uninformative, consisting of random numbers from 0 to 1 (Fig. 5-2). Additionally, noise is added to the simulated 200-variables spectra, consisting of random numbers in the range between 0 and 0.005, i.e. small compared to the pure spectral profiles. Each subset is split into a training set of 100 and a test set of 20 samples using the duplex method and 10-fold cross validation is conducted.

Table 5-2 Results of variable reduction methods for the Corn data set

Model	Response	Method characteristics	Full Spectrum	UVE-GA-PLS	UVE-iPLS	Method SVR-1	Method SVR-2	Method RCAM	Method FCAM	Method ICAM
7	**Moisture**	PLS complexity selecting best set	15	7	11	15	15	4	2	2
		Number of variables, K_{Best}	700	10	11	15	21	4	2	2
		$RMSECV_{Best}$	$1.12\cdot10^{-2}$	$2.83\cdot10^{-4}$	$2.51\cdot10^{-4}$	$3.24\cdot10^{-4}$	$3.70\cdot10^{-4}$	$3.03\cdot10^{-4}$	$3.04\cdot10^{-4}$	$3.04\cdot10^{-4}$
		PLS complexity best set	15	7	11	15	15	4	2	2
		$RMSEP$	$1.19\cdot10^{-2}$	$3.20\cdot10^{-4}$	$3.27\cdot10^{-4}$	$3.54\cdot10^{-4}$	$4.16\cdot10^{-4}$	$3.41\cdot10^{-4}$	$3.00\cdot10^{-4}$	$3.00\cdot10^{-4}$
		R^2_{Test}	0.999	1.000	1.000	1.000	1.000	1.000	1.000	1.000
8	**Oil**	PLS complexity selecting best set	11	10	3	11	11	10	11	7
		Number of variables, K_{Best}	700	25	5	18	40	17	18	11
		$RMSECV_{Best}$	0.061	0.028	0.081	0.022	0.058	0.023	0.022	0.047
		PLS complexity best set	11	10	3	10	8	11	10	7
		$RMSEP$	0.060	0.052	0.110	0.021	0.066	0.020	0.021	0.069
		R^2_{Test}	0.869	0.895	0.637	0.983	0.837	0.984	0.983	0.842
9	**Protein**	PLS complexity selecting best set	14	9	14	14	14	12	14	7
		Number of variables, K_{Best}	700	21	20	28	104	20	28	19
		$RMSECV_{Best}$	0.103	0.118	0.055	0.043	0.049	0.048	0.043	0.078
		PLS complexity best set	14	9	14	12	13	10	12	10
		$RMSEP$	0.090	0.114	0.070	0.072	0.040	0.063	0.072	0.069
		R^2_{Test}	0.968	0.950	0.983	0.982	0.994	0.988	0.982	0.982
10	**Starch**	PLS complexity selecting best set	15	7	15	15	15	15	15	11
		Number of variables, K_{Best}	700	26	19	26	68	26	26	35
		$RMSECV_{Best}$	0.222	0.247	0.202	0.085	0.129	0.085	0.085	0.083
		PLS complexity best set	15	7	15	10	10	10	10	11
		$RMSEP$	0.170	0.319	0.465	0.125	0.139	0.125	0.125	0.130
		R^2_{Test}	0.962	0.865	0.719	0.979	0.974	0.979	0.979	0.977

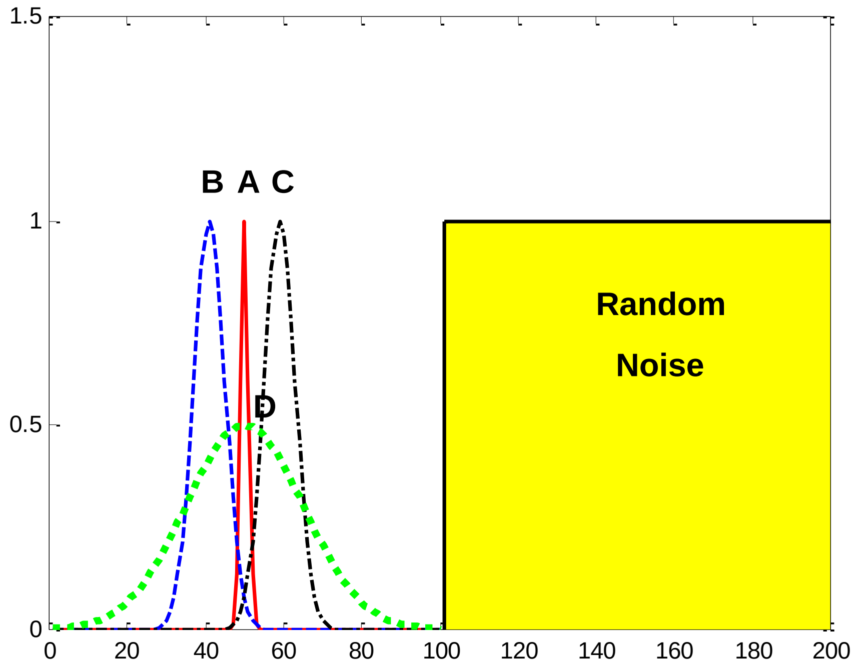


Fig. 5-2 Spectral/Chromatographic profiles used for simulated data

5.5.4 Software

All calculations are made with in-house programs developed in Matlab (V. 6.5) (The Math Works, Natick, MA, USA) [58]. The Uninformative Variable Elimination procedures and the duplex algorithm are from ChemoAC Standard Functions Toolbox for MATLAB [59]. Variable selection using genetic algorithm and interval PLS is conducted with the PLS-Toolbox V5.2 [60]. Statistical tests are conducted with the Statistics Toolbox of Matlab.

5.6 Results and discussion

First, for the 16 models (Table 5-1 - Table 5-3), the optimal factor number of the PLS1 models was determined by cross validation as described in section 5.3.3, and *RMSECV* and *RMSEP* are calculated for the full spectrum models. Variable reduction is then applied on all **X-y** sets by the five PPRV(R) methods, with variables ranked on the magnitude of absolute PLS regression coefficients REG_k, and by the two hybrid methods, UVE-GA-PLS and UVE-iPLS.

For all methods, one PLS1 model is selected for each response. The variables and responses are pre-processed by mean centring.

Table 5-3 Results of variable reduction methods for the Simulated data set

Model	Mixtures	Method characteristics	Full spectrum	UVE-GA-PLS	UVE-iPLS	Method SVR-1	Method SVR-2	Method RCAM	Method FCAM	Method ICAM
11	**A**	PLS complexity selecting best set	11	3	3	11	11	3	3	3
		Number of variables, K_{Best}	200	8	10	11	11	3	3	3
		$RMSECV_{Best}$	0.075	$1.06\cdot10^{-3}$	$1.02\cdot10^{-3}$	$1.04\cdot10^{-3}$	$1.08\cdot10^{-3}$	$1.07\cdot10^{-3}$	$1.07\cdot10^{-3}$	$1.07\cdot10^{-3}$
		PLS complexity best set	11	3	3	5	5	1	1	1
		$RMSEP$	0.069	$1.16\cdot10^{-3}$	$1.22\cdot10^{-3}$	$1.30\cdot10^{-3}$	$1.28\cdot10^{-3}$	$1.12\cdot10^{-3}$	$1.12\cdot10^{-3}$	$1.12\cdot10^{-3}$
		R^2_{Test}	0.957	1.000	1.000	1.000	1.000	1.000	1.000	1.000
12	**A,B**	PLS complexity selecting best set	14	4	4	14	14	4	4	4
		Number of variables, K_{Best}	200	12	13	14	14	4	4	4
		$RMSECV_{Best}$	0.065	$9.50\cdot10^{-4}$	$9.83\cdot10^{-4}$	$1.00\cdot10^{-3}$	$1.48\cdot10^{-3}$	$1.04\cdot10^{-3}$	$1.03\cdot10^{-3}$	$1.03\cdot10^{-3}$
		PLS complexity best set	14	4	4	3	8	2	2	2
		$RMSEP$	0.052	$1.38\cdot10^{-3}$	$1.41\cdot10^{-3}$	$1.27\cdot10^{-3}$	$1.38\cdot10^{-3}$	$1.35\cdot10^{-3}$	$1.33\cdot10^{-3}$	$1.33\cdot10^{-3}$
		R^2_{Test}	0.964	1.000	1.000	1.000	1.000	1.000	1.000	1.000
13	**A,D**	PLS complexity selecting best set	12	2	4	12	12	7	7	2
		Number of variables, K_{Best}	200	35	30	12	12	7	7	9
		$RMSECV_{Best}$	0.088	$9.75\cdot10^{-4}$	$9.39\cdot10^{-4}$	$1.09\cdot10^{-3}$	$1.47\cdot10^{-3}$	$1.20\cdot10^{-3}$	$1.20\cdot10^{-3}$	$1.34\cdot10^{-3}$
		PLS complexity best set	12	2	4	12	5	2	2	2
		$RMSEP$	0.081	$1.23\cdot10^{-3}$	$1.58\cdot10^{-3}$	$1.67\cdot10^{-3}$	$1.17\cdot10^{-3}$	$1.59\cdot10^{-3}$	$1.59\cdot10^{-3}$	$1.23\cdot10^{-3}$
		R^2_{Test}	0.932	1.000	1.000	1.000	1.000	1.000	1.000	1.000
14	**A,B,C**	PLS complexity selecting best set	13	4	3	13	13	5	5	3
		Number of variables, K_{Best}	200	17	16	13	37	5	5	5
		$RMSECV_{Best}$	0.076	$1.08\cdot10^{-3}$	$1.09\cdot10^{-3}$	$1.16\cdot10^{-3}$	$1.50\cdot10^{-3}$	$1.13\cdot10^{-3}$	$1.13\cdot10^{-3}$	$1.15\cdot10^{-3}$
		PLS complexity best set	13	4	3	3	12	3	3	3
		$RMSEP$	0.042	$1.42\cdot10^{-3}$	$1.22\cdot10^{-3}$	$1.14\cdot10^{-3}$	$1.31\cdot10^{-3}$	$1.21\cdot10^{-3}$	$1.21\cdot10^{-3}$	$1.24\cdot10^{-3}$
		R^2_{Test}	0.986	1.000	1.000	1.000	1.000	1.000	1.000	1.000
15	**A,B,D**	PLS complexity selecting best set	16	3	4	16	16	6	11	3
		Number of variables, K_{Best}	200	30	36	16	18	8	11	12
		$RMSECV_{Best}$	0.090	$9.12\cdot10^{-4}$	$9.12\cdot10^{-4}$	$9.82\cdot10^{-4}$	$1.22\cdot10^{-3}$	$1.08\cdot10^{-3}$	$1.10\cdot10^{-3}$	$1.15\cdot10^{-3}$
		PLS complexity best set	16	3	4	5	6	3	3	3
		$RMSEP$	0.050	$1.55\cdot10^{-3}$	$1.55\cdot10^{-3}$	$1.84\cdot10^{-3}$	$1.51\cdot10^{-3}$	$1.98\cdot10^{-3}$	$1.67\cdot10^{-3}$	$1.62\cdot10^{-3}$
		R^2_{Test}	0.974	1.000	1.000	1.000	1.000	1.000	1.000	1.000
16	**A,B,C,D**	PLS complexity selecting best set	15	4	4	15	15	7	11	4
		Number of variables, K_{Best}	200	38	40	15	15	11	11	15
		$RMSECV_{Best}$	0.102	$1.27\cdot10^{-3}$	$1.20\cdot10^{-3}$	$1.37\cdot10^{-3}$	$1.80\cdot10^{-3}$	$1.46\cdot10^{-3}$	$1.47\cdot10^{-3}$	$1.52\cdot10^{-3}$
		PLS complexity best set	15	4	4	4	7	4	4	4
		$RMSEP$	0.085	$1.40\cdot10^{-3}$	$1.67\cdot10^{-3}$	$1.82\cdot10^{-3}$	$1.70\cdot10^{-3}$	$1.64\cdot10^{-3}$	$1.84\cdot10^{-3}$	$1.21\cdot10^{-3}$
		R^2_{Test}	0.935	1.000	1.000	1.000	1.000	1.000	1.000	1.000

5.6.1 Application of UVE-GA-PLS and UVE-iPLS

The number of variables eliminated by UVE is variable. To get reliable results, for each **X-y** set, the UVE method was repeated five times, starting with the full spectrum. Further variable selection by both GA and iPLS was applied on the resulting reduced variable sets, after mean centring. Default parameter settings for GA and iPLS are used [49].

Because GAs show a large variability in variable selection, five times repeated GA runs are conducted on each UVE reduced variable set. The GA run with the lowest RMSECV was selected as best. As the results of iPLS are constant, forward iPLS is applied only once after UVE. For the variables selected by GA and iPLS, the complexity of the PLS model was determined by SCV (see 5.3.3). The results of the sets with the median number of retained variables are shown in Table 5-1-Table 5-3.

5.6.2 Application of the PPRV(R) methods on the Diesel data set and response viscosity

The PPRV(R) methods consist of four steps, for which the first and last are common. First, the data set is split into a training and a test set. The X matrix contains all variables. The optimal number of PLS factors *A* is determined by SCV. In the fourth step, using the reduced variable set, the PLS model is externally validated (RMSEP) using a test set, after a renewed determination of the optimal number of PLS factors *A* by SCV. Further details on the other steps are given below. In Fig. 5-3 flow charts are given for the new PPRVR-CAM methods.

As representative example of variable reduction by the five PPRV(R) methods, the results for the Diesel data set with response viscosity are discussed below. The PLS complexity selecting the best set, the number of remaining variables in the best set K_{Best}, RMSECV of the best set $RMSECV_{Best}$, the number of PLS factors after renewed determination of the optimal number of factors for the best set, RMSEP and the squared correlation coefficient for prediction with the test, R^2_{Test}, are shown in Table 5-1.

5.6.2.1 Application of SVR-1

This method has a constant PLS model complexity *A* during variable reduction.
In step 2 a proper PLS model is developed and the RMSECV determined by SCV. REG_k is calculated for all variables and ranked. The variable with the lowest REG_k is deleted. Step 2 is repeated at constant PLS model complexity *A* until the number of remaining variables is equal to *A* (Fig. 5-1A). In the third step, $RMSECV_{Best}$ is determined as described in section 5.3.4 and the corresponding subset of variables selected.
In the example, the optimal number of PLS factors is 11 in the full spectrum model for viscosity (Table 5-1). In the SVR-1 method, variable reduction is thus conducted with model complexity *A*=11. Variable reduction stops at 11 variables. Fig. 5-4A shows the RMSECV curve as a function of the number of remaining variables. The best variable set has $RMSECV_{Best}$= 0.105 and contains 13 variables. Using this remaining variable set, the best model complexity becomes 7 and for the test set, *RMSEP*=0.099 and the squared correlation coefficient $R^2_{Test} = 0.938$.

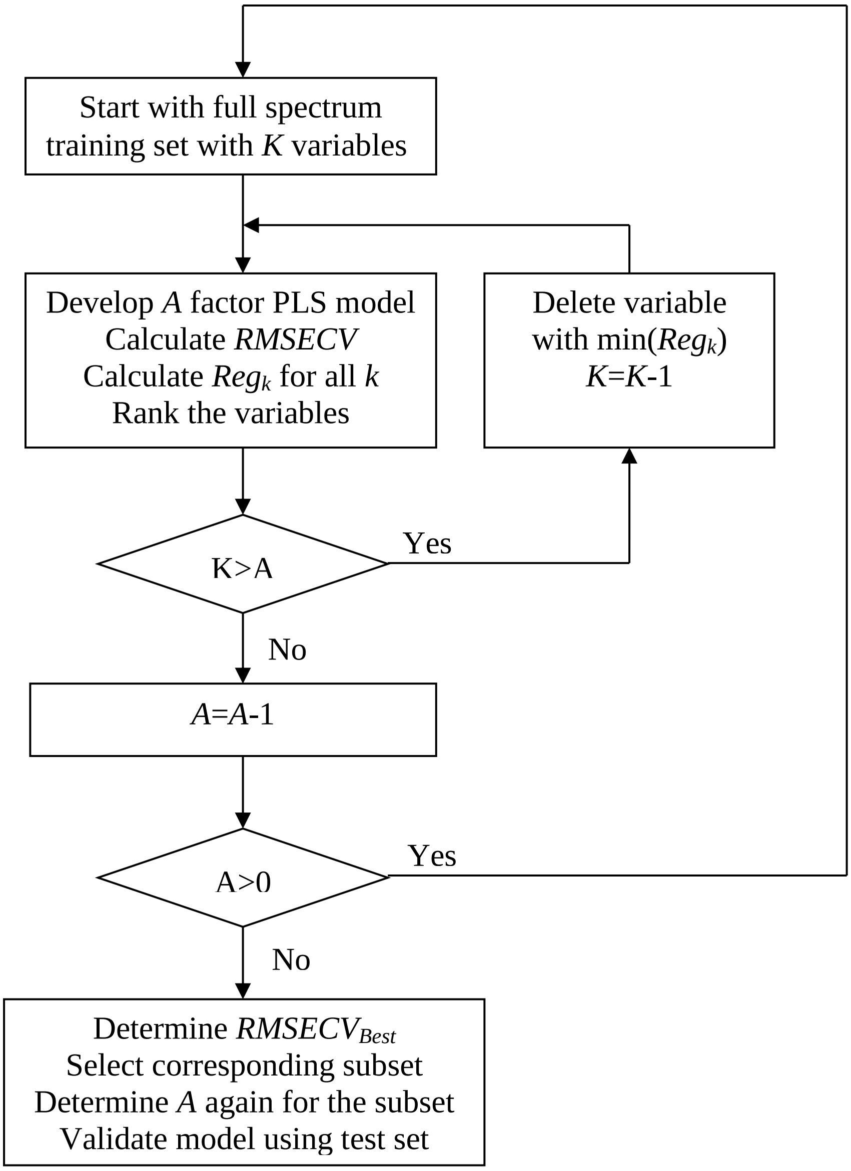

Start with full spectrum training set with *K* variables
Develop *A* factor PLS model
Calculate *RMSECV*
Calculate Reg_k for all *k*
Rank the variables
Delete variable with min(Reg_k)
K=*K*-1
K>A
Yes
No
A=*A*-1
A>0
Yes
No
Determine $RMSECV_{Best}$
Select corresponding subset
Determine *A* again for the subset
Validate model using test set

Fig. 5-3A Flow chart of the RCAM method

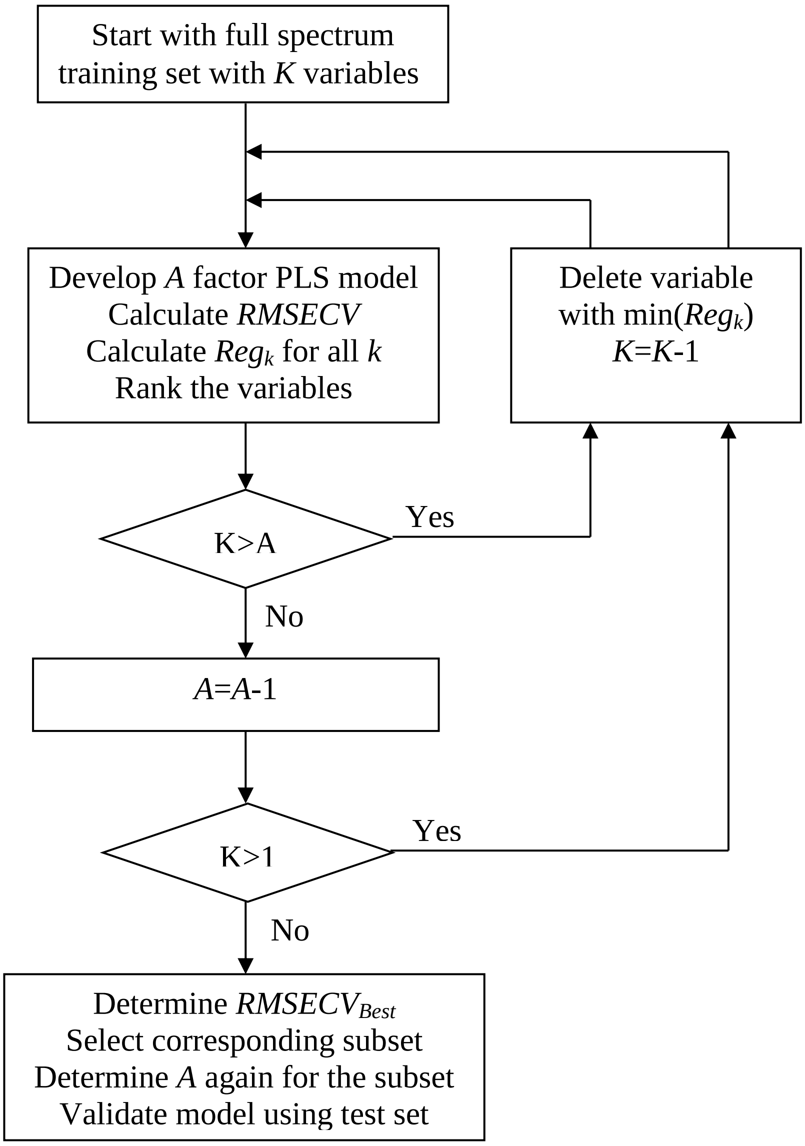

Start with full spectrum training set with *K* variables
Develop *A* factor PLS model
Calculate *RMSECV*
Calculate *Reg*$_k$ for all *k*
Rank the variables
Delete variable with min(*Reg*$_k$)
K=*K*-1
K>A
Yes
No
A=*A*-1
K>1
Yes
No
Determine *RMSECV*$_{Best}$
Select corresponding subset
Determine *A* again for the subset
Validate model using test set

Fig. 5-3B Flow chart of the FCAM method

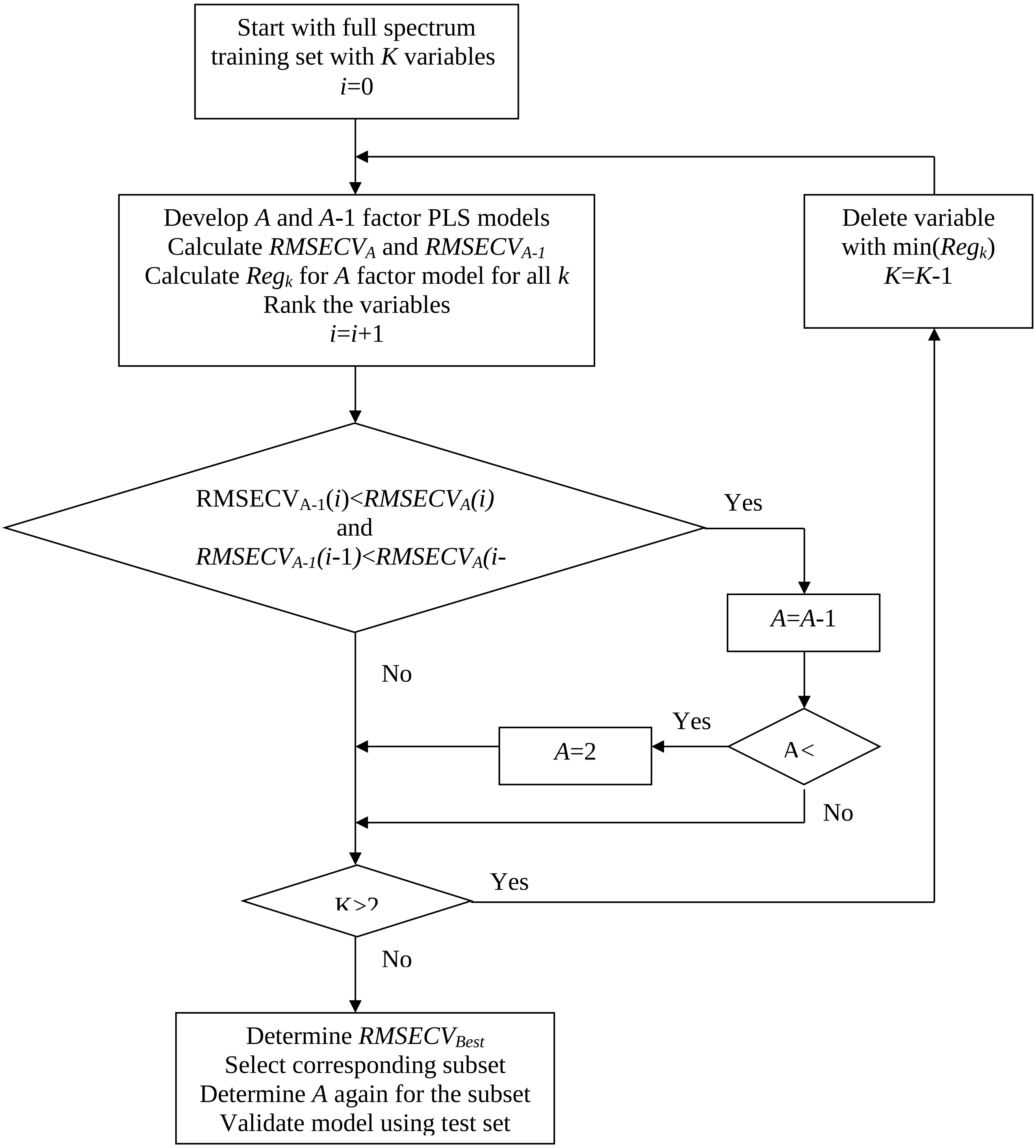

Start with full spectrum training set with K variables i=0
Develop A and A-1 factor PLS models
Calculate $RMSECV_A$ and $RMSECV_{A-1}$
Calculate Reg_k for A factor model for all k
Rank the variables
i=i+1
Delete variable with min(Reg_k)
K=K-1
$\mathrm{RMSECV}_{\mathrm{A-1}}(i)<RMSECV_A(i)$
and
$RMSECV_{A-1}(i\text{-}1)<RMSECV_A(i\text{-}$
Yes
A=A-1
No
Yes
A=2
A<
No
Yes
K>2
No
Determine $RMSECV_{Best}$
Select corresponding subset
Determine A again for the subset
Validate model using test set

Fig. 5-3C Flow chart of the ICAM method

5.6.2.2 Application of SVR-2

SVR-2 has a constant PLS complexity *A*, but different from SVR-1, variables are ranked only once. In step 2 REG_k is calculated for all variables, based on the full spectrum model, and ranked. The variable with the lowest REG_k is deleted. A PLS model is developed and the corresponding RMSECV determined. Step 2 is repeated, but without renewed ranking, until the number of remaining variables equals the complexity *A* of the PLS model (Fig. 5-1A).
In the third step, $RMSECV_{Best}$ is determined and the corresponding subset of variables selected.
In the example, variable reduction is conducted with constant model complexity *A*=11 until 11 remaining variables. Variables are ranked once on the absolute regression coefficients REG_k of the full spectrum PLS model. Fig. 5-4B shows the RMSECV curve as a function of the number of remaining variables. $RMSECV_{Best}$=0.119, located at 34 remaining variables. Using this remaining variable set, the model complexity is 9 and for the test set *RMSEP*=0.116 and $R^2_{Test} = 0.914$.

5.6.2.3 Application of RCAM

RCAM is an extended version of SVR-1. In Fig. 5-3A a flow chart is given for the method. In step 2, variable reduction is repeated with stepwise descending complexities *A*, *A*-1, …, 1. At each model complexity, variable reduction stops when the number of remaining variables equals the model complexity (Fig. 5-1B). In the third step, the determination of the critical *RMSECV* is different from the procedure described in section 5.3.4. The minimum *RMSECV* is the global minimum of all RMSECV curves (Fig. 5-4C). This results in $RMSECV_{Crit}$ (see equation 6). The smallest variable set with $RMSECV<RMSECV_{Crit}$ is then selected. This smallest set can be selected either at the considered complexity curve with $RMSECV_{Min}$ or on a curve of lower complexity. This allows selecting a smaller variable set when $RMSECV<RMSECV_{Crit}$ is fulfilled.
In the example, variable reduction starts with complexity *A*=11 and is repeated with complexities *A*=10, 9, …, 1, until the number of remaining variables is 10, 9, …, 1.
The resulting RMSECV curves are shown in Fig. 5-4C. The global minimum of the RMSECV curves, $RMSECV_{Min}$=0.103, is located on the curve for *A*=8 at 12 remaining variables and $RMSECV_{Crit}$=0.119 (calculated with equation (6)). The best variable set on the same curve has *RMSECV*= 0.118 and contains 8 variables. On the less complex curve *A*=7 a smaller set of variables with *RMSECV*< 0.119 i.e. with $RMSECV_{Best}$=0.118 is found for a set with only 7 variables which is thus preferred. Using this remaining variable set, the best model complexity becomes 5 and for the test set *RMSEP*=0.131 and $R^2_{Test} = 0.891$.

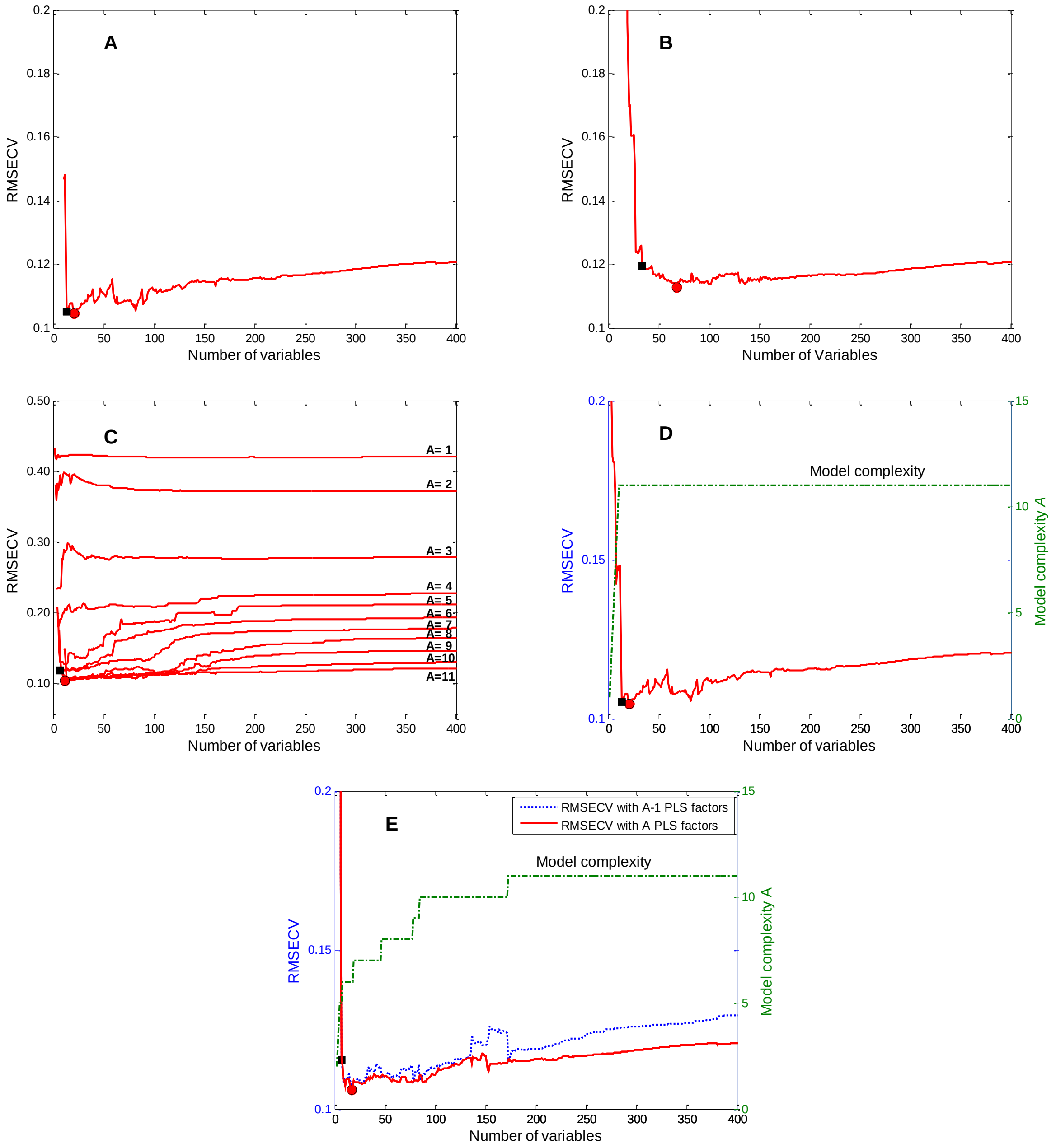


Fig. 5-4 RMSECV curves of the PPRV(R) methods for the Diesel data set with response viscosity: (A) SVR-1, (B) SVR-2, (C) RCAM, (D) FCAM, (E) ICAM; — RMSECV-curve; -·- PLS model complexity; (●) Minimum RMSECV; (■) RMSECV best set

5.6.2.4 Application of FCAM

Method FCAM consist of a first part with constant PLS model complexity *A* until the selection of *A* variables, and a second part with stepwise decreasing PLS model complexity *A*-1, *A*-2, …,1. In Fig. 5-3B a flow chart is given for the method. In step 2 a PLS model is developed with the best model complexity *A*, and the corresponding RMSECV is determined by SCV. REG_k is calculated for all variables and ranked. The variable with the lowest REG_k is deleted. When the number of remaining variables is *A*, the model complexity is decreased by one. Step 2 is repeated until the number of remaining variables and the PLS model complexity are equal to 1 (Fig. 5-1C). In the third step, $RMSECV_{Best}$ is determined and the corresponding subset of variables selected.
In the example, the FCAM method consists of a first part, identical to SVR-1, with constant model complexity *A*=11 between 401 and 11 remaining variables. In the final part, from 10 till 1 variable, the model complexity decreases stepwise from 10 to 1 after each variable removal. The resulting curves of *RMSECV* and of the PLS model complexity *A* are shown in Fig. 5-4D. $RMSECV_{Best}$=0.105, located at 13 remaining variables, identical to that of the SVR-1 method. The second part of variable reduction does in this case not result in a lower number of remaining variables. Using this remaining variable set, the model complexity is 7 and for the test set *RMSEP*=0.099 and $R^2_{Test} = 0.938$.

5.6.2.5 Application of ICAM

In ICAM, the possibility of decreasing the PLS model complexity is built in from the beginning. In Fig. 5-3C a flow chart is given for the method. In step 2, two PLS models with complexities *A* and *A*-1 are developed and the corresponding RMSECV values are calculated after each removal of a variable, one for the model complexity *A*, $RMSECV_A$, and one for a complexity *A*-1, $RMSECV_{A-1}$. REG_k is calculated for all variables, based on the PLS model with complexity *A* and ranked. The variable with the lowest REG_k is deleted. The model complexity *A* is decreased by one if $RMSECV_{A-1} < RMSECV_A$ holds twice in a row (Fig. 5-1D). Because the minimal value for *A*-1=1, the complexity *A* is not decreased below 2. Step 2 is repeated until the number of remaining variables is 2 (Fig. 5-1D). In the third step, $RMSECV_{Best}$ is determined and the corresponding subset of variables selected.
In the example, the ICAM procedure starts with *A*=11. The resulting curves of $RMSECV_A$ and $RMSECV_{A-1}$ and of the PLS model complexity *A* are shown in Fig. 5-4E. $RMSECV_{Best}$=0.116, located at 7 remaining variables, corresponding to *A*=5. Using this remaining variable set, the model complexity is 4 and for the test set *RMSEP*=0.124 and $R^2_{Test} = 0.905$.

In Fig. 5-5 the spectra of the Diesel data set are shown in the top window, and for the remaining variables for both methods, SVR-1 and FCAM, for the response viscosity, the absolute PLS regression coefficients REG_k are given in the bottom window. The remaining variables have high absolute regression coefficients, as expected. Fig. 5-6 shows the experimental and predicted viscosities for the PLS model with complexity 7, developed with the variables selected by methods SVR-1 and FCAM, for both the training and test sets. The squared correlation coefficients for calibration with the training set and prediction with the test set are $R^2_{Cal} = 0.950$ and $R^2_{Test} = 0.938$ respectively.

Analogously, all variable reduction methods were applied on all responses of the different data sets. The results are shown in Table 5-2 and Table 5-3.

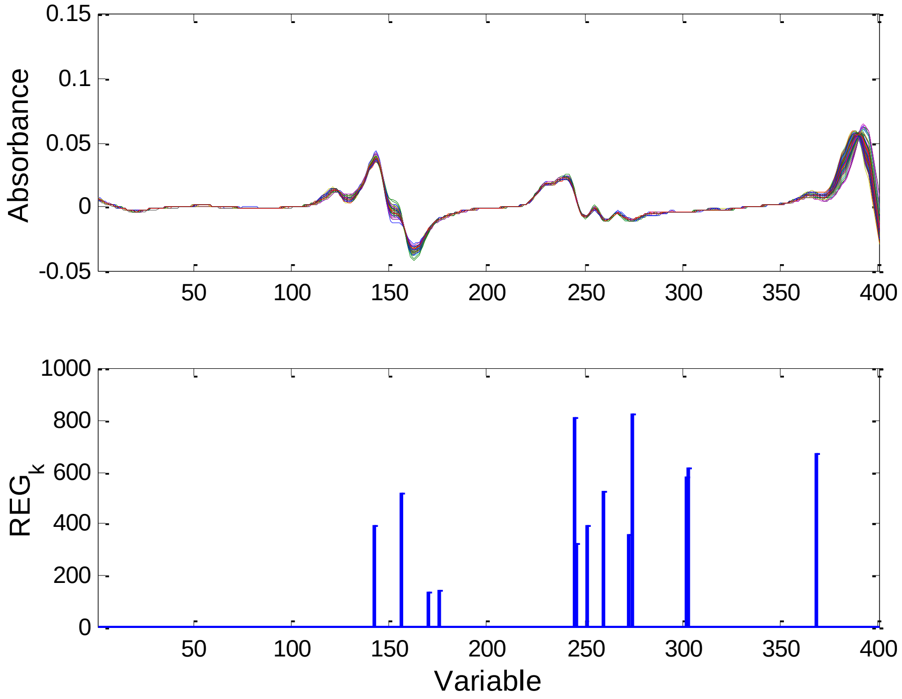


Fig. 5-5 Diesel data set with response viscosity: (top) spectra; (bottom) *REG$_k$* for the variables retained by methods SVR-1 and FCAM

5.6.3 Comparison of the predictive and selective performances of the methods

For all 16 **X-y** combinations, the predictive and selective performances of the variable reduction methods are compared with FCAM, because the latter often combines a low number of retained variables with a good predictive performance. Differences between pairs of methods are statistically tested, using the Wilcoxon signed rank test [50,61], for (*i*) RMSEP's of PLS1 models developed after variable reduction, to compare the predictive ability, and for (*ii*) numbers of retained variables, to compare the selective ability of the methods. In Table 5-4, two tailed *p* values of the test statistic are given for the pair wise comparison of method FCAM with the other methods.

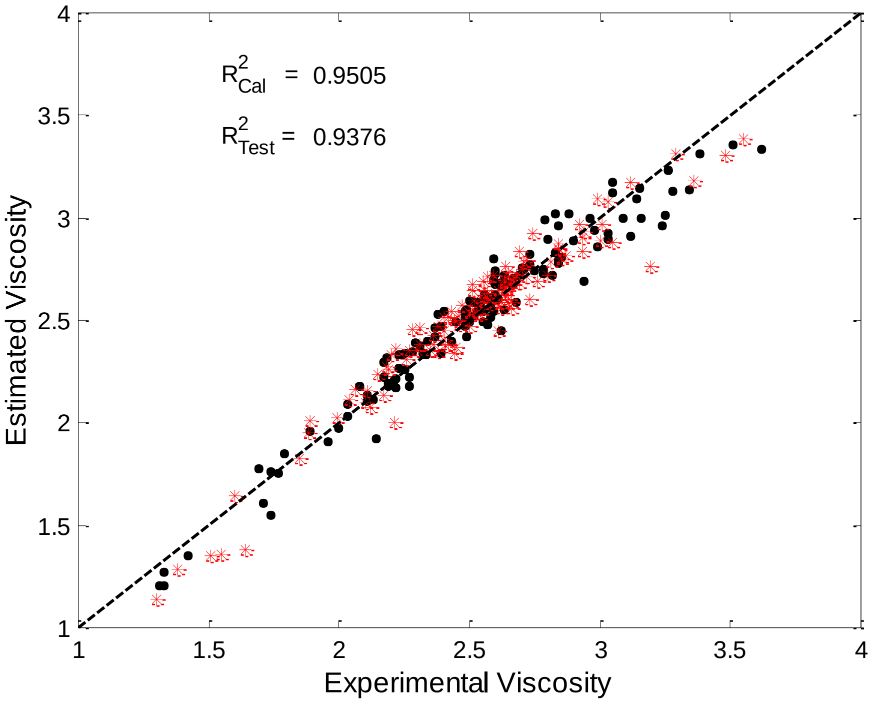


Fig. 5-6 Estimation of viscosity after variable reduction with methods SVR-1 and FCAM with the variable set corresponding to the best RMSECV for data set Diesel; (•) training set, (✳) test set

5.6.4 Comparison of predictive performances

The majority of the resulting models for all methods are better than the full spectrum models though this is data set dependent. Usually three groups are observed. For the responses 7 (Corn, moisture) and 11-16, of the simulated data set, the new models based on reduced variable sets result in large prediction improvement relative to the full spectrum method. For a second group, responses 1-6, i.e. the Diesel data set, mostly a similar performance is observed. Some other responses, 8-10 of the Corn data set (oil, protein, starch), show a large variability in improvement or worsening of their prediction, depending on the applied method.

Table 5-4 shows that the statistical tests, for differences in RMSEP's, confirm that the predictive capabilities of all methods are similar to those of method FCAM.

Table 5-4 Comparison of methods with FCAM by the Wilcoxon signed rank test

	Two-tailed probabilities p for test on differences in RMSEP	**Two-tailed probabilities p for test on differences in numbers of selected variables**
UVE-GA-PLS	0.163	**0.0002**
UVE-iPLS	0.762	**0.0292**
SVR-1	0.098	**0.0039**
SVR-2	0.088	**0.0004**
RCAM	0.320	0.109
ICAM	0.340	0.898

Significant two-tailed p-values (p<0.05) are in bold

Table 5-5 Number of random variables (x=101-200) retained in Simulated data set

Model	Mixtures	UVE-GA-PLS	UVE-iPLS	Method SVR-1	Method SVR-2	Method RCAM	Method FCAM	Method ICAM
11	A	2	2	6	6	0	0	0
12	A,B	2	2	1	9	0	0	0
13	A,D	0	3	0	4	0	0	0
14	A,B,C	1	0	0	28	0	0	0
15	A,B,D	0	1	0	3	0	0	0
16	A,B,C,D	0	0	0	3	0	0	0

5.6.5 Comparison of selective performances

Table 5-4 shows that the statistical tests for differences in the number of retained variables confirm that the simplicity of the final data sets of method FCAM are similar to those of methods RCAM and ICAM, but that the size of the final variable sets differ from those of the modified existing SVR methods (1 and 2) and of the reference methods UVE-GA-PLS and UVE-iPLS. Combined with the results in Table 5-1 - Table 5-3, it can be concluded that the CAM methods result in finding variable sets of a similar size, while the other methods retain more variables.

In method SVR-2, contrary to the other PPRV(R) methods, variables are ranked only once, after development of a PLS model based on the full spectrum. Method SVR-2 provides larger remaining data sets than the other PPRV(R) methods. Therefore, it seems that renewed ranking of variables after remodelling is beneficial for the selection of small variable sets with low RMSECV values and good predictive capability.

Both UVE and GAs show a large variability in variable selection [16,25]. Therefore, the number of variables selected by the hybrid methods UVE-GA-PLS and UVE-iPLS, is also variable. An advantage of the PPRV(R) methods is that there is no variability in the number of retained variables.

The three newly proposed CAM methods combine good selective and predictive abilities. They outperform methods SVR-1 and 2, UVE-GA-PLS and UVE-iPLS regarding the number of selected variables, while the corresponding RMSEP's are not significantly different. However, the methods RCAM and ICAM are computationally more intensive than FCAM. In RCAM the variable reduction procedure is repeated with stepwise descending complexities, and in ICAM two PLS models with complexities *A* and *A*-1 need to be calculated simultaneously. Of all seven methods, we consider method FCAM as the preferred variable reduction method, based on computational intensity, predictive and selective capabilities.

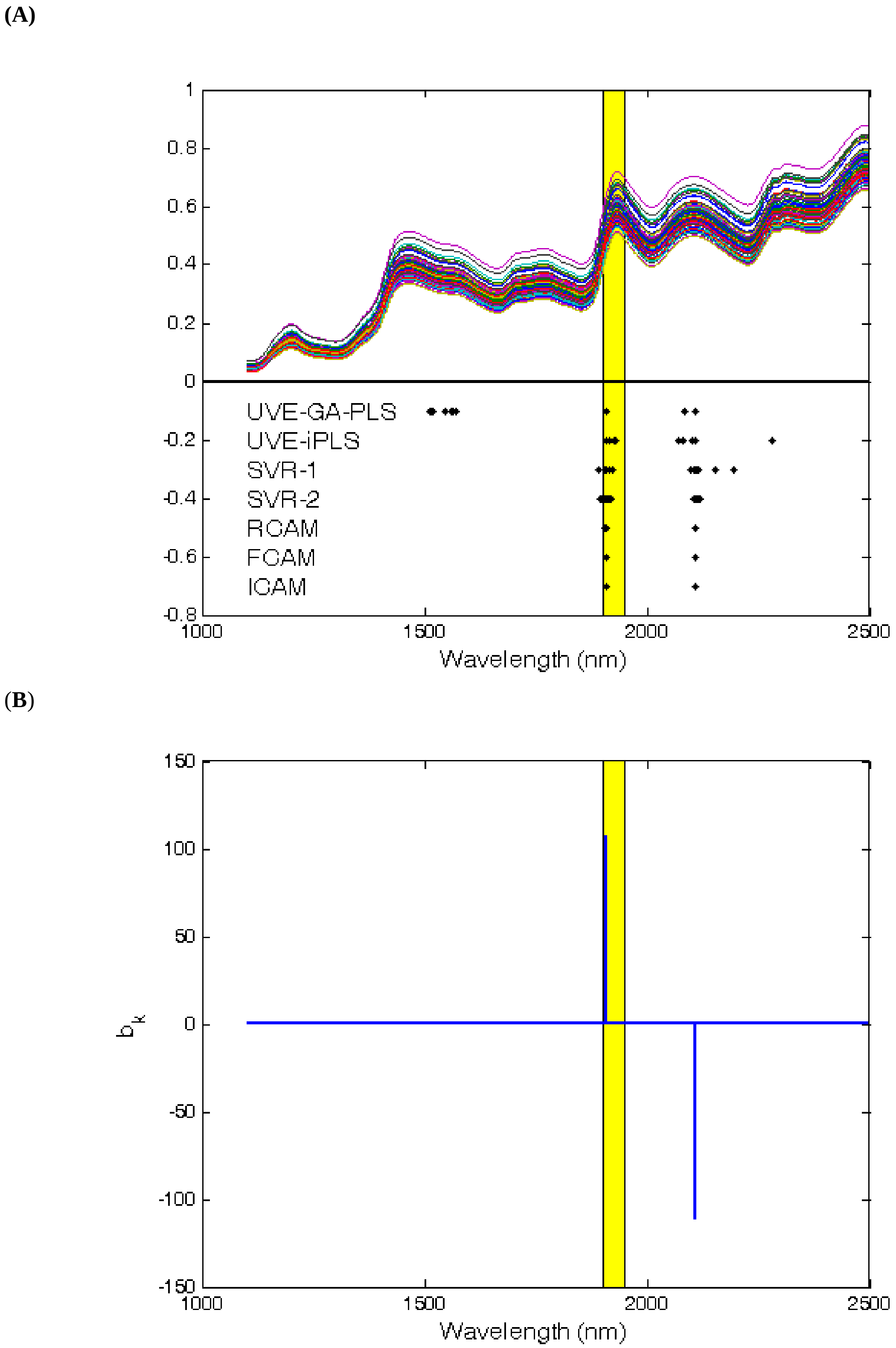


Fig. 5-7 Data set Corn, response moisture; (A) Spectra en selected wavelengths for all methods; (B) PLS regression coefficients of wavelengths selected by the FCAM method; the yellow column represents the water absorption band (see text)

5.6.6 Quality of selected variable sets

The ability of the methods to select predictors with a chemical meaning relevant to the response, is demonstrated for response 7 (Corn set, moisture) and response 11 (simulated set). This is not possible for the Diesel set because no wavelength information is provided.

Dry food samples such as corn show a strong absorption band for water near 1900 to 1950 nm which is often used for the quantitative analysis of water contents [62]. Fig. 5-7A shows the spectra, the selected wavelengths for all methods and the water absorption band. Both methods FCAM and ICAM are very selective because only two wavelengths at 1908 and 2108 nm are retained, with very good predictive properties (Table 5-2). Wavelength 1908 lies inside the water band and has a large positive regression coefficient, see Fig. 5-7B. Wavelength 2108, outside the water band, has a large negative regression coefficient and is probably due to an interferent. All other methods have these two key wavelengths in their selection, with large positive and negative regression coefficients. Other wavelengths are selected around 1908 and 2108 nm, with relatively low absolute regression coefficients.
An increased spread in the selected variables is observed for the methods SVR-1, UVE-iPLS and UVE-GA-PLS. Like FCAM and ICAM, method RCAM is also very selective, because only 4 variables are retained with very good predictive properties (Table 5-2).

In the simulated data set the selective abilities of the methods are investigated by using in all analyte mixtures as response vector the concentrations of analyte A, i.e. the substance with the narrowest Gaussian peak profile. Table 5-3 shows for the pure analyte A in model 11 that the three CAM methods are very selective, because sets with only 3 variables are selected with good predictivity, $R^2_{Test} = 1.000$.

Fig. 5-8 shows the simulated spectra and the selected wavelengths for all methods for response 11. The new CAM methods are very selective because only 3 informative variables (49, 50 and 51) are retained, below the top and the inflection points of the narrow Gaussian peak. All other methods are less selective because more variables are retained, both below the peak and inside the uninformative noise area of x=101-200.

In addition to that, for the six simulated mixtures, the new CAM-methods do not select uninformative random variables from the range x=101-200, while between 1 and 28 of these random variables are retained by the other methods, see Table 5-5.

It is concluded that the capability of the new CAM methods to select low numbers of informative variables is better than that of the other methods. It is also observed that, for the new CAM methods, important variables, with a chemical meaning relevant to the response, are not excluded in the stepwise backward variable selection procedures.

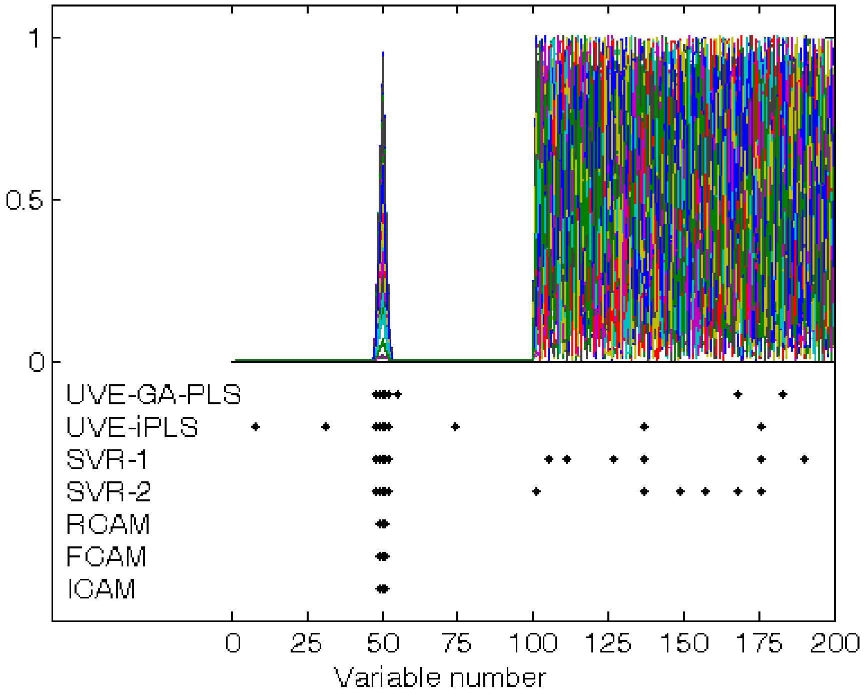


Fig. 5-8 Simulated set, model 11; Spectra and selected variables for all methods

5.7 Conclusions

The aim of this work was to investigate and test the predictive and selective abilities of three new stepwise variable reduction methods, using predictive-property-ranked variables. In the new CAM methods it is accounted for the fact that predictive-variable properties may change during the variable reduction process. A possibility for decreasing the PLS1 model complexity *A* is built in differently for each method. After variable reduction, *A* is determined again for the remaining sets. Therefore, a lot of flexibility is built in regarding the adaptation of model complexity *A*.

It has been demonstrated that the newly developed CAM methods are able to retain smaller numbers of variables by adapting the PLS model complexity with improved or similar predictability as the original data set. They provide significantly lower numbers of retained informative variables than the modifications of the existing methods, SVR-1 and 2, and the reference methods, UVE-GA-PLS and UVE-iPLS. Important variables, with a chemical meaning relevant to the response, are not excluded by the CAM methods in the stepwise backward variable reduction procedure.

Renewed ranking of variables, after deletion of a variable, followed by remodelling, is beneficial. The prediction abilities of all methods are similar. Contrary to UVE-GA-PLS and UVE-iPLS, there is no variability in the number of retained variables of the PPRV(R) methods.

The three PPRVR-CAM methods combine good selective and predictive abilities. Because the RCAM and ICAM method are computationally more intensive, FCAM is our preferred variable reduction method.

The results from this study indicate that variable reduction in PLS modelling can be improved by the application of the proposed new PPRVR-CAM methods.

Acknowledgements

Henk Claessens and Govert Somsen of the Avans Hogeschool are gratefully thanked for their managerial support of this project.

References

[1] S. Wold, M. Sjöström, L. Eriksson, Chemom. Intell. Lab. Syst. 58 (2001) 109.

[2] H. Martens, T. Næs, Multivariate Calibration, (2nd edn), Wiley, NewYork, 1993.

[3] M. Forina, S. Lanteri, M.C. Cerrato Oliveros, C. Pizarro Millan, Anal. Bioanal. Chem. 380 (2004) 397.

[4] M. Forina, S. Lanteri, M. Casale, J. Chromatogr. A. 1158 (2007) 61.

[5] C.H. Spiegelman, M.J. McShane, M.J. Goetz, M. Motamedi, Q.L. Yue, G.L. Coté, Anal. Chem. 70 (1998) 35.

[6] S.P. Reinikainen, A. Höskuldsson, J. Chemom. 17 (2003) 130.

[7] A. Höskuldsson, J. Chemom. 22 (2008) 150.

[8] L. Xu, I. Schechter, Anal. Chem. 68 (1996) 2392.

[9] B. Nadler, R.R. Coifman, J. Chemom. 19 (2005) 107.

[10] R.F. Teófilo, J.P.A. Martins, M.M.C. Ferreira, J. Chemom. 23 (2009) 32.

[11] J.A. Hageman, M. Streppel, R. Wehrens, L.M.C. Buydens, J. Chemom. 17 (2003) 427.

[12] A.S. Bangalore, R.E. Shaffer, G.W. Small, M.A. Amold, Anal. Chem. 68 (1996) 4200.

[13] A. Garrido Frenich, D. Jouan-Rimbaud, D.L. Massart, S. Kuttatharmmakul, M. Martinez Galera, J.L. Martinez Vidal, Analyst 120 (1995) 2787.

[14] H.J. Kubinyi, J Chemometr. 10 (1996) 119.

[15] W. Cai, Y. Li, X. Shao, Chemom. Intell. Lab. Syst. 90 (2008) 188.

[16] V. Centner, D.L. Massart, O.E. de Noord, S. de Jong, B.G.M. Vandeginste, C. Sterna, Anal. Chem. 68 (1996) 3851.

[17] T. Hancock, R. Put, D. Coomans, Y. Vander Heyden, Y. Everingham, Chemom. Intell. Lab. Syst. 76 (2005) 185.

[18] R. Put, Y. Vander Heyden, Proteomics 7 (2007) 1664.

[19] X. Shao, F. Wang, D. Chen, Q Su, Anal. Bioanal. Chem. 378 (2004) 1382.

[20] J. Moros, J. Kuligowski, G. Quintás, S. Garrigues, M. de la Guardia, Anal. Chim. Acta 630 (2008) 150.

[21] R. Leardi, A.L. Gonzalez, Chemom. Intell. Lab. Syst 41 (1998) 195.

[22] H.C. Goicoechea, A.C. Olivieri, J. Chem. Inf. Comput. Sci. 42 (2002) 1146.

[23] L. Norgaard, A. Saudland, J. Wagner, J.P. Nielsen, L. Munck, S.B. Engelsen, Appl. Spectrosc. 54 (2000) 413.

[24] R. Leardi, L. Norgaard, J. Chemom. 18 (2004) 486.

[25] J.P. Gauchi, P. Chagnon, Chemom. Intell. Lab. Syst. 58 (2001) 171.

[26] A. Lazraq, R. Cléroux, J.P. Gauchi, Chemom. Intell. Lab. Syst. 66 (2003) 117.

[27] H. Xu, Z. Liu, W. Cai, X. Shao, Chemom. Intell. Lab. Syst. 97 (2009) 189.

[28] S.A. Dodds, W.P. Heath, Chemom. Intell. Lab. Syst. 76 (2005) 37.

[29] M.J. Anzanello, S.L. Albin, W.A. Chaovalitwongse, Chemom. Intell. Lab. Syst. 97 (2009) 111.

[30] M. Forina, C. Casolino, C.P. Millán, J Chemometrics 13 (1999) 165.

[31] A. Höskuldsson, Chemom. Intell. Lab. Syst. 55 (2001) 23.

[32] S. Wold, E. Johansson, M. Cocchi, 3D QSAR in Drug Design; Theory, Methods, and Applications, ESCOM, Leiden, Holland, 1993.

[33] I.G. Chong, C.H. Jun, Chemom. Intell. Lab. Syst. 78 (2005) 103.

[34] R. Gosselin, D. Rodrigue, C. Duchesne, Chemom. Intell. Lab. Syst. 100 (2010) 12.

[35] F. Westad, H. Martens, J. Near Infrared Spectrosc. 8 (2000) 117.

[36] C. Abrahamsson, J. Johansson, A. Sparén, F. Lindgren, Chemom. Intell. Lab. Syst. 69 (2003) 3.

[37] T. Rajalahti, R. Arneberg, A.C. Kroksveen, M. Berle, K.M. Myhr, O.M. Kvalheim, Anal. Chem. 81 (2009) 2581.

[38] T. Rajalahti, R. Arneberg, F.S. Berven, K.M. Myhr, R.J. Ulvik, O.M. Kvalheim, Chemom. Intell. Lab. Syst. 95 (2009) 35.

[39] H. Swierenga, F. Wülfert, O.E. de Noord, A.P. de Weijer, A.K. Smilde, L.M.C. Buydens, Anal. Chim. Acta 411 (2000) 121.

[40] S. Caetano, C. Krier, M. Verleysen, Y. Vander Heyden, Anal. Chim. Acta 602 (2007) 37.

[41] D.A. Konovalov, N. Sim, E. Deconinck, Y. Vander Heyden, D. Coomans, J. Chem. Inf. Model. 48 (2008) 370.

[42] H. Li, Y. Liang, Q. Xu, D. Cao, Anal. Chim. Acta 648 (2009) 77.

[43] F. Westad, N.K. Afseth, R. Bro, Anal. Chim. Acta 595 (2007) 323.

[44] M. Blanco, I. Villarroya, Trends Anal. Chem. 21 (2002) 240.

[45] Z. Xiaobo, Z. Jiewen, M.J.W. Povey, M. Holmes, M. Hanpin, Anal. Chim. Acta 667 (2010) 14.

[46] P. Geladi, B.R. Kowalski, Anal. Chim. Acta 185 (1986) 1.

[47] B. Li, J. Morris, E.B. Martin, Chemom. Intell. Lab. Syst. 64 (2002) 79.

[48] S. Wold, Technometrics 24 (1978) 397.

[49] B.M. Wise, N.B. Gallagher, R.Bro, J.M. Shaver, W. Windig, R.Scott Koch, PLS_Toolbox Version 4.0, Eigenvector Research, Wenatchee.

[50] D.L. Massart, B.G.M. Vandeginste, L.M.C. Buydens, S. de Jong, P.J. Lewi, J. Smeyers-Verbeke, Handbook of Chemometrics and Qualimetrics, Part A, Elsevier, Amsterdam, 1997.

[51] J.A.F. Pierna, O. Abbas, V. Baeten, P. Dardenne, Anal. Chim. Acta 642 (2009) 89.

[52] R. Leardi, J. Chromatogr. A 1158 (2007) 226.

[53] C.B. Lucasius, G. Kateman, Chemom. Intell. Lab. Syst. 19 (1993) 1.

[54] C.B. Lucasius, G. Kateman, Chemom. Intell. Lab. Syst. 25 (1994) 99.

[55] R. Wehrens, L.M.C. Buydens, Trends Anal. Chem. 17 (1998) 193.

[56] R. Leardi, M.B. Seasholtz, R.J. Pell, Anal. Chim. Acta 461 (2002) 189.

[57] R.D. Snee, Technometrics 19 (1977) 415.

[58] http://www.mathworks.com/ (accessed on February 25, 2011).

[59] CHEMOAC Standard Function Toolbox, http://www.vub.ac.be/fabi/publiek/index.html (accessed on February 25, 2011).

[60] http://software.eigenvector.com/ (accessed on February 25, 2011).

[61] J. Nyström, P. Geladi, B. Lindholm-Sethson, J. Larson, A.C. Svensk, L. Franzén, Chemom. Intell. Lab. Syst. 90 (2008) 43.

[62] H. Büning-Pfaue, Food Chemistry 82 (2003) 107.

6 Predictive-Property-Ranked Variable Reduction in Partial Least Squares Modelling with Final Complexity Adapted Models: Comparison of Properties for Ranking[3]

6.1 Abstract

The calibration performance of Partial Least Squares regression for one response (PLS1) can be improved by eliminating uninformative variables. Many variable-reduction methods are based on so-called predictor-variable properties or predictive properties, which are functions of various PLS-model parameters, and which may change during the steps of the variable-reduction process. Recently, a new Predictive-Property-Ranked Variable Reduction method with Final Complexity Adapted Models, denoted as PPRVR-FCAM or simply FCAM, was introduced. It is a backward variable elimination method applied on the predictive-property-ranked variables. The variable number is first reduced, with constant PLS1 model complexity *A*, until *A* variables remain, followed by a further decrease in PLS complexity, allowing the final selection of small numbers of variables.

In this study for three data sets the utility and effectiveness of six individual and nine combined predictor-variable properties are investigated, when used in the FCAM method. The individual properties include the absolute value of the PLS1 regression coefficient (REG), the significance of the PLS1 regression coefficient (SIG), the norm of the loading weight vector (NLW), the variable importance in the projection (VIP), the selectivity ratio (SR), and the squared correlation coefficient of a predictor variable with the response **y** (COR). The selective and predictive performances of the models resulting from the use of these properties are statistically compared using the one-tailed Wilcoxon signed rank test.

The results indicate that the models, resulting from variable reduction with the FCAM method, using individual or combined properties, have similar or better predictive abilities than the full spectrum models. After mean-centring of the data, REG and SIG, provide low numbers of informative variables, with a meaning relevant to the response, and lower than the other individual properties, while the predictive abilities are similar or better. SIG has the best selective ability of all individual and combined properties, while the predictive ability is similar. REG is faster than SIG. This means that variable reduction with the FCAM method is preferably conducted with properties REG or SIG. The selective ability of REG can be improved by combining it with NLW or VIP.

Keywords: Variable reduction, PLS1, predictor-variable properties, PPRVR-FCAM

[3] **Jan P.M. Andries**, Yvan Vander Heyden, Lutgarde M.C. Buydens
Published in Analyica Chimica Acta 760 (2013) 34-45

6.2 Introduction

Partial Least Squares (PLS) is a commonly used multivariate regression technique, which is able to deal with a large number of noisy and correlated variables, and small numbers of samples [1-3]. However, both theoretical [4-8] and experimental evidence [3,9,10-16] exist that elimination of uninformative variables improves the performance of PLS calibration.

For PLS1, with one response y, many variable-elimination methods are based on so-called predictor-variable properties or predictive properties, which are functions of various model parameters [9-11,17,18]. In these methods variable reduction is made on the variables ranked in descending order of a given property. This ranking reflects their importance for the PLS1 model. The higher the magnitude of the property, the more important the variable.

In the Stepwise Variable Reduction methods using Predictive-Property-Ranked Variables, denoted as SVR-PPRV methods [9], iteratively, the variable with the smallest property value is eliminated and a new PLS1 model calculated. The predictive abilities of the PLS1 models are assessed by the root mean squared error of cross validation (RMSECV). The set of variables, resulting in the optimal model, is then selected. The goal is to obtain small sets of variables with improved or similar predictability, as the original data set. A low number of variables can be beneficial with regard to a better understanding of the model and selection of a viable set of sensors in process control.

Properties, such as weights, loadings and PLS regression coefficients, are functions of the parameters of the PLS1 algorithm, and are dependent on each other [1]. In the stepwise variable-reduction process the data matrix is changing continuously and therefore the parameters of the PLS algorithm can change. The optimal number of PLS factors, i.e. the best PLS model complexity, can change as well. If the same PLS model complexity is used during the variable reduction procedure, RMSECV values may become overoptimistic [19], since the best model complexity decreases due to the elimination of uninformative variables [20].

In a previous study [9] a new backward variable-reduction method was introduced, based on the variables ranked in descending order of a predictor-variable property. In this method, the fact that both the properties for the remaining variables and the best PLS1 model complexity may change during the variable-reduction process, is taken into account. The method was called Predictive-Property-Ranked Variable Reduction with Final Complexity Adapted Models, denoted as PPRVR-FCAM and abbreviated to FCAM. In the FCAM method, iteratively, the variable with the smallest property is eliminated, a new PLS model calculated, properties redetermined and variables reranked. In the final part of variable reduction, the model complexity is adapted to the number of remaining variables. The FCAM method combines good selective and predictive abilities because it is able to reduce to small numbers of variables with improved or similar predictability as the full spectrum model.

Common examples of predictive properties used for variable reduction are: (*i*) magnitude of PLS1 regression coefficients [11,14,16,17,21-24], (*ii*) magnitude of PLS regression coefficients multiplied [3,25] or divided [14] by the standard deviation of the predictor variable, (*iii*) correlation coefficients between predictor variables and the response [3,11,26], (*iv*) variable importance in the projection (VIP) score of a variable [9,11,18,21,24,27,28], (*v*) reliability, uncertainty or significance of PLS1 regression coefficients assessed by the Student t value, calculated from the ratio of the PLS1 regression coefficient and its standard deviation, estimated by jack knifing [3,16,20,29,30], (*vi*) selectivity ratio (SR) [9,31,32]. In [9], the absolute value of the PLS regression coefficients was used. To our knowledge these different properties have not yet been compared.

In the actual study the utility and effectiveness of six individual and nine combined properties are investigated when used in the FCAM method on near-infrared (NIR) spectra and on simulated data. NIR spectroscopy is chosen as application field because PLS1 is extensively used in this domain [33,34]. Two NIR and one simulated data set were investigated. The latter is used to test the general applicability of the selected property. The data sets contain a total of 16 responses (see Table 6-1-Table 6-6). With this high number of responses, more reliable results were obtained for the statistical tests, carried out for the comparison of the predictive and selective performance of the FCAM method when using different properties.

6.3 Theory

6.3.1 PLS model

The aim of PLS is to model the relationship between a data matrix **X** and a response vector **y** by using a set of latent variables that maximize the explained covariance between them. The PLS1 model for one response is developed from a calibration set of *N* objects or observations with one response or dependent variable in the **y** vector and *K* predictor variables in the **X** matrix. The **y**($N \times 1$) vector consist of the *N* responses of the observations denoted by y_i (i=1, …, N). The **X**($N \times K$) matrix consist of *K* column vectors of independent predictor variables denoted by $\mathbf{x}_k$ (k=1, …, K). The objective of PLS is to select the optimal number *A* ($A \leq K$) of latent variables or PLS factors, which are linear combinations of the original variables $\mathbf{x}_k$. The PLS model is given by Eqs. (**1**) and (**2**).

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{\mathbf{T}} + \mathbf{E}_{\mathbf{A}} \quad \mathbf{(1)}$$

$$\mathbf{y} = \mathbf{T}\mathbf{q}^{\mathbf{T}} + \mathbf{f}_{\mathbf{A}} \quad (\mathbf{2})$$

where **T**($N \times A$) is a score matrix, **P**($K \times A$) a matrix with the x-loading vectors $\mathbf{p}_a$ (a=1, 2, …, A) as columns, **q**($1 \times A$) the y-loading vector, $\mathbf{E_A}$ and $\mathbf{f_A}$ the residual matrix for **X** and the residual **y**-vector, respectively, after the extraction of *A* factors. The optimal number of PLS factors, *A*, can be determined using cross-validation (CV). Further details on PLS can be obtained in Refs. [1,2,35]. The model-dependent predictor-variable properties are calculated from various parameters of the PLS model.

6.3.2 Predictor-variable properties

In this section six individual predictor-variable properties, that were used for variable reduction in PLS1 modelling, are discussed. All variable properties, except for the correlation coefficients between predictor variables and the response, are dependent on the *A* factor PLS1 model. It is assumed that influential variables have high property values.

6.3.2.1 PLS regression coefficient (REG)

The variable reduction may be based on the PLS1 regression coefficients b_k, which are elements of the regression vector $\mathbf{b}(K \times 1)$, calculated with,

$$\mathbf{b} = \mathbf{W}\left(\mathbf{P}^T\mathbf{W}\right)^{-1}\mathbf{q} \tag{3}$$

where $\mathbf{W}(K \times A)$ is the **X** weight matrix, $\mathbf{P}(K \times A)$ the **X--**loading matrix and $\mathbf{q}(1 \times A)$ the y-loading vector [1]. The PLS1 regression coefficients b_k are interdependent unless A equals K [2]. Influential variables have large positive or negative regression coefficients. The absolute value of the PLS1 regression coefficient of variable k, denoted as REG_k, is used in this study as one of the predictor-variable properties for variable reduction.

$$REG_k = \left|b_k\right| \tag{4}$$

6.3.2.2 Significance of PLS regression coefficient (SIG)

Influential predictor variables have low uncertainties in the model parameters of multivariate regression models [29]. Therefore, the significance of the regression coefficients, and of the property REG_k will also be large. This significance can be estimated by jack-knifing.

The significance of the PLS regression coefficient b_k of variable k, denoted as SIG_k, is defined as the Student t value, calculated as

$$SIG_k = t_k = \left|\frac{b_k}{s_{b_k}}\right| \tag{5}$$

with t_k the Student t value for variable k, b_k the PLS regression coefficient of variable k calculated with Eq. (3), and s_{b_k} the standard deviation of the estimates of b_k calculated from n fold jack-knifing with Eq. (6).

$$s_{b_k} = \sqrt{\frac{n-1}{n}\sum_{j=1}^{n}\left(b_{k(-j)} - \bar{b}_{k(-j)}\right)^2} \tag{6}$$

where $b_{k(-j)}$ is the estimate of coefficient b_k based on the calibration with all objects, except for the objects in the left out segment j [36].

$$\bar{b}_{k(-j)} = \frac{\sum_{j=1}^{n} b_{k(-j)}}{n} \tag{7}$$

with $\bar{b}_{k(-j)}$ the mean of the $b_{k(-j)}$. Influential variables have large PLS regression coefficients and low standard deviations, and therefore large SIG_k values.

6.3.2.3 Norm loading weights (NLW)

In the PLS1 algorithm a loading weight vector $\mathbf{w}_a$ is sought which maximizes the covariance between the linear combination $\mathbf{X}_{a-1}\mathbf{w}_a$ and the response vector $\mathbf{y}$ under the constraint $\mathbf{w}_a^T\mathbf{w}_a=1$ [1]. The influence of a variable k of matrix $\mathbf{X}$ on the a^{th} PLS factor in the model is determined by the value of the k^{th} element in the loading weight vector $\mathbf{w}_a$, w_{ka}, and is considered to be large if the loading weight w_{ka} is large [2]. Large loading weights w_{ka} of variable k for the A PLS factors will result in a high norm of the loading weight vector $\mathbf{w}_k$. The norm of the loading weight vector $\mathbf{w}_k$, NLW_k, with weights of variable k on each of the A PLS factors in the model is defined by

$$NLW_k = \sqrt{\sum_{a=1}^{A} w_{ka}^2} \qquad \textbf{(8)}$$

A scaled version of NLW is used for variable selection in [24].

6.3.2.4 Variable importance in the projection (VIP)

The variable importance in the projection (VIP) score was first published in [27]. VIP is a measure for the importance of a predictor variable for both $\mathbf{X}$ and $\mathbf{y}$ [2]. For each variable k a weighted sum of y-variance, VIP_k, is calculated [28], applying $\|\mathbf{w_k}\| = 1$, with:

$$VIP_k = \sqrt{\frac{K\sum_{a=1}^{A} w_{ka}^2 q_a^2 \mathbf{t}_a^T \mathbf{t}_a}{\sum_{a=1}^{A} q_a^2 \mathbf{t}_a^T \mathbf{t}_a}} \qquad \textbf{(9)}$$

where K is the number of predictor variables, q_a the a-the element of the y-loading vector $\mathbf{q}(1 \times A)$, $\mathbf{t}_a$ the a-th column vector of score matrix $\mathbf{T}$. VIP_k weighs the contribution of each variable k according to the variance explained by each PLS component [28]. VIP_k is large for influential variables. The criterion $VIP_k>1$ for influential variables is often used for variable selection [18,28]. However, in this study, variables are ranked in descending order of VIP scores and the threshold is not used.

6.3.2.5 Selectivity ratio (SR)

The selectivity ratio of predictor variable k can be calculated after developing an A factor PLS model, after reconstruction of the $\mathbf{X}$ matrix by

$$\hat{\mathbf{X}}_A = \mathbf{T}\mathbf{P}^{\mathrm{T}} \qquad \textbf{(10)}$$

From Eqs. (1) and (10) it follows that for the residual $\mathbf{X}$ matrix $\mathbf{E}_A$ holds that

$$\mathbf{E}_A = \mathbf{X} - \hat{\mathbf{X}}_A \qquad \textbf{(11)}$$

The explained variance of predictor variable *k*, $s^2_{\mathrm{expl},k}$, in the reconstructed $\hat{\mathbf{X}}_A$ matrix is

$$s^2_{\mathrm{expl},k} = \frac{\sum_{i=1}^{N}\left(x_{A,ik} - \bar{x}_{A,k}\right)^2}{N-1} \tag{12}$$

where x_A are the elements of the reconstructed $\hat{\mathbf{X}}_A$ matrix.
The residual variance of predictor variable *k* is calculated from the residual matrix $\mathbf{E}_\mathrm{A}$ (Eq. (11)) with

$$s^2_{\mathrm{res},k} = \frac{\sum_{i=1}^{N}\left(e_{A,ik} - \bar{e}_{A,k}\right)^2}{N-1} \tag{13}$$

where e_A are the elements of the residual matrix $\mathbf{E}_\mathrm{A}$.

The selectivity ratio SR_k is defined in Ref. [31] as the ratio of the explained variance from Eq. (12) to the residual variance from Eq. (13):

$$SR_k = \frac{s^2_{\mathrm{expl},k}}{s^2_{res,k}} \tag{14}$$

6.3.2.6 Squared correlation coefficient between variables of X and y (COR)

The squared correlation coefficient of predictor variable $\mathbf{x}_\mathrm{k}$ with the response **y,** R_k^2**,** is considered as a measure for influential variables. R_k is calculated with

$$R_k = \frac{\sum_{i=1}^{N}\left(x_{ik} - \bar{x}_k\right)\left(y_i - \bar{y}\right)}{\left(N-1\right)\sqrt{s_k s_y}} \tag{15}$$

where x_{ik} is the i^{th} value of variable *k*, $\bar{x}_k$ the mean of $\mathbf{x}_\mathrm{k}$, y_i the i^{th} response, $\bar{y}$ the mean of **y**, s_k and s_y the standard deviations of $\mathbf{x}_\mathrm{k}$ and **y,** respectively. R_k^2 is denoted as COR_k. COR_k is a model independent variable property [3,11,26].

6.3.2.7 Combinations of predictor-variable properties

Combinations of properties can also be used for variable reduction. Wold et al. [27] recommend a combination of REG and VIP which states that both should be small for a variable to be excluded. In [18] it was observed that REG and VIP might be complementary and in [11] products of absolute values of predictor variable properties were used for variable selection.

Table 6-1 Results of variable reduction using individual predictive variable properties for the Diesel data set. Abbreviations: see text.

Model	Response	Method characteristics	Full spectrum	Predictor-variable properties					
				REG	SIG	NLW	VIP	SR	COR
1	**Viscosity**	PLS complexity	11	7	5	13	14	12	11
		Number of variables, K_{Best}	401	13	7	59	46	141	353
		RMSEP	0.102	0.099	0.117	0.103	0.099	0.093	0.101
		R^2_{Test}	0.934	0.938	0.914	0.932	0.937	0.944	0.935
2	**BP50**	PLS complexity	11	7	6	11	13	11	11
		Number of variables, K_{Best}	401	11	8	266	87	285	326
		RMSEP	3.605	3.485	3.989	3.656	3.264	3.693	3.749
		R^2_{Test}	0.956	0.959	0.946	0.954	0.965	0.953	0.952
3	**CN**	PLS complexity	5	4	4	4	5	5	13
		Number of variables, K_{Best}	401	4	4	5	21	13	258
		RMSEP	2.106	2.076	2.103	2.191	2.174	2.179	2.001
		R^2_{Test}	0.654	0.661	0.655	0.624	0.629	0.632	0.691
4	**D4052**	PLS complexity	15	10	12	15	15	15	15
		Number of variables, K_{Best}	401	17	13	111	134	99	326
		RMSEP	$9.20{\cdot}10^{-4}$	$1.07{\cdot}10^{-3}$	$1.14{\cdot}10^{-3}$	$9.32{\cdot}10^{-4}$	$9.10{\cdot}10^{-4}$	$1.14{\cdot}10^{-3}$	$9.23{\cdot}10^{-4}$
		R^2_{Test}	0.991	0.989	0.987	0.991	0.992	0.987	0.991
5	**Freeze**	PLS complexity	9	7	4	10	8	9	8
		Number of variables, K_{Best}	401	7	6	33	25	256	278
		RMSEP	2.490	2.685	2.532	2.599	2.735	2.438	2.638
		R^2_{Test}	0.624	0.570	0.623	0.590	0.549	0.642	0.581
6	**Total**	PLS complexity	14	9	9	15	13	15	13
		Number of variables, K_{Best}	401	15	12	90	47	72	64
		RMSEP	0.592	0.617	0.622	0.577	0.664	0.616	0.633
		R^2_{Test}	0.991	0.990	0.990	0.991	0.988	0.990	0.989

The combined predictor-variable properties were made by taking the unweighted sum of two individual variable properties. Because of the good selective abilities of REG and SIG found in this study (see section 6.7.1), only combinations of REG or SIG with the other individual properties were investigated. Combined properties are denoted with a plus sign between the individual properties.

6.3.3 Model validation and selection criterion for the preferred variable set

The predictive ability of the models is both assessed by internal validation in the training set, using segmented (*n*-fold) cross validation, and external validation with a test set, resulting in the root mean squared error of cross validation (RMSECV) and the root mean squared error of prediction (RMSEP), respectively. After variable reduction, using the reduced variable set, the best PLS model complexity is redetermined by segmented cross validation (SCV), which is then used for the external validation. The best complexity of a PLS model is determined by SCV [9]. In order to avoid overfitting an adjusted Wold's R criterion, $R_{adj} < 0.98$, is applied [37-39].

RMSECV values are plotted as a function of the number of remaining variables. The model with the global minimal value, $RMSECV_{Min}$, corresponds to the variable set with optimal predictive capability. However, a smaller variable set, with K_{Best} variables, and with *RMSECV* not significantly higher than that corresponding to the global minimal value, $RMSECV_{Min}$, and smaller than or equal to the *RMSECV* of the full spectrum (FS) model, $RMSECV_{FS}$, is selected as the best set [9].

For prediction with a test set, squared values of the correlation coefficient between estimated and experimental properties (R^2_{Test}) are calculated with the retained variable sets, using the model complexity, redetermined after variable reduction.

Further details about model validation and the selection of the preferred variable set are described in Ref. [9].

6.4 FCAM method

The FCAM method is a backward stepwise variable-reduction method based on predictive-property-ranked variables. Variables are reduced with constant PLS1 model complexity *A* until *A* variables remain. Then, the PLS model complexity is stepwise decreased, *A*-1, *A*-2, …,1, after each removal of a variable, allowing reduction to small numbers of variables.
The method consists of four steps. First, the data set is split into a training and a test set. The predictive ability of the full spectrum PLS1 models is assessed by internal validation with the training set, using SCV. The optimal number of PLS1 factors *A*, is determined by the application of the adjusted Wold's R criterion R_{adj} <0.98, see section 6.3.3. Based on the *A* factor PLS1 model, a given property is calculated for all variables and ranked.
In step 2, iteratively, the variable with the smallest property value is eliminated, a new PLS1 model, *RMSECV*, and new property values calculated, and variables reranked. When the number of remaining variables is *A*, the model complexity is decreased by one. Step 2 is repeated until the number of remaining variables and the PLS model complexity equals 1.

Table 6-2 **Results of variable reduction using combined predictive variable properties for the Diesel data set. Abbreviations: see text.**

Model	**Response**	**Method characteristics**	**Predictor-variable properties**								
			REG+ SIG	**REG+ NLW**	**REG+ VIP**	**REG+ SR**	**REG+ COR**	**SIG+ NLW**	**SIG+ VIP**	**SIG+ SR**	**SIG+ COR**
1	**Viscosity**	PLS complexity selecting best set	5	7	5	12	7	7	7	12	6
		Number of variables, K_{Best}	10	7	8	83	13	8	8	140	9
		RMSEP	0.107	0.100	0.110	0.106	0.099	0.110	0.133	0.093	0.109
		R^2_{Test}	0.926	0.935	0.925	0.927	0.938	0.922	0.889	0.944	0.925
2	**BP50**	PLS complexity selecting best set	7	7	7	8	7	7	8	11	8
		Number of variables, K_{Best}	11	11	11	27	11	8	24	290	9
		RMSEP	3.485	3.485	3.485	3.939	3.485	3.771	3.972	3.673	3.595
		R^2_{Test}	0.959	0.959	0.959	0.947	0.959	0.951	0.946	0.954	0.957
3	**CN**	PLS complexity selecting best set	4	4	4	5	4	4	4	5	2
		Number of variables, K_{Best}	4	4	4	13	4	4	4	13	3
		RMSEP	2.076	2.076	2.076	2.179	2.076	2.110	2.103	2.179	2.229
		R^2_{Test}	0.661	0.661	0.661	0.632	0.661	0.653	0.655	0.632	0.611
4	**D4052**	PLS complexity selecting best set	11	11	12	15	11	12	11	15	11
		Number of variables, K_{Best}	13	18	18	99	16	13	13	99	12
		RMSEP	$1.06\cdot10^{-3}$	$1.06\cdot10^{-3}$	$1.09\cdot10^{-3}$	$1.14\cdot10^{-3}$	$1.10\cdot10^{-3}$	$1.14\cdot10^{-3}$	$1.12\cdot10^{-3}$	$1.14\cdot10^{-3}$	$1.13\cdot10^{-3}$
		R^2_{Test}	0.989	0.989	0.988	0.987	0.988	0.987	0.987	0.987	0.987
5	**Freeze**	PLS complexity selecting best set	6	6	6	8	7	4	5	9	4
		Number of variables, K_{Best}	6	6	6	24	7	6	7	167	6
		RMSEP	2.672	2.672	2.672	2.551	2.685	2.454	2.469	2.492	2.445
		R^2_{Test}	0.576	0.576	0.576	0.609	0.570	0.648	0.642	0.627	0.660
6	**Total**	PLS complexity selecting best set	9	9	9	13	19	10	10	15	10
		Number of variables, K_{Best}	15	15	15	58	15	11	10	72	11
		RMSEP	0.617	0.617	0.617	0.657	0.617	0.646	0.596	0.616	0.641
		R^2_{Test}	0.990	0.990	0.990	0.988	0.990	0.989	0.990	0.990	0.989

In the third step, $RMSECV_{Best}$ is determined (see [9]) and the corresponding subset of variables selected. In the fourth step, using the reduced variable set, the PLS model is externally validated (RMSEP) using a test set, after a renewed determination of the optimal number of PLS factors *A* by SCV and the application of the criterion R_{adj} <0.98.
To illustrate the results of steps 1 and 2, Fig. 6-1 shows the RMSECV curve and the PLS model complexity as a function of the number of remaining variables for the FCAM method on one of the studied data sets (Corn set, response moisture) with REG as predictive variable property.

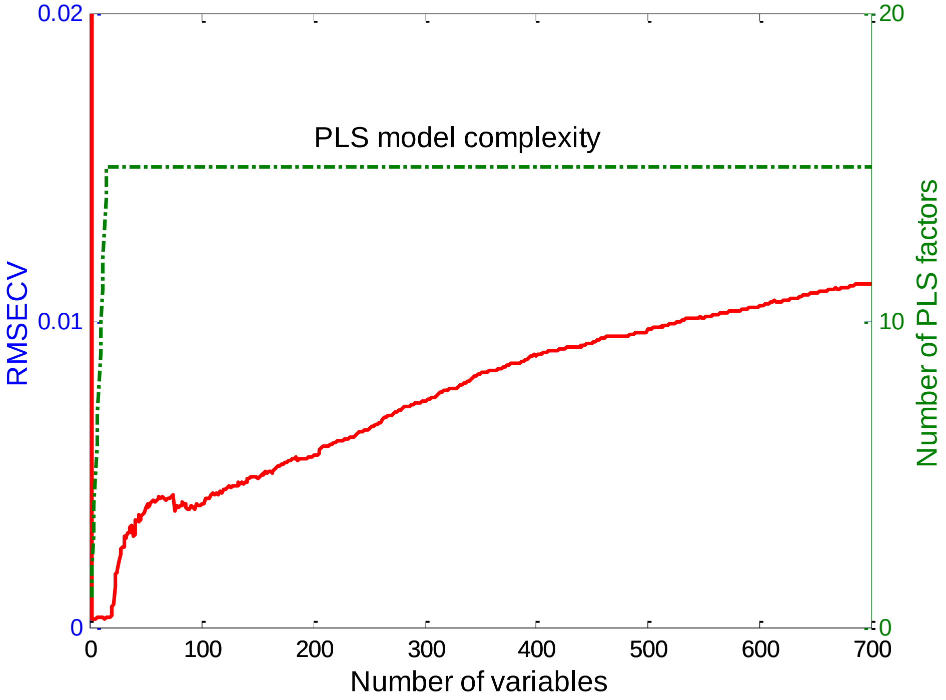


Fig. 6-1 RMSECV curve and PLS model complexity as a function of the number of remaining variables for the FCAM method using REG for response moisture of the Corn data set; — RMSECV-curve

6.5 Wilcoxon signed rank test

The results of the predictor-variable properties are compared using the one-tailed Wilcoxon signed rank test. This is a robust and sensitive non-parametric statistical test for two groups of paired samples. It is used just as the paired t-test, without any distributional assumptions [40-42]. The null hypothesis is accepted if the originating populations of the paired samples have the same median. Both the direction and the magnitude of the difference between the results of two methods for each subset are considered in the test.

The absolute differences $|d_i|$, between results of paired samples for two properties, are given a rank R_i in ascending order. Thereafter, each rank R_i is attributed with the same sign as the original difference d_i, and the sum of all positive ranks T_+ and of all negative ranks T_- is determined. The minimum of T_+ and T_- is the test statistic. The test statistic is small if there is no true difference between the two paired samples. For a one-tailed Wilcoxon signed rank test, the direction of the differences between the paired samples is determined from the maximum of T_+ and T_- and the one-tailed probability *p* of the test statistic is calculated [42].

Table 6-3 Results of variable reduction using individual predictive variable properties for the Corn data set. Abbreviations: see text.

Model	Response	Method characteristics	Full	Predictor-variable properties					
			spectrum	REG	SIG	NLW	VIP	SR	COR
7	Moisture	PLS complexity	15	2	2	12	15	15	15
		Number of variables, K_{Best}	700	2	2	12	113	51	376
		RMSEP	$1.19\cdot10^{-2}$	$3.00\cdot10^{-4}$	$3.00\cdot10^{-4}$	$3.08\cdot10^{-4}$	$3.55\cdot10^{-3}$	$2.02\cdot10^{-3}$	$6.56\cdot10^{-3}$
		R^2_{Test}	0.999	>0.9995	>0.9995	>0.9995	>0.9995	>0.9995	>0.9995
8	Oil	PLS complexity	11	10	7	11	11	10	9
		Number of variables, K_{Best}	700	18	7	12	29	11	110
		RMSEP	0.060	0.021	0.017	0.078	0.070	0.056	0.085
		R^2_{Test}	0.869	0.983	0.990	0.789	0.850	0.879	0.728
9	Protein	PLS complexity	14	12	11	14	14	14	15
		Number of variables, K_{Best}	700	28	11	16	582	20	238
		RMSEP	0.090	0.071	0.071	0.094	0.082	0.065	0.092
		R^2_{Test}	0.968	0.982	0.984	0.966	0.974	0.985	0.971
10	Starch	PLS complexity	15	10	8	14	15	15	15
		Number of variables, K_{Best}	700	26	8	14	109	44	691
		RMSEP	0.170	0.125	0.129	0.235	0.204	0.100	0.170
		R^2_{Test}	0.962	0.979	0.978	0.927	0.946	0.987	0.962

Table 6-4 Results of variable reduction using combined predictive variable properties for the Corn data set. Abbreviations: see text.

Model	Response	Method characteristics	Predictor-variable properties								
			REG+ SIG	REG+ NLW	REG+ VIP	REG+ SR	REG+ COR	SIG+ NLW	SIG+ VIP	SIG+ SR	SIG+ COR
7	Moisture	PLS complexity selecting best set	2	2	2	15	2	2	2	15	2
		Number of variables, K_{Best}	2	2	2	51	2	2	2	51	2
		RMSEP	$3.00\cdot10^{-4}$	$3.00\cdot10^{-4}$	$3.00\cdot10^{-4}$	$2.02\cdot10^{-3}$	$3.00\cdot10^{-4}$	$3.00\cdot10^{-4}$	$3.00\cdot10^{-4}$	$2.02\cdot10^{-3}$	$3.00\cdot10^{-4}$
		R^2_{Test}	>0.9995	>0.9995	>0.9995	>0.9995	>0.9995	>0.9995	>0.9995	>0.9995	>0.9995
8	Oil	PLS complexity selecting best set	11	10	9	10	10	7	4	10	8
		Number of variables, K_{Best}	23	14	15	11	15	7	4	11	8
		RMSEP	0.018	0.022	0.024	0.056	0.023	0.075	0.087	0.056	0.018
		R^2_{Test}	0.987	0.983	0.978	0.879	0.981	0.816	0.757	0.879	0.988
9	Protein	PLS complexity selecting best set	13	12	12	14	11	10	11	14	10
		Number of variables, K_{Best}	24	26	23	20	18	13	18	20	11
		RMSEP	0.046	0.075	0.066	0.066	0.064	0.067	0.073	0.066	0.073
		R^2_{Test}	0.993	0.983	0.985	0.985	0.987	0.985	0.981	0.985	0.982
10	Starch	PLS complexity selecting best set	10	14	10	10	15	13	13	15	9
		Number of variables, K_{Best}	16	46	16	44	26	13	13	44	9
		RMSEP	0.120	0.131	0.105	0.100	0.125	0.121	0.153	0.100	0.108
		R^2_{Test}	0.984	0.978	0.987	0.987	0.979	0.980	0.969	0.987	0.985

6.6 Data and methodology

6.6.1 Data sets

Three data sets were investigated. The first data set is the Diesel set from Eigenvector Research http://software.eigenvector.com/. The Diesel set consists of first derivative NIR data at 401 wavelengths (no wavelengths provided), of 252 diesel samples, and six physical properties as responses. The physical properties viscosity (Visc), boiling point (BP50), cetane number (CN), density (D4052), freezing temperature (Freeze) and total aromatics (Total) are the responses. The set was split into a training and a test set with 136 and 116 samples, respectively. Ten-fold cross validation is conducted during model building.

The second data set is the Corn set from Eigenvector Research, consisting of NIR spectra of 80 corn samples from the "m5" spectrometer with a wavelength range of 1100–2498 nm at 2 nm intervals, resulting in 700 predictor variables. Moisture, oil, protein and starch contents of the samples are the responses. The Corn set was split into a training and a test set using the duplex method [43], with 60 and 20 samples, respectively. Eight fold cross validation is conducted during model building.

The third simulated data set consists of six subsets. They represent the spectra or chromatograms of mixtures with one to four compounds (A, B, C or D), see Table 6-5 and Table 6-6. The response vector **y** contains the concentrations of compound A. The pure spectral/chromatographic profiles of the compounds were Gaussian peaks g(μ,σ), with mean μ and standard deviation σ, i.e. $g_A(50,1)$, $g_B(41,4)$, $g_C(59,4)$ and $g_D(50,15)$ for the respective components. The maximum heights of the Gaussian peaks are 1 for compounds A, B and C, and 0.5 for compound D. The first 100 variables in these profiles are informative, with *x* values used for the calculation of the analyte profiles in the mixtures. The last 100 variables are uninformative, consisting of random numbers between 0 and 1. These uninformative variables have a high signal level, comparable to that of the informative variables in the range x=1-100. This is to investigate if the FCAM method is capable to find informative variables with a chemical meaning in profiles containing many uninformative variables at a similar signal level. The simulated subsets were split using the duplex method, into a training and a test set with 100 and 20 samples, respectively, and ten-fold cross validation is conducted during model building.

Further details about the data sets are described in Ref. [9].

6.6.2 Methodology

The data sets contain a total of 16 **X**-**y** combinations. The variables and responses of all **X**-**y** combinations are pre-processed by mean-centring. This may affect the results of the variable reduction applied [10]. For each **X**-**y** combination, the FCAM method is applied using one of the individual or combined properties. The numbers of retained variables and the resulting RMSEP's are used to investigate the effectiveness of these properties. The numbers of retained variables are compared using box plots. Pairwise differences in numbers of retained variables and in RMSEP's are statistically tested, using the one-tailed Wilcoxon signed rank test. The ability of the FCAM method to retain variables with a chemical meaning, using the applied variable properties, is also investigated.

6.6.3 Software

All calculations are made with in-house programs developed in Matlab (V. 6.5) (The Math Works, Natick, MA, USA) (http://www.mathworks.com/). The procedure for the duplex splitting algorithm is from ChemoAC Standard Functions Toolbox for MATLAB (http://www.vub.ac.be/fabi/publiek/index.html). Statistical tests are conducted with SPSS V20 (http://www-01.ibm.com/software/analytics/spss/products/statistics).

6.7 Results and discussion

First, for each of the 16 responses (Table 6-1,3,5) a PLS1 model is developed. The optimal model complexities were determined for the full spectra, by segmented cross validation, and *RMSECV* and *RMSEP* calculated for the full spectrum models. Variable reduction is then applied on the **X**-**y** sets by the FCAM method, considering an individual or combined property, using the optimal PLS model complexity determined for the full spectrum. After variable reduction the optimal PLS model complexity is redetermined for the remaining best variable set.

In Table 6-1, 3 and 5, for the full spectrum models, the optimal PLS complexity, the number of variables, *RMSEP* and the squared correlation coefficient for prediction with the test set, R^2_{Test} are given. For each individual and combined property, the redetermined optimal PLS complexity, the number of remaining variables K_{Best}, *RMSEP* and R^2_{Test}, are shown in Table 6-1, 3, 5 and Table 6-2, 4, 6 respectively. For the simulated sets, the relatively high PLS model complexities of the full spectrum models are noteworthy. These complexities are strongly increased by the addition of the 100 uninformative variables to the informative variables at the applied high signal level, to compensate for change correlations with the response.

In this section, for the 16 **X**-**y** combinations, the predictive and selective performances of the models resulting from variable reduction using individual or combined predictor-variable properties are compared with those resulting from REG or SIG.
It is statistical tested if (*i*) RMSEP's of the PLS models after variable reduction, and (*ii*) numbers of retained variables K_{Best} , are significantly lower than those resulting from REG or SIG. In addition, it is also tested if RMSEP's of the PLS models developed for the retained variable sets are significantly lower than those of the full spectrum models. The RMSEP's and the numbers of retained variables are compared to test the predictive and selective performances, respectively. Therefore, pairwise differences for RMSEP's and numbers of remaining variables are statistically tested, using a one-tailed Wilcoxon signed rank test.
The tests are conducted such that differences d_{ij} between RMSEP's or numbers of remaining variables of paired samples, are calculated as $d_{ij}=h_{property\,j}-h_{property\,i}$; h=RMSEP or K_{Best}, *i* refers to a property in the first column of Table 6-7, and *j* to the full spectrum, REG or SIG. This results in negative differences if $RMSEP_j$ or K_{Best_j} is lower than the equivalent i property. Then, the sum of the negative ranks T_- will be larger than the sum of the positive ranks T_+.

Table 6-5 Results of variable reduction using individual predictive variable properties for the Simulated data set. Abbreviations: see text.

Model	**Mixtures**	**Method characteristics**	**Full**	**Predictor-variable properties**					
			spectrum	**REG**	**SIG**	**NLW**	**VIP**	**SR**	**COR**
11	**A**	PLS complexity	11	3	2	6	3	11	11
		Number of variables, K_{Best}	200	3	2	6	3	11	11
		RMSEP	0.069	$1.12\cdot10^{-3}$	$1.23\cdot10^{-3}$	$1.49\cdot10^{-3}$	$1.12\cdot10^{-3}$	$1.25\cdot10^{-3}$	$1.33\cdot10^{-3}$
		R^2_{Test}	0.957	>0.9995	>0.9995	>0.9995	>0.9995	>0.9995	>0.9995
12	**A,B**	PLS complexity	14	2	2	13	5	8	2
		Number of variables, K_{Best}	200	4	4	32	7	20	2
		RMSEP	0.052	$1.33\cdot10^{-3}$	$1.35\cdot10^{-3}$	$1.62\cdot10^{-3}$	$1.33\cdot10^{-3}$	$1.38\cdot10^{-3}$	$1.57\cdot10^{-2}$
		R^2_{Test}	0.964	>0.9995	>0.9995	>0.9995	>0.9995	>0.9995	>0.9995
13	**A,D**	PLS complexity	12	2	5	9	2	12	2
		Number of variables, K_{Best}	200	7	14	18	6	17	2
		RMSEP	0.081	$1.59\cdot10^{-3}$	$1.66\cdot10^{-3}$	$1.88\cdot10^{-3}$	$1.65\cdot10^{-3}$	$1.25\cdot10^{-3}$	$4.24\cdot10^{-3}$
		R^2_{Test}	0.932	>0.9995	>0.9995	>0.9995	>0.9995	>0.9995	>0.9995
14	**A,B,C**	PLS complexity	13	3	3	9	10	9	12
		Number of variables, K_{Best}	200	5	5	21	31	29	71
		RMSEP	0.042	$1.21\cdot10^{-3}$	$1.33\cdot10^{-3}$	$2.40\cdot10^{-3}$	$1.35\cdot10^{-3}$	$1.25\cdot10^{-3}$	$1.62\cdot10^{-3}$
		R^2_{Test}	0.986	>0.9995	>0.9995	>0.9995	>0.9995	>0.9995	>0.9995
15	**A,B,D**	PLS complexity	16	3	5	7	4	10	10
		Number of variables, K_{Best}	200	11	11	8	11	25	16
		RMSEP	0.050	$1.67\cdot10^{-3}$	$1.73\cdot10^{-3}$	$2.26\cdot10^{-3}$	$1.97\cdot10^{-3}$	$1.72\cdot10^{-3}$	$2.24\cdot10^{-3}$
		R^2_{Test}	0.974	>0.9995	>0.9995	>0.9995	>0.9995	>0.9995	>0.9995
16	**A,B,C,D**	PLS complexity	15	4	7	11	7	9	15
		Number of variables, K_{Best}	200	11	16	18	11	48	35
		RMSEP	0.085	$1.84\cdot10^{-3}$	$1.67\cdot10^{-3}$	$1.87\cdot10^{-3}$	$1.99\cdot10^{-3}$	$1.47\cdot10^{-3}$	$3.50\cdot10^{-3}$
		R^2_{Test}	0.935	>0.9995	>0.9995	>0.9995	>0.9995	>0.9995	>0.9995

In Table 6-7, the probabilities *p* and the direction of differences are given, for the one-tailed Wilcoxon signed rank test. $T_- > T_+$ is indicated by (T_-) and $T_+ > T_-$ by (T_+). The results are discussed below. In the comparison of properties, significantly lower results are found for the properties j than for i if p<0.05, and $T_- > T_+$. The results for the properties i are significantly lower than those for j if p<0.05, and $T_+ > T_-$.

6.7.1 Comparison of the individual properties

To compare the selective performances of the FCAM method with all investigated predictive-variable properties, box plots are made for the numbers of retained variables (see Fig. 6-2). The box plots in Fig. 6-2A show much smaller numbers of retained variables for REG and SIG than for the other individual properties.

The Wilcoxon signed rank test (Table 6-7) shows that the RMSEP's of the full spectrum models are significantly higher than those resulting from the individual variable properties REG, SIG and SR (*p*=0.0008, *p*=0.049, *p*=0.044 and $T_+ > T_-$). For REG and SIG, the numbers of the retained variables are significantly lower than those of the other individual properties (*p*<0.05 and $T_- > T_+$).

The RMSEP's of models resulting from REG or SIG are similar (*p*=0.106) and the numbers of retained variables for SIG are significantly lower than for REG (*p*=0.046 and $T_- > T_+$). That means that SIG has better selective abilities than REG, while the predictive abilities are similar.

The RMSEP's of models resulting from REG, are significantly lower than those of NLW, VIP and COR (*p*=0.017, *p*=0.015, *p*=0.022, respectively, and $T_- > T_+$), and similar to those of SIG and SR. The RMSEP's of models resulting from SIG are similar to those of the other individual properties (*p*≥0.05).

Therefore, it is concluded that the predictive performance of the models is significantly improved after variable reduction with the FCAM method, using either the individual property REG, SIG or SR. From the individual variable properties, REG and SIG have the best selective abilities, while the predictive abilities are better than or similar to those of the other individual properties. SIG has the best selective abilities. REG is faster because no jack-knifing is needed to calculate standard deviations of the regression coefficients.

Because REG and SIG have the best predictive and selective abilities, only combinations of REG or SIG with the other individual properties are investigated.

Table 6-6 Results of variable reduction using combined predictive variable properties for the Simulated data set

Model	Mixtures	Method characteristics	Predictor-variable properties								
			REG+ SIG	REG+ NLW	REG+ VIP	REG+ SR	REG+ COR	SIG+ NLW	SIG+ VIP	SIG+ SR	SIG+ COR
11	**A**	PLS complexity selecting best set	2	2	2	11	3	3	3	11	3
		Number of variables, K_{Best}	2	2	2	12	3	3	3	13	3
		RMSEP	$1.23\cdot10^{-3}$	$1.23\cdot10^{-3}$	$1.29\cdot10^{-3}$	$1.39\cdot10^{-3}$	$1.12\cdot10^{-3}$	$1.12\cdot10^{-3}$	$1.12\cdot10^{-3}$	$1.30\cdot10^{-3}$	$1.12\cdot10^{-3}$
		R^2_{Test}	>0.9995	>0.9995	>0.9995	>0.9995	>0.9995	>0.9995	>0.9995	>0.9995	>0.9995
12	**A,B**	PLS complexity selecting best set	2	2	5	8	2	2	2	8	2
		Number of variables, K_{Best}	4	3	7	23	7	4	4	23	4
		RMSEP	$1.33\cdot10^{-3}$	$1.43\cdot10^{-3}$	$1.33\cdot10^{-3}$	$1.36\cdot10^{-3}$	$1.34\cdot10^{-3}$	$1.35\cdot10^{-3}$	$1.33\cdot10^{-3}$	$1.43\cdot10^{-3}$	$1.33\cdot10^{-3}$
		R^2_{Test}	>0.9995	>0.9995	>0.9995	>0.9995	>0.9995	>0.9995	>0.9995	>0.9995	>0.9995
13	**A,D**	PLS complexity selecting best set	4	2	2	12	2	4	4	7	3
		Number of variables, K_{Best}	15	3	4	18	7	15	11	20	7
		RMSEP	$1.66\cdot10^{-3}$	$1.65\cdot10^{-3}$	$1.86\cdot10^{-3}$	$1.72\cdot10^{-3}$	$1.78\cdot10^{-3}$	$1.39\cdot10^{-3}$	$1.48\cdot10^{-3}$	$1.47\cdot10^{-3}$	$1.66\cdot10^{-3}$
		R^2_{Test}	>0.9995	>0.9995	>0.9995	>0.9995	>0.9995	>0.9995	>0.9995	>0.9995	>0.9995
14	**A,B,C**	PLS complexity selecting best set	3	3	7	9	3	3	3	9	3
		Number of variables, K_{Best}	5	5	18	28	10	5	5	29	5
		RMSEP	$1.33\cdot10^{-3}$	$1.24\cdot10^{-3}$	$1.26\cdot10^{-3}$	$1.34\cdot10^{-3}$	$1.18\cdot10^{-3}$	$1.22\cdot10^{-3}$	$1.25\cdot10^{-3}$	$1.35\cdot10^{-3}$	$1.25\cdot10^{-3}$
		R^2_{Test}	>0.9995	>0.9995	>0.9995	>0.9995	>0.9995	>0.9995	>0.9995	>0.9995	>0.9995
15	**A,B,D**	PLS complexity selecting best set	4	4	3	15	3	5	7	15	5
		Number of variables, K_{Best}	10	4	5	28	7	10	13	27	13
		RMSEP	$1.75\cdot10^{-3}$	$2.38\cdot10^{-3}$	$1.74\cdot10^{-3}$	$2.04\cdot10^{-3}$	2.03e-003	$1.93\cdot10^{-3}$	$2.06\cdot10^{-3}$	$1.92\cdot10^{-3}$	$1.95\cdot10^{-3}$
		R^2_{Test}	>0.9995	>0.9995	>0.9995	>0.9995	>0.9995	>0.9995	>0.9995	>0.9995	>0.9995
16	**A,B,C,D**	PLS complexity selecting best set	5	4	4	10	5	7	5	9	8
		Number of variables, K_{Best}	12	4	7	48	11	14	12	48	16
		RMSEP	$2.07\cdot10^{-3}$	$1.77\cdot10^{-3}$	$1.78\cdot10^{-3}$	$1.58\cdot10^{-3}$	$3.17\cdot10^{-3}$	$2.41\cdot10^{-3}$	$1.96\cdot10^{-3}$	$1.68\cdot10^{-3}$	$1.74\cdot10^{-3}$
		R^2_{Test}	>0.9995	>0.9995	>0.9995	>0.9995	>0.9995	>0.9995	>0.9995	>0.9995	>0.9995

Table 6-7 Probabilities *p* and direction of effects for pair-wise differences in RMSEP and in numbers of retained variables, applied in one-tailed Wilcoxon signed rank tests

	Test on differences in RMSEP			**Test on differences in numbers of retained variables**	
Models resulting from	**Comparison with full spectrum model**	**Comparison with model resulting from REG**	**Comparison with model resulting from SIG**	**Comparison with model resulting from REG**	**Comparison with model resulting from SIG**
REG	**0.008** (T_+)	-	0.106 (T_+)	-	**0.046** (T_-)
SIG	**0.049** (T_+)	0.106 (T_-)	-	**0.046** (T_+)	-
NLW	0.418 (T_+)	**0.017** (T_-)	0.067 (T_-)	**0.021** (T_-)	**<0.0005** (T_-)
VIP	0.163 (T_+)	**0.015** (T_-)	0.064 (T_-)	**0.001** (T_-)	**0.002** (T_-)
SR	**0.044** (T_+)	0.459 (T_+)	0.096 (T_+)	**0.005** (T_-)	**<0.0005** (T_-)
COR	0.219 (T_+)	**0.022** (T_-)	0.090 (T_-)	**0.0005** (T_-)	**<0.0005** (T_-)
REG+SIG	**0.012** (T_+)	0.395 (T_+)	0.085 (T_+)	0.206 (T_+)	**0.037** (T_-)
REG+NLW	**0.008** (T_+)	**0.042** (T_-)	0.475 (T_+)	**0.040** (T_+)	0.270 (T_-)
REG+VIP	**0.012** (T_+)	0.464 (T_-)	0.167 (T_+)	0.051 (T_+)	0.136 (T_-)
REG+SR	0.190 (T_+)	0.074 (T_-)	0.213 (T_-)	**0.0005** (T_-)	**<0.0005** (T_-)
REG+COR	0.075 (T_+)	0.146 (T_-)	0.140 (T_+)	0.264 (T_+)	0.054 (T_-)
SIG+NLW	0.067 (T_+)	0.099 (T_-)	0.254 (T_+)	**0.028** (T_+)	0.214 (T_-)
SIG+VIP	0.054 (T_+)	0.050 (T_-)	0.438 (T_-)	0.101 (T_+)	0.252 (T_-)
SIG+SR	**0.042** (T_+)	0.438 (T_-)	0.257 (T_+)	**0.0005** (T_-)	**<0.0005** (T_-)
SIG+COR	**0.022** (T_+)	0.191 (T_-)	0.352 (T_+)	**0.025** (T_+)	0.280 (T_-)

Direction of effect is indicated by (T_+) or (T_-); (T_+) =T_+ >T_- ; (T_-)= T_- >T_+ ;
Significant one-tailed *p* values (*p*<0.05) are in bold

6.7.2 Comparison of the combined properties

The box plots in Fig. 6-2B show small numbers of retained variables for REG and SIG and for all combinations of REG or SIG with the other properties, except for REG+SR and SIG+SR for which large spreads are seen.
Table 6-7 shows that the RMSEP's of models resulting from all combined properties are significantly or borderline significantly lower than those of the full spectrum models, except for REG+SR. RMSEP's of models resulting from REG or SIG are similar to those of all combined properties, except for REG+NLW with higher RMSEP's than for REG.
The numbers of the retained variables are significantly lower for REG+NLW (*p*=0.040 and T_+ >T_-), and borderline significantly lower for REG+VIP (*p*=0.051 and T_+ >T_-) than for REG, but similar to SIG. That means that the selective ability of REG is improved by combining it with NLW or VIP.

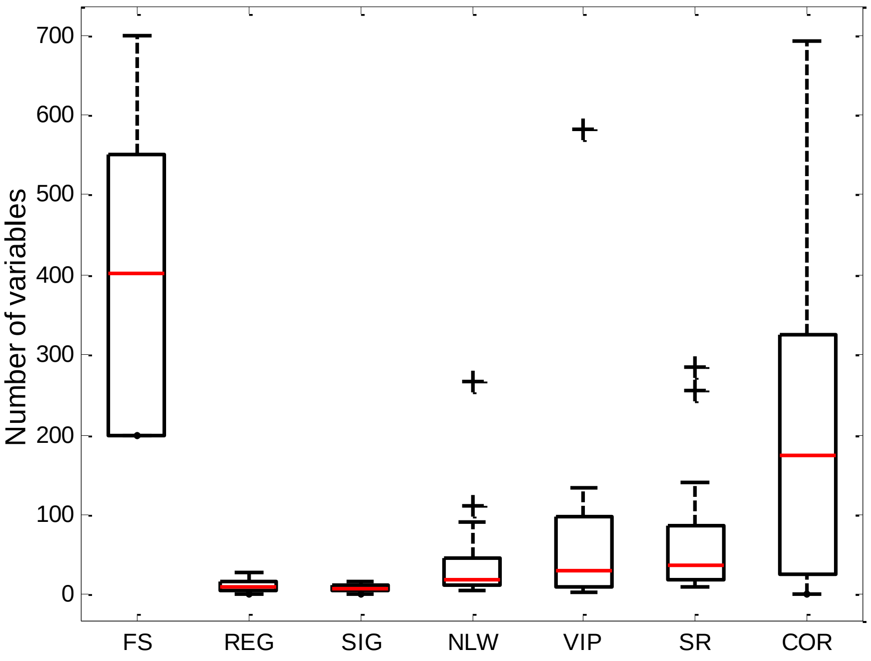
In this study only equal contributions of the individual properties in combined properties are investigated. The contribution of the individual properties in the combinations REG+NLW and REG+VIP with improved selective abilities compared to REG, may further be optimised in future studies.

For SIG, (*i*) the numbers of the retained variables are either similar or significantly lower than for all individual and combined properties, and (*ii*) the RMSEP's are similar for all individual and combined properties. SIG combines a good predictive with the best selective ability.

While the individual property SR has significantly better predictive abilities than the full spectrum models its selective ability is lower than those of REG and SIG and of combinations of SR with REG or SIG. Therefore, the use of SR for variable reduction will mostly be disadvantageous.

In combined properties, both the sums of normalised and of auto scaled individual properties were also investigated, but the results were worse than for the individual properties and are therefore not reported.

(A)



(B)

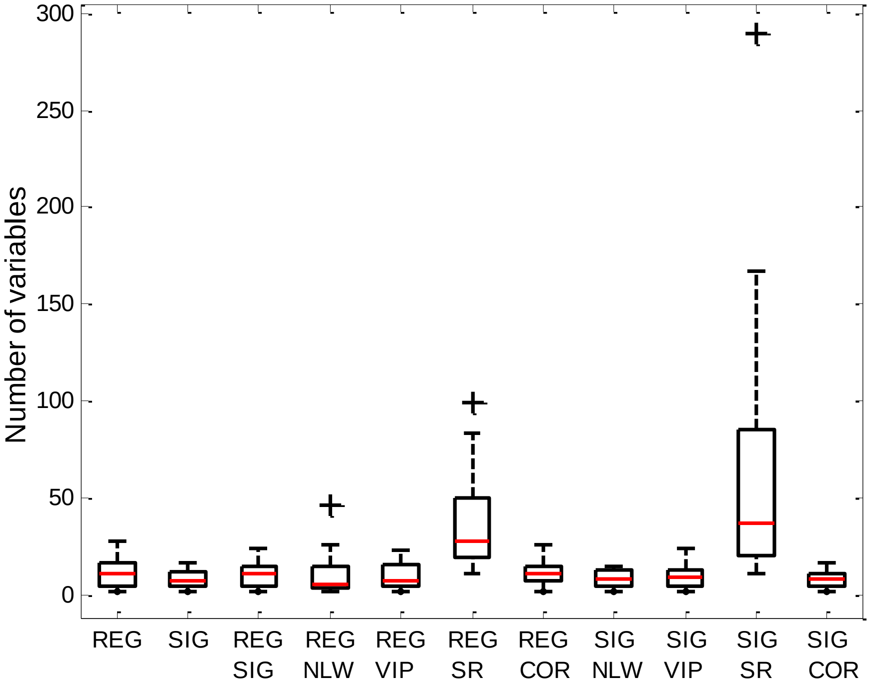


Fig. 6-2 Box plots for the numbers of retained variables by the FCAM method; (A) using the individual predictor-variable properties, (B) using REG, SIG and combined predictor-variable properties

In the FCAM method, variable selection starts with complexity *A*, determined for full spectrum modelling. The PLS model complexity is adapted to the decreasing numbers of selected variables at the end of the variable reduction process. From Table 6-1 - Table 6-6 it is seen that, for REG 9 out of 16 and for SIG 14 out of 16, best-set models have a lower complexity than the starting complexity *A*. For the combined properties REG+SIG, REG+COR, SIG+NLW, SIG+VIP, and SIG+COR, 10, 9, 15, 14 and 15 best-set models, respectively, are simpler than *A*. This demonstrates again that the adaptation of the PLS1 model complexity to decreasing numbers of selected variables is advantageous.

6.7.3 Quality of the selected variable sets

The ability of the FCAM method, using individual or combined predictor-variable properties to select features with a chemical meaning relevant to the response, is demonstrated for responses 7 (moisture from Corn set) and 14 (A, B and C from simulated set). However, this is not possible for the Diesel set because no wavelength information is provided.

Dry food samples, such as corn, show a strong absorption band for water from 1900 to 1950 nm, which is often used for the quantitative analysis of water contents [44]. Fig. 6-3A shows the corn spectra, the selected wavelengths for all individual and combined properties, and the water absorption band. The individual properties REG and SIG are very selective because only two wavelengths at 1908 and 2108 nm, having large values for REG and SIG, are retained, with very good predictive abilities (Table 6-3). Wavelength 1908 nm lies inside and 2108 nm outside the water band. The latter signal is possibly due to an interferent. All other individual and combined properties have also these two key wavelengths in their selection. Other wavelengths around 1908 and 2108 nm are also selected. An increased spread in the selected variables is observed for properties NLW, VIP, SR and COR.
All combinations of REG and SIG, except for REG+SR and SIG+SR, have only two variables in their selections. REG and SIG have already good selective abilities with two retained variables. Therefore, in this case, it is not possible to further improve the variable selection by making combinations of properties.

In the simulated data set the selective abilities of the FCAM method, are investigated for all mixtures by using as response the concentrations of analyte A, i.e. the substance with the narrowest peak profile. Table 6-5 shows for model 14, the model for analyte A in mixtures of A with interferents B and C, that REG and SIG are very selective, because sets with only 5 variables are selected with good predictivety, $R_{Test}^{2} > 0.9995$.

Fig. 6-3B shows the simulated profiles and the retained variables for all properties for model 14. REG and SIG are very selective because only 5 informative variables are retained, inside a region below the tops of analyte A and of the two interferents B and C, at positions 41, 50, and 59, respectively. The other properties NLW, VIP, SR and COR are again less selective because more variables are retained, also inside the uninformative noise area between variables 101-200. Using the combinations REG+VIP, REG+SR and SIG+SR, also uninformative random variables are selected.

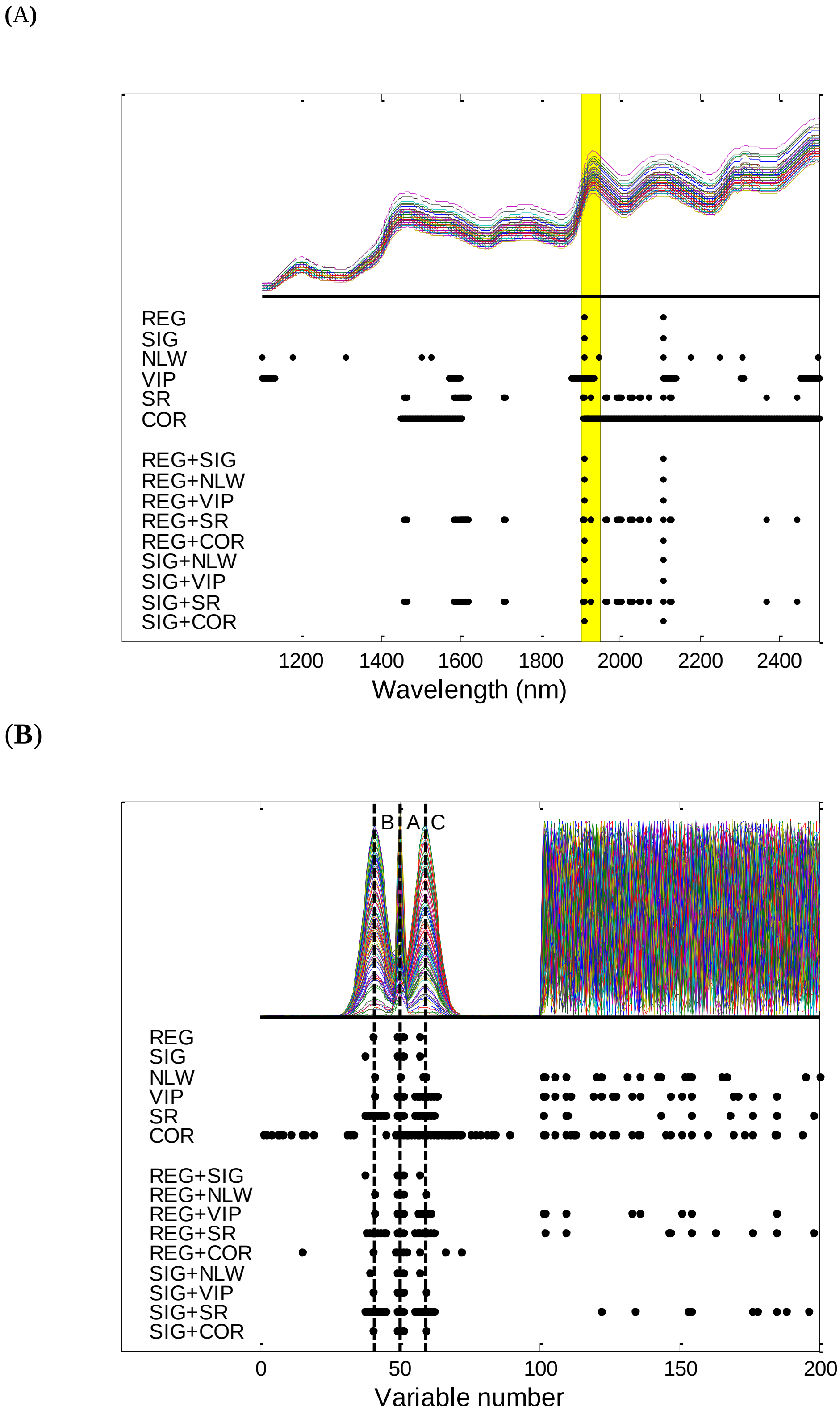


Fig. 6-3 (A) Spectra from data set Corn, response moisture, and retained wavelengths using individual and combined predictor-variable properties; the yellow column represents the water band (see text), (B) Profiles of the simulated set, model 14, and retained variables using individual and combined predictor-variable properties

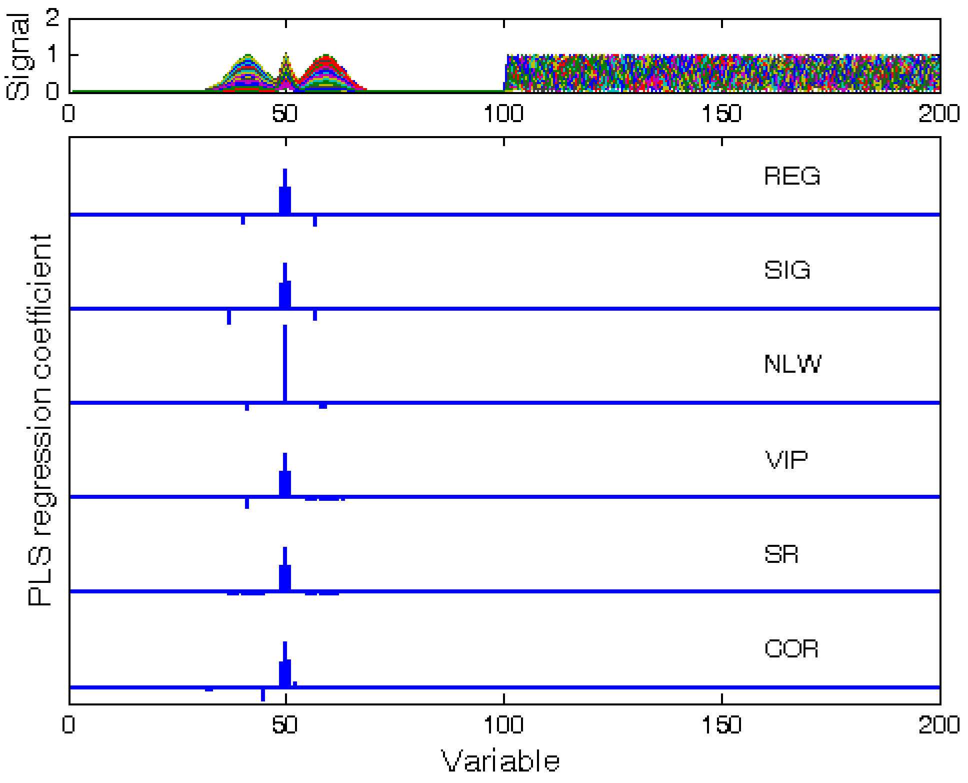


Fig. 6-4 (Top) Profiles of the simulated set, for model 14; (Bottom) PLS regression coefficients of retained variables using individual predictor-variable properties

Table 6-8 Simulated data set: number of random variables (x=101-200) retained

(A)

Model	Components	Predictor-variable properties					
		REG	SIG	NLW	VIP	SR	COR
11	A	0	0	5	0	9	2
12	A,B	0	0	27	3	11	0
13	A,D	0	2	15	0	9	0
14	A,B,C	0	0	17	18	9	25
15	A,B,D	0	2	5	1	12	10
16	A,B,C,D	0	3	13	3	10	26

(B)

Model	Components	Predictor-variable properties								
		REG+ SIG	REG+ NLW	REG+ VIP	REG+ SR	REG+ COR	SIG+ NLW	SIG+ VIP	SIG+ SR	SIG+ COR
11	A	0	0	0	10	0	0	0	10	0
12	A,B	0	0	3	11	0	0	0	11	0
13	A,D	0	0	0	9	0	2	1	9	1
14	A,B,C	0	0	8	9	0	0	0	9	0
15	A,B,D	1	0	0	12	0	2	5	12	2
16	A,B,C,D	0	0	0	8	0	2	1	10	3

Mostly, regression coefficients of uninformative variables are small. This is demonstrated in Fig. 6-4. PLS regression coefficients of the retained variables are shown, together with the spectral profiles, for model 14, and for all individual properties. Large positive regression coefficients are observed below the central peak in the spectra of compound A, which is modelled. Regression coefficients of retained uninformative variables after using NLW, VIP, SR and COR (see Fig. 6-3B) are so small that they cannot be seen in the graph (Fig. 6-4).

In addition to that, for the six simulated sets 11-16 , the FCAM method using REG or SIG, selects none or only few uninformative random variables, while many of these are retained by the other individual properties (Table 6-8A). Using the combined properties, REG+VIP, REG+SR, SIG+VIP and SIG+SR, also many random variables are retained (Table 6-8B).

It is concluded that the capability of the properties REG and SIG, to select low numbers of informative variables, with a meaning relevant to the response, is better than that of the other individual properties considered. All combinations of REG and SIG with the other properties, except REG+SR, REG+VIP, SIG+VIP and SIG+SR, are capable to select low numbers of informative variables, with a meaning relevant to the response.

6.8 Conclusions

The PPRVR-FCAM method is a backward stepwise variable-reduction method based on predictive-property-ranked variables, in which variables are first reduced at constant PLS1 model complexity *A*, until the selection of *A* variables, followed by further variable reduction and a stepwise decrease in PLS complexity (*A*-1, *A*-2, …,1), after each removal of a variable, allowing the selection of small numbers of variables.
The aim of this work was to investigate and to compare the utility and effectiveness of six individual (REG, SIG, COR, NLW, VIP and SR) and nine combined properties, in variable reduction by the PPRVR-FCAM method. The predictive and selective abilities of the different PLS1 models developed after variable reduction were statistically compared using the one-tailed Wilcoxon signed rank test.

Variable reduction with the FCAM method, using the properties REG and SIG, based on the PLS regression coefficients, after mean-centring of the data, provides low numbers of informative variables, with a meaning relevant to the response, and lower than the other individual properties. The resulting models have similar or better predictive abilities than the full spectrum models.
REG and SIG have better selective abilities than the other individual properties, while the predictive abilities are similar or better. SIG has the best selective ability of all individual and combined properties, while the predictive ability is similar. REG is faster than SIG. This means that variable reduction with the FCAM method is preferably conducted with property REG or SIG. The selective ability of REG can be improved by combining it with NLW or VIP.

Acknowledgements

Prof. Dr. G.J. de Jong and Dr. G.W. Somsen of the Avans Hogeschool are gratefully thanked for their managerial support of this project.

References

[1] H. Martens, T. Næs, Multivariate Calibration, (2nd edn), Wiley, New York, 1993.

[2] S. Wold, M. Sjöström, L. Eriksson, Chemom. Intell. Lab. Syst. 58 (2001) 109.

[3] M. Forina, S. Lanteri, M.C. Cerrato Oliveros, C. Pizarro Millan, Anal. Bioanal. Chem. 380 (2004) 397.

[4] C.H. Spiegelman, M.J. McShane, M.J. Goetz, M. Motamedi, Q.L. Yue, G.L. Coté, Anal. Chem. 70 (1998) 35.

[5] S.P. Reinikainen, A. Höskuldsson, J. Chemom. 17 (2003) 130.

[6] A. Höskuldsson, J. Chemom. 22 (2008) 150.

[7] L. Xu, I. Schechter, Anal. Chem. 68 (1996) 2392.

[8] B. Nadler, R.R. Coifman, J. Chemom. 19 (2005) 107.

[9] J.P.M. Andries, Y. Vander Heyden, L.M.C. Buydens, Anal. Chim. Acta, 705 (2011) 292.

[10] C. M. Andersen, R. Bro, J. Chemom. 24 (2010) 728.

[11] R.F. Teófilo, J.P.A. Martins, M.M.C. Ferreira, J. Chemom. 23 (2009) 32.

[12] J.A. Hageman, M. Streppel, R. Wehrens, L.M.C. Buydens, J. Chemom. 17 (2003) 427.

[13] A.S. Bangalore, R.E. Shaffer, G.W. Small, M.A. Amold, Anal. Chem. 68 (1996) 4200.

[14] A. Garrido Frenich, D. Jouan-Rimbaud, D.L. Massart, S. Kuttatharmmakul, M. Martinez Galera, J.L. Martinez Vidal, Analyst 120 (1995) 2787.

[15] H.J. Kubinyi, J. Chemom. 10 (1996) 119.

[16] W. Cai, Y. Li, X. Shao, Chemom. Intell. Lab. Syst. 90 (2008) 188.

[17] J.P. Gauchi, P. Chagnon, Chemom. Intell. Lab. Syst. 58 (2001) 171.

[18] I.G. Chong, C.H. Jun, Chemom. Intell. Lab. Syst. 78 (2005) 103.

[19] F. Westad, N.K. Afseth, R. Bro, Anal. Chim. Acta 595 (2007) 323.

[20] V. Centner, D.L. Massart, O.E. de Noord, S. de Jong, B.G.M. Vandeginste, C. Sterna, Anal. Chem. 68 (1996) 3851.

[21] A. Lazraq, R. Cléroux, J.P. Gauchi, Chemom. Intell. Lab. Syst. 66 (2003) 117.

[22] H. Xu, Z. Liu, W. Cai, X. Shao, Chemom. Intell. Lab. Syst. 97 (2009) 189.

[23] S.A. Dodds, W.P. Heath, Chemom. Intell. Lab. Syst. 76 (2005) 37.

[24] M.J. Anzanello, S.L. Albin, W.A. Chaovalitwongse, Chemom. Intell. Lab. Syst. 97 (2009) 111.

[25] M. Forina, C. Casolino, C.P. Millán, J. Chemom. 13 (1999) 165.

[26] A. Höskuldsson, Chemom. Intell. Lab. Syst. 55 (2001) 23.

[27] S. Wold, E. Johansson, M. Cocchi, 3D QSAR in Drug Design; Theory, Methods, and Applications, ESCOM, Leiden, Holland, 1993.

[28] R. Gosselin, D. Rodrigue, C. Duchesne, Chemom. Intell. Lab. Syst. 100 (2010) 12.

[29] F. Westad, H. Martens, J. Near Infrared Spectrosc. 8 (2000) 117.

[30] C. Abrahamsson, J. Johansson, A. Sparén, F. Lindgren, Chemom. Intell. Lab. Syst. 69 (2003) 3.

[31] T. Rajalahti, R. Arneberg, F.S. Berven, K.M. Myhr, R.J. Ulvik, O.M. Kvalheim, Chemom. Intell. Lab. Syst. 95 (2009) 35.

[32] T. Rajalahti, R. Arneberg, A.C. Kroksveen, M. Berle, K.M. Myhr, O.M. Kvalheim, Anal. Chem. 81 (2009) 2581.

[33] M. Blanco, I. Villarroya, Trends Anal. Chem. 21 (2002) 240.

[34] Z. Xiaobo, Z. Jiewen, M.J.W. Povey, M. Holmes, M. Hanpin, Anal. Chim. Acta 667 (2010) 14.

[35] P. Geladi, B.R. Kowalski, Anal. Chim. Acta 185 (1986) 1.

[36] B. Efron, G. Gong, The American Statistician, 37 (1983) 36.

[37] B. Li, J. Morris, E.B. Martin, Chemom. Intell. Lab. Syst. 64 (2002) 79.

[38] S. Wold, Technometrics 24 (1978) 397.

[39] B.M. Wise, N.B. Gallagher, R.Bro, J.M. Shaver, W. Windig, R.Scott Koch, PLS_Toolbox Version 4.0, Eigenvector Research, Wenatchee.

[40] D.L. Massart, B.G.M. Vandeginste, L.M.C. Buydens, S. de Jong, P.J. Lewi, J. Smeyers-Verbeke, Handbook of Chemometrics and Qualimetrics, Part A, Elsevier, Amsterdam, 1997.

[41] H.R. Cederkvist, A.H. Aastveit, T. Naes: J. Chemom. 19 (2005) 500.

[42] G.W. Corder, D.I. Foreman, Nonparametric Statistics for Non-Statisticians: A Step-By-Step Approach, John Wiley & Sons, Hoboken, 2009.

[43] R.D. Snee, Technometrics 19 (1977) 415.

[44] H. Büning-Pfaue, Food Chemistry 82 (2003) 107.

7 Predictive-Property-Ranked Variable Reduction with Final Complexity Adapted Models in Partial Least Squares Modelling for Multiple Responses[4]

7.1 Abstract

For partial least-squares regression with one response (PLS1), many variable-reduction methods have been developed. However, only few address the case of multiple-response partial-least-squares (PLS2) modelling. The calibration performance of PLS1 can be improved by elimination of uninformative variables. Many variable-reduction methods are based on various PLS-model-related parameters, called predictor-variable properties. Recently, an important adaptation, in which the model complexity is optimised, was introduced in these methods. This method was called Predictive-Property-Ranked Variable Reduction with Final Complexity Adapted Models, denoted as PPRVR-FCAM or simply FCAM.

In this study, variable reduction for PLS2 models, using an adapted FCAM method, FCAM-PLS2, is investigated. The utility and effectiveness of four new predictor-variable properties, derived from the multiple response PLS2 regression coefficients, are studied for six data sets consisting of ultraviolet-visible (UV-VIS) spectra, near-infrared (NIR) spectra, NMR spectra and two simulated sets, one with correlated and one with uncorrelated responses. The four properties include the mean of the absolute values as well as the norm of the PLS2 regression coefficients and their significances.

The four properties were found to be applicable by the FCAM-PLS2 method for variable reduction. The predictive abilities of models resulting from the four properties are similar. The norm of the PLS2 regression coefficients has the best selective abilities, low numbers of variables with an informative meaning to the responses are retained. The significance of the mean of the PLS2 regression coefficients is found to be the least-selective property.

Keywords: Variable selection, Partial least squares, PLS2, predictor-variable properties, FCAM-PLS2

[4] **Jan P.M. Andries**, Yvan Vander Heyden, Lutgarde M.C. Buydens
Published in Analytical Chemistry 85 (2013) 5444-5453

7.2 Introduction

Partial least squares (PLS) is a commonly used multivariate regression technique, able to deal to a certain extent with large numbers of noisy and correlated variables and small numbers of samples [1-3]. PLS calibration of multiple response data can be performed in two ways, either building multiple models each with one response (PLS1) or constructing one model with several responses (PLS2). PLS2 has a few advantages. First, there is one common set of PLS factors for all responses. This simplifies the procedure and interpretation, and it allows a simultaneous graphical inspection. Second, when the responses are strongly correlated, one may expect the PLS2 model to be more robust than with separate PLS1 models. Finally, when the number of responses is large, the development of a single PLS2 model is performed much faster than that of many separate PLS1 models. Practical experience, however, indicates that PLS1 calibration usually performs equally well or better in terms of predictive accuracy [4].

Both theoretical [5-9] and experimental evidence [3,10-18] exist that elimination of uninformative variables improves the performance of PLS calibration. For PLS1, many variable-reduction methods have been developed [3,12,19-22]. However, only a few address PLS2 modelling [23-26]. In this study, a new variable reduction method for PLS2 modelling is proposed and evaluated.

For PLS1, many variable-elimination methods are based on so-called predictor-variable properties, which are functions of various PLS1-model parameters, and which may change during the variable-reduction process. In these methods reduction is made on the variables ranked in descending property magnitude. This ranking reflects their importance for the model. The higher its magnitude, the more important the variable.

In the Stepwise Variable Reduction methods using Predictive-Property-Ranked Variables, denoted as SVR-PPRV methods, iteratively, the variable with the smallest property value is eliminated and a new PLS1 model calculated [10]. The predictive abilities of the models are assessed by the root-mean-squared error of cross validation (RMSECV). The set of variables resulting in the optimal model is then selected. The goal is to obtain models from small sets of variables with improved or similar predictability relative to that of the original data set. A low number of variables can also be beneficial with regard to (*i*) a better understanding of the model, and (*ii*) selection of a viable set of sensors in process control.

Properties such as weights, loadings, and PLS regression coefficients are functions of the parameters of the PLS1 algorithm, and they are interdependent [1]. In the stepwise variable reduction process the data matrix changes continuously and therefore the parameters of the PLS algorithm can also change. The optimal number of PLS factors, i.e. the best PLS model complexity, can change as well. If the same PLS model complexity is used during the entire variable reduction procedure, as is done often, RMSECV values may become overoptimistic [27], since the best model complexity decreases due to the elimination of uninformative variables [28].

In a previous study [10], a new backward variable-reduction method for PLS1 was introduced, based on variables ranked in descending order of a predictor-variable property. The method accounts for the facts that both the property values of the remaining variables and the best model complexity change during the variable-reduction process. The method was called Predictive-Property-Ranked Variable Reduction with Final Complexity Adapted

Models, denoted as PPRVR-FCAM and abbreviated to FCAM, or FCAM-PLS1. With the use of a fixed PLS1 model complexity *A*, from all to *A* variables, iteratively, the variable with the smallest property value is eliminated, a new model calculated, and the variables re-ranked. In this part of the procedure, the model complexity is not re-optimized because computationally it would slow down the method considerably. In the final part of variable reduction, the PLS model complexity is stepwise decreased, *A*-1, *A*-2, etc. after each removal of a variable.
The FCAM-PLS1 method combines good selective and predictive abilities because it is able to retain small numbers of variables with improved or similar predictability compared to the full spectrum model.

In a second study evaluating different properties [11], the best predictive and selective models resulted from variable reduction using either the absolute values of the PLS1 regression coefficients (*REG*) or their significance (*SIG*) as predictor-variable properties.

In this study, the FCAM method is adapted for variable reduction with PLS2 models. The method is called FCAM-PLS2. Because of its computational efficiency and the good results obtained with FCAM-PLS1, in FCAM-PLS2 again a fixed model complexity *A* is used from all to *A* variables. This is followed by a stepwise decreased complexity, *A*-1, *A*-2, etc. after each removal of a variable.
In the proposed FCAM-PLS2 method, only the relative values of predictive-variable properties are again important. Therefore, no threshold for the predictor-variable properties is used to remove uninformative variables. Four new predictor-variable properties, derived from the PLS2 regression coefficients, are investigated. The FCAM-PLS2 method is tested using six data sets from different sources, consisting of UV-VIS spectra, normal and second-derivative NIR spectra, NMR spectra and of simulated data, one set with correlated and one with noncorrelated responses. The simulated data sets are used to test the general applicability of the method for PLS2 models.

7.3 Theory

7.3.1 PLS2 regression coefficients

The variable reduction may be based on predictor-variable dependent properties derived from the matrix of PLS2 regression coefficients, **B**(*K* × *M*), calculated as,

$$\mathbf{B} = \mathbf{W}\left(\mathbf{P}^{T}\mathbf{W}\right)^{-1}\mathbf{Q} \qquad \textbf{(1)}$$

where **W**(*K* × *A*) is the **X** weight matrix, **P**(*K* × *A*) the **X-**loading matrix and **Q**(M × *A*) the **Y**-loading matrix. *K* is the number of predictor variables in the **X**(*N* × *K*) matrix, *M* the number of responses in the **Y**(N × M) matrix, *A* the number of PLS2 factors and *N* the number of objects. Further details of PLS2 can be found in refs [1,2,29].

Four predictor-variable properties, the mean and the norm of the PLS2 regression coefficients, and their significances, derived from **B**, are used for variable reduction in order to find an optimal set of variables for PLS2 modelling. They are described below. These properties are dependent of the *A*-factor PLS2 model.

7.3.2 Mean and norm of PLS2 regression coefficients

Predictor variables influential for a response have large positive or negative regression coefficients in the corresponding row of the **B** matrix. Therefore, both the mean of the absolute values and the norm of the PLS2 regression coefficients $b_{k1}, b_{k2}, \ldots, b_{km}$ of predictor variable k for the responses $y_1, y_2, \ldots, y_m$ in the **Y** matrix, denoted as $M_{REG,k}$ and $N_{REG,k}$, respectively, are considered as measures for the influence of k on the PLS2 model.

$$M_{REG,k} = \sum_{m=1}^{M} |b_{km}| / M = \overline{|b_k|} \quad \textbf{(2)}$$

$$N_{REG,k} = \sqrt{\sum_{m=1}^{M} b_{km}^2} \quad \textbf{(3)}$$

Influential variables have large $M_{REG,k}$ and $N_{REG,k}$ values.

7.3.3 Significance of mean and norm of PLS2 regression coefficients

Influential predictor variables have low uncertainties in the model parameters of multivariate regression models [28,30]. Therefore, the significance of the properties $M_{REG,k}$ and $N_{REG,k}$ will also be high. These significances are also considered as measures for the influence of variables k on the PLS2 model. They can be estimated by jack knifing. Influential variables will have large $M_{REG,k}$ and $N_{REG,k}$ values, combined with low standard deviations.

The significance of $M_{REG,k}$, denoted as $SIG(M_{REG,k})$, is defined as the student t value calculated from n fold jack knifing by

$$SIG(M_{REG,k}) = t_k = \frac{M_{REG,k}}{s(M_{REG,k})} \quad \textbf{(4)}$$

t_k is the student t value for variable k; $M_{REG,k}$ is calculated by eq 2 and $s(M_{REG,k})$ is the standard deviation of the estimates of $M_{REG,k}$, calculated from n fold jack knifing with eq 5.

$$s(M_{REG,k}) = \sqrt{\frac{n-1}{n} \sum_{j=1}^{n} \left(M_{REG,k(-j)} - \overline{M}_{REG,k(-j)}\right)^2} \quad \textbf{(5)}$$

$M_{REG,k(-j)}$ is the estimate of $M_{REG,k}$ based on the calibration of all objects except for the objects in the left-out segment j [31].

$$\overline{M}_{REG,k(-j)} = \frac{\sum_{j=1}^{n} M_{REG,k(-j)}}{n} \quad \textbf{(6)}$$

with $\overline{M}_{REG,k(-j)}$ the mean of $M_{REG,k(-j)}$.
Similar equations are used for $SIG(N_{REG,k})$, the significance of $N_{REG,k}$.

7.3.4 Internal model validation

The predictive ability of the PLS2 models is assessed by internal validation with the training set, using venetian blinds segmented (*n*-fold) cross validation (SCV), resulting in the root-mean-squared error of cross validation (RMSECV),

$$RMSECV = \sqrt{\frac{1}{N_{cal}M}\sum_{i=1}^{N_{cal}}\sum_{j=1}^{M}\left(y_{ij} - \hat{y}_{ij}\right)^2} \qquad \textbf{(7)}$$

where y_{ij} and $\hat{y}_{ij}$ are the experimental and predicted responses for the training set, respectively, of the j^{th} response in the i^{th} calibration sample when situated in a left-out segment, N_{cal} is the number of calibration samples in the training set, and *M* is the number of responses.

7.3.5 Model complexity

Before and after variable reduction, the best complexity *A* of a PLS2 model is determined by venetian blinds *n*-fold SCV. In order to avoid over-fitting an adjusted Wold's *R* criterion, R_{adj}, is applied [32,33].

First, the minimum in the RMSECV versus model-complexity curve is determined. Thereafter, to select a model with an as low as possible number of factors, the additional criterion $R_{adj} < 0.98$ is applied. The idea is that an additional factor should be only included in the model if the RMSCEV is improved with at least 2% [34], (i.e. if $R_{adj} < 0.98$). Models with complexities less than the one giving the smallest RMSECV are pairwise compared in eq 8.

$$R_{adj} = \frac{RMSECV_A}{RMSECV_{A-1}} \qquad \textbf{(8)}$$

The maximal complexity *A*, for which $R_{adj} < 0.98$, is then considered as the best model complexity.

7.3.6 External model validation

Before and after variable reduction, the predictive ability of the PLS2 models, developed with the training set, is also assessed by external validation with a test set, resulting in the root-mean-squared error of prediction (RMSEP),

$$RMSEP = \sqrt{\frac{1}{N_{test}M}\sum_{i=1}^{N_{test}}\sum_{j=1}^{M}\left(y_{ij} - \hat{y}_{ij}\right)^2} \qquad \textbf{(9)}$$

where y_{ij} and $\hat{y}_{ij}$ are the experimental and predicted responses for the test set, N_{test} the number of samples in the test set, and *M* the number of responses.

As another measure for external validation of each response, R^2_{Test}, the squared values of the correlation coefficient *R* between estimated (from the PLS2 models developed for the reduced variable sets) and experimental responses are also calculated for the test set samples.

7.3.7 Selection criterion for the preferred variable set

After variable reduction, RMSECV values are plotted as a function of the number of remaining variables. The model with the global minimal value, $\text{RMSECV}_{\text{Min}}$, corresponds to the variable set with optimal predictive capability. However, a smaller variable set with RMSECV not significantly higher than $\text{RMSECV}_{\text{Min}}$ is preferred. Its maximal value, $\text{RMSECV}_{\text{Crit}}$ is defined from the one-tailed F-test stochastic [35],

$$RMSECV^2_{Crit} = F_{(\alpha, N_{cal}M, N_{cal}M)} RMSECV^2_{Min} \qquad \textbf{(10)}$$

with significance level $\alpha = 0.05$, N_{cal} the number of calibration samples in the training set, *M* the number of responses, and $N_{cal}M$ is the degrees of freedom for both numerator and denominator.

Thus, the remaining variable set with a smaller number of variables, K_{Best}, than in the variable set corresponding to $\text{RMSECV}_{\text{Min}}$, and a $\text{RMSECV}_{\text{Best}}$, not significantly higher than $\text{RMSECV}_{\text{Min}}$, is considered the best.

7.4 Data and methodology

7.4.1 Metal ions data set

The first data set contains 130 samples and consists of ultraviolet/visible absorption spectra involving three-component mixtures of metal ions (Cr^{3+}, Ni^{2+}, Co^{2+}). The data set was downloaded from the Web site of the Chemometrics Group of the Dalhousie University, http://myweb.dal.ca/pdwentze/downloads.html (accessed on October 16, 2012). Details are both found on the Web site and described in ref [36]. After deletion of noisy signals at low and high wavelengths, the range of 394-590 nm with 2 nm intervals was used, resulting in 94 predictor variables. The molar concentrations of the three metal ions in the samples are used as responses. The data set is split into a training set of 100 and a test set of 30 samples using the duplex method [37]. A 10-fold cross validation is conducted during model building.

7.4.2 Corn data set

The second data set consists of NIR spectra of 80 corn samples with a wavelength range of 1100–2498 nm with 2 nm intervals, resulting in 700 predictor variables. This data set, labelled corn from the “m5” spectrometer, is provided by Eigenvector Research, http://software.eigenvector.com/ (accessed on October 16, 2012). Moisture, oil, protein and starch contents of the samples are the responses. The data set is split into a training set of 60 and a test set of 20 samples using the duplex method. An 8-fold cross validation is conducted during model building.

7.4.3 Sugars data set

The third data set consists of second derivative NIR spectra of sugar samples with a wavelength range of 1100–2498 nm with 2 nm intervals, resulting in 700 predictor variables. This data set, labelled sugars, is downloaded from http://www.blackwellpublishing.com/rss/Volumes/Bv64p3_read1.htm (accessed on October 16, 2012). Details are described in refs [38,39]. The concentrations of sucrose, glucose and fructose are the responses. The data set is provided with 125 samples in the training set and 21 samples in the test set. A 10-fold cross validation is conducted during model building.

7.4.4 Alcohols data set

The fourth data set is composed of 231 samples and contains ^{1}H NMR spectra of mixtures of the alcohols propanol, butanol and pentanol, with chemical shifts from 3.85 to 0.65 ppm, resulting in 14000 predictor variables. The data set was downloaded from http://www.models.kvl.dk/datasets (accessed on October 16, 2012). Details are described in ref [40]. The alcohol percentages in the mixtures are used as responses. The data set is split into a training set of 171 and a test set of 60 samples using the duplex method. A 10-fold cross validation is conducted during model building.

7.4.5 Simulated data sets

Because PLS2 models perform better with correlated responses [2,4], variable reduction is also investigated with two simulated sets, I and II, having a correlated and uncorrelated response matrix **Y,** respectively. Both simulated sets represent the spectra or chromatograms of mixtures containing three compounds, A, B and C. Samples of mixtures ABC are created. The pure profiles of the compounds were formed by Gaussian peaks g(μ,σ), with mean μ and standard deviation σ [$g_A(50,1)$, $g_B(41,4)$, $g_C(59,4)$], and equal maximum heights 1, measured within the first 100 variables of the global profile.

For set I, correlated responses $Y'(i,j)$, ($i = 1\ldots120$, $j = 1..3$), of the three compounds were generated with mean 0, standard deviation 1 and a predefined covariance matrix, using the Matlab function mvnrnd for the creation of random vectors from the multivariate normal distribution. The generated responses $Y'(i,j)$ are rescaled between 0 and 1 using $Y(i,j) = [Y'(i,j) - \min(\boldsymbol{y'_j})]/[\max(\boldsymbol{y'_j}) - \min(\boldsymbol{y'_j})]$. $Y(i,j)$ is the rescaled response of profile i and compound j, and $\min(\boldsymbol{y'_j})$ and $\max(\boldsymbol{y'_j})$ the minimum and maximum, respectively, of column vector j of the originally generated unscaled responses of compound j. The sample profiles i in the mixtures were generated using these rescaled correlated responses as weight factors, and the above mentioned Gaussian peaks, by $Y(i,1)\cdot g_A+ Y(i,2)\cdot g_B+ Y(i,3)\cdot g_C$

For set II, with uncorrelated random responses $Z(i,j)$, ($i = 1\ldots120$, $j = 1..3$), these responses were randomly generated between 0 and 1, using the Matlab function for uniformly distributed pseudorandom numbers, rand. The sample profiles i in the mixtures were generated analogously as described above but now using the uncorrelated random responses $Z(i,j)$ instead of $Y(i,j)$, without rescaling.

Both simulated sets consist of 120 samples each with 200 predictor variables. The first 100 variables are informative, representing the sample profiles in the mixtures. The last 100 variables are uninformative, consisting of random numbers from 0 to 1. These uninformative variables have a high signal level, comparable to that of the informative variables in the range x = 1-100. This is to investigate if the FCAM-PLS2 method can be used to find informative variables with a chemical meaning in spectra containing many uninformative variables with a similar signal level. Additionally, noise is added to the simulated spectra, consisting of random numbers in the range between 0 and 0.005, i.e. small compared to the pure signals. Each set is split into a training set of 100 and a test set of 20 samples, using the duplex method and a10-fold cross validation is conducted during model building.

7.4.6 FCAM-PLS2 method

In the FCAM-PLS2 method, variables are reduced with constant PLS model complexity *A*, until *A* variables remain. Then, the model complexity is stepwise decreased, *A*-1, *A*-2, …, *m*, after each removal of a variable. The minimal number of remaining variables is equal to the number of independent responses, *m*. In order to obtain useful predictions of all responses, the minimal PLS2 complexity is $A = m$, and therefore at least *m* variables are needed.
The FCAM method described in ref [10] is adapted to the PLS2 modelling regarding (i) the calculation of RMSECV and RMSEP (eqs 7 and 8), and (ii) the PLS2-related predictor-variable properties as described above in the theory section.
The FCAM-PLS2 method consists of four steps. First, the data set is split into a training and test set. The predictive ability of the full spectrum PLS2 models is assessed by internal validation with the training set, using SCV. The optimal number of PLS2 factors *A*, is determined by the application of the adjusted Wold's *R* criterion $R_{adj} < 0.98$. On the basis of the *A* factor PLS2 model, the values for a given property are calculated for all variables and ranked. In step 2, iteratively, the variable with the smallest property value is eliminated; a new PLS2 model, RMSECV, and new property values are calculated, and the variables are re-ranked. When the number of remaining variables becomes *A*, the model complexity is decreased by one until the number of remaining variables and the PLS2 model complexity equal *m*. In the third step, $\text{RMSECV}_{\text{Best}}$ and the corresponding set of remaining variables is determined. In the fourth step, using the reduced variable set, the PLS2 model is externally validated (RMSEP) using a test set, after a renewed determination of the optimal number of PLS2 factors by SCV and the application of the criterion $R_{adj} < 0.98$.

7.4.7 PLS2 algorithm

The PLS2 algorithm, according to ref [1], is implemented, with the modification that, if no covariance of **Y** with **X** is left, the extraction of factors is aborted.

7.4.8 Software

All calculations are made with in-house programs developed in Matlab (V. 7.14) (The Math Works, Natick, MA, USA. The procedure for the duplex splitting algorithm is from ChemoAC Standard Functions Toolbox for MATLAB, CHEMOAC Standard Function Toolbox, http://www.vub.ac.be/fabi/publiek/index.html. The correlated responses in simulated set I were generated using the statistic toolbox of Matlab.

7.5 Results and discussion

Variable reduction is conducted after pre-processing the x variables and y responses by mean centring. Variable reduction was also investigated with mean-centred x variables and auto-scaled y responses. However, the results were worse and therefore not reported.

First, for the six data sets, the optimal factor number *A* of the PLS2 models was determined by segmented cross validation and the application of the criterion $R_{adj} < 0.98$. The RMSECV and RMSEP are calculated for the full spectrum models. Variable reduction is then applied on the **X**-**Y** sets by the FCAM-PLS2 method, using one of the four predictor-variable properties mentioned in the theory section. One PLS2 model is selected for each response matrix **Y**. With the use of the selection criterion described in the theory section, the best variable set with K_{Best} variables is selected. Thereafter, the optimal number of PLS2 factors is determined for the best variable set by SCV and the application of the criterion $R_{adj} < 0.98$.

For the six data sets, for property N_{REG}, the resulting curves of the RMSECVs and the corresponding model complexities as a function of the number of remaining variables are shown in Fig. 7-1A-F. For the full spectrum models and for those with the variable sets reduced based on one of the four predictor-variable properties, the optimal PLS2 complexity, the number of (remaining) variables K_{Best}, RMSEP and R^2_{Test} of the *M* components, are shown in Table 7-1. The ability of the FCAM-PLS2 method, using one of the four predictor-variable properties, to select predictors with a meaning relevant to the responses is discussed below in more detail for all data sets.

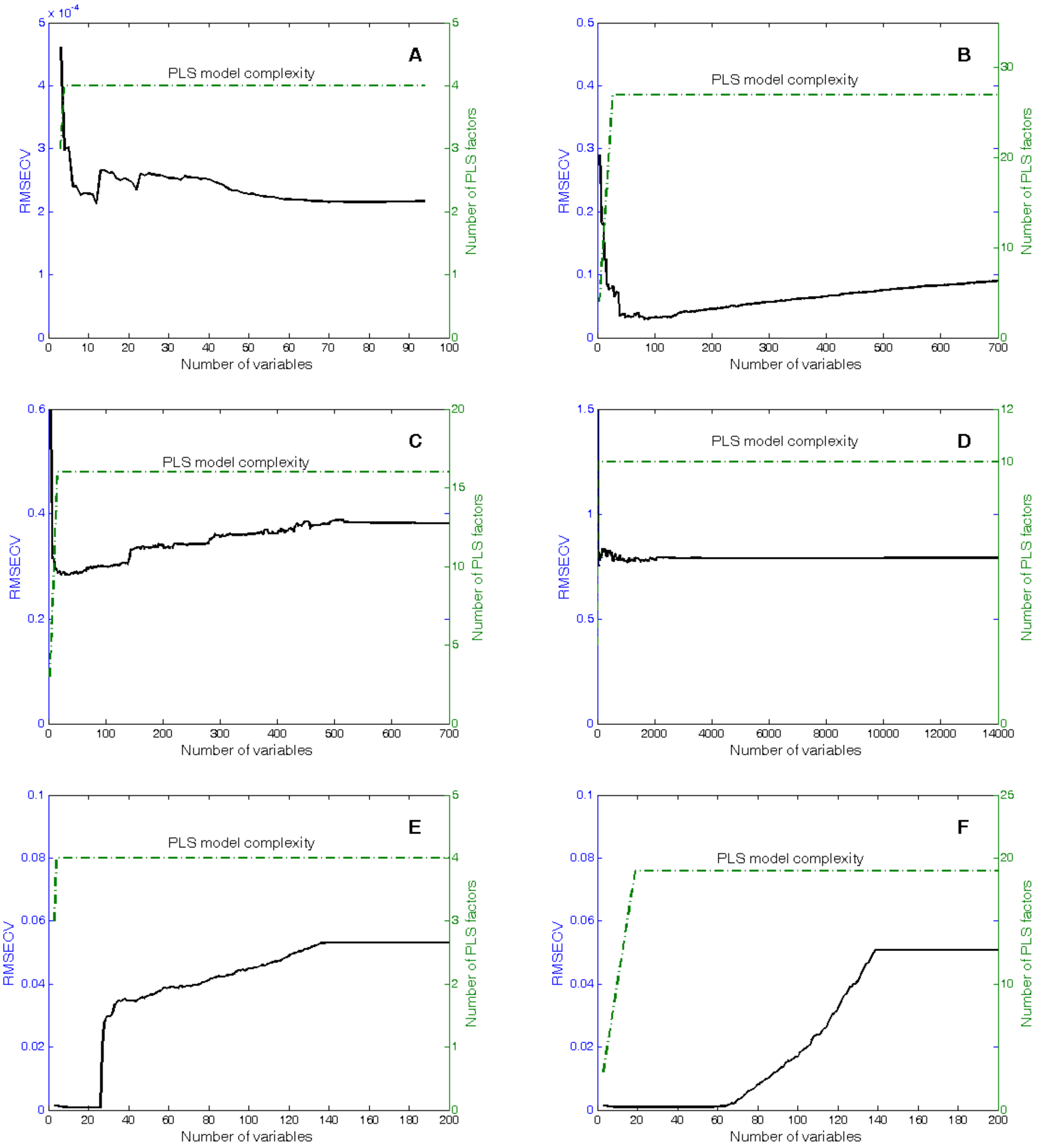


Fig. 7-1 RMSECV curve and PLS model complexity for variable reduction with the FCAM-PLS2 method using N_{REG} : (A) Metal ion set, (B) Corn set, (C) Sugars set , (D) Alcohols set, (E) Simulated set I, (F) Simulated set II; — RMSECV-curve; —·— PLS model complexity

Table 7-1 Results of the FCAM-PLS2 method for

(A) Metal ions set

Method characteristics	Full spectrum	Predictor-variable properties			
		M_{REG}	N_{REG}	SIG(M_{REG})	SIG(N_{REG})
PLS2 complexity	4	4	4	4	4
Number of variables K_{best}	94	48	8	90	14
RMSEP	1.89×10^{-4}	2.01×10^{-4}	2.26×10^{-4}	2.02×10^{-4}	2.46×10^{-4}
R^2_{Test} Cr	>0.9995	>0.9995	0.999	>0.9995	0.999
R^2_{Test} Ni	>0.9995	>0.9995	>0.9995	>0.9995	>0.9995
R^2_{Test} Co	>0.9995	>0.9995	>0.9995	>0.9995	>0.9995

(B) Corn set

Method characteristics	Full spectrum	Predictor-variable properties			
		M_{REG}	N_{REG}	SIG(M_{REG})	SIG(N_{REG})
PLS2 complexity	27	24	23	24	24
Number of variables K_{best}	700	80	48	45	62
RMSEP	0.077	0.056	0.051	0.079	0.068
R^2_{Test} moisture	0.999	0.996	0.998	>0.9995	>0.9995
R^2_{Test} oil	0.961	0.986	0.978	0.974	0.977
R^2_{Test} protein	0.976	0.991	0.993	0.979	0.986
R^2_{Test} starch	0.981	0.988	0.993	0.976	0.984

(C) Sugars set

Method characteristics	Full spectrum	Predictor-variable properties			
		M_{REG}	N_{REG}	SIG(M_{REG})	SIG(N_{REG})
PLS2 complexity	16	6	7	12	12
Number of variables K_{best}	700	8	11	86	91
RMSEP	1.738	0.771	0.687	1.843	0.895
R^2_{Test} sucrose	0.967	0.998	0.998	0.960	0.996
R^2_{Test} glucose	0.983	0.996	0.996	0.996	0.999
R^2_{Test} fructose	0.991	0.997	0.999	0.977	0.997

(D) Alcohols set

Method characteristics	Full spectrum	Predictor-variable properties			
		M_{REG}	N_{REG}	SIG(M_{REG})	SIG(N_{REG})
PLS2 complexity	10	10	8	9	8
Number of variables K_{best}	14,000	19	21	117	61
RMSEP	0.753	1.104	1.090	0.945	1.053
R^2_{Test} propanol	0.999	0.998	0.998	0.998	0.998
R^2_{Test} butanol	0.999	0.999	0.999	0.999	0.999
R^2_{Test} pentanol	0.999	0.998	0.998	0.999	0.998

(E) Simulation set I

Method characteristics	Full spectrum	Predictor-variable properties			
		M_{REG}	N_{REG}	SIG(M_{REG})	SIG(N_{REG})
PLS2 complexity	6	3	3	3	3
Number of variables K_{best}	200	16	18	16	16
RMSEP	0.046	7.97×10^{-4}	7.71×10^{-4}	7.60×10^{-4}	8.02×10^{-4}
R^2_{Test} A	0.932	>0.9995	>0.9995	>0.9995	>0.9995
R^2_{Test} B	0.944	>0.9995	>0.9995	>0.9995	>0.9995
R^2_{Test} C	0.966	>0.9995	>0.9995	>0.9995	>0.9995

(F) Simulation set II

Method characteristics	Full spectrum	Predictor-variable properties			
		M_{REG}	N_{REG}	SIG(M_{REG})	SIG(N_{REG})
PLS2 complexity	19	3	3	7	3
Number of variables K_{best}	200	11	9	17	15
RMSEP	0.033	9.89×10^{-4}	1.07×10^{-3}	9.34×10^{-4}	8.77×10^{-4}
R^2_{Test} A	0.985	>0.9995	>0.9995	>0.9995	>0.9995
R^2_{Test} B	0.997	>0.9995	>0.9995	>0.9995	>0.9995
R^2_{Test} C	0.995	>0.9995	>0.9995	>0.9995	>0.9995

7.5.1 Metal ion set

From Fig. 7-1A, for the metal ion set, it is seen that the RMSECVs for the remaining variable sets, until 13 variables, are similar to that of the full spectrum model. The selected best set has eight variables and the best PLS2 complexity is 4. R^2_{Test} for Cr^{3+}, Ni^{2+} and Co^{2+} are >0.999. Table 7-1A shows that the least variables are retained using N_{REG}. Large numbers of variables remain from M_{REG} and $SIG(M_{REG})$. The predictive abilities of all retained variable sets are similar. They are very good because, in all cases, for the three components, $R^2_{Test} \geq 0.999$.

Fig. 7-2 shows the spectra of the pure components and the selected wavelengths for the metal ion data set. The eight variables selected using N_{REG} correspond to the maxima in the pure spectra and the isobestic point of the Co^{2+} and Ni^{2+} spectra. $SIG(N_{REG})$ and M_{REG} select 14 and 48 variables, respectively, which are usually also located around the maxima and the isobestic point, while with $SIG(M_{REG})$, hardly any variable reduction is realised. For the metal ion data, it is concluded that with N_{REG} or $SIG(N_{REG})$, low numbers of variables are selected with a chemically relevant meaning.

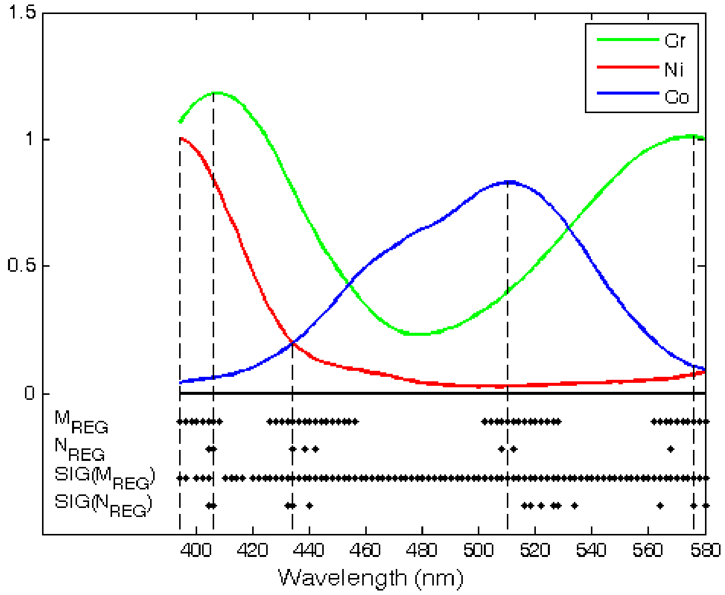


Fig. 7-2 Metal ions set: Spectra of pure components and selected wavelengths for PLS2 models using one of the predictor-variable properties

7.5.2 Corn set

Fig. 7-1B shows, for the Corn set, that the RMSECVs decrease almost steadily with the number of remaining variables, while the model complexity remains 27. With the use of the best set with 48 variables, the best PLS2 model complexity becomes 23. R^2_{Test} for moisture, oil, protein and starch are 0.998, 0.978, 0.993 and 0.993, respectively. Table 7-1B shows that the number of retained variables is higher for M_{REG} than for the other properties. The predictive ability of the model resulting from N_{REG} is best because of the lowest RMSEP. For all four components, $R^2_{Test} \geq 0.98$. The predictive abilities of the other retained variable sets are either slightly better or similar to that of the full spectrum model.

For each of the four components considered in corn, strong absorption bands in NIR are reported. Dry food samples, such as corn, show a strong absorption band for water from 1900 to 1950 nm [41]. Absorption bands for oil are at 1650-1780 nm and 2100-2200 nm [42,43], for protein they are at 1610-1760 nm and 2130-2320 nm [44] and for starch they are at 1700-1800 nm [45]. Fig. 7-3 shows the corn spectra, the selected wavelengths, and the combined absorption bands of the four components. For all properties, variables are selected from or close to the combined absorption bands. For the corn set, it is also concluded that, for the four properties, variables are selected relevant to the responses.

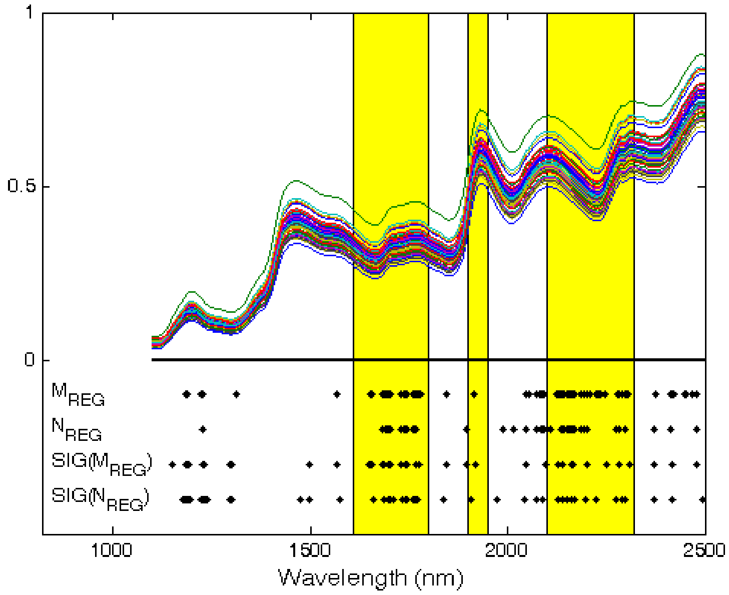


Fig. 7-3 Corn set: Spectra and selected wavelengths for PLS2 models using one of the predictor-variable properties as variable-reduction criterion; the yellow columns represent specific absorption bands (see text)

7.5.3 Sugars set

Fig. 7-1C shows, for the Sugars set, that the RMSECVs decrease with the number of remaining variables. With the use of the best set with 11 variables, the best PLS2 model complexity becomes 7. R^2_{Test} for sucrose, glucose and fructose is 0.998, 0.996 and 0.999, respectively. Table 7-1C shows that the numbers of retained variables are much higher for $SIG(M_{REG})$ and $SIG(N_{REG})$ than for M_{REG} and N_{REG}. The predictive ability of the model resulting from N_{REG} is best because of the lowest RMSEP, which is much lower than for the full spectrum model, while for the three responses, $R^2_{Test} \geq 0.996$. The predictive abilities of the other retained variable sets are lower but still good because, in all cases, for the three components, R^2_{Test} is similar or better than for the full spectrum model.

Fig. 7-4 shows the second derivative NIR spectra of the sugars dataset. Strong absorption bands for sugar molecules are at 2050-2380 nm [46,47]. For all properties, variables are retained from this absorption band. The selectivity of N_{REG} and M_{REG} is best, because only variables from or close to the absorption band remain.

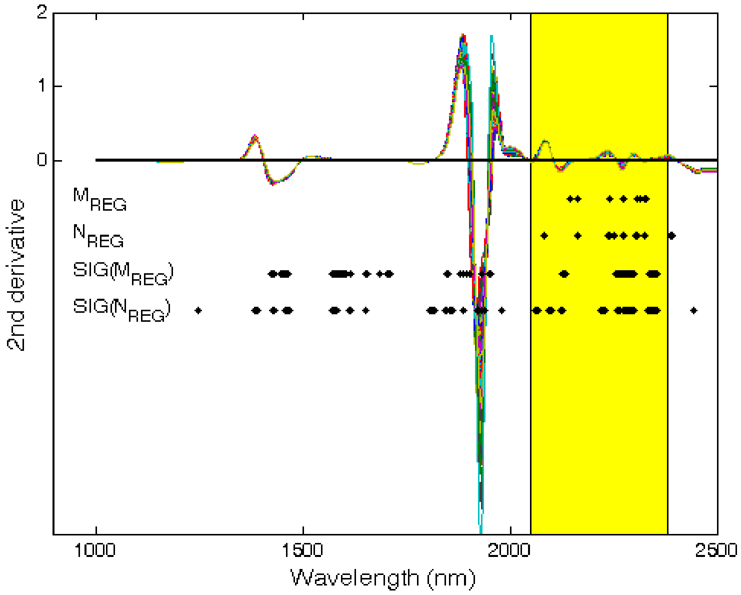


Fig. 7-4 Sugars set: Spectra and selected wavelengths for PLS2 models using one of the predictor-variable properties; the yellow columns represent specific absorption bands (see text)

7.5.4 Alcohols set

Fig. 7-1D shows, for the Alcohols set, that the RMSECVs remain rather constant as well as the model complexity (remains 10). With the use of the best set with 21 variables, the best PLS2 model complexity becomes 8. For the three responses $R^2_{Test} \geq 0.998$.

Table 7-1D shows that for all four predictor-variable properties, the number of variables is strongly reduced from 14000 to 117 or less. The predictive abilities of the models resulting from the reduced variable sets are slightly worse than that of the full spectrum model because the RMSEPs are slightly higher. However, in all cases, the three alcohol components are predicted well because, for all components, $R^2_{Test} \geq 0.998$.

Fig. 7-5 shows the NMR spectra and the variables selected from the alcohols data set. Pure propanol yields a triplet at 0.90 ppm from CH_3, a quintet at 1.55 ppm from CH_2 and a triplet at 3.57 ppm from CH_2 next to an OH group. Similar assignments apply to butanol and pentanol, but they also contain aliphatic CH_2's with a chemical shift in the range of 1.30–1.35 ppm [40]. For all properties, variables remain around the three multiplets 0.90, 1.30-1.35 and 3.57 ppm. Variables around the quintet of 1.55 ppm are only retained in the large set of *SIG*(*M*$_{REG}$), possibly because they belong to the weakest common signal group in the NMR spectra of the pure alcohols [40].

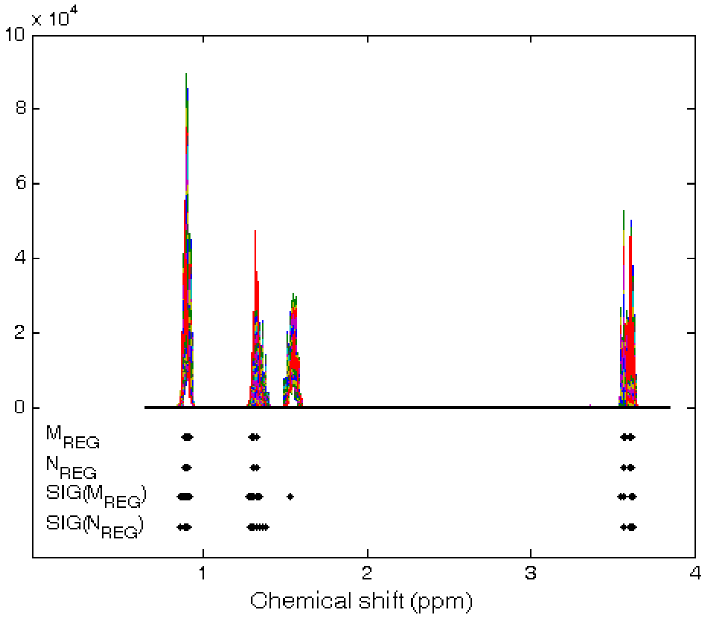


Fig. 7-5 Alcohols set: Spectra and selected variables for PLS2 models using one of the predictor-variable properties

7.5.5 Simulated set I

In the simulated set I, the concentration vectors $\mathbf{y}_1$, $\mathbf{y}_2$ and $\mathbf{y}_3$ of the three compounds in the **Y** matrix are correlated. The correlation coefficients between the vectors are: $R_{1,2}$ = 0.966, $R_{1,3}$ = 0.933 and $R_{2,3}$ = 0.896. In Fig. 7-1E, the RMSECVs for the reduced variable sets decrease slowly after about 140 and strongly after 30 remaining variables. For the best set with 18 variables, the best PLS2 model complexity becomes 3. R^2_{Test} for the three compounds are all >0.9995.

Table 7-1E shows that the number of retained variables and the RMSEPs are similar for all four predictor-variable properties. The selectivity of the FCAM-PLS2 method for all properties is good because low numbers of variables are retained. The predictive abilities of the models from all retained variable sets are similar and much better than for the full spectrum model because the RMSEPs are smaller, while for the three components $R^2_{Test} > 0.9995$.

Fig. 7-6A shows the simulated signals and the selected variables for all properties. Only small sets of variables are retained, situated in the informative area underneath the Gaussian peaks.

7.5.6 Simulated set II

In simulated set II, the concentration vectors in the **Y** matrix are uncorrelated. The correlation coefficients between the concentration vectors are: $R_{1,2}$ = -0.106, $R_{1,3}$ = -0.178, $R_{2,3}$ = -0.057. In Fig. 7-1F, the RMSECVs of the PLS2 models for the reduced variable sets steadily decrease after about 140 remaining variables, until 65 variables. The best set contains 9 variables, and the best PLS2 model complexity becomes 3. R^2_{Test} for the three compounds are all >0.9995, which is better than for the full-spectrum model. Table 7-1F shows that similar results are found as described for Simulated Set I. Also similar conclusions can be drawn.

Fig. 7-6B shows the simulated signals and the selected variables for all properties. Mostly, variables are selected in the informative area, underneath the Gaussian peaks. Only for property *SIG*(M_{REG}), uninformative variables are selected. In all cases, the selected sets are small.

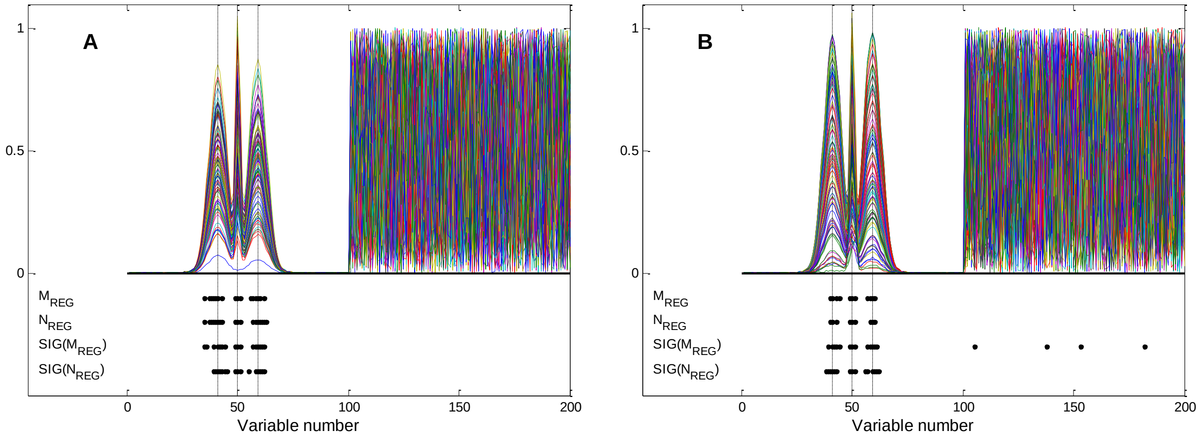


Fig. 7-6 Signals and selected variables for PLS2 models using one of the predictor-variable properties; (A) simulated set I, (B) simulated set II

For both simulated sets, the capability of the properties N_{REG}, M_{REG}, and *SIG*(N_{REG}), to select low numbers of informative variables, with a meaning relevant to the response, is good and better than that of *SIG*(M_{REG}).

For the simulated sets, I with correlated and II with uncorrelated responses**,** the PLS2 model complexity and the shape of RMSECV-curves during variable reduction are different, but the results of variable reduction**,** measured by K_{Best} and RMSEP, are similar for all predictive-properties. The RMSECV curves in Fig. 7-1E and F show also a strong reduction in the prediction error of the PLS2 models after variable reduction. Therefore, it is concluded that variable reduction by the FCAM-PLS2 method, using each of the four predictor-variable properties, works equally well for correlated and uncorrelated responses in the **Y** matrix. However, this result should be considered as a preliminary indicative, because it is only based on two data sets.

7.5.7 Comparison of the predictive properties

The selectivity of M_{REG} and N_{REG} is better than that of *SIG*(M_{REG}) and *SIG*(N_{REG}). N_{REG} is the most selective property because, for all data sets, the minimum or a similar number of variables are retained. The predictive abilities of the resulting models are mostly similar or better than for the models resulting from the other properties. Therefore, the curves for RMSECV and the PLS2 complexity in Fig. 7-1 are drawn for this property. In general, *SIG*(M_{REG}) is the least selective property, because for three out of six data sets most variables are retained while moreover uninformative variables are selected for simulation set II. The finding that the selectivity of M_{REG} and N_{REG} is better than that of *SIG*(M_{REG}) and *SIG*(N_{REG}), should further be investigated. In a future study, we will compare also the outcome of the best FCAM-PLS2 method, using N_{REG}, with those of existing variable reduction methods for PLS2.

The influence of important predictor variables with large absolute regression coefficients seems lower on the estimation of the mean than on the norm of the PLS2 regression coefficients. Probably, for important predictor variables, the corresponding quadratation in eq 3 has a larger influence on the norm than the absolute value used in eq 2 has on the mean.

7.5.8 Final adaptation of the PLS2 model complexity in the FCAM-PLS2 method

Especially the results of the selected sets from the Sugars data, using M_{REG} or N_{REG}, and for the Simulation set II, using all properties, demonstrate the benefits of the adaptation of the PLS2 model complexity in the FCAM-PLS2 method in the final part of the variable reduction process. Variable reduction is started with PLS2 complexity A = 16 (Sugars) or A = 19 (Simulation set II). After having 16, then 19 remaining variables, respectively, less variables are then selected, simpler models are build and finally found better.

7.6 Conclusions

The FCAM-PLS2 method, is a backward stepwise variable reduction method based on ranked predictor-variable properties. The variable number is first reduced with constant PLS2 model complexity A, until the selection of A variables, followed by a further variable reduction with a stepwise decrease in PLS2 complexity, A-1, A-2, …,m, after each removal of a variable.

The aim of this work was to investigate the utility and effectiveness of four predictor-variable properties, derived from the multiple-response PLS2 regression, on variable reduction by the FCAM-PLS2 method. The four variable properties include the mean of the absolute values and the norm of the PLS2 regression coefficients, and their significances.

It is found that the four predictor-variable properties can be used for variable reduction by the FCAM-PLS2 method. The predictive abilities of the four properties are similar. N_{REG} has the best selective abilities, and low numbers of variables with an informative meaning to the responses are retained. $SIG(M_{REG})$ is the least selective property.

Summarized, this study indicates that variable reduction in PLS2 modelling can be performed by the application of the FCAM-PLS2 method, using one of the proposed predictor-variable properties as reduction criterion.

Acknowledgements

Prof. Dr. G.J. de Jong and Dr. G.W. Somsen of the Avans Hogeschool are gratefully thanked for their managerial support of this project.

References

[1] H. Martens, T. Næs, Multivariate Calibration, (2^{nd} edn), Wiley, New York, 1993.

[2] S. Wold, M. Sjöström, L. Eriksson, Chemometr. Intell. Lab. Syst. 58 (2001) 109.

[3] M. Forina, S. Lanteri, M.C. Cerrato Oliveros, C. Pizarro Millan, Anal. Bioanal. Chem. 380 (2004) 397.

[4] B.G.M. Vandeginste, D.L. Massart, L.M.C. Buydens, S. de Jong, P.J. Lewi, J. Smeyers-Verbeke, Handbook of Chemometrics and Qualimetrics, Part B, Elsevier, Amsterdam, 1998.

[5] C.H. Spiegelman, M.J. McShane, M.J. Goetz, M. Motamedi, Q.L. Yue, G.L. Coté, Anal. Chem. 70 (1998) 35.

[6] S.P. Reinikainen, A. Höskuldsson, J. Chemometr. 17 (2003) 130.

[7] A. Höskuldsson, J. Chemometr. 22 (2008) 150.

[8] L. Xu, I. Schechter, Anal. Chem. 68 (1996) 2392.

[9] B. Nadler, R.R. Coifman, J. Chemometr. 19 (2005) 107.

[10] J.P.M. Andries, Y. Vander Heyden, L.M.C. Buydens, Anal. Chim. Acta, 705 (2011) 292.

[11] J.P.M. Andries, Y. Vander Heyden, L.M.C. Buydens, Anal. Chim. Acta, 760 (2013) 34.

[12] C. M. Andersen, R. Bro, J. Chemometr. 24 (2010) 728.

[13] R.F. Teófilo, J.P.A. Martins, M.M.C. Ferreira, J. Chemometr. 23 (2009) 32.

[14] J.A. Hageman, M. Streppel, R. Wehrens, L.M.C. Buydens, J. Chemometr. 17 (2003) 427.

[15] A.S. Bangalore, R.E. Shaffer, G.W. Small, M.A. Amold, Anal. Chem. 68 (1996) 4200.

[16] A. Garrido Frenich, D. Jouan-Rimbaud, D.L. Massart, S. Kuttatharmmakul, M. Martinez Galera, J.L. Martinez Vidal, Analyst 120 (1995) 2787.

[17] H.J. Kubinyi, J. Chemometr. 10 (1996) 119.

[18] W. Cai, Y. Li, X. Shao, Chemometr. Intell. Lab. Syst. 90 (2008) 188.

[19] Z. Xiaobo, Z. Jiewen, M.J.W. Povey, M. Holmes, M. Hanpin, Anal. Chim. Acta 667 (2010) 14.

[20] R.M. Balabin, S.V. Smirnov, Anal. Chim. Acta 692 (2011) 63.

[21] J.P. Gauchi, P. Chagnon, Chemometr. Intell. Lab. Syst. 58 (2001) 171.

[22] R. Leardi, J. Chemometr. 15 (2001) 559.

[23] Z. Ramadan, X.H. Song, P.K. Hopke, M.J. Johnson, K.M. Scow, Anal. Chim. Acta, 446 (2001) 233.

[24] M. De Luca, F. Oliverio, G. Ioele, G. Ragno, Chemometr. Intell. Lab. Syst. 96 (2009) 14.

[25] J.M. Roger, B. Palagos, D. Bertrand, E. Fernandez-Ahumada, Chemometr. Intell. Lab. Syst. 106 (2011) 216.

[26] B.K. Alsberg, D.B. Kell, R. Goodacre, Anal. Chem. 70 (1998) 4126.

[27] F. Westad, N.K. Afseth, R. Bro, Anal. Chim. Acta 595 (2007) 323.

[28] V. Centner, D.L. Massart, O.E. de Noord, S. de Jong, B.G.M. Vandeginste, C. Sterna, Anal. Chem. 68 (1996) 3851.

[29] S. de Jong, Chemometr. Intell. Lab. Syst. 18 (1993) 251.

[30] F. Westad, H. Martens, J. Near Infrared Spectrosc. 8 (2000) 117.

[31] B. Efron, G. Gong, The American Statistician, 37 (1983) 36.

[32] B. Li, J. Morris, E.B. Martin, Chemometr. Intell. Lab. Syst. 64 (2002) 79.

[33] S. Wold, Technometrics 24 (1978) 397.

[34] B.M. Wise, N.B. Gallagher, R.Bro, J.M. Shaver, W. Windig, R.Scott Koch, PLS_Toolbox Version 4.0, Eigenvector Research, Wenatchee.
[35] D.L. Massart, B.G.M. Vandeginste, L.M.C. Buydens, S. de Jong, P.J. Lewi, J. Smeyers-Verbeke, Handbook of Chemometrics and Qualimetrics, Part A, Elsevier, Amsterdam, 1997.
[36] D. W. Osten, B. R. Kowalski, Anal. Chem. 57 (1985) 908.
[37] R.D. Snee, Technometrics 19 (1977) 415-428.
[38] P.J. Brown, J. Chemometr. 6 (1992) 151.
[39] P.J. Brown, M. Vannucci, T. Fearn, J. Chemometr. 12 (1998) 173.
[40] H. Winning, F.H. Larsen, R. Bro, S.B. Engelsen, Journal of Magnetic Resonance 190 (2008) 26.
[41] H. Büning-Pfaue, Food Chemistry 82 (2003) 107.
[42] T. Sato, S. Kawano, M. Iwamoto, JAOCS 68 (1991) 827.
[43] J.A. Panford, J.M. deMan, JAOCS, 67 (1990) 473.
[44] S. Wrang Bruun, I. Sondergaard, S. Jacobsen, J. Agric. Food Chem. 55 (2007) 7234.
[45] C.C. Fertig, F. Podczeck, R.D. Jee, M.R. Smith, European Journal of Pharmaceutical Sciences 21 (2004) 155.
[46] F.J. Rambla, S. Garrigues, M. de la Guardia, Anal. Chim. Acta 344 (1997) 41.
[47] L. Xie, X. Ye, D. Liu, Y. Ying, Food Chemistry 114 (2009) 1135.

8 Summary, conclusions, discussion and future perspectives

8.1 Summary and conclusions

In the introduction is stated that new or improved chemometric methods should be developed to master the data flood, generated by the wide application of modern highly sophisticated instrumental analysis techniques in analytical chemistry, life sciences, bio-informatics, and metabolomics [1]. The goal of the research presented in this thesis was to develop new or improved chemometric methods both for sample and variable selection, to help mastering the data flood.

The development of a new method for sample selection is focussed on classical Quantitative Structure-Retention Relationships (QSRRs) for the widely used Reversed-Phase Liquid Chromatography (RPLC). Sample selection is used for the construction of reduced calibration sets for the development of classical QSRRs, based on linear regression or multiple linear regression models. Efficient and cost effective sample selection for RPLC can reduce the number of experiments, and hence also less data will be generated.

In chapter 1 an introduction is given to his thesis, and chapter 2 gives an introduction to sample selection for RPLC. RPLC columns can be characterised either by empirical methods or based on QSRR models. For empirical methods, generally, a low number of test components is used, while for QSRR based methods, with four to six components per descriptor, the number of components is much larger.

In chapter 3, a strategy is presented for the construction of reliable reduced calibration sets that are useful for three types of classical QSRR models containing small numbers (1-5) of descriptors:

$\log k_w = \beta_0 + \beta_1 \log P$,

$\log k_w = \beta_0 + \beta_1 \delta_{\min} + \beta_2 \mu^2 + \beta_3 A_{WAS}$,

and $\log k_w = \gamma + \varepsilon E + \sigma S + \alpha A + \beta B + \nu V$.

The analytes in the reduced calibration sets were selected using the Kennard-Stone algorithm, applied on the independent variables in the molecular-descriptor space, before the experimental determination of retentions in the chromatographic system at hand.

The proposed strategy works very well. The calibration and prediction errors of the QSRR models, developed with the reduced calibration sets, are similar to the calibration errors of the corresponding QSRR models developed with all available calibration samples. Both the dependent and independent variable spaces are covered well by the QSRR models, developed with the reduced calibration sets. For each type of classical QSRR model, the required minimal number of calibration samples in the reduced sets is determined. With the use of the proposed strategy, a substantial reduction of the number of analytes for the calibration sets is realised, allowing the reduction of the number of RPLC experiments.

The development of new variable-selection methods is focussed on PLS modelling because it dominates multivariate modelling in chemometrics. With the use of variable selection, the data flood becomes manageable by the elimination of noisy and uninformative variables.

Subsets containing only informative variables are obtained, which can be used for the development of simple, robust and interpretable PLS models. These PLS models can be used for both qualitative and quantitative analysis in many different application fields, such as food chemistry, pharmaceutical analysis, agriculture, environment, industrial and clinical chemistry, bio-informatics and metabolomics.

Following the strategy described in section 4.15, three new backward variable-selection methods for PLS1 with good selective and predictive abilities are developed. They select individual variables and are therefore generally applicable, both for continuous and non-continuous data. These methods are described in chapter 5.

The new methods are iterative, and predictive-variables are ranked on the size of a specified property. They are so-called Predictive-Property-Ranked Variables based methods, denoted as PPRV methods. These methods use Complexity Adapted Models (CAM), meaning that, during the variable-reduction process, the PLS1 model complexity can be adapted. Three new CAM methods are developed. They include Repetitive Complexity Adapted Models (RCAM), Final Complexity Adapted Models (FCAM), and Integral Complexity Adapted Models (ICAM). These methods are different in the way the PLS model complexity is adapted.

The selective and predictive abilities of the new CAM methods were investigated, using the absolute PLS1 regression coefficient as predictive-variable property. They were compared with two modifications of existing related iterative PPRV methods, using a constant PLS1 model complexity, and with two reference methods: Uninformative Variable Elimination, followed by either a Genetic Algorithm for PLS or interval PLS. It was found that the three new CAM methods combine good selective and predictive abilities. They are similar for the three CAM methods. The selectivities of the CAM methods are significantly better than those of the two modifications of existing related iterative PPRV methods, and both reference methods, while the predictive abilities are similar. Important variables, with a chemical meaning relevant to the response, are retained by the CAM methods.

RCAM is the least attractive new method. It is based on a computer intensive brute force technique where variable reduction is conducted repeatedly, starting with all variables, with stepwise descending complexities. FCAM is the preferred variable-selection method, seen from computational intensity, predictive and selective capabilities. ICAM is an attractive method for future developments in variable selection. Its predictive and selective capabilities are similar to those of FCAM and its computational intensity is only slightly higher than that of FCAM.

The preferred FCAM method was used for further development of the variable-selection methods for PLS1. In chapter 6 the utility and effectiveness of six individual and nine combined predictor-variable properties are investigated, when used in the FCAM method. It was found that the models resulting from variable reduction have similar or better predictive abilities than the models developed with all available variables. The individual properties *absolute value of the PLS1 regression coefficient* and *significance of the PLS1 regression coefficient*, have the best selective abilities. They provide lower numbers of informative variables, with a meaning relevant to the response, than the other individual properties, while the predictive abilities are similar or better. The *significance of the PLS1 regression coefficient* has the best selective ability while the *absolute value of the PLS1 regression coefficient* is computationally faster.

The preferred FCAM method for PLS1 (FCAM-PLS1) was also used as starting point for the development of a variable-selection method for PLS with multiple responses (PLS2). In chapter 7 four new predictor-variable properties, derived from the multiple response PLS2 regression coefficients, were proposed and investigated. They include the *mean of the absolute values of the PLS2 regression coefficients* as well as the *norm of the PLS2 regression coefficients*, and their *significances*. It was found that these four new properties are applicable by the adapted FCAM method for variable reduction with PLS2 models (FCAM-PLS2). The predictive abilities of models resulting from the four properties are similar. The *norm of the PLS2 regression coefficient* has the best selective abilities, and low numbers of variables with an informative meaning to the responses retained. The *significance of the mean of the PLS2 regression coefficients* is the least selective property.

Summarized, in this PhD project, five new chemometric methods are developed and tested which can help mastering the data flood. The methods include (*i*) one for sample selection to construct reduced calibration sets for classical QSRR modelling for Reversed-Phase Liquid Chromatography, (*ii*) three generally applicable variable-selection methods for PLS1 (RCAM-PLS1, FCAM-PLS1 and ICAM-PLS1), and (*iii*) one generally applicable variable-selection method for PLS2 (FCAM-PLS2). These methods form a good starting point for a new research line dedicated to the mastering of the data flood in chemometrics, as discussed below.

8.2 Discussion and future perspectives

The variable-selection methods developed in this project and summarized above, can further be extended and improved in future research. Extension of the methods can be realised by the adaptation of the ICAM method to PLS2. Improvement is possible by acceleration of the variable-selection methods, resulting in faster method modifications, allowing shorter calculation times. Additionally, the (modified) methods can be applied in combination with other methodologies, such as PLS-DA, QSRRs and Quantitative Structure-Activity Relationships (QSARs), and in new application fields, such as metabolomics. Finally, the application of the sample selection method and variable-selection methods can be integrated, for instance in QSRR and QSAR. This is explained below.

From the results of the studies presented in chapter 5 is found that ICAM is a variable-selection method for PLS1 with good selective and predictive abilities. Like the FCAM method, the ICAM method can also be adapted to PLS2, using the new predictor-variable properties, derived from the multiple response PLS2 regression coefficients, as proposed in chapter 7. This will result in a new ICAM-PLS2 method. After this development, the selective and predictive abilities of the ICAM-PLS2 method can be investigated and compared with those of FCAM-PLS2. Although it would also be possible to adapt the RCAM method to PLS2, it will not be very attractive, because of the repeated variable reduction iterations in the RCAM method, resulting in long data-analysis times. Adaptation of ICAM to PLS2 will finally result in four variable-selection methods with reasonable computation times, which can be used for PLS1 or PLS2: FCAM-PLS1, FCAM-PLS2, ICAM-PLS1, and ICAM-PLS2.

Although the calculation times for the FCAM and ICAM methods are reasonable, especially on the fast modern computer systems, it would be advantageous to still accelerate these methods. In the FCAM and ICAM methods, iteratively, the variable with the smallest predictive property value is eliminated, a new PLS model with the retained variables calculated, and its predictive ability assessed by the RMSECV. These iterative methods are rather time consuming [2]. However, they are effective because their selective and predictive abilities are good [3].

The FCAM and ICAM may be accelerated, both for PLS1 and PLS2, by a group-wise elimination of the variables in an iterative process. In each iteration step, the variables with a predictor-variable property below a pre-defined upper limit, could be eliminated. After each iteration step, the predictive ability of the PLS model, built with the remaining variable set, can be assessed by the RMSECV. Variable elimination by the use of thresholds is fast and easy to compute. However, the selection of a good upper limit will be important [2].

The variable-selection methods developed in this project have been used only for quantitative tasks in multivariate calibration and prediction. However, it is also possible to use them in the future in combination with other methodologies, such as (*i*) qualitative classification tasks in the form of Partial Least Squares Discrimination Analysis [4,5], (*ii*) for modelling with wide QSRR [6] or QSAR-data [7], both containing large numbers of theoretical molecular descriptors generated by calculation chemistry [8], and (*iii*) for biomarker discovery in metabolomics [9].

Additionally, the new variable-selection methods may also be used for quantitative tasks in multivariate calibration and prediction in application fields such as food chemistry, pharmaceutical analysis, agriculture, environment, industrial and clinical chemistry.

Finally, the new sample-selection method for the construction of reduced calibration sets and the new variable-selection methods can be integrated for quantitative tasks in multivariate calibration and prediction for both wide QSRR and QSAR data sets. First, with the use of the new variable-selection methods, informative molecular descriptors can be selected based on PLS models. Then the strategy proposed in chapter 3 can be used to construct reduced calibration sets for either PLS or MLR modelling.

As described above, the results of this thesis project form a sound basis to set up a new promising research line, dedicated to the mastering of the data flood with chemometrics.

References

[1] L. Buydens, The Analytical Scientist, 1 (2013) 24.
[2] T. Mehmood, K.H. Liland, L. Snipen, S. Sæbø, Chemom. Intell. Lab. Syst. 118 (2012) 62.
[3] R.F. Teófilo, J.P.A. Martins, M.M.C. Ferreira, J. Chemometr. 23 (2009) 32.
[4] M. Barker, W. Rayens, J. Chemometr. 17 (2003) 166.
[5] M. Bylesjö, M. Rantalainen, O. Cloarec, J.K. Nicholson, E. Holmes, J. Trygg, J. Chemometr. 20 (2006) 341.
[6] R. Put, Y. Vander Heyden, Anal. Chim. Acta 602 (2007) 164.
[7] M. Goodarzi, S. Funar-Timofei, Y. Vander Heyden, Trends Anal. Chem. 42 (2013) 49.
[8] R. Todeschini, V. Consonni, Handbook of Molecular Descriptors, Wiley, Weinheim, 2000.
[9] T. Rajalahti, R. Arneberg, F.S. Berven, K.M. Myhr, R.J. Ulvik, O.M. Kvalheim, Chemom. Intell. Lab. Syst. 95 (2009) 35.

Appendix

A. List of abbreviations

BiPLS	Backward interval PLS
CAM	Complexity Adapted Models
CARS	Competitive Adaptive Reweighted Sampling
CE-MS	Capillary Electrophoresis-Mass Spectrometry
COR	Squared correlation coefficient
CovProc	Covariance Procedures
CovSel	Covariance Selection
CSMWPLS	Changeable Size Moving Window Partial Least Squares
CV	Cross-Validation
Eq(s).	Equation(s)
FCAM	Final Complexity Adapted Models
FCAM-PLS1	Final Complexity Adapted Models for PLS1
FCAM-PLS2	Final Complexity Adapted Models for PLS2
FiPLS	Forward interval PLS
FS	Full Spectrum
FTIR	Fourier Transform Infrared Spectroscopy
GA(s)	Genetic Algorithm(s)
GA-PLS	Genetic Algorithm for PLS
GC-MS	Gas Chromatography-Mass Spectrometry
HPLC	High Performance Liquid Chromatography
ICAM	Integral Complexity Adapted Models
ICAM-PLS1	Integral Complexity Adapted Models for PLS1
ICAM-PLS2	Integral Complexity Adapted Models for PLS2
iPLS	Interval PLS
KS	Kennard and Stone
LASSO	Least Absolute Shrinkage and Selection Operator
LC-MS	Liquid Chromatography-Mass Spectrometry
LDA	Linear Discriminant Analysis
LR	Linear Regression
LSER(s)	Linear Solvation Energy Relationship(s)
MCSMWPLS	Modified Changeable Size Moving Window Partial Least Squares
MCUVE	Monte-Carlo UVE
MLR	Multiple Linear Regression
MSC	Multiplicative Scatter Correction
MWPLS	Moving Window PLS
NIPALS	Nonlinear Iterative Partial Least Squares
NIR	Near Infrared
NLW	Norm of the loading weight vector
NMR	Nuclear Magnetic Resonance
OSC	Orthogonal Signal Correction
PLS	Partial Least Squares
PLS1	PLS model with one response
PLS2	PLS model with multiple responses
PLS-DA	Partial Least Squares Discrimination Analysis
PPRV	Predictive-Property-Ranked Variables
PPRVR	Predictive-Property-Ranked Variables Reduction
PPRVR-CAM	Predictive-Property-Ranked Variable Reduction with Complexity Adapted Models
PPRVR-FCAM	Predictive-Property-Ranked Variable Reduction with Final Complexity Adapted Models
PPRVR-ICAM	Predictive-Property-Ranked Variable Reduction with Integral Complexity Adapted Models
PPRVR-RCAM	Predictive-Property-Ranked Variable Reduction with Repetitive Complexity Adapted Models
QCI	Quantum Chemical Indices
QSAR(s)	Quantitative Structure-Activity Relationship(s)
QSRR(s)	Quantitative Structure-Retention Relationship(s)
RCAM	Repetitive Complexity Adapted Models
REG	Absolute value of the PLS1 regression coefficient
RM	Replacement Method
RMSEC	Root Mean Squared Errors of Calibration

RMSECV	Root Mean Squared Errors of Cross-Validation
RMSEP	Root Mean Squared Errors of Prediction
RPLC	Reversed-Phase Liquid Chromatography
SCMWPLS	Searching Combination Moving Window Partial Least Squares
SCV	Segmented Cross-Validation
SG	Savitzky and Golay
SIG	Significance of the PLS1 regression coefficient
siPLS	Synergy interval PLS
SNV	Standard Normal Variate
SPA	Successive Projections Algorithms
SR	Selectivity Ratio
SSRs	Sums of Squared Residues
SVR	Stepwise Variable Reduction
SVR-PPRV	Stepwise Variable Reduction methods using Predictive-Property-Ranked Variables
UVE	Uninformative Variable Elimination
UVE-GA-PLS	Uninformative Variable Elimination followed by a Genetic Algorithm for PLS
UVE-iPLS	Uninformative Variable Elimination followed by interval PLS
UVE-PLS	Uninformative Variable Elimination for PLS
UV-VIS	Ultraviolet-Visible
VIP	Variable Importance in the Projection

B. List of publications

1. **J.P.M. Andries**, H.A. Claessens, Y. Vander Heyden, L. M.C. Buydens: *Strategy for reduced calibration sets to develop quantitative structure–retention relationships in high-performance liquid chromatography,* Anal. Chim. Acta 652 (2009) 180-188.
2. **J.P.M. Andries**, Y. Vander Heyden, L. M.C. Buydens: *Improved Variable Reduction in partial least squares modelling based on Predictive-Property-Ranked Variables and adaptation of partial least squares complexity,* Anal. Chim. Acta 705 (2011) 292-305.
3. Y. Vander Heyden, **J.P.M. Andries**, M. Goodarzi: *Variable selection and reduction in multivariate calibration and modelling,* LC-GC Europe, 24 (2011) 642-644.
4. **J.P.M. Andries**, Y. Vander Heyden, L. M.C. Buydens: *Predictive-Property-Ranked Variable Reduction in Partial Least Squares Modelling with Final Complexity Adapted Models: Comparison of Properties for Ranking,* Anal. Chim. Acta 760 (2013) 34-45.
5. **J.P.M. Andries**, Y. Vander Heyden, L. M.C. Buydens: *Predictive-Property-Ranked Variable Reduction with Final Complexity Adapted Models in Partial Least Squares Modelling for Multiple Responses,* Anal. Chem. 85 (2013) 5444-5453.

Samenvatting

Door de brede toepassing van instrumentele analysetechnieken in de analytische chemie, life sciences, bio-informatica en metabolomics is er een overvloed aan data ontstaan. Om deze te kunnen beheersen en analyseren zijn nieuwe of verbeterde chemometrische methoden nodig. Het doel van het onderzoek dat wordt gepresenteerd in dit proefschrift was om nieuwe of verbeterde chemometrische methoden te ontwikkelen voor zowel monster- als variabelenselectie die kunnen helpen bij de beheersing van deze overvloed aan data.

De ontwikkeling van een nieuwe monsterselectiemethode is gericht op het gebruik bij klassieke Quantitatieve Structuur-Retentie Relaties (QSRRs) voor de veel toegepaste omkeerfase vloeistofchromatografie ofwel Reversed-Phase Liquid Chromatography (RPLC). Met de nieuwe monsterselectiemethode worden gereduceerde kalibratiesets samengesteld voor de ontwikkeling van klassieke QSRRs die zijn gebaseerd op lineaire of multipele lineaire regressiemodellen. Door efficiënte en kosteneffectieve monsterselectie voor RPLC kan het aantal experimenten worden beperkt.

Hoofdstuk 1 vormt een inleiding op dit proefschrift. Hoofdstuk 2 geeft een inleiding op monsterselectie voor RPLC. Daarin wordt beschreven dat RPLC-kolommen kunnen worden gekarakteriseerd met behulp van empirische methoden of op basis van QSRR modellen. Voor de empirische methoden wordt in het algemeen een klein aantal test componenten gebruikt. Voor de methoden die zijn gebaseerd op QSRR modellen is het aantal componenten, met vier tot zes componenten per descriptor, veel groter.

In hoofdstuk 3 wordt een strategie gepresenteerd voor de constructie van betrouwbare gereduceerde kalibratiesets voor drie soorten klassieke QSRR modellen met een klein aantal (1-5) descriptoren:
$\log k_w = \beta_0 + \beta_1 \log P$,
$\log k_w = \beta_0 + \beta_1 \delta_{\min} + \beta_2 \mu^2 + \beta_3 A_{WAS}$,
en $\log k_w = \gamma + \varepsilon E + \sigma S + \alpha A + \beta B + \nu V$.
De chemische verbindingen in de gereduceerde kalibratiesets werden geselecteerd met behulp van het Kennard-Stone algoritme, dat wordt toegepast op de onafhankelijke variabelen in de moleculaire descriptorruimte. Deze selectie vindt plaats vóór de experimentele bepaling van de retenties in een chromatografische systeem.

De ontwikkelde strategie werkt naar behoren. De kalibratie- en predictiefouten van de QSRR modellen die zijn ontwikkeld met de gereduceerde kalibratiesets zijn van dezelfde grootte-orde als de kalibratiefouten van de QSRR modellen die zijn ontwikkeld met alle beschikbare kalibratiecomponenten. Zowel de afhankelijke als onafhankelijke variabelenruimtes worden goed afgedekt door de gereduceerde kalibratiesets, en bijgevolg ook door de QSRR modellen die ermee zijn ontwikkeld. Voor elk van de drie QSRR modellen is het vereiste minimum aantal componenten voor de gereduceerde kalibratiesets bepaald. Door toepassing van deze strategie kan een substantiële reductie van het aantal verbindingen in de kalibratiesets worden gerealiseerd en kan ook het aantal RPLC experimenten worden beperkt.

In hoofdstuk 4 wordt een inleiding gegeven op variabelenselectie. De ontwikkeling van nieuwe variabelen-selectiemethoden is gelinkt aan Partial Least Squares (PLS) omdat deze techniek in de chemometrie het meest wordt toegepast bij multivariate modellering. Door toepassing van variabelenselectie worden variabelen, die slechts ruis voorstellen en/of die niet-informatief zijn, geëlimineerd en wordt de overvloed aan data beter beheersbaar. Er worden sub-sets verkregen die uitsluitend informatieve variabelen bevatten, die kunnen worden gebruikt voor de ontwikkeling van eenvoudige robuuste en interpreteerbare PLS modellen. Deze PLS modellen kunnen worden toegepast voor kwantitatieve en kwalitatieve analyses in veel domeinen zoals levensmiddelenchemie, farmaceutische analyse, landbouw, milieukunde, industriële en klinische chemie, bio-informatica en metabolomics.

Er zijn drie nieuwe variabelen-selectiemethoden ontwikkeld voor PLS modellen met één afhankelijke variabele (PLS1), met goede selectieve en voorspellende eigenschappen. De methoden zijn algemeen toepasbaar, zowel voor continue als discontinue data, omdat er individuele variabelen mee worden geselecteerd. Ze worden beschreven in hoofdstuk 5.

De nieuwe methoden zijn iteratief en de variabelen worden gerangschikt in volgorde van grootte van een eigenschap van voorspellende variabelen. Het zijn zogenaamde “Predictive-Property-Ranked Variables” methoden, aangeduid als PPRV-methoden. Bij deze methoden wordt bovendien de PLS1 modelcomplexiteit aangepast gedurende het variabelen-reductie proces door het gebruik van zogenaamde “Complexity Adapted Models”, afgekort als CAM. Er zijn drie nieuwe CAM-methoden ontwikkeld: “Repetitive Complexity Adapted Models (RCAM)”, “Final Complexity Adapted Models (FCAM)”, en “Integral Complexity Adapted Models (ICAM)”. De methoden verschillen van elkaar in de wijze waarop de PLS1 modelcomplexiteit wordt aangepast.

Het selectieve en voorspellende vermogen van de nieuwe CAM methoden werd onderzocht met de PLS1-regressiecoefficient als eigenschap van de voorspellende (onafhankelijke) variabelen. De resultaten werden vergeleken met die van twee modificaties van bestaande verwante iteratieve PPRV-methoden met een constante PLS1-complexiteit, en met twee referentiemethoden: “Uninformative Variable Elimination”, gevolgd door een Genetisch Algoritme of door interval PLS. Gebleken is dat de selectiviteit van de CAM methoden significant beter is dan die van de twee modificaties van bestaande verwante PPRV-methoden en van beide referentiemethoden, terwijl de voorspellende vermogens vergelijkbaar zijn. Met de CAM methoden worden belangrijke variabelen geselecteerd die een chemisch relevante betekenis hebben voor de respons.

De RCAM-methode is de minst aantrekkelijke omdat deze gebaseerd is op een rekenintensieve domme kracht techniek. De variabelenreductie wordt daarbij herhaald uitgevoerd, steeds opnieuw beginnend met alle variabelen, maar met stapsgewijs afnemende modelcomplexiteiten. De FCAM-methode heeft de voorkeur, gelet op de benodigde rekenkracht en het selectieve en voorspellend vermogen. De ICAM-methode kan mogelijk ook voor toekomstige ontwikkelingen op het gebied van variabelenselectie worden gebruikt omdat het selectieve en voorspellende vermogen vergelijkbaar is met die van FCAM, terwijl de benodigde rekenkracht maar weinig hoger is.

De geprefereerde FCAM-methode werd bij dit onderzoek gebruikt voor de verdere ontwikkeling van variabelenselectiemethoden voor PLS1. In hoofdstuk 6 zijn de bruikbaarheid en effectiviteit van zes individuele en negen gecombineerde eigenschappen van voorspellende variabelen onderzocht, in combinatie met de FCAM-methode. Het bleek dat de modellen die ontwikkeld waren na variabelenreductie een vergelijkbaar of beter voorspellend vermogen hadden dan de modellen ontwikkeld met alle beschikbare variabelen.

De individuele eigenschappen “*absolute waarde van de PLS1 regressiecoëfficiënt*” en “*significantie van de PLS1 regressiecoëfficiënt*” zijn het meest selectief. Met behulp van deze eigenschappen worden kleinere aantallen voor de respons relevante informatieve variabelen geselecteerd dan met de andere individuele eigenschappen, terwijl het voorspellend vermogen vergelijkbaar of beter is. De “*significantie van de PLS1 regressiecoëfficiënt*” is het meest selectief, terwijl de “*absolute waarde van de PLS1 regressiecoëfficiënt*” rekentechnisch sneller is.

De FCAM-methode voor PLS1 (FCAM-PLS1) is ook als startpunt gekozen voor de ontwikkeling van een variabelen-selectiemethode voor PLS met meerdere responsen (PLS2). In hoofdstuk 7 werden vier nieuwe eigenschappen van voorspellende variabelen gedefinieerd en onderzocht, die zijn afgeleid van PLS2 regressiecoëfficiënten. Het betreft het “*gemiddelde van de absolute waarden van de PLS2 regressiecoëfficiënten*”, de “*norm van de PLS2 regressiecoëfficiënten*”, en de *significanties* ervan. Het bleek dat de vier nieuwe eigenschappen geschikt zijn om te gebruiken bij de FCAM-methode die is aangepast voor variabelenreductie met PLS2-modellen (FCAM-PLS2).
Het voorspellend vermogen van de modellen die zijn ontwikkeld na variabelenselectie op basis van deze vier eigenschappen is gelijkwaardig. De “*norm van de PLS2 regressie-coëfficiënten*” is het meest selectief. Daarbij worden kleine aantallen informatieve variabelen geselecteerd die chemisch relevant zijn voor de respons. De “*significantie van het gemiddelde van de absolute waarden van de PLS2 regressiecoëfficiënten*” is het minst selectief.

Samengevat. Bij het onderzoek dat wordt gepresenteerd in dit proefschrift zijn vijf nieuwe chemometrische methoden ontwikkeld en getest die kunnen helpen bij de beheersing van grote data sets. Het betreft (*i*) een monsterselectiemethode voor het samenstellen van gereduceerde data sets voor klassieke QSRR modellen voor RPLC, (*ii*) drie algemeen toepasbare variabelen-selectiemethoden voor PLS1 (RCAM-PLS1, FCAM-PLS1 en ICAM-PLS1), en (*iii*) een algemeen toepasbare variabelen-selectiemethode voor PLS2 (FCAM-PLS2). Deze methoden vormen een stevige basis voor het opzetten van een nieuwe onderzoekslijn die gericht is op de beheersing van grote data sets met chemometrie.

Curriculum Vitae

Johannes (Jan) Petrus Maria Andries werd geboren op 20 november 1945 in Tilburg. In 1963 behaalde hij het diploma HBS-B aan het Sint Odulphuslyceum in Tilburg. Daarna combineerde hij werken met studeren. In 1963 begon hij aan de deeltijdopleiding MO-A Natuurkunde en Scheikunde aan de Katholieke Leergangen in Tilburg. Het diploma werd behaald in 1967. De studie werd daarna onderbroken voor het vervullen van de dienstplicht als onderofficier bij de militaire administratie. In 1969 startte hij in deeltijd met de opleiding MO-B Scheikunde aan de Universiteit van Utrecht. Het diploma, met keuzevak fysische chemie, werd behaald in 1973. In 1980 slaagde hij aan de Universiteit van Utrecht voor het doctoraal examen Scheikunde met specialisatie Chemische thermodynamica en bijvak Pedagogiek en didactiek van de scheikunde.

Zijn werkzame leven begon in 1963. Gedurende twee jaar werkte hij als analist in het bedrijfsleven. In 1965 begon hij als docent natuur- en scheikunde aan de Apothekers-assistentenopleiding van de Katholieke Leergangen in Tilburg, een van de rechtsvoorgangers van de Avans Hogeschool. In 1968 stapte hij als docent over naar de Brabantse Medische Analistenschool (BMAS) te Breda. De BMAS vormde een onderdeel van de Katholieke Leergangen, is in 1979 van naam veranderd in Dr. Struycken-Instituut, en in 1986 opgegaan in de hogeschool. Vanaf 1984 was hij hoofd van de opleiding Laboratorium Informatica en Automatisering (LIA), vanaf 1990 hoofd van de opleidingen LIA en Chemie, en vanaf 2001 hoofd van de cluster laboratoriumopleidingen. In 2002 heeft hij de functie hoofd laboratoriumopleidingen om gezondheidsredenen ter beschikking gesteld. Sinds 2003 is hij medewerker en tevens coördinator van het lectoraat Analysetechnieken in de Life Sciences (ALS) van de Avans Hogeschool in Breda.

Hij heeft ruim 25 jaar het vak chemometrie gedoceerd[5] in de laboratoriumopleidingen van de Avans Hogeschool en is betrokken geweest bij de ontwikkeling van een leerplan chemometrie en bijbehorend lesmateriaal voor het Hoger Laboratorium Onderwijs (HLO) in Nederland[6]. Hij is co-auteur van het boek Chemometrie dat gebruikt wordt in het HLO[7].

Tijdens zijn werkzaamheden voor het lectoraat ALS is hij in deeltijd begonnen aan een promotie-onderzoek op het gebied van chemometrie. Het is een samengevoegd promotieonderzoek van het Departement Analytische Chemie van de Radboud Universiteit in Nijmegen onder supervisie van Prof. Dr. L.M.C. Buydens, en het Departement Analytische Scheikunde en Farmaceutische Technologie van de Vrije Universiteit Brussel in België onder supervisie van Prof. Dr. Y Vander Heyden. De resultaten van dit onderzoek staan beschreven in dit proefschrift.

[5] J. Andries : Chemometrie in de studierichting Laboratorium Informatica en Automatisering; Chemisch Magazine, juni/juli 1989, blz. 404-405

[6] J. Andries en A. de Vries: Chemometrie in het HBO; Chemisch Magazine, maart 1990, blz. 153

[7] J.P.M. Andries, A.B. de Vries: Chemometrie, 3^{e} druk, Syntax Media, Arnhem, 2007