UNIVERSITAT ROVIRA I VIRGILI
EXPERIMENTAL DESIGN APPLIED TO THE SELECTION OF SAMPLES AND SENSORS IN MULTIVARIATE CALIBRATION
Joan Ferré Baldrich
ISBN:978-84-691-1875-7/DL: T-337-2008

Experimental De

Samples and Sen ànica

Tesi Doctoral

UNIVERSITAT ROVIRA I VIRGILI

UNIVERSITAT ROVIRA I VIRGILI
EXPERIMENTAL DESIGN APPLIED TO THE SELECTION OF SAMPLES AND SENSORS IN MULTIVARIATE CALIBRATION
Joan Ferré Baldrich
ISBN:978-84-691-1875-7/DL: T-337-2008

Experimental Design Applied to the Selection of Samples and Sensors in Multivariate Calibration

Tesi Doctoral

UNIVERSITAT ROVIRA I VIRGILI

UNIVERSITAT ROVIRA I VIRGILI

Departament de Química Analítica i Química Orgànica

Àrea de Química Analítica

EXPERIMENTAL DESIGN APPLIED TO THE SELECTION OF SAMPLES AND SENSORS IN MULTIVARIATE CALIBRATION

Memòria presentada per

JOAN FERRÉ BALDRICH

per assolir el grau de

Doctor en Ciències Químiques

Tarragona, 1997

UNIVERSITAT ROVIRA I VIRGILI

DEPARTAMENT DE QUÍMICA ANALÍTICA
I QUÍMICA ORGÀNICA

Plaça Imperial Tàrraco, 1
43005 Tarragona
Tel. 34 (9) 77 55 81 37
Fax 34 (9) 77 55 95 63
e-mail: secqaqo@quimica.urv.es

Dr. FRANCESC XAVIER RIUS i FERRÚS, Catedràtic del Departament de Química Analítica i Química Orgànica de la Facultat de Química de la Universitat Rovira i Virgili,

CERTIFICA: Que la present memòria que duu per títol: "**EXPERIMENTAL DESIGN APPLIED TO THE SELECTION OF SAMPLES AND SENSORS IN MULTIVARIATE CALIBRATION**", ha estat realitzada per en JOAN FERRÉ BALDRICH sota la meva direcció a l'Àrea de Química Analítica del Departament de Química Analítica i Química Orgànica d'aquesta Universitat i que tots els resultats presentats són fruit de les experiències realitzades per l'esmentat doctorant.

Tarragona, octubre de 1997

Prof. F. Xavier Rius i Ferrús

UNIVERSITAT ROVIRA I VIRGILI
EXPERIMENTAL DESIGN APPLIED TO THE SELECTION OF SAMPLES AND SENSORS IN MULTIVARIATE CALIBRATION
Joan Ferré Baldrich
ISBN:978-84-691-1875-7/DL: T-337-2008

Als fulls que segueixen es resumeixen quasi quatre anys de feina davant d'un ordinador. Havia pensat paraules d'agraïment per a tots els que, d'una forma o altra, han estat per aquí tot aquest temps. Alguns han estat força transcendents i segurament han deixat la seva empremta en mi. Però he decidit resumir. Els que busqueu més avall el vostre nom, encara que no el trobeu escrit segur que hi sou. Així, doncs, voldria expressar la meva profunda gratitud ...

... en primer lloc al professor F. Xavier Rius, per haver acceptat dirigir aquest treball, per la seva dedicació i recolzament, per haver la seva confiança i ànims quan les coses no sortien i pel molt que m'ha ensenyat en tots els aspectes (no només científics) i que m'agradaria que continués ensenyant-me.

... al professor Roger Phan-Tan-Luu de la Université d'Aix-Marseille III, a qui considero un professor i una persona excepcional, per haver-me tractat sempre tant bé i haver acceptat estar al tribunal que jutja aquesta tesi. Ell em va transmetre la seva passió per la metodologia de la recerca experimental i a les seves ensenyances dec una part important d'aquest treball. Desitjo haver-les sabut reflectir prou bé.

... als doctors D.L. Massart, Romà Tauler, Jaume Puy i Itziar Ruisánchez, per haver-me fet l'honor d'acceptar jutjar aquest treball.

... als molts i bons companys i companyes, del grup de Quimiometria i de fora d'ell, pels molts moments que hem compartit aquests anys. I a algú molt especial per a mi... a tu, Montse (tu i jo ja sabem perquè).

... a tots aquells i aquelles de l'Àrea de Química Analítica que han tingut alguna cosa a veure, de prop o de lluny, amb la realització d'aquest treball.

... i sobretot moltes gràcies als de casa, pel seu constant suport i tots els i tots els sacrificis que desinteressadament han fet per mi.

UNIVERSITAT ROVIRA I VIRGILI
EXPERIMENTAL DESIGN APPLIED TO THE SELECTION OF SAMPLES AND SENSORS IN MULTIVARIATE CALIBRATION
Joan Ferré Baldrich
ISBN:978-84-691-1875-7/DL: T-337-2008

Als de casa...

UNIVERSITAT ROVIRA I VIRGILI
EXPERIMENTAL DESIGN APPLIED TO THE SELECTION OF SAMPLES AND SENSORS IN MULTIVARIATE CALIBRATION
Joan Ferré Baldrich
ISBN:978-84-691-1875-7/DL: T-337-2008

CONTENTS

UNIVERSITAT ROVIRA I VIRGILI
EXPERIMENTAL DESIGN APPLIED TO THE SELECTION OF SAMPLES AND SENSORS IN MULTIVARIATE CALIBRATION
Joan Ferré Baldrich
ISBN:978-84-691-1875-7/DL: T-337-2008

UNIVERSITAT ROVIRA I VIRGILI
EXPERIMENTAL DESIGN APPLIED TO THE SELECTION OF SAMPLES AND SENSORS IN MULTIVARIATE CALIBRATION
Joan Ferré Baldrich
ISBN:978-84-691-1875-7/DL: T-337-2008

UNIVERSITAT ROVIRA I VIRGILI
EXPERIMENTAL DESIGN APPLIED TO THE SELECTION OF SAMPLES AND SENSORS IN MULTIVARIATE CALIBRATION
Joan Ferré Baldrich
ISBN:978-84-691-1875-7/DL: T-337-2008

Chapter 1

Introduction and Objectives

UNIVERSITAT ROVIRA I VIRGILI
EXPERIMENTAL DESIGN APPLIED TO THE SELECTION OF SAMPLES AND SENSORS IN MULTIVARIATE CALIBRATION
Joan Ferré Baldrich
ISBN:978-84-691-1875-7/DL: T-337-2008

1.1 Introduction

The aim of the chapter is to present this thesis: the objectives, the structure of the thesis and the work reported. A short bibliographic revision is used to justify the objectives.

The present introductory chapter has been structured in several sections. Section §1.1 introduces the importance of multivariate calibration in the present chemical analysis. This section serves, together with the bibliographic revision in §1.2, as a background to situate this thesis and its objectives in the context of the chemical analysis. The section §1.3 contains the objectives of this thesis and §1.4 presents the structure of this thesis. Finally, §1.5 contains the references cited in this chapter.

1.2 Multivariate calibration in quantitative analysis

Analytical chemistry plays an important role in our society. Chemical analyses can be performed for almost any substance of interest. Although many different problems of varied nature are presented to fulfill the needs of the present society, many situations in chemical analysis consist of identifying some constituents of a sample (qualitative analysis) or determining their concentration (quantitative analysis).

Quantitative analysis assumes that the measurands, usually concentrations of the constituents of interest in a sample, are related to the quantities measured from the technique used for analyzing the sample. These measured quantities can be a volume, a weight, or signals (e.g. spectra, chromatograms or voltages) from instruments such as spectrophotometers, chromatographs, potentiometers, etc.

In some procedures the analyst may only be interested in the raw numeric result of this measurement that is either compared to previous measured values

UNIVERSITAT ROVIRA I VIRGILI
EXPERIMENTAL DESIGN APPLIED TO THE SELECTION OF SAMPLES AND SENSORS IN MULTIVARIATE CALIBRATION
Joan Ferré Baldrich
ISBN:978-84-691-1875-7/DL: T-337-2008

(e.g. in quality control) or used to detect relative changes in the data (e.g. inflexion points in a potentiometric curve). However, the usual situation when an instrumental technique is used to quantitatively analyze samples is to build a mathematical model relating the instrumental responses of a set of calibration samples to the quantities of chemical or physical variables such as analyte concentrations or indexes such as the octane number in fuels or the impact index in polymers. This relationship is used to predict these quantities from the instrumental response data of new *unknown* samples* measured in the same manner (e.g. spectra of test mixtures). Calibration is the process of building that mathematical model.

The motivation underlying multivariate calibration is to relate two types of measurements in a sample under study: one is easy to obtain and the other may either require expensive equipment, be time consuming, inaccurate or more difficult to obtain. The statistical model that establishes a relationship between the two series of measurements can be used for statistical inference of the unknown value of the difficult variable after observing the easy-obtainable variable. An example is the quantitative analysis using spectral data. Spectra are the *easy* measurements that can be related to the concentration of a determined analyte. Otherwise, the analyte should be obtained with the more costly reference or well-established method. If the instrumental method is faster and less costly than the reference method we could save time and money.

A variety of mathematical methods can be used to establish appropriate relationships between instrumental responses and chemical or physical measurands. Quantitative analysis has traditionally used univariate calibration based on a single measured signal and converting it to concentration via a calibration line. Two examples are the measurement of the pH and the single element atomic absorption. The major difficulty of this model is that the instrumental response must depend only on the concentration of the analyte of interest and all selectivity problems must be removed before the measurement. In these cases the analysis of mixture samples requires either a method to separate the analyte from the interferents or using a

* *Unknown sample* [1-5] will be used in this thesis to designate a sample to be analyzed (the problem sample). The usually objective is to predict the analyte concentration in this sample using a calibration model. The recommended name *test sample*[6] is not used to avoid confusion with a sample from the *test set* used in the modeling stage (see chapter 2).

highly selective instrument. The drawback is that the sample manipulation in the laboratory increases the cost of the analysis. On the other hand, the cost for making additional measurements on the unknown sample is decreasing due to the availability of analytical instruments that can perform many measurements per sample (multivariate signals) such as diode-array spectrophotometers and mass spectrometers. Hence, there is a tendency to make less sample manipulation and fewer experiments but to obtain more data in each of them. The simultaneous use of multiple responses and multivariate calibration can overcome the limitations of univariate calibration employed on methods which give *single point* measurements. Since many measurements are made in each sample, multivariate calibration can separate analytes from interferences without the need of highly selective measurements for the analyte and enables concentration determinations from non-selective measurements, thus improving the applicability of quantitative spectral analysis. The so called *inverse models* can calibrate for individual constituents in samples with very complex compositions, provided that the future unknown samples exhibit the same behavior as the calibration samples.

One advantage of multivariate signals combined with chemometric techniques is the considerable reduction in cost and time of the analysis. This is due to the reduction in the number of steps in the sample manipulation since it is not necessary to achieve complete separation or selectivity. Moreover, the concentration of more than one analyte can be determined at a time. However, more complex mathematical expressions than the simple univariate calibration are required. Sometimes, these models give a less precise or accurate result than the obtained with the traditional method of analysis but more rapid and cheaper. Examples of that are the analysis of protein in wheat or of water in meat. For these reasons, multivariate calibration methods are increasingly used in laboratories and spectrometers are becoming measurement devices of preferred choice in many current applications. Several reviews and textbooks are available on the subject[1,3,4,7-12].

The quality of the result of an analysis is influenced by all the steps involved in the analysis. Since prediction with multivariate calibration models is becoming a common step of the analytical procedure, the necessity of building these models that offer the guarantee of precise and unbiased predictions is clear. Actually, the development and improvement of these models is one of the focuses of chemometric research at the moment. Recent examples of this development are

new regression methods[13,14] and new equations that enable to understand the effect of the measurement errors and their propagation through the model on the error of the predicted concentrations[5]. The selection of the most adequate samples and sensors for calibration is also of primary concern for economical and statistical reasons. The economical reasons involve the cost of analyzing the calibration samples with a well-established or reference method and the cost of the instrumentation to measure the responses used in the analysis. The statistical reasons involve the ability of the quantitative calibration model to ensure good predictions for new samples. This is significantly influenced by the samples and sensors used for calibration.

Although the selection of sensors is regularly investigated, the proper selection of calibration samples seems more forgotten in the analytical literature and not many papers refer to this problem (see §3.1.3 for a bibliographic revision). In addition, the methods proposed for sample selection have some drawback about their real applicability such as not a clear mathematical criteria to decide the optimal number of samples to be used and to discard the samples that simply have redundant information. New criteria and methodologies must be developed to meet the quality criteria in multivariate calibration methods. The solution could be offered by the experimental design theory, until now used in multiple linear regression (MLR) but rarely employed in multivariate calibration models such as principal component regression (PCR) or partial least squares (PLS) regression.

Concerning the wavelength selection in multivariate calibration, the usual problem is how to identify the best range of wavelengths or individual sensors for prediction. Different methods and criteria for optimal wavelength selection have been proposed, specially in classical least squares regression (CLS) and the experimenter interested in using a selection criterion must review (and understand!) a large volume of literature. In addition, there is some discrepancies about the performance of these criteria and whether if they improve the precision, the trueness or the accuracy of the result. Many times the authors do no define what they consider to be accuracy or precision so that their results are difficult to compare. The usefulness of these criteria and their effect on the modern concepts of precision and trueness of the results should be clarified.

Moreover, a software to calculate the criteria and methodologies should be elaborated to facilitate its application in the laboratory.

UNIVERSITAT ROVIRA I VIRGILI
EXPERIMENTAL DESIGN APPLIED TO THE SELECTION OF SAMPLES AND SENSORS IN MULTIVARIATE CALIBRATION
Joan Ferré Baldrich
ISBN:978-84-691-1875-7/DL: T-337-2008

1.3 Objectives of the thesis

The objective of this doctoral thesis is to study the sample and sensor selection criteria for the optimization of the multivariate calibration models used in analytical methods using the concepts from the experimental design. The criteria found in the scientific literature are considered and, when possible, improved procedures for sensor and sample selection are proposed.

More specifically, we contribute to the study of:

1. A new sample selection method in principal components regression (PCR) based on the application of the D-optimality criterion and the Fedorov's exchange algorithm. The method only uses the instrumental responses of the candidate samples to select the minimum number of calibration samples to build the model. The analyte concentration is only determined for the selected samples. Complementing this method, a new procedure for the fast selection of the relevant factors in PCR is devised. The Fedorov's algorithm was also applied to select samples to check if model standardization is necessary in PCR models. The performance of the sets selected using the D-criterion or the Kennard-Stone algorithm in MLR was compared.

2. New guidelines for sensor selection in CLS based on the experimental design theory. The different criteria for wavelength selection are critically reviewed in the modern terms of the precision , accuracy and trueness. They are interpreted from the point of view of the experimental design theory using the confidence hyperellipsoid of the predicted concentrations. The effects of collinearity in CLS and its effects in the quality of the predicted concentrations have been studied.

3. A new method to detect and reduce bias in future test samples in multicomponent analysis (CLS) .

4. The methodologies and algorithms have been written in m-files Matlab. In the future they will be implemented in Toolbox for their application in the laboratory.

UNIVERSITAT ROVIRA I VIRGILI
EXPERIMENTAL DESIGN APPLIED TO THE SELECTION OF SAMPLES AND SENSORS IN MULTIVARIATE CALIBRATION
Joan Ferré Baldrich
ISBN:978-84-691-1875-7/DL: T-337-2008

1.4 Structure of the work presented

The thesis has been structured in five chapters. Each chapter is divided in sections. The first section of each chapter is the introduction and contains the aim, the summary of the contents of the chapter and the justification of the objectives with a bibliographic revision. This section ends with the references used in the bibliographic revision. The next sections in the chapter contain the experimental work written as papers (either published, submitted or in preparation). Each paper contains the following parts: introduction, theoretical background, experimental part, discussion and results, conclusions and references. The conclusions of each chapter are, together with the general conclusions of the thesis, in the chapter 5. The contents of each chapter are the following:

- Chapter 1 contains the introduction to the work reported in this thesis, the structure of the thesis and the major aims of the work that is developed in the following chapters.

- Chapter 2 contains six sections of theoretical (not experimental) nature. Section §2.1 is the introduction to the chapter. Section §2.2 has the notation and list of symbols used in this thesis. The notation used in the published papers is adequately indicated in the paper and is very close to the indicated in §2.2. Section §2.3 introduces the multiple linear regression (MLR) and the least squares solution, quite used in this thesis. Some concepts of the experimental design theory, focused on the optimality criteria to select samples in MLR are reviewed. These ideas are then used in the chapters §3 and §4. Section §2.4 deals with the basic concepts of the multivariate calibration methods used in this thesis. A review of their theoretical basis, that is disperse in many papers and monographs, is considered important to understand the need of sample and wavelength selection. This part also presents some new concepts related to the net analyte signal in classical least-squares (CLS) calibration that are later used in the section §4.3 and §4.7. The section §2.5 contains guidelines for applying experimental design to the calibration models. These ideas are used for sample and wavelength selection in chapters 3 and 4. Chapter 2 ends with the references (§2.6) of this chapter.

UNIVERSITAT ROVIRA I VIRGILI
EXPERIMENTAL DESIGN APPLIED TO THE SELECTION OF SAMPLES AND SENSORS IN MULTIVARIATE CALIBRATION
Joan Ferré Baldrich
ISBN:978-84-691-1875-7/DL: T-337-2008

• Chapter 3 deals mainly with the selection of the best calibration sample subset for PCR. The procedures proposed in the literature for sample selection are critically reviewed (§3.1.3) A new methodology to select samples from the instrumental responses of a large subset when the samples cannot be synthesized is presented in *Selection of best calibration sample subset for multivariate regression.* Joan Ferré, F. Xavier Rius *Anal. Chem.* 68, (1996) 1565-1571 (section §3.2). The next paper *Determination of ethylene content in poly(propylene-ethylene) copolymers using near-infrared spectra (NIR) and multivariate calibration* Villagrasa C., Ferré J., Larrechi M.S., Rius F.X., García C. (*in preparation*) (section §3.3) compares the predictive ability of PLS and of PCR with the factors selected according to the methodology described in §3.2 using near-infrared (NIR) data of industrial copolymers. Section §3.4 is the paper *Constructing D-optimal designs from a list of candidate samples.* Joan Ferré, F. Xavier Rius *Trends Anal. Chem.* 16 (1997) 70-73 and is a comparative study of the Fedorov's algorithm, the popular Kennard-Stone algorithm and the random division of samples into calibration and validation sets for the selection of samples for a MLR model. §3.5 is the paper *Selection of calibration points for PCR in QSAR studies.* Joan Ferré, F. X. Rius (*in preparation*). It deals with the selection of calibration samples in quantitative structure-activity relationship (QSAR) studies based on the principal components using the Fedorov's algorithm. The last paper, *Assessing the validity of principal component regression models in different analytical conditions.* Rius A.; Callao M.P., Ferré J.; Rius F.X., *Anal. Chim. Acta* 337 (1997) 287-296 presents a procedure for selecting samples for assessing if a PCR model is still valid before using the piecewise direct standardization (PDS) technique when the working conditions are different from those used for modeling. The contribution to this work consists on applying the D-optimality criterion for selecting, from a large set, the minimum number of samples that must be analyzed in the new conditions.

• Chapter 4 deals with the selection of the best sensor subset for multicomponent analysis. In the paper *A graphical criterion to examine the quality of multicomponent analysis. Implications for wavelength selection.* J. Ferré and F.X. Rius *Trends Anal. Chem.* 16 (1997) 155-162 (section §4.2) the basic concepts used in sensor selection in CLS are explained on the basis of their effect on the volume, shape and orientation of the confidence region (an ellipsoid) of the predicted concentrations. This paper is complemented with *Further considerations on the sensitivity and selectivity of multicomponent systems.* J. Ferré and F.X. Rius. *In preparation* (section §4.3). Here, the

mathematical expressions of sensitivity, selectivity, variance-proportion decompositions and condition number are discussed and interpreted using the confidence ellipsoid. The effect of adding a new sensor to the calibration matrix on the confidence ellipsoid is shown. The section §4.4 is the paper *Equivalence between Selectivity and Variance Inflation factors in multicomponent analysis.* J.Ferré, F.X. Rius *Química Analítica* 15 (1996) 259-262, where the mathematical equivalence between selectivity and variance inflation factors is shown. Both can be used as measures of collinearity in CLS. The section §4.5 is a tutorial where the ISO definitions of accuracy, trueness and precision are reviewed and their relationships with the wavelength selection criteria in CLS are revisited and clarified. This is motivated by the confusion in the literature about the effect of these criteria on the accuracy, trueness and precision of the results. The section §4.6 is the paper *Figures of Merit in Multivariate Calibration. Determination of Four Pesticides in Water by FIA and Spectrophotometric Detection.* J. Ferré, R. Boqué, B. Fernández-Band, M.S. Larrechi and F.X. Rius. *Anal. Chim. Acta* 348 (1997) 167-175. This paper studies the variance proportion decompositions and the effects of the selectivity and sensitivity in the prediction error of four pesticides analyzed with a FIA system and CLS. The whole published paper has been included although the part concerning the detection limits is not a subject of this thesis. The last part of chapter 4, §4.7, is the paper *Detection and correction of biased results of individual analytes in multicomponent spectroscopic analysis* J.Ferré, F.X. Rius, *Submitted for publication.* This work was motivated by the fact that, in the paper in §4.6, the large selectivity and sensitivity values of the analytes did not agree with the calculated prediction errors. It was supposed that the large inexplicable errors were not due to the instability of the system of equations but to an erroneous preparation of the validation samples with a deficient assigned value of the concentration. In the present paper a tool for internal validation of the standards and the validation samples based on the net analyte signal is developed. It enables the bias in CLS models to be detected taking advantage of the multivariate signal. A wavelength selection procedure is used to select the wavelength with less prediction error.

• Chapter 5 contains the conclusions of the chapters 3 and 4 and the general conclusions of the thesis. The advantages and drawbacks of the proposed methodologies are commented and the trends for future work are devised. It must be reminded that each paper already has its particular conclusions.

The format of the published papers in this thesis. The published papers have been edited to give a uniform format to the thesis but the contents have not been changed. The only change in nomenclature with respect to what has been published can be found in paper §4.6 *Equivalence between Selectivity and Variance Inflation factors in multicomponent analysis.* J.Ferré, F.X. Rius *Química Analítica* 15 (1996) 259-262 where the sign ' , used in the paper to indicate transposition, has been changed in the thesis for a ^{T}.

1.5 References

1. Haaland D.M., Thomas E.V. *Anal. Chem.* 60 (1988) 1193-1202.
2. Lorber A., Kowalski B.R. *J. Chemom.* 2 (1988) 93-109.
3. Booksh K.S., Kowalski B.R *Anal. Chem.* 66 (1994) 782A-804A.
4. Beebe K.R., Kowalski B.R. *Anal. Chem.* 59 (1987) 1007A-1017A.
5. Faber K., Kowalski B.R. *J. Chemom.* 11 (1997) 181-238.
6. ISO 3534-1 (1993) *Statistics-Vocabulary and symbols.* International Organization for Standardization, Geneva, Switzerland.
7. Martens H., Naes T. *Multivariate Calibration,* Wiley: New York, 1989.
8. Sanchez E., Kowalski B.R. *J. Chemom.* 2 (1988) 247-263.
9. Kowalski B.R., Seasholtz M.B. *J. Chemom.* 5 (1991)129-145.
10. Thomas E.V. *Anal. Chem.* 66 (1994) 795A-804A.
11. Geladi P., Kowalski B.R. *Anal. Chim. Acta.* 185 (1986) 1-17.
12. Martens H., Karstang T., Naes T. *J. Chemom.* 1 (1987) 201-219.
13. Vigneau E., Devaux M.F., Qannari E.M., Robert P. *J. Chemom.* 11 (1997) 239-249.
14. Wentzell P.D., Andrews D.T., Kowalski B.R. *Anal. Chem.* 69 (1997) 2299-2311.

UNIVERSITAT ROVIRA I VIRGILI
EXPERIMENTAL DESIGN APPLIED TO THE SELECTION OF SAMPLES AND SENSORS IN MULTIVARIATE CALIBRATION
Joan Ferré Baldrich
ISBN:978-84-691-1875-7/DL: T-337-2008

2.1. Introduction

2.1.1 Aim of the chapter

The aim of this chapter is to introduce the concepts about experimental design and multivariate calibration used in this thesis. It is also a compendium of the large amount of information disperse in many papers and books.

2.1.2 Structure of the chapter

This chapter has six parts: introduction (§2.1), notation and definitions (§2.2), concepts of experimental design theory in general statistical terms (§2.3), the theoretical background of the multivariate calibration models used in this thesis and their advantages and limitations (§2.4), the connection points between the experimental design theory and the multivariate calibration models (§2.5) and the collinearity problem in the calibration models (§2.6).

2.2 Notation and definitions

Multivariate calibration can be used with any type of suitable multivariate data to model any property. However, the terminology used here associates instrumental responses with sample spectra and the property of interest with the analyte concentration since the major part of this thesis deals with this type of data.

Matrices are represented by bold capital letters, e.g. **R**, bold lowercase letters denote column vectors, e.g. **c** (row vectors are transposed column vectors) and italic characters represent scalars, e.g. c_k. True values are indicated by Greek characters or

the subscript $_{true}$. Calculated or measured values are indicated by Roman characters. The *hat* (^), used in the literature to indicate *calculated,* has been dropped from the symbols to simplify the notation; if the magnitude is measured or calculated can be deduced from the context. The running indexes in multivariate calibration are: $k = 1$ to K analytes are present in $i = 1$ to I calibration samples whose instrumental responses are measured using j=1 to J sensors. The following is the list of symbols used in this thesis:

Symbols beginning with a Roman letter

A-	criterion of the trace of the dispersion matrix
D-	criterion of the determinant of the dispersion matrix
G-	criterion of the maximal variance function
a	index for factors
A	dimensionality (number of factors)of PCR or PLS models
$\mathbf{a}_k=[a_{1,k},\ldots\ a_{j,k}\ \ldots\ a_{J,K}]^T$	(J×1) spectrum of the analyte k at J wavelengths at concentration c_k^0
$\mathbf{a}_k^*$	(J×1) net analyte signal for the analyte k in a pure analyte spectrum at concentration c_k^0
$\mathbf{A}=[\mathbf{a}_1,\ldots\ \mathbf{a}_k\ \ldots\ \mathbf{a}_K]$	(J×K) spectra of the K components at J wavelengths at concentration c_k^0
$\mathbf{A}_k=[\mathbf{a}_1,\ldots,\ \mathbf{a}_{k-1},\ \mathbf{a}_{k+1},\ldots\mathbf{a}_K]$	(J×(K-1)) matrix **A** without the *kth* column $\mathbf{a}_k$
b_j	estimated coefficient of the variable j
$b_{j,k}$	estimated coefficient of the variable j for the analyte k
$\mathbf{b}=[(b_0),\ b_1,\ \ldots,\ b_j,\ \ldots,\ b_J]^T$	((J+1)×1 or J×1) estimated coefficients
$\mathbf{b}_k=[b_{1,k},\ \ldots,\ b_{j,k},\ \ldots,\ b_{J,k}]^T$	(J×1) estimated coefficients for the analyte k
$\mathbf{b}_{k,CLS}\ \mathbf{b}_{k,ILS}\ \mathbf{b}_{k,PCR}\ \mathbf{b}_{k,PLS}$	vectors of coefficients for the analyte k estimated in the CLS, ILS, PCR and PLS models respectively
$\mathbf{B}=\{b_{j,k}\}=[\mathbf{b}_1,\ \ldots,\ \mathbf{b}_k,\ \ldots,\ \mathbf{b}_K]$	(J×K) estimated coefficients for K analytes and J variables
$c_{i,k}$	concentration of the analyte k in the sample i
$c_{un,k}$	concentration of the analyte k in the unknown sample
$\overline{c}_k$	mean value of $\mathbf{c}_k$
$\mathbf{c}_k=[c_{1,k},\ \ldots,\ c_{i,k},\ \ldots,\ c_{I,k}]^T$	(I×1) concentration of the analyte k in I calibration samples
$\mathbf{c}$	(K×1) concentration of K analytes in a sample
$\mathbf{c}_{un}$	(K×1) concentration of K analytes in an unknown sample
$\mathbf{C}=\{c_{i,k}\}=[\mathbf{c}_1,\ \ldots,\ \mathbf{c}_k,\ \ldots,\ \mathbf{c}_K]$	(I×K) concentrations of K analytes in I calibration samples. A column is the concentration of one analyte in the I calibration samples. A row is the concentration of the K analytes in one

UNIVERSITAT ROVIRA I VIRGILI
EXPERIMENTAL DESIGN APPLIED TO THE SELECTION OF SAMPLES AND SENSORS IN MULTIVARIATE CALIBRATION
Joan Ferré Baldrich
ISBN:978-84-691-1875-7/DL: T-337-2008

	calibration sample
$\mathbf{C}_0$	(K×K) diagonal matrix whose diagonal elements are c_k^0
$cov(x_1,x_2)$	covariance of the random variables x_1 and x_2
$d(\mathbf{x}_{un})$	variance function at the point $\mathbf{x}_{un}$
$\mathbf{D}$	(dimensions depend on the equation) diagonal matrix of singular values of **R** or **S**
$\mathbf{D}_A$	(A×A) diagonal matrix of A singular values of **R** or **S**
$\mathbf{E}$	(dimensions depend on the equation) residual matrix
$\mathbf{f}(\mathbf{x}^i)$	(Q×1) regression functions evaluated at the point $\mathbf{x}^i$. Transposed row of the **X** matrix
$F_{\nu 1,\ \nu 2,\alpha}$	upper α-percentage point of F-distribution with ν_1 and ν_2 degrees of freedom
$\mathbf{H}$	(I×I) orthogonal projection matrix. Hat matrix.
i	index for experiments, calibration samples, objects or points
I	number of experiments or samples in the calibration matrix
I_P	number of samples in the validation set
$\mathbf{I}$	appropriately dimensioned identity matrix
j	index for independent variables or wavelengths or sensors
J	number of independent variables or wavelengths in a spectrum
k	index for analytes
K	number of analytes or components or constituents in a sample
$\mathbf{M}$	(Q×Q) normalized information matrix or matrix of moments
$\max(\lambda_j)$	the largest eigenvalue of $(\mathbf{X}^T\mathbf{X})^{-1}$
$\min(\lambda_j)$	the lowest eigenvalues of $(\mathbf{X}^T\mathbf{X})^{-1}$
N	number of candidate points
$N(\mu,\sigma^2)$	normal distribution with mean μ and variance σ^2
$\mathbf{P}$	(J×J) eigenvectors of $\mathbf{R}^T\mathbf{R}$
$\mathbf{p}_a$	(J×1) ath column of **P**, eigenvector of the ath factor
$\mathbf{P}_A$	(J×A) selected factors
Q	number of coefficients in a MLR model
r_j	instrumental response measured at the sensor j
$r_{i,j}$	instrumental response of the sample i measured at the sensor j
$r_{un,j}$	instrumental response of the unknown sample measured at the sensor j
$r_{un,j,k}^*$	net analyte signal of the analyte k at the sensor j in the unknown sample
$\mathbf{r}=[r_1, ..., r_j, ..., r_J]^T$	(J×1) instrumental responses measured at J sensors (e.g. absorbances at J wavelengths)
$\mathbf{r}_i=[r_{i,1}, ..., r_{i,j}, ..., r_{i,J}]^T$	(J×1) instrumental responses of the calibration sample i measured at J sensors
$\mathbf{r}_{un}=[r_{un,1}, ..., r_{un,j}, ..., r_{un,J}]^T$	(J×1) instrumental responses of the unknown sample measured at J sensors

$\mathbf{r}_{un,k}^*$	(J×1) net analyte signal for the analyte k in the unknown sample
$\bar{\mathbf{r}}$	(J×1) column means of $\mathbf{R}$
$\mathbf{R}=\{r_{i,j}\}$	(I×J) instrumental responses of I calibration samples measured at J sensors
s^2	estimate of σ^2
$s_{j,k}$	partial sensitivity of the analyte k in the sensor j, defined as the slope of the analytical calibration plot of the response of the sensor j to the concentration of the analyte k (responses divided by the concentration of the analyte in a pure sample). It is usually the molar absorptivity coefficient of the pure component at unit pathlenght
$s_{j,k}^*$	net analyte signal of the analyte k in the sensor j of the spectrum of the analyte k pure
$\mathbf{s}_k=[s_{1,k},\ldots s_{j,k} \ldots s_{J,K}]^T$	(J×1) partial sensitivities of the analyte k at J wavelengths
$\mathbf{s}_k^*$	(J×1) net analyte signal for the analyte k in a pure analyte spectrum
$\mathbf{S}=[\mathbf{s}_1,\ldots \mathbf{s}_k \ldots \mathbf{s}_K]$	(J×K) partial sensitivities of the K components at J wavelengths
$\mathbf{S}_k=[\mathbf{s}_1,\ldots, \mathbf{s}_{k-1}, \mathbf{s}_{k+1},\ldots \mathbf{s}_K]$	(J×(K-1)) matrix $\mathbf{S}$ *without the kth column* $\mathbf{s}_k$
$t_{i,a}$	ath score associated with the ith sample
$\mathbf{t}_a$	(I×1) sample scores for the ath factor
$\mathbf{t}_A^i$	(A×1) scores of the sample i for A factors
$\mathbf{t}_{un,A}$	(A×1) scores of the unknown sample for A factors
$\mathbf{T}_A$	(I×A) scores of $\mathbf{R}$ for A factors
$\mathbf{U}_A$	(I×A) normalized PCA scores for A factors
$UVIF_j$	jth diagonal element of $(\mathbf{X}^T\mathbf{X})^{-1}$ called the unscaled variance inflation factor (or just variance coefficient) of the coefficient j
$var()$	variance of a scalar quantity
$\mathrm{var}(\mathbf{v})$	symmetric variance-covariance matrix of the vector $\mathbf{v}$
$\mathbf{v}_a$	ath eigenvector of $(\mathbf{X}^T\mathbf{X})^{-1}$
VIF_j	variance inflation factor of the coefficient j
x	independent variable or input variable
$x_{i,j}$	value of the variable j in the experiment (point) i
$\mathbf{x}_j$	column j of $\mathbf{X}$
$\mathbf{x}^i$	row i of $\mathbf{X}$, a point of the experimental domain
$\bar{\mathbf{x}}$	column means of $\mathbf{X}_{-1}$
$\mathbf{X}=(\mathbf{x}_j)=(\mathbf{x}^i)$	(I×Q) (usually I×J or I×J+1) model matrix
$\mathbf{X}_{-1}$	(I×Q-1) model matrix with the column of ones deleted
$\mathbf{X}^T\mathbf{X}$	(Q×Q) information matrix
$(\mathbf{X}^T\mathbf{X})^{-1}$	(Q×Q) dispersion matrix
$(\mathbf{X}^T\mathbf{X})^{-1}\sigma^2$	(Q×Q) variance-covariance matrix
y	dependent variable, output variable, response
y_i	measured output variable in the experiment i

UNIVERSITAT ROVIRA I VIRGILI
EXPERIMENTAL DESIGN APPLIED TO THE SELECTION OF SAMPLES AND SENSORS IN MULTIVARIATE CALIBRATION
Joan Ferré Baldrich
ISBN:978-84-691-1875-7/DL: T-337-2008

y_{un}	predicted output variable in the experiment the point x_{un}
$\bar{y}$	mean value of the elements of **y**
y	(I×1) measured output variable in I experiments
$\mathbf{y}_{pred}$	(I×1) predicted values of the output variable in I experiments

Symbols beginning with a Greek letter

α	significance level of a statistical test
β_j	true regression coefficient for the variable j
$\beta=[(\beta_0), \beta_1,...,\beta_j,...,\beta_J]^T$	(J×1 or (J+1)×1) true regression coefficients of the model
$\beta_k=[\beta_{1,k},...,\beta_{j,k},...,\beta_{J,k}]^T$	(J×1) true regression coefficients for the analyte k
ε_i	unknown error in the experiment i
ε	(I×1) unknown errors in I experiments
η_i	true value of the output variable in the experiment i
η	(I×1) true values of the output variable in I experiments
λ_a	ath eigenvalue of $(\mathbf{X}^T\mathbf{X})^{-1}$
λ_j	jth eigenvalue of $(\mathbf{S}^T\mathbf{S})^{-1}$
σ_a	ath singular value of **R**
σ_i	ith singular value of **S**
σ	standard deviation
σ^2	variance of a scalar quantity
ξ_N	(N×J) matrix of candidate points
ξ_I	(I×J) matrix of experiments (points)
ξ^{*C}	matrix of experiments optimal for the C-criterion
χ	experimental domain of interest
Ξ_I	set of matrices of I experiments (points)
$\theta_{k,a}$	true regression coefficient for scores of factor a
θ_k	true regression vector with respect to scores

Other symbols

, (coma)	adds columns in matrices or vectors: $\mathbf{P}=[\mathbf{p}_1, \mathbf{p}_2]=[\mathbf{p}_1\ \mathbf{p}_2]$
; (semicolon)	adds rows in matrices or vectors: $\mathbf{X}=[\mathbf{x}_1^T; \mathbf{x}_2^T] = \begin{bmatrix}\mathbf{x}_1^T \\ \mathbf{x}_2^T\end{bmatrix}$
*	(superscript) net analyte signal
$\|\cdot\|$	2-norm: Euclidean norm
1	appropriately dimensioned vector of ones
Cond(**X**)	condition number of the matrix **X**
Det(**X**)	determinant of the matrix **X**

UNIVERSITAT ROVIRA I VIRGILI
EXPERIMENTAL DESIGN APPLIED TO THE SELECTION OF SAMPLES AND SENSORS IN MULTIVARIATE CALIBRATION
Joan Ferré Baldrich
ISBN:978-84-691-1875-7/DL: T-337-2008

$E[\cdot]$	expected value
eq eqs	equation equations
$i \times j$	(subscript) dimensions of a vector or matrix
k-row k-col	(subscript) kth row or column of the matrix
max	maximum value in a list
min	minimum value in a list
Tr(X)	trace of the matrix **X**
$\mathbf{X}^T$ $\mathbf{x}^T$	(superscript) transposition of a matrix **X** or a vector **x**
$\mathbf{X}^+$	(superscript) Moore-Penrose pseudo-inverse of the matrix **X**
$\mathbf{X}^{-1}$	(superscript) the inverse of a matrix of the matrix **X**
un	(subscript) magnitude related to the unknown sample

Definitions

Orthonormal. If **X** is orthonormal then $\mathbf{X}^T\mathbf{X}=\mathbf{X}\mathbf{X}^T=\mathbf{I}$

Singular-value decomposition (SVD). Decomposes a matrix **X** ($I \times Q$ of rank k) into three matrices $\mathbf{X}=\mathbf{U}\mathbf{D}\mathbf{P}^T$ with $\mathbf{U}^T\mathbf{U}=\mathbf{P}^T\mathbf{P}=\mathbf{I}$ ($k \times k$) and **D** diagonal with the k positive singular values on the diagonal.

Moore-Penrose pseudo-inverse [1,2]: used for non-square or overdetermined matrices, where the inverse cannot be calculated. The pseudo-inverse of a matrix **X** ($I \times Q$ of rank k) is calculated as $\mathbf{X}^+=\mathbf{P}\mathbf{D}^{-1}\mathbf{U}^T$ where **P**, **U** and **D** result from the SVD of **X**. The pseudo-inverse is also $\mathbf{X}^+=(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ if **X** has linearly independent columns. If **y** is an I-vector of responses, the ordinary least squares (OLS) estimator of the regression coefficients is $\mathbf{b}_{OLS}=(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}=\mathbf{X}^+\mathbf{y}$. When some rows or columns of **X** are linearly dependent or **X** has more columns than rows (i.e. **X** is not of full rank, $k<Q$), the determinant of $\mathbf{X}^T\mathbf{X}$ is zero, $(\mathbf{X}^T\mathbf{X})^{-1}$ does not exist and there is no unique least squares estimator. However, the pseudo-inverse can still be calculated and $\mathbf{b}_{MLLS}=\mathbf{X}^+\mathbf{y}$ is the unique minimum length least squares (MLLS) solution, which means that the solution **b** minimizes $\mathbf{b}^T\mathbf{b}$ in the case that $\mathbf{X}^T\mathbf{X}$ is singular[3]. The OLS and MLLS estimator coincide in the full rank case[4].

UNIVERSITAT ROVIRA I VIRGILI
EXPERIMENTAL DESIGN APPLIED TO THE SELECTION OF SAMPLES AND SENSORS IN MULTIVARIATE CALIBRATION
Joan Ferré Baldrich
ISBN:978-84-691-1875-7/DL: T-337-2008

2.3 Experimental design in multiple linear regression

2.3.1 The study of chemical systems

Chemical systems can be represented as in Figure 2.1 where one or more outputs y, also called *dependent variables* or *responses* can be measured. Their result depends of one or more input variables x *, also called *independent variables* [5]:

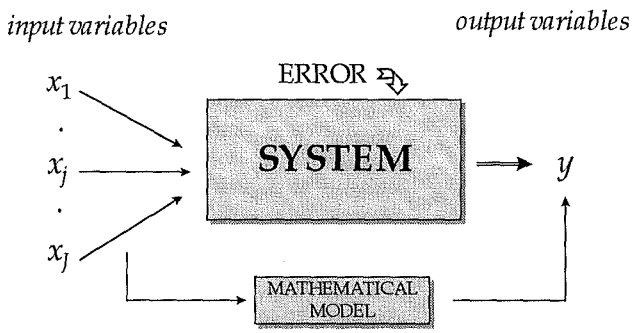


Figure 2.1. Representation of a system under study.

The aim of research in many chemistry areas is to understand and improve the system under study by finding the dependence between the input and the output variables. This requires conducting *experiments* by varying the values (*levels*) of one or more input variables and studying the changes in the response. The comparison of the obtained data puts into evidence that dependence and enables to identify the optimal conditions, the variables that most influence the results and those that do not, the presence of interactions, etc. The range of values that can take each variable is the *domain* and the combination of the domain of all the variables is the *domain of the variables.* This contains all the feasible experiments (the *possible experimental domain*) from where the *experimental domain of interest* (χ) may be a part

* The word *factor* used in experimental design as the cause of the studied phenomenon is not used here to avoid confusion with the *factors* from PCA and PLS.

of it. The variables are usually expressed in coded units. The set of I experiments to be performed, expressed in coded units, constitutes the *matrix of experiments* (ξ_I), where each row represents one experiment and each column one of the J variables. The *experimental plan* is the matrix of experiments expressed in raw variables.

To study such dependences, the design of the experimental runs is many times based on the experimenter's intuition and it is not rare to use the sequential method of changing one-variable-at-a-time experiments. However, this has been shown to be inefficient[6-8]: it may require too many experiments if the number of variables and levels to study is large and it provides no information on interactions among the variables. A more efficient way of conducting the study consists of changing simultaneously several variables in the same experiment. The important decisions derived from the experimental results and the non-negligible cost of the experimentation may advise using a methodological approach to establish the optimal organization of the experiments. Statistical experimental designs provide the mathematical framework for studying several variables simultaneously in a small number of experiments and obtaining information on interaction behavior.

2.3.2 Aims of experimental design

Experimental design, also called *design of experiments* (DOE), stands for a systematic way of planning the experiments aimed at efficiently extracting information employing statistical tools. The DOE helps the experimenter to select an optimal experimental strategy to achieve the proposed objectives. This includes:

1. To devise a reduced-cost set of experiments necessary to obtain the answer to a problem (e.g. the influence of the variables on the system or optimization of a response through a mathematical model among others[9]) by varying systematically multiple variables in a single experiment. This ensures the maximum efficiency of the experimental work and avoids random assays, redundant information and the pitfalls of the one-component-at-a-time method. This affects the cost of the experimentation.

UNIVERSITAT ROVIRA I VIRGILI
EXPERIMENTAL DESIGN APPLIED TO THE SELECTION OF SAMPLES AND SENSORS IN MULTIVARIATE CALIBRATION
Joan Ferré Baldrich
ISBN:978-84-691-1875-7/DL: T-337-2008

2. To assure that the selected experiments will yield a correct and reliable information about the influence of each variable on the system by minimizing the variance of estimated coefficients obtained through regression[10]. In addition the experiments must enable to interpret the obtained information and the generalization of the conclusions in the domain of interest. Statistical analysis methods are employed to determine whether a treatment is statistically significant in influencing the system response. This affects the quality of the result.

2.3.3 The linear model of multiple independent variables

2.3.3.1 The multiple linear regression (MLR) model

Many times the dependece between the input and output variables is expressed as a *mathematical model* * :

$$\eta = F(x_1, \ldots x_j, \ldots x_J, \beta_1, \ldots \beta_q, \ldots \beta_Q) \qquad (2.1)$$

where η is the true value of the response and β_q (q=1, ..., Q) are the true coefficients of the model, supposed constant in the domain under study. The model can be used to describe experimental results, to interpret the influence of the independent variables in the response or for prediction purposes in the experimental domain. The mathematical expression depends on the phenomenon being described and on the domain (usually, the smaller the domain, the simpler the model can be). The models considered here are linear with respect to coefficients since they usually represent a sufficient approximation to the reality of the domain. They can be written as:

$$\eta = \mathbf{f}^T(x_1, \ldots x_j, \ldots x_J)\beta \qquad (2.2)$$

where β is the vector of coefficients and $\mathbf{f}(x_1, \ldots x_j, \ldots x_J)$ is a vector of x variables that represents the equation of the model (it may include, for example, the

* Since regression actually is a statistical problem, along §2.3 the common symbols x and y are used for the independent and dependent variables respectively instead of a specialized chemical notation.

UNIVERSITAT ROVIRA I VIRGILI
EXPERIMENTAL DESIGN APPLIED TO THE SELECTION OF SAMPLES AND SENSORS IN MULTIVARIATE CALIBRATION
Joan Ferré Baldrich
ISBN:978-84-691-1875-7/DL: T-337-2008

measured variables, cross-products and powers e.g. in $\eta=\beta_0 + \beta_1 x_1 + \beta_{11} x_1^2$, the vectors are $\mathbf{f}^T=[1\ x_1\ x_1^2]$ and $\beta=[\beta_0\ \beta_1\ \beta_{11}]^T$). The models considered in this thesis are first degree polynomia for J independent variables:

$$\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_J x_J \qquad (2.3)$$

so that $\mathbf{f}^T=[1, x_1, x_2, ... x_J]$ and $\beta=[\beta_0, \beta_1, \beta_2, ..., \beta_J]^T$. Since some unknown error ε_i (random and/or systematic) is always present in any experimentally obtained quantity, the measured response value in the experiment i is not the true value η_i but

$$y_i = \eta_i + \varepsilon_i \qquad (2.4)$$

For this reason, the same experiment performed several times in conditions as similar as possible will never give the same response. Here ε is assumed to be independent (for two experiments i and i' $cov(\varepsilon_i, \varepsilon_{i'})=0$) and normally distributed with mean 0 and variance σ^2. Eq 2.4 is then written as:

$$y_i = \mathbf{f}^T(\mathbf{x}^i)\,\beta + \varepsilon_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + ... + \beta_J x_{i,J} + \varepsilon_i \qquad (2.5)$$

where $\mathbf{f}^T(\mathbf{x}^i)$ is the vector of the variable settings of the ith experiment $\mathbf{x}^i=[x_{i,1}, x_{i,2}, ... x_{i,J}]$. The equations necessary to estimate the coefficients are obtained by conducting the experiments described in the *matrix of experiments* ξ_I (where I must not be inferior to the number of coefficients in the model).The matrix notation for the I experiments is (Figure 2.2)

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon \qquad (2.6)$$

$$\underset{I}{\begin{bmatrix} y_1 \\ \cdot \\ y_i \\ \cdot \\ y_I \end{bmatrix}} = \underset{I \times Q}{\begin{bmatrix} 1 & x_{11} & \cdots & x_{1j} & \cdots & x_{1J} \\ \cdot & \cdot & \cdots & \cdot & \cdots & \cdot \\ 1 & x_{i1} & \cdots & x_{ij} & \cdots & x_{iJ} \\ \cdot & \cdot & \cdots & \cdot & \cdots & \cdot \\ 1 & x_{I1} & \cdots & x_{Ij} & \cdots & x_{IJ} \end{bmatrix}} \underset{Q}{\begin{bmatrix} \beta_0 \\ \beta_1 \\ \cdot \\ \beta_j \\ \cdot \\ \beta_J \end{bmatrix}} + \underset{I}{\begin{bmatrix} \varepsilon_1 \\ \cdot \\ \varepsilon_i \\ \cdot \\ \varepsilon_I \end{bmatrix}}$$

Figure 2.2 Matrix representation on eq 2.6.

UNIVERSITAT ROVIRA I VIRGILI
EXPERIMENTAL DESIGN APPLIED TO THE SELECTION OF SAMPLES AND SENSORS IN MULTIVARIATE CALIBRATION
Joan Ferré Baldrich
ISBN:978-84-691-1875-7/DL: T-337-2008

where $\mathbf{y}_{I\times1}$ is the vector of observed responses, $\mathbf{X}_{I\times Q} = [\mathbf{f}^T(\mathbf{x}^1); ...; \mathbf{f}^T(\mathbf{x}^i); ... ; \mathbf{f}^T(\mathbf{x}^I)]$ is called the *model matrix* and $\varepsilon_{I\times1}$ is the vector of non-observable errors supposed $E(\varepsilon)=0$ and variance-covariance matrix $var(\varepsilon)=\sigma^2\mathbf{I}$. If the model is given by eq 2.5, the ith row of $\mathbf{X}_{I\times Q}$ is $\mathbf{f}^T(\mathbf{x}^i) =[1\ x_{i,1}, x_{i,2}, ... x_{i,J}]$.

2.3.3.2 Estimation of the model. The least-squares solution

Since ε in eq 2.6 is unknown, only an estimation **b** of β can be found. Different approaches[2] exit for estimatingβ. The least-squares solution is given by :

$$\mathbf{b}= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \qquad (2.7)$$

where $\mathbf{b}=[b_0, ... b_j, ... , b_J]^T$ is the vector of estimated regression coefficients. The properties of this solution depend on the validity of some requirements (see references 11,12). Notice that to evaluate $(\mathbf{X}^T\mathbf{X})^{-1}$ (the *dispersion matrix*) the experiments must be chosen in such a manner that $\mathbf{X}^T\mathbf{X}$ (the *information matrix*) is non-singular. If **X** is a full rank matrix, **b** in eq 2.7 can also be calculated as (see §2.2):

$$\mathbf{b}= \mathbf{X}^+\mathbf{y} \qquad (2.8)$$

By defining $\mathbf{X}_{-1}$ as the matrix **X** with the column of ones deleted ($\mathbf{X}=[\mathbf{1},\mathbf{X}_{-1}]$) and $\mathbf{X}_c$ as $\mathbf{X}_{-1}$ after column-centering , the same solution can be found using column-centered data [2 page 369, 12 page 192, 13 page 43]:

$$\mathbf{b}_1= (\mathbf{X}_c^T\mathbf{X}_c)^{-1}\mathbf{X}_c^T\mathbf{y}_c = (\mathbf{X}_c^T\mathbf{X}_c)^{-1}\mathbf{X}_c^T\mathbf{y} \qquad (2.9)$$

$$b_0 = \bar{y} - \bar{\mathbf{x}}^T\mathbf{b}_1 \qquad (2.10)$$

where $\mathbf{b} =[b_0; \mathbf{b}_1]$, $\mathbf{y}_c$ is the column-centered vector **y**, $\bar{y}$ is the mean value of the elements of **y** and $\bar{\mathbf{x}}$ is the vector of the means of the columns of $\mathbf{X}_{-1}$. Centered data is frequently used in multivariate calibration where the preferred notation for the model equation is (see § 2.4.2):

$$\mathbf{y} = \mathbf{1}\beta_0 + \mathbf{X}\beta + \varepsilon \qquad (2.11)$$

UNIVERSITAT ROVIRA I VIRGILI
EXPERIMENTAL DESIGN APPLIED TO THE SELECTION OF SAMPLES AND SENSORS IN MULTIVARIATE CALIBRATION
Joan Ferré Baldrich
ISBN:978-84-691-1875-7/DL: T-337-2008

In this case, **X** corresponds to the matrix $\mathbf{X}_{-1}$ defined above. It must always be clear if either eq 2.6 or eq. 2.11 is being used. Below, the **X** of eq 2.6 is used.

The variance-covariance matrix of **b** is:

$$\text{var}(\mathbf{b}) = (\mathbf{X}^T\mathbf{X})^{-1}\,\sigma^2 = \begin{pmatrix} \text{var}(b_0) & \text{cov}(b_0,b_1) & \cdots & \text{cov}(b_0,b_J) \\ \text{cov}(b_0,b_1) & \text{var}(b_1) & \cdots & \text{cov}(b_1,b_J) \\ \vdots & \vdots & \ddots & \vdots \\ \text{cov}(b_J,b_0) & \text{cov}(b_J,b_1) & \cdots & \text{var}(b_J) \end{pmatrix} \quad (2.12)$$

The variance of the coefficient b_j is

$$var(b_j) = \text{UVIF}_j\,\sigma^2 \quad (2.13)$$

where the *variance coefficient*[10] UVIF_j is the jth element in the diagonal of $(\mathbf{X}^T\mathbf{X})^{-1}$. The off-diagonal elements of $(\mathbf{X}^T\mathbf{X})^{-1}\sigma^2$ are the covariances between the coefficients. The expression for column-centered data is:

$$\text{var}(\mathbf{b}_1) = (\mathbf{X}_c^T\mathbf{X}_c)^{-1}\,\sigma^2 \quad (2.14)$$

The boundary of the 100(1-α) per cent confidence region for all the coefficients is[11,14-16]:

$$(\beta - \mathbf{b})^T\,\mathbf{X}^T\mathbf{X}\,(\beta - \mathbf{b}) = p^2 \quad (2.15)$$

where $p^2 = Q\,s^2\,F_{Q,df,\alpha}$, s^2 is an estimate of σ^2 with df degrees of freedom and $F_{Q,df,\alpha}$ is the α per cent point of the F distribution with Q and df degrees of freedom. In the Q-dimensional space of the coefficients, eq 2.15 defines an hyperellipsoid centered in **b** (an ellipse for designs with two coefficients). The true values of the coefficients are supposed to be inside that hyperellipsoid with a probability α to commit a type I error. A plot of the confidence region for the case of two and three coefficients is shown in the Figure 2.3. Many characteristics of the ellipsoid are related to the

UNIVERSITAT ROVIRA I VIRGILI
EXPERIMENTAL DESIGN APPLIED TO THE SELECTION OF SAMPLES AND SENSORS IN MULTIVARIATE CALIBRATION
Joan Ferré Baldrich
ISBN:978-84-691-1875-7/DL: T-337-2008

eigenvalues (λ_a) and eigenvectors (v_a) of $(X^TX)^{-1}$. The ath axis is oriented in the direction of v_a and its half-length is $p\sqrt{\lambda_a}$. In addition, $\text{Det}(X^TX)^{-1} = \prod_a \lambda_a$ and $\text{Tr}(X^TX)^{-1} = \sum_a \lambda_a$.

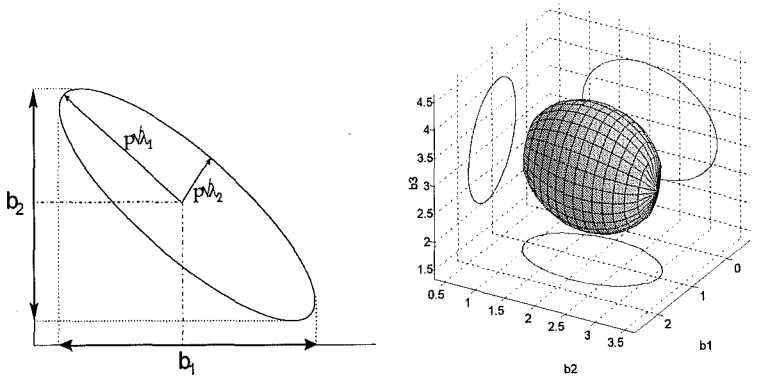


Figure 2.3. Confidence ellipses of a two-coefficient model (left) and a three-coefficient model (right).

2.3.3.3 Prediction

The predicted response value for the input variables $x_{un} = [x_{un,1}, x_{un,2}, \ldots x_{un,J}]^T$ is:

$$y_{un} = \mathbf{f}^T(x_{un})\,\mathbf{b} \qquad (2.16)$$

and has an associated variance (due to the propagation of the uncertainty of the coefficients):

$$var(y_{un}) = \mathbf{f}^T(x_{un})\,(X^TX)^{-1}\,\mathbf{f}\,(x_{un})\,\sigma^2 = d(x_{un})\,\sigma^2 \qquad (2.17)$$

UNIVERSITAT ROVIRA I VIRGILI
EXPERIMENTAL DESIGN APPLIED TO THE SELECTION OF SAMPLES AND SENSORS IN MULTIVARIATE CALIBRATION
Joan Ferré Baldrich
ISBN:978-84-691-1875-7/DL: T-337-2008

where $d(\mathbf{x}_{un})= \mathbf{f}^T(\mathbf{x}_{un}) (\mathbf{X}^T\mathbf{X})^{-1} \mathbf{f}(\mathbf{x}_{un})$ is called the *variance function*. For the model in eq 2.11, the prediction is:

$$y_{un}= b_0 + \mathbf{f}^T(\mathbf{x}_{un})\, \mathbf{b} \qquad (2.18)$$

For the particular case of a first degree polynomium :

$$y_{un} = b_0 + b_1 x_{un,1} + b_2 x_{un,2} + \ldots + b_J x_{un,J} = b_0 + \mathbf{x}_{un}{}^T \mathbf{b} \qquad (2.19)$$

2.3.3.4 Geometrical interpretation of the least squares solution

The geometrical interpretation of the least-squares solution is the underlying basis for the calculation of the net analyte signal[17] and of the quantification using the classical least squares (CLS) model (see § 2.4.1.1). Each column of $\mathbf{X}$ can be regarded as a vector in a Euclidean space so that all the columns define an (hyper)plane Π. The vectors $\eta=\mathbf{X}\beta$ and $\mathbf{y}_{pred}=\mathbf{X}\mathbf{b}$ are linear combinations of the columns of $\mathbf{X}$ and lie in the (hyper)plane. However the vector $\mathbf{y}=\mathbf{X}\beta+\varepsilon$ does not lie in this (hyper)plane since ε may be different from zero. The least-squares estimation of β is the vector $\mathbf{b}$ that gives a residual vector $\mathbf{e}=\mathbf{y}-\mathbf{y}_{pred}$ of minimal length (i.e. $\mathbf{e}^T\mathbf{e}$ = minimal, which is the sum of the squared errors). This happens when $\mathbf{y}_{pred}$ is the orthogonal projection of $\mathbf{y}$ onto the (hyper)plane and then $\mathbf{e}$ is orthogonal to this (hyper)plane. Since $\mathbf{b}= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$, then $\mathbf{y}_{pred} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = \mathbf{H}\mathbf{y}$ and $\mathbf{e}=(\mathbf{I}-\mathbf{H})\mathbf{y}$ where $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ is a *projection matrix*. Figure 2.4 represents this situation in a three-dimensional space.

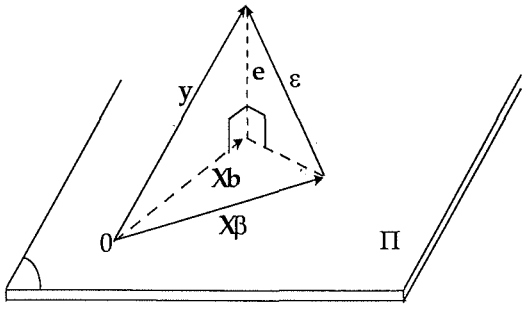


Figure 2.4. Geometrical interpretation of the least-squares solution.

2.3.3.5 Interpretation of the estimated coefficients

The coefficients of the models considered give the rate of change of the dependent variable for a unit change in the independent variable when all of the other variables are held constant. By properly scaling the independent variables (see §3.2), the coefficients enable the effect in the response of each independent variable (main effect) or combination of variables (interaction effects) to be compared[18]. A treatment is judged statistically significant if the variation in the response caused by changing the variable setting (or combination of variable settings) is larger than the experimental error in the measurement of the response. Otherwise, the variable is judged as statistically insignificant. This approach is used in experimental plans based on matrices such as Hadamard, factorial, fractional factorial, etc. [19,20].

2.3.4 Optimal design in multiple linear regression

Eqs 2.12 and 2.17 show that the precision of the estimated coefficients and of the predicted response depends on the matrix **X** but not on the values of y. The important consequence is that, before carrying out any experiment, the settings of the independent variables in **X** (called the *design of experiments*) can be decided to achieve the quality of the properties demanded to the model (e.g. the best global precision of **b** or an acceptable precision of y_{un}). The cost of the model can be reduced by finding the design with the minimum number of experiments that contains the required information (see Scheme 2.1 at the end of §2.3). Several criteria (see §2.3.4.2) enable to evaluate the quality of a design using only the **X** matrix.

Experimental designs are routinely used in many areas of scientific investigation. Catalogs of the most appropriate designs linked to different types of regression models are available [5,16,21]. Some examples are the factorial, Hadamard and Doehlert designs. The importance of this kind of designs (called *classical designs*) lies in two points: they give the maximum information with a small number of experiments by varying several variables simultaneously; and they have the ability of uncorrelating the estimated coefficients (e.g. in factorial and fractional factorial

UNIVERSITAT ROVIRA I VIRGILI
EXPERIMENTAL DESIGN APPLIED TO THE SELECTION OF SAMPLES AND SENSORS IN MULTIVARIATE CALIBRATION
Joan Ferré Baldrich
ISBN:978-84-691-1875-7/DL: T-337-2008

designs the experiments are chosen so that the columns of **X** are orthogonal to each other). One of their limitations is that the variable settings must be independently adjusted in the same experiment according to the design. This can not always be accomplished. Sometimes the variables are correlated by necessity or some experiments cannot be performed due to an irregularly shaped experimental domain. In addition, models that deviate from the usual first or second order ones may have no classical design. Moreover the number of experiments in the classical design may be too large for the experimenter's purposes. However, using non-designed experiments has the danger of leading to highly collinear columns in **X** and even situations where $\mathbf{X}^T\mathbf{X}$ is not of full rank.

When the requirements of classical experimental designs cannot be met, the permitted design space can be examined to search for the optimal experiments that achieves the necessary quality of the properties of the model (e.g. precise estimates of the coefficients of the model). This requires:

1. A matrix of N candidate points (experiments) (ξ_N) where each row represents one experiment and each column a variable. Sets of I candidate points (ξ_I) are chosen from this matrix. The set of all possible ξ_I is designed by Ξ.

2. A criterion to quantify the quality of a proposed set ξ_I. The optimal set optimizes this criterion among all other possible sets of the same number of experiments.

3. An algorithm for finding the optimal set and avoid checking all possible combinations of experiments in Ξ.

4. A criterion to decide the optimal number of experiments I.

These requirements are general for optimization problems of combinatorial nature and are considered in the next sections applied to the DOE. They are valid for sample and wavelength selection considered in the chapters 3 and 4.

UNIVERSITAT ROVIRA I VIRGILI
EXPERIMENTAL DESIGN APPLIED TO THE SELECTION OF SAMPLES AND SENSORS IN MULTIVARIATE CALIBRATION
Joan Ferré Baldrich
ISBN:978-84-691-1875-7/DL: T-337-2008

2.3.4.1 List of candidate points

The quality of all the candidate points is of primary concern for the quality of the optimal set selected from this list. If the candidate points do not meet the requirements to built the model, the selected points may not be suitable despite being optimal for the given list. In multivariate calibration this implies that the preparation or selection of calibration samples must span all known sources of variation present in analysis samples, and the quality of the selected sets must be checked to assess that the sensors (or samples) selected as optimal are good enough for the experimenter's purposes. Some algorithms can search the design space and do not need a list of candidate points[22].

2.3.4.2 Optimality criteria for experimental designs

To base optimization decisions on, the *goodness* of each design is evaluated by a mathematical function called a *criterion function*. Criterion functions for optimal parameter estimation, model discrimination, robust regression, design augmentation, etc. have been classified and explained in MLR [23-28]. However, many publications usually employ mathematical and statistical language that requires a large effort for a chemist to understand it. Some easily readable texts are references: 9, 14, 15, 16, 21, 29, 30, 31 and the review by Steinberg[32]. Among the many criteria available, the most popular consider the quality of the estimated coefficients (which are related to the volume, shape, and orientation of the confidence ellipsoid (Figure 2.3)) and to the variance of the predicted response[32,33] . These optimality criteria are the following *:

1. *D-criterion.* A design is *D-optimal* if it minimizes $\mathrm{Det}(\mathbf{X}^T\mathbf{X})^{-1}$ over all other possible designs in Ξ. Since $\mathrm{Det}(\mathbf{X}^T\mathbf{X})=1/\mathrm{Det}\ (\mathbf{X}^T\mathbf{X})^{-1}$, a D-optimal design also maximizes $\mathrm{Det}(\mathbf{X}^T\mathbf{X})$. It is said that this design minimizes the generalized variance of the least squares estimate of β (low values for the variances and covariances of the coefficients). The volume of the confidence ellipsoid is

* The value of σ^2 ,assumed constant during the experimentation, is not considered in the comparison of experimental designs since it is the same for all of them.

UNIVERSITAT ROVIRA I VIRGILI
EXPERIMENTAL DESIGN APPLIED TO THE SELECTION OF SAMPLES AND SENSORS IN MULTIVARIATE CALIBRATION
Joan Ferré Baldrich
ISBN:978-84-691-1875-7/DL: T-337-2008

proportional to the product of the length of the half-axes [14-16,26,27,34], thus to $\Pi\sqrt{\lambda_a}=[\mathrm{Det}(\mathbf{X}^T\mathbf{X})]^{-1/2}$. A D-optimal design minimizes the volume of the confidence region: the larger the determinant, the smaller the volume and the larger the precision of the coefficients. D-optimality is a very popular criterion. Its drawback is that the volume of the ellipsoid of a D-optimal design can be small because it is "narrow but long"[26]. This corresponds to a list of candidate samples that are quite collinear. Therefore the optimal subset should be always checked for its quality.

2. *A-criterion.* A design is *A-optimal* if it minimizes $\mathrm{Tr}(\mathbf{X}^T\mathbf{X})^{-1}$ over all other possible designs in Ξ . This is related to the shape of the ellipsoid and to the average variance of the coefficients. The ellipsoid of the A-optimal set is the smallest hypersphere of radius r given by $r^2=\mathrm{Tr}[\sigma^2(\mathbf{X}^T\mathbf{X})^{-1}]$.

3. *E-criterion.* A design is *E-optimal* if it minimizes the maximal eigenvalue of $(\mathbf{X}^T\mathbf{X})^{-1}$ over all other possible designs in Ξ, thus the largest principal axis of the uncertainty ellipsoid has minimum length. This is related to the shape of the ellipsoid. The ellipsoid of the E-optimal set is the smallest hypersphere of radius r given by $r^2=$ maximum eigenvalue of $[\sigma^2(\mathbf{X}^T\mathbf{X})^{-1}]$.

4. *G-criterion.* A design is G-optimal if it minimizes the maximum value of the variance of the predicted response at $\mathbf{x}_{un}$ $var(y_{un})$ given by eq 2.17 over all possible $\mathbf{x}$ in the design space. D- and G-optimal designs are equivalent in continuous designs [32] (designs were the distribution of experiments is given as a measure[15,27]). However, this does not always hold for exact designs (designs for a specified number of experiments I) where the D- and G-optimal may be not the same[15,29].

5. *Condition number of* $(\mathbf{X}^T\mathbf{X})^{-1}$. It is defined as $\mathrm{Cond}(\mathbf{X}^T\mathbf{X})^{-1}=\max(\lambda_a)/\min(\lambda_a)$. This a measure of eccentricity of the ellipsoid, and does not depend either on the volume or on the orientation of the ellipsoid.

6. *Variance inflation factors.* This criterion is related to the orientation of the ellipsoid. It is considered in §2.6.

UNIVERSITAT ROVIRA I VIRGILI
EXPERIMENTAL DESIGN APPLIED TO THE SELECTION OF SAMPLES AND SENSORS IN MULTIVARIATE CALIBRATION
Joan Ferré Baldrich
ISBN:978-84-691-1875-7/DL: T-337-2008

2.3.4.3 Algorithms for experiment selection

Searching the best subset in a design space made of a large list of candidate experiments is a combinatorial problem that may be time-prohibitive if all possible combinations must be checked. Optimization algorithms can find the optimum (or an acceptable solution) in a reasonable time. Among the algorithms specialized for optimizing a specific criterion[26], the ones for D-optimal designs are very popular [15, 31, 32] ,specially Fedorov's exchange algorithm[16,31,35]. Fedorov's algorithm uses the list of candidate points expressed as a model matrix. Among the general optimization procedures, genetic algorithms (GA) and generalized simulated annealing (GSA) [36] have been used to find optimal designs. GA has advantages over Fedorov's algorithm in the search of D-optimal matrices with many experimental points and a large list of candidates[22]: less computation time and no need of a list of candidate points. These algorithms compete with others based on spanning the experimental design as much as possible such as clustering[37] or the Kennard-Stone algorithm[38,39]. The Fedorov's algorithm is considered in §3.2.

2.3.4.4 Criterion to decide the optimal number of experiments

The optimal number of experiments depends on the constraints imposed by the cost of the experiments, the time and the difficulty to perform them, the precision required, etc. Only the criterion derived from the D-optimality is considered here. $\mathrm{Det}(\mathbf{X}^T\mathbf{X})$ measures the information of the design and always increases when a new experiment (a new row) is added to $\mathbf{X}$[22]. Thus, the more experiments are performed, the more information the design has and the precision of the coefficients increases. However $\mathrm{Det}(\mathbf{X}^T\mathbf{X})$ is not useful for comparing designs with a different number of experiments: a relatively small increase in the precision may not justify carrying out an extra experiment. Instead, the *normalized information matrix*, (or *matrix of moments*[35]) $\mathbf{M}=(\mathbf{X}^T\mathbf{X})/I$ is used. Then $\mathrm{Det}(\mathbf{M})=\mathrm{Det}(\mathbf{X}^T\mathbf{X})/I^Q$, where Q is the number of coefficients in the model, is independent of the number of experiments and gives a measure of the amount of information *per experiment*. This enables designs with a different number of points to be compared. The design with maximum $\mathrm{Det}(\mathbf{M})$ is selected as optimal. However, M-optimality has selected

UNIVERSITAT ROVIRA I VIRGILI
EXPERIMENTAL DESIGN APPLIED TO THE SELECTION OF SAMPLES AND SENSORS IN MULTIVARIATE CALIBRATION
Joan Ferré Baldrich
ISBN:978-84-691-1875-7/DL: T-337-2008

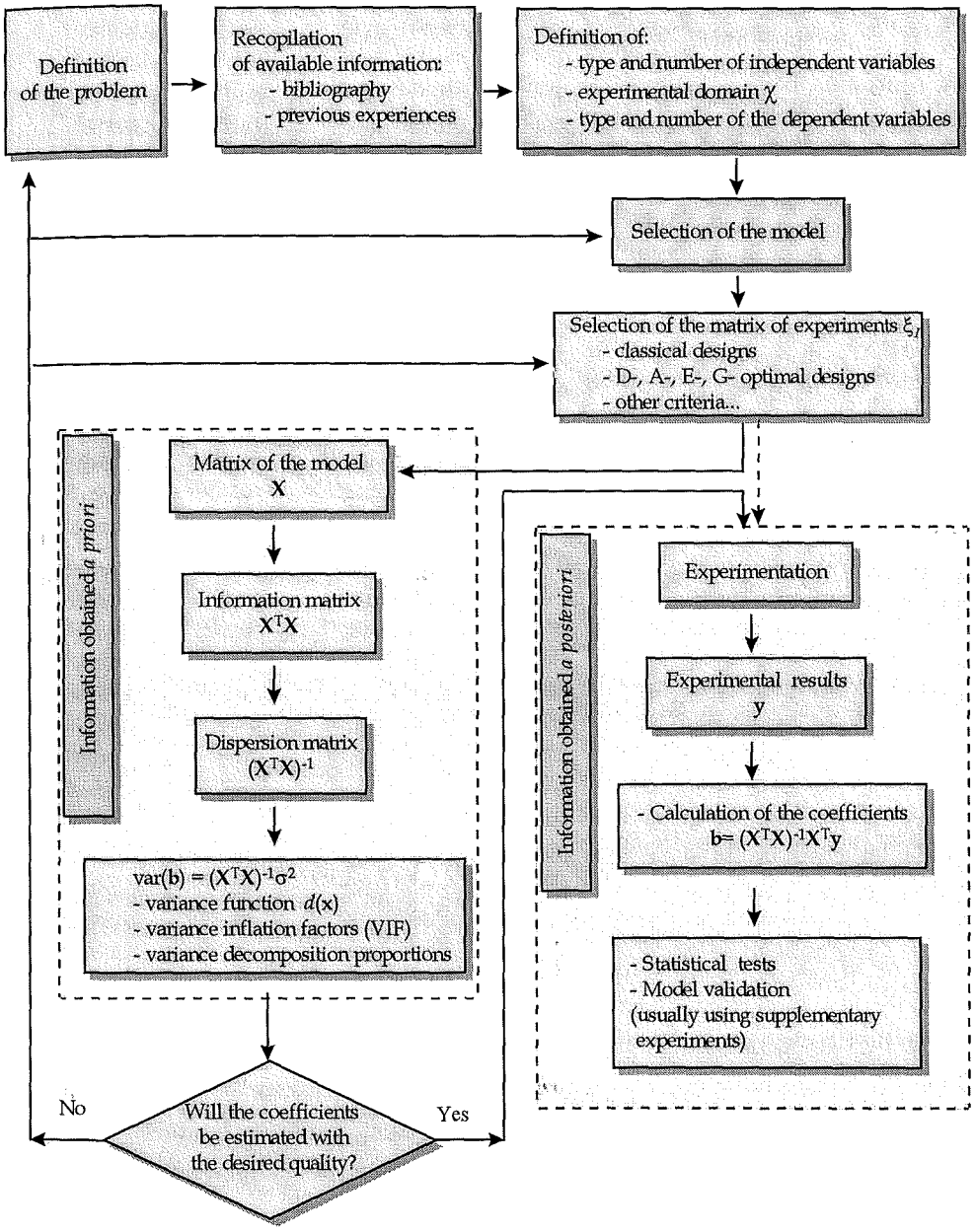
the best points from the list of candidates. If the list is made of highly collinear points, the M-optimal solution can be unacceptable. For this reason, the variance inflation factors (VIFs) should be calculated for the M-optimal subset to assess that it can be used for regression.

2.3.4.5 Objections to some experimental design criteria

The main objection to using the cited optimal experimental designs is their dependence on a model that must be postulated completely in advance[37,40]. Consequently, the selected points can be unsuitable for another model and provide no means for testing the structure of the model or to estimate possible additional effects. Zemroch[40] discussed model sensitivity of an experimental design and said that single- and multifactor designs with evenly spread points over the actual region will not have this drawback. However, optimal designs have some excellent statistical properties and enable to obtain the maximum information with the minimum number of experimental runs. They are the best option if some *a priori* knowledge of the model is assumed.

2.3.4.6 Summary of optimal experimental design

Scheme 2.1 summarizes the ideas of the methodology for optimal experimental design: using the necessary references (bibliography, needs of the experimenter, purposes of the experimentation, etc.) the experimental domain of interest χ is decided and a mathematical model is postulated. Then, the matrix of experiments ξ_I is selected, either using catalogs of designs or algorithms that optimize the appropriate selection criterion. *A priori* criteria of quality are evaluated from the model matrix **X**. If the matrix contains the necessary information to provide a model with the desired qualities the experiments are performed and the coefficients of the model calculated. Finally, statistical tests are made on the coefficients and/or the model is validated usually with supplementary experiments. The information on the left hand column in the scheme is known before experimentation and it helps to choose the best experimental conditions.



Scheme 2.1 Summary of the procedure for optimal experimental design.

UNIVERSITAT ROVIRA I VIRGILI
EXPERIMENTAL DESIGN APPLIED TO THE SELECTION OF SAMPLES AND SENSORS IN MULTIVARIATE CALIBRATION
Joan Ferré Baldrich
ISBN:978-84-691-1875-7/DL: T-337-2008

2.4.Multivariate calibration models

Multivariate calibration models are used in chemical analysis for predicting the concentration of the analyte k in an unknown sample ($c_{un,k}$) from the vector of measured responses of this sample at J sensors $\mathbf{r}_{un}^T=[r_{un,1}, ..., r_{un,J}]$. The most common prediction model is the linear equation (Figure 2.5):

$$c_{un,k} = b_{0,k} + r_{un,1}b_{1,k} + ... + r_{un,J}b_{J,k} = b_{0,k} + \mathbf{r}_{un}^T \mathbf{b}_k \qquad (2.20)$$

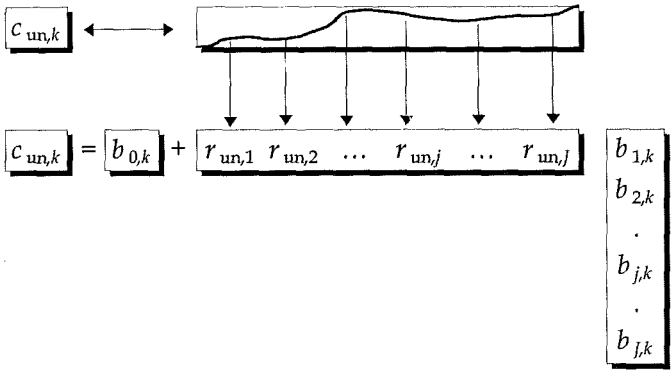


Figure 2.5. Prediction of the concentration in an unknown sample from its digitalized instrumental responses.

where $\mathbf{b}_k=[b_{1,k}, ... , b_{J,k}]^T$. Using the responses and known analyte concentrations of the *calibration samples*, the coefficients $b_{j,k}$ are estimated in a way that distinguishes the different regression methods[41-53]. The multivariate calibration techniques considered in this thesis are: classical least squares (CLS), inverse least squares (ILS), principal component regression (PCR) and partial least squares (PLS). They are also called first-order calibration models since a response vector is used for each sample[42,43,52]. Their underlying theory, advantages and limitations are explained in this section.

2.4.1 Direct models

2.4.1.1 Classical least squares (CLS)

This quantification method, also known as K-matrix calibration[47,51,54-56], is based on the Beer's law applied to many wavelengths. The model assumes that the measured absorbance at each wavelength is a linear additive function of the concentrations of the chemical constituents and that there are no interaction terms in the spectrum between the various components of the sample. The model equation for a sample i of K analytes is:

$$r_{i,j} = r_{0,j} + \sum_{k=1}^{K} s_{j,k}\, c_{i,k} + e_{i,j} \qquad (2.21)$$

where $r_{i,j}$ is the response measured at the jth sensor, $s_{j,k}$ is the sensitivity (the response divided by the concentration of the analyte in a pure sample), $r_{0,j}$ is the background contribution, $c_{i,k}$ is the concentration of the analyte k in the mixture and $e_{i,j}$ is the measurement error that is supposed independent and normally distributed $e_{i,j} \sim N(0,\sigma^2)$. The model error is assumed to derive from the measurement of the absorbances. The model for J measured wavelengths after accounting for the background (for example by substracting from each measured spectrum the response of a blank containing only the sample matrix, without the analytes of interest) is (Figure 2.6) :

$$\mathbf{r} = \mathbf{S}\, \mathbf{c}_{true} + \boldsymbol{\varepsilon} \qquad (2.22)$$

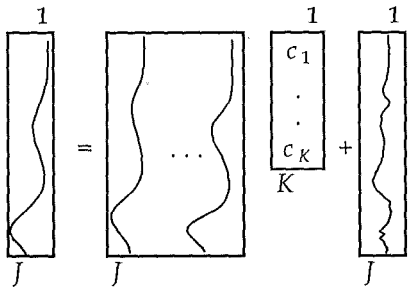


Figure 2.6 Matrix representation of the CLS model.

where $\mathbf{r}_{J\times1}$ is the response vector of the sample, $\mathbf{c}_{true}$ is the $K\times1$ vector of all analyte concentrations in the sample that give response, $\mathbf{S}_{J\times K}=[\mathbf{s}_1, ..., \mathbf{s}_k, ..., \mathbf{s}_K]$ is a matrix of sensitivities whose columns are the pure-component spectra at unit concentration and unit path length (absorptivity-path length products[51]) and $\varepsilon_{J\times1}$ is a vector of errors.

2.4.1.1.1 Calibration

Calibration consists of calculating **S**. Eq 2.22 for I calibration samples is:

$$\mathbf{R}^T=\mathbf{S}\mathbf{C}^T + \mathbf{E} \tag{2.23}$$

and **S** is estimated as:

$$\mathbf{S} = (\mathbf{C}^+\mathbf{R})^T \tag{2.24}$$

where $\mathbf{R}_{I\times J}$ contains the measured absorbances at J wavelengths for I calibration solutions of individual components or mixtures, $\mathbf{C}_{I\times K}$ are the known concentrations of the K analytes in the calibration samples and $\mathbf{E}_{I\times J}$ is the matrix of spectral errors. If the number of calibration samples and constituents is the same **C** is square ($K\times K$) and **S** can be estimated as:

$$\mathbf{S} = (\mathbf{C}^{-1}\mathbf{R})^T \tag{2.25}$$

In addition, if each calibration sample contains only one analyte, $\mathbf{C}_{K\times K}$ is diagonal and eq 2.25 just calculates **S** by dividing the spectrum of each calibration sample by its analyte concentration. To obtain a better estimation of **S** the number of calibration samples is usually larger than the number of components. To calculate **S** in the above equations, the relative amounts of constituents in at least K calibration samples must change from one sample to another. That means that only dilutions of one concentrated calibration sample cannot be used (see § 2.6.5 for the collinearity problem). In addition, to have the necessary linearly independent equations in the prediction step, the number of wavelengths must be equal or larger than the number of constituents in the mixtures ($J\geq K$). Usually the entire spectrum is used.

UNIVERSITAT ROVIRA I VIRGILI
EXPERIMENTAL DESIGN APPLIED TO THE SELECTION OF SAMPLES AND SENSORS IN MULTIVARIATE CALIBRATION
Joan Ferré Baldrich
ISBN:978-84-691-1875-7/DL: T-337-2008

2.4.1.1.2 Prediction

The least-squares estimation of the concentrations of the K analytes in the unknown sample is:

$$\mathbf{c}_{un} = \mathbf{S}^{+}\mathbf{r}_{un} \tag{2.26}$$

In ordinary least squares $\mathbf{S}^{+}$ is calculated as $(\mathbf{S}^{T}\mathbf{S})^{-1}\mathbf{S}^{T}$. The concentration of the analyte k (the kth element of $\mathbf{c}_{un}$) is obtained by multiplying the kth row of $\mathbf{S}^{+}$ by $\mathbf{r}_{un}$:

$$c_{un,k} = \mathbf{S}^{+}_{k\text{-row}}\,\mathbf{r}_{un} = \mathbf{r}_{un}^{T}\,\mathbf{S}^{T+}_{k\text{-col}} \tag{2.27}$$

since $(\mathbf{S}^{+})^{T} = \mathbf{S}^{T+}$. The expressions for the variance of the predicted concentrations and related measures can be found in §4.3. Comparing the last term in eq 2.27 with eq 2.20, it can be seen that, considering $b_{0,k}=0$, the kth-column of $\mathbf{S}^{T+}$ is the vector of coefficients:

$$\mathbf{b}_{k,CLS} = \mathbf{S}^{T+}_{k\text{-col}} \tag{2.28}$$

(also $\mathbf{B}_{CLS}=[\mathbf{b}_{1,CLS}, \mathbf{b}_{2,CLS}, ..., \mathbf{b}_{K,CLS}]=\mathbf{S}^{T+}$ and $\mathbf{B}_{CLS}^{T}\mathbf{B}_{CLS} = (\mathbf{S}^{T}\mathbf{S})^{-1}$). This shows that, to quantify, it is not necessary to know $\mathbf{S}$ but only $\mathbf{S}^{T+}$ that can be estimated as[1,57]:

$$\mathbf{S}^{T+} = (\mathbf{C}^{+}\mathbf{R})^{+} = \mathbf{R}^{+}\mathbf{C} \tag{2.29}$$

and the vector of coefficients for the analyte k is:

$$\mathbf{b}_{k,ILS} = \mathbf{R}^{+}\mathbf{c}_{k} \tag{2.30}$$

where $\mathbf{c}_k$ is the column in $\mathbf{C}$ of the concentrations of the analyte k in all the calibration samples. These coefficients correspond to the P-matrix calibration (§2.4.2.1). Eqs 2.28 and 2.30 have been used to justify that the K-matrix and P-matrix approaches are basically the same model[1], and that the K-matrix calibration is a particular case of the P-matrix calibration where the concentration of all the analytes that give response are simultaneously predicted. However, the equality $\mathbf{b}_{k,CLS} = \mathbf{b}_{k,ILS}$ is only true when $\mathbf{R}^{T}=\mathbf{S}\mathbf{C}^{T}$, that is to say, when there is no error term

in eq 2.23. Since errors are always present because **R** is made of measured quantities $\mathbf{b}_{k,CLS}$ and $\mathbf{b}_{k,ILS}$ are normally slightly different.

Prediction using the net analyte signal

The *net analyte signal* (NAS)[17], is the part of the spectrum of a component in a mixture that is orthogonal to the spectra of the other components in that mixture. In CLS the net signal of the analyte k in the pure component spectra and in the spectrum of the unknown sample are given by:

$$\mathbf{s}_k^* = (\mathbf{I} - \mathbf{S}_k\mathbf{S}_k^+)\mathbf{s}_k \quad (2.31)$$

$$\mathbf{r}_{un,k}^* = (\mathbf{I} - \mathbf{S}_k\mathbf{S}_k^+)\mathbf{r}_{un} \quad (2.32)$$

where $\mathbf{S}_k = [\mathbf{s}_1, \ldots, \mathbf{s}_{k-1}, \mathbf{s}_{k+1}, \ldots \mathbf{s}_K]$ is the **S** matrix without the kth column. $\mathbf{s}_k^*$ is the vector of residuals of the regression of $\mathbf{s}_k$ versus $\mathbf{S}_k$ ($\mathbf{s}_k = \mathbf{S}_k\mathbf{b}^* + \mathbf{s}_k^*$) since:

$$\mathbf{s}_k^* = (\mathbf{I} - \mathbf{S}_k\mathbf{S}_k^+)\mathbf{s}_k = \mathbf{s}_k - \mathbf{S}_k\mathbf{S}_k^+\mathbf{s}_k = \mathbf{s}_k - \mathbf{S}_k\mathbf{b}^* = \mathbf{s}_k - \mathbf{s}_{k,proj} = \text{residuals}$$

This is related to the geometrical interpretation of the least-square solution (see §2.3.3.4). The spectra of the other pure components ($\mathbf{S}_k$) define a (hyper)plane (Figure 2.7) so that $\mathbf{s}_{k,proj} = \mathbf{S}_k\mathbf{b}^*$ is the part of $\mathbf{s}_k$ that lies in that (hyper)plane because it is a linear combination of the columns of $\mathbf{S}_k$. The remaining part of the signal, the residuals $\mathbf{s}_k - \mathbf{s}_{k,proj}$, are orthogonal to the (hyper)plane and can be used for quantification. $\mathbf{S}_k\mathbf{S}_k^+$ is a *projection matrix*[1] of the vector onto the space spanned by $\mathbf{S}_k$ and $(\mathbf{I} - \mathbf{S}_k\mathbf{S}_k^+)$ is projection matrix onto a space orthogonal to that spanned by $\mathbf{S}_k$. The same stands for $\mathbf{r}_{un,k}^*$. The vector $\mathbf{s}_k^* / \|\mathbf{s}_k^*\|^2$ contains the regression coefficients and is different for each analyte but the same for all the unknown samples. The $\mathbf{s}_k^*$ of the different analytes are not necessarily orthogonal among them. $\mathbf{r}_{un,k}^*$ is different for each analyte and unknown sample. In absence of the

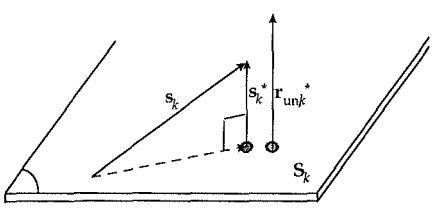


Figure 2.7. The net analyte signal of the analyte k and of the unknown sample spectrum for a model without error.

UNIVERSITAT ROVIRA I VIRGILI
EXPERIMENTAL DESIGN APPLIED TO THE SELECTION OF SAMPLES AND SENSORS IN MULTIVARIATE CALIBRATION
Joan Ferré Baldrich
ISBN:978-84-691-1875-7/DL: T-337-2008

error term in eq 2.22 the vectors $\mathbf{r}_{un,k}^*$ and $\mathbf{s}_k^*$ are parallel and the proportionality constant is the concentration of the analyte k (see §4.7):

$$\mathbf{r}_{un,k}^* = c_{un,k,true}\,\mathbf{s}_k^* \tag{2.33}$$

so that $c_{un,k,true}$ can be calculated using either the net analyte signal at any wavelength j or the norm of the net analyte signals:

$$c_{un,k,true} = \mathbf{r}_{un,j,k}^* / \mathbf{s}_{j,k}^* = \| \mathbf{r}_k^* \| / \| \mathbf{s}_k^* \| \tag{2.34}$$

The vectors in eq 2.33 are represented in Figure 2.7. Since the error term in eq 2.22 is present the relationship is:

$$\mathbf{r}_{un,k}^* = c_{un,k,true}\,\mathbf{s}_k^* + \boldsymbol{\varepsilon}_k^* \tag{2.35}$$

where $\boldsymbol{\varepsilon}_k^*=(\mathbf{I}-\mathbf{S}_k\mathbf{S}_k^+)\boldsymbol{\varepsilon}$ is the net analyte signal of the error, which is unknown. Then $c_{un,k}$, the least squares estimation of $c_{un,k,true}$, is found as:

$$\begin{aligned}
&\mathbf{r}_{un,k}^* = c_{un,k}\,\mathbf{s}_k^* + \mathbf{e}_k^* \\
&\mathbf{r}_{un,k}^{*T} = c_{un,k}\,\mathbf{s}_k^{*T} + \mathbf{e}_k^{*T} \\
&\mathbf{r}_{un,k}^{*T}\mathbf{s}_k^* = c_{un,k}\,\mathbf{s}_k^{*T}\mathbf{s}_k^* + \mathbf{e}_k^{*T}\mathbf{s}_k^* \\
&c_{un,k} = \mathbf{r}_{un,k}^{*T}\mathbf{s}_k^* / \| \mathbf{s}_k^* \|^2
\end{aligned}$$

where $\mathbf{e}_k^*$ is the part of $\mathbf{r}_{un,k}^*$ not explained by $c_{un,k}\,\mathbf{s}_k^*$ and orthogonal to $\mathbf{s}_k^*$ and $\mathbf{e}_k^{*T}\mathbf{s}_k^*=0$. Using the idempotency property of the projection matrices:

$$\mathbf{r}_{un,k}^{*T}\mathbf{s}_k^* = \mathbf{r}_{un}^T(\mathbf{I}-\mathbf{S}_k\mathbf{S}_k^+)^T(\mathbf{I}-\mathbf{S}_k\mathbf{S}_k^+)\mathbf{s}_k = \mathbf{r}_{un}^T(\mathbf{I}-\mathbf{S}_k\mathbf{S}_k^+)\mathbf{s}_k = \mathbf{r}_{un}^T\mathbf{s}_k^* \tag{2.36}$$

so that the concentration of the analyte k in the unknown sample is given by:

$$c_{un,k} = \mathbf{r}_{un}^T\mathbf{s}_k^* / \| \mathbf{s}_k^* \|^2 \tag{2.37}$$

Comparing eq 2.20, 2.28 and 2.35, the vector of regression coefficients of the CLS model for the analyte k is :

$$\mathbf{b}_{k,CLS} = \mathbf{S}^{T+}_{k\text{-col}} = [\mathbf{S}^{+}_{k\text{-row}}]^T = \mathbf{s}_k^* / \|\mathbf{s}_k^*\|^2 \quad (2.38)$$

So that the prediction equation for CLS is:

$$c_{un,k} = \mathbf{S}^{+}_{k\text{-row}}\, \mathbf{r}_{un} = \mathbf{b}_{k,CLS}^T\, \mathbf{r}_{un} = \mathbf{r}_{un}^T \mathbf{b}_{k,CLS} \quad (2.39)$$

The mathematical expressions of the variance of the predicted concentrations, the selectivity and sensitivity in CLS can be found in §4.2.3. The following equalities can be deduced from the equations above [59]:

$$\|\mathbf{b}_{k,CLS}\|^2 = (\mathbf{S}^T\mathbf{S})^{-1}_{kk} = \|\mathbf{S}^{+}_{k\text{-row}}\|^2 = 1/\|\mathbf{s}_k^*\|^2 \quad (2.40)$$

$$\mathbf{b}_{k,CLS} / \|\mathbf{b}_{k,CLS}\| = (\mathbf{s}_k^*/\|\mathbf{s}_k^*\|^2)/(1/\|\mathbf{s}_k^*\|) = \mathbf{s}_k^*/\|\mathbf{s}_k^*\| \quad (2.41)$$

$$\mathbf{b}_{k,CLS}^T \mathbf{s}_k^* = 1 \quad (2.42)$$

where $(\mathbf{S}^T\mathbf{S})^{-1}_{kk}$ is the kth diagonal element of $(\mathbf{S}^T\mathbf{S})^{-1}$.

The expression in CLS can also be formulated[17] using $\mathbf{a}_k$, the spectrum of the analyte k at concentration c_k^0. This vector is related to the sensitivities for k analyte as: $\mathbf{a}_k = c_k^0 \mathbf{s}_k$ and therefore the norm of the vectors follows $\|\mathbf{a}_k\| = c_k^0 \|\mathbf{s}_k\|$. The NAS of $\mathbf{a}_k$ is $\mathbf{a}_k^* = (\mathbf{I} - \mathbf{A}_k\mathbf{A}_k^+)\mathbf{a}_k$ where $\mathbf{A}_k$ is the matrix $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_k \ldots, \mathbf{a}_K]$ with the ath column deleted ($\mathbf{A} = \mathbf{S}\mathbf{C}_0$, $\mathbf{S} = \mathbf{A}\mathbf{C}_0^{-1}$ and $\mathbf{S}^+ = \mathbf{C}_0\mathbf{A}^+$ where $\mathbf{C}_0$ is a $K \times K$ diagonal matrix whose diagonal elements are c_k^0). The NAS is related to $\mathbf{s}_k^*$ with $\mathbf{a}_k^* = c_k^0\, \mathbf{s}_k^*$ (and also $\mathbf{a}_k^*/\|\mathbf{a}_k^*\| = \mathbf{s}_k^*/\|\mathbf{s}_k^*\|$). The net analyte signal of the unknown sample is calculated as $\mathbf{r}_{un,k}^* = (\mathbf{I} - \mathbf{A}_k\mathbf{A}_k^+)\mathbf{r}_{un} = (\mathbf{I} - \mathbf{S}_k\mathbf{S}_k^+)\mathbf{r}_{un}$ and it is related to the concentration as: $\mathbf{r}_{un,k}^* = c_{un,k}\, \mathbf{s}_k^* = c_{un,k}\, \mathbf{a}_k^*/c_k^0$. Therefore, the predicted concentration in the unknown sample can be calculated as $c_{un,k} = c_k^0\, \mathbf{r}_{un,k}^{*T} \mathbf{a}_k^*/\|\mathbf{a}_k^*\|^2 = c_k^0\, \mathbf{r}_{un,k}^T \mathbf{a}_k^*/\|\mathbf{a}_k^*\|^2$

UNIVERSITAT ROVIRA I VIRGILI
EXPERIMENTAL DESIGN APPLIED TO THE SELECTION OF SAMPLES AND SENSORS IN MULTIVARIATE CALIBRATION
Joan Ferré Baldrich
ISBN:978-84-691-1875-7/DL: T-337-2008

2.4.1.1.3 Advantages of CLS

1. Unlike the univariate calibration, CLS does not require selective measurements.

2. The Beer's Law provides a sound foundation for the predictive model.

3. This model can be used for mixtures of known qualitative composition. (e.g., gas phase spectroscopy, some process monitoring or pharmaceutical samples).

4. The model can use a large number of wavelengths to gain a signal averaging effect[43] beneficial for the precision of the predicted concentration, making it less susceptible to noise in the spectra.

5. It may provide a reasonable basis for extrapolation and understanding of the uncertainty in predicted values of the analyte.

2.4.1.1.4 Limitations of CLS [51,60]

1. CLS can only be applied to systems where the spectrum of all the pure constituents giving rise to a signal is known since equations assume the response at a wavelength is due entirely to the calibrated constituents. If the response of the unknown sample contains signal from constituents that have not been included in the calibration matrix **S** in addition to background problems and baseline effects, biased predicted concentrations can be obtained. A CLS model built using second-derivative spectra has been reported to solve partially this problem[61].

2. CLS is not useful for mixtures with interaction between constituents or deviations from Beer's law (nonlinear calibration curves).

3. A severe overlap of spectral bands, quite usual in ultraviolet and visible (UV-vis) spectra, can introduce large uncertainty in the estimated concentrations. This is considered more extensively in the section §2.6.

2.4.2 Inverse Models

The limitations of CLS would make this method fail in the analysis of many common samples such as natural products (water, flour, meat,...) whose complex compositional chemistry makes it impossible to know the spectra of all the pure components that give response. Moreover, sometimes only the quantities of not all but only some of the constituents are of interest. The spectroscopic quantitative analysis of samples with complex matrices can be made using the *inverse* calibration:

$$c_k = f(r_1, r_2, \ldots r_J) + \text{residual}\, f_k$$

where the concentration of the analyte of interest, c_k, is modeled as a function of the instrumental measurements $r_1, r_2, \ldots r_J$ (e.g. absorbances at selected wavelengths) following a empirical relationship, without a theoretical underlying such as the Beer's law. These methods, which include ILS, PCR or PLS, can build calibration models without knowledge of the concentrations of all the constituents in the calibration set. The concentration of only the analyte of interest in each calibration sample (or matrix of concentration of the constituents of interest, C) is needed. Thus, so that unknown interferents can be present in the calibration samples. This important advantage, common for ILS, PCR and PLS, will not be mentioned again. These models have been applied successfully to both natural and manufactured products such as wheat, meat, gasoline or plastics. Although samples with known interfering species do not need to be prepared, the samples must contain the analytes and interferences which contribute to the response so that all the causes of variation can be considered in the model.

Calibration of inverse models

In the ILS, PCR and PLS models considered here the concentration of the analyte of interest k in the sample i is regressed as a linear combination of the instrumental measurements at J selected sensors[3,44,46,62]:

$$c_{i,k} = \beta_{0,k} + \beta_{1,k} r_{i,1} + \ldots + \beta_{J,k} r_{i,J} + \varepsilon_{i,k} = \beta_{0,k} + \mathbf{r}_i^T \beta_k + \varepsilon_{i,k} \qquad (2.43)$$

UNIVERSITAT ROVIRA I VIRGILI
EXPERIMENTAL DESIGN APPLIED TO THE SELECTION OF SAMPLES AND SENSORS IN MULTIVARIATE CALIBRATION
Joan Ferré Baldrich
ISBN:978-84-691-1875-7/DL: T-337-2008

which for I calibration samples ($\mathbf{R}_{I\times J}$) is:

$$\mathbf{c}_k = \mathbf{1}\beta_{0,k} + \mathbf{R}\beta_k + \varepsilon_k \qquad (2.44)$$

where $\mathbf{c}_k$ contains the concentration of the analyte k in these samples. By column-centering both $\mathbf{R}$ and $\mathbf{c}_k$ (substract the vector of the means of the columns of $\mathbf{R}$, $\bar{\mathbf{r}}$, from each sample response and substract the mean of the values of $\mathbf{c}_k$, $\bar{c}_k$, from all the calibration sample concentrations) the term $\beta_{0,k}$ is zero [3,45,63] and the model becomes (Figure 2.8):

$$\mathbf{c}_k = \mathbf{R}\beta_k + \varepsilon_k \qquad (2.45)$$

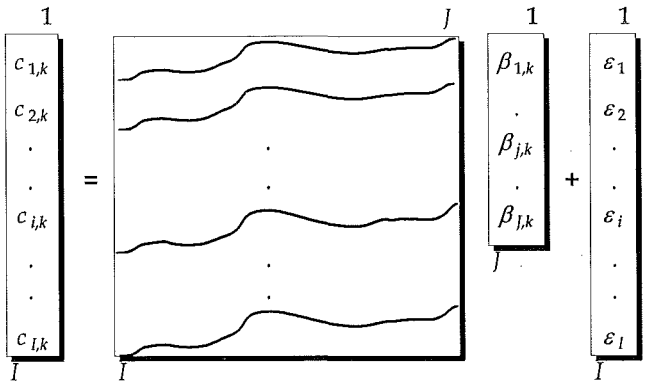


Figure 2.8. Matrix representation of eq 2.40

The estimation of β_k is:

$$\mathbf{b}_k = \mathbf{R}^+\mathbf{c}_k \qquad (2.46)$$

Each regression method calculates the pseudo-inverse matrix $\mathbf{R}^+$ in a different way which produces different estimated coefficients $\mathbf{b}_{k,\mathrm{ILS}}$, $\mathbf{b}_{k,\mathrm{PCR}}$ and $\mathbf{b}_{k,\mathrm{PLS}}$. To calculate $\mathbf{R}^+$, $\mathbf{R}$ is first decomposed into three matrices [3,43,47,49,64,65]:

$$\mathbf{R} = \mathbf{U}\mathbf{D}\mathbf{W}^T \qquad (2.47)$$

where the column vectors of $\mathbf{U}_{I\times Q}$ and $\mathbf{W}_{J\times Q}$ are orthonormal and $\mathbf{D}_{Q\times Q}$ can be either bidiagonal (PLS1) or diagonal (ILS,PCR) and Q is the rank of $\mathbf{R}$. For ILS and PCR, eq 2.47 is the singular-value decomposition (SVD) of $\mathbf{R}$ (see §2.4.2.3). PLS uses other algorithms for the decomposition [3,43,57,65]. The pseudo-inverse is calculated as [43,64]:

UNIVERSITAT ROVIRA I VIRGILI
EXPERIMENTAL DESIGN APPLIED TO THE SELECTION OF SAMPLES AND SENSORS IN MULTIVARIATE CALIBRATION
Joan Ferré Baldrich
ISBN:978-84-691-1875-7/DL: T-337-2008

$$\mathbf{R}^{+} = \mathbf{W}\mathbf{D}^{-1}\mathbf{U}^{T} \qquad (2.48)$$

Before evaluating eq 2.48, PCR and PLS (not ILS) truncate the three matrices so that a low dimension approximation of the data is obtained which retains the relevant information and has less noise (see §2.4.2.2).

Prediction with inverse models

The predicted concentration of the analyte k in an unknown sample can be calculated either using raw data (eq 2.49), $\mathbf{r}_{un}$ centered (eq 2.50) or using both $\mathbf{r}_{un}$ and $c_{un,k}$ centered (eq 2.51):

$$c_{un,k} = b_{0,k} + b_{1,k} r_{un,1} + \ldots + b_{J,k} r_{un,J} = \bar{c}_k - \bar{\mathbf{r}}^{T}\mathbf{b}_k + \mathbf{r}_{un}^{T}\mathbf{b}_k \qquad (2.49)$$

$$c_{un,k} = \bar{c}_k + \mathbf{r}_{un,c}^{T}\mathbf{b}_k \qquad (2.50)$$

$$c_{un,k,c} = \mathbf{r}_{un,c}^{T}\mathbf{b}_k \qquad (2.51)$$

where $\mathbf{r}_{un,c} = \mathbf{r}_{un} - \bar{\mathbf{r}}$ and $c_{un,k,c} = c_{un,k} - \bar{c}_k$. The eq 2.50 is frequently found in the literature [42,43,62,66] and corresponds to models built with centered data and $\mathbf{r}_{un}$ centered before prediction. Eq 2.51 can be easily converted into eq 2.49 by uncentering the data: $c_{un,k} - \bar{c}_k = (\mathbf{r}_{un} - \bar{\mathbf{r}})^{T}\mathbf{b}_k$. If the constant term in eq 2.43 and 2.44 is not present [1,43,49,51,57,64,65,67,68], eqs 2.45 to 2.48 are used but without centering the data, and the prediction is given by eq 2.49 without the $b_{0,k}$.

The mathematical expressions for the variance of the estimated coefficients and of the predicted concentration in ILS, PCR and PLS can be found in Ref. 41,47, 58,62,.

2.4.2.1 Inverse Least Squares (ILS)

Inverse least squares (ILS), also known as P-matrix calibration, is a least-squares method that assumes the inverse calibration model given in eq 2.38. The error $\varepsilon_{i,k}$ is assumed to derive from uncertainties in the determination of the concentration in the calibration samples whereas no error is assumed in the absorbance values.

2.4.2.1.1 Calibration

Calibration consists of solving the system of I equations in eq 2.45 or eq 2.46. For both $\mathbf{R}$ and $\mathbf{c}_k$ column-centered, the coefficients are calculated as [43]

$$\mathbf{b}_{k,\mathrm{ILS}} = \mathbf{R}^{+}\mathbf{c}_k \tag{2.52}$$

$$b_{0,k} = \bar{c}_k - \bar{\mathbf{r}}^{\mathrm{T}}\mathbf{b}_k \tag{2.53}$$

The same $\mathbf{b}_{k,\mathrm{ILS}}$ is obtained in eq 2.52 if $\mathbf{c}_k$ contains the raw values[62]. $\mathbf{R}^{+}$ can be calculated as $(\mathbf{R}^{\mathrm{T}}\mathbf{R})^{-1}\mathbf{R}^{\mathrm{T}}$ or using the SVD of $\mathbf{R}$ (eqs 2.42 and 2.43). For $\mathbf{R}^{\mathrm{T}}\mathbf{R}$ to be invertible, it must be full rank. That is, all the columns and at least J rows (equal to the number of coefficients) must be linearly independent. Usually a higher number of calibration samples are analyzed to improve the precision of the estimated coefficients.

2.4.2.1.2 Prediction

Prediction from the response of the unknown sample $\mathbf{r}_{\mathrm{un}}$ is given by eqs 2.49, 2.50 or 2.51 depending on the preprocessing of the data.

Simultaneous ILS modeling of several components

The calibration and prediction can also be simultaneously done for K analytes as:

$$\mathbf{C} = \mathbf{1}\beta_0^T + \mathbf{R}\beta + \mathbf{E} \tag{2.54}$$

where $\mathbf{C}=[\mathbf{c}_1, \ldots, \mathbf{c}_K]$, $\beta=[\beta_1, \ldots, \beta_K]$, $\mathbf{E}=[\mathbf{e}_1, \ldots, \mathbf{e}_K]$ and $\beta_0=[\beta_{0,1}, \ldots, \beta_{0,K}]^T$. The least-squares solution for centered data is:

$$\mathbf{B} = \mathbf{R}^+\mathbf{C} \tag{2.55}$$

$$\mathbf{b}_0^T = \bar{\mathbf{c}}^T - \bar{\mathbf{r}}^T\mathbf{B} \tag{2.56}$$

where $\bar{\mathbf{c}}$ is the vector of the means of the columns of C. The predicted concentration of several analytes in the unknown sample is:

$$\mathbf{c}_{un}^T = \bar{\mathbf{c}}^T - \bar{\mathbf{r}}^T\mathbf{B} + \mathbf{r}_{un,c}^T\mathbf{B} \tag{2.57}$$

or, if the mean of the calibration set responses is substracted from $\mathbf{r}_{un}$ before prediction:

$$\mathbf{c}_{un}^T = \bar{\mathbf{c}}^T + \mathbf{r}_{un,c}^T\mathbf{B} \tag{2.58}$$

Prediction using the net analyte signal

Recently, Lorber *et al.*[69] described that the NAS of the calibration and unknown samples in ILS can be evaluated respectively as:

$$\mathbf{r}_{i,k}^* = (\mathbf{I} - \mathbf{R}_k^T(\mathbf{R}_k^T)^+)\mathbf{r}_i \tag{2.59}$$

$$\mathbf{r}_{un,k}^* = (\mathbf{I} - \mathbf{R}_k^T(\mathbf{R}_k^T)^+)\mathbf{r}_{un} \tag{2.60}$$

where $\mathbf{R}_k = \mathbf{R} - [\mathbf{c}_k\mathbf{r}^T/(\mathbf{r}^T\mathbf{R}^+\mathbf{c}_k)]$ and $\mathbf{r}$ in this expression is a linear combination of the rows of $\mathbf{R}$ which must include information about the spectrum of the analyte k. Then, the vector of sensitivities for the calibration sample i is: $\mathbf{s}_{i,k}^* = \mathbf{r}_{i,k}^*/c_{i,k}$, where $c_{i,k}$ is the

UNIVERSITAT ROVIRA I VIRGILI
EXPERIMENTAL DESIGN APPLIED TO THE SELECTION OF SAMPLES AND SENSORS IN MULTIVARIATE CALIBRATION
Joan Ferré Baldrich
ISBN:978-84-691-1875-7/DL: T-337-2008

concentration in the *i*th calibration sample. In an errorless situation, all calibration samples should produce the same vector $\mathbf{s}_{i,k}^*$. This is not the case in real situations and the different $\mathbf{s}_{i,k}^*$ may be combined to form an estimate $\mathbf{s}_k^*$ (e.g. the mean value) that is representative of all the calibration set. Then, the concentration of the analyte *k* in the unknown sample is $c_{un,k}=\|\mathbf{r}_{un,k}^*\| / \|\mathbf{s}_k^*\|$.

2.4.2.1.3 Advantages of ILS

1. ILS may have some advantages over the methods based on latent variables such as PCR or PLS: ILS is easier to chemically interpret and it enables the computation of a higher number of statistical parameters, such as the statistical determination of confidence limits for the concentrations.

2.4.2.1.4 Limitations of ILS

1. The number of measured wavelengths is restricted to a subset of the available spectral wavelengths since the number of calibration samples must be equal or superior to the number of coefficients to be estimated. Using many wavelengths would require analyzing a large number of samples with the reference or well-established method which could be costly and tedious.

2. Collinear wavelengths must be avoided since they cause large variances for $\mathbf{b}_k$ and $c_{un,k}$ (see §2.6).

3. To build accurate ILS models, the constituent of interest must absorb at the selected wavelengths. The prediction improves if wavelengths related to constituent of interest are added to the model. However, too many wavelengths may include spectral noise which is unique to the training set and degrade the prediction accuracy for unknown samples that are more unlikely to vary in exactly the same manner (the *overfitting* problem). Ideally, there is a crossover point between selecting enough wavelengths to compute an accurate least squares solution and selecting few enough so that the calibration is not affected by the collinearity of the spectral data.

UNIVERSITAT ROVIRA I VIRGILI
EXPERIMENTAL DESIGN APPLIED TO THE SELECTION OF SAMPLES AND SENSORS IN MULTIVARIATE CALIBRATION
Joan Ferré Baldrich
ISBN:978-84-691-1875-7/DL: T-337-2008

4. ILS has a more reduced signal averaging effect than CLS since just a few responses are considered[43].

Selecting the appropriate set of wavelengths is critically important for the final quality of the ILS model. This can be done using the chemical and spectral knowledge about the analyte of interest and the interferents. In absence of this information, empirical selection methods such as stepwise regression [70] or genetic algorithms[71,72] based on some quality criterion can be used for this purpose. Other alternatives are the regression techniques which can handle collinear data, such as ridge regression (RR) (although RR may solve the collinearity problem but does not reduce the number of wavelengths used) or factor-based methods such PCR and PLS that use linear combinations of all the wavelengths. A recent study compared the PCR, RR and PLS[73].

2.4.2.2 Factor-based regression methods (PCR and PLS)

Principal component regression (PCR) and partial least squares (PLS) are factor-based regression methods that solve some limitations of CLS and ILS. Both PCR and PLS express the correlated information in the many measured variables in a new coordinate system of a few "latent" variables (*factors*) that are a linear combination of the original variables. This is done by decomposing the matrix of instrumental responses of the I calibration samples $\mathbf{R}_{I\times J}$ (often column-centered or autoscaled) into the product of two smaller matrices[51]:

$$\mathbf{R}_{I\times J} = \mathbf{t}_1 \mathbf{p}_1^T + \mathbf{t}_2 \mathbf{p}_2^T + \ldots + \mathbf{t}_A \mathbf{p}_A^T + \mathbf{E} = \mathbf{T}\mathbf{P}^T + \mathbf{E} \qquad (2.54)$$

where $\mathbf{T}_{I\times A}=[\mathbf{t}_1,\ldots, \mathbf{t}_A]$ (*scores*) and $\mathbf{P}_{J\times A}=[\mathbf{p}_1,\ldots, \mathbf{p}_A]$ (*loadings*) are full rank matrices, A is the number of factors and $\mathbf{E}_{I\times J}$ is the part of the data that is not modeled (Figure 2.9). The ith row of $\mathbf{T}_{I\times A}$, $\mathbf{t}^i$, contains the coordinates of the ith calibration sample in this new system, called *scores*. The ath column $\mathbf{t}_a$ contains the scores of all the samples in the factor a.. Each score is a linear combination of the instrumental measurements:

UNIVERSITAT ROVIRA I VIRGILI
EXPERIMENTAL DESIGN APPLIED TO THE SELECTION OF SAMPLES AND SENSORS IN MULTIVARIATE CALIBRATION
Joan Ferré Baldrich
ISBN:978-84-691-1875-7/DL: T-337-2008

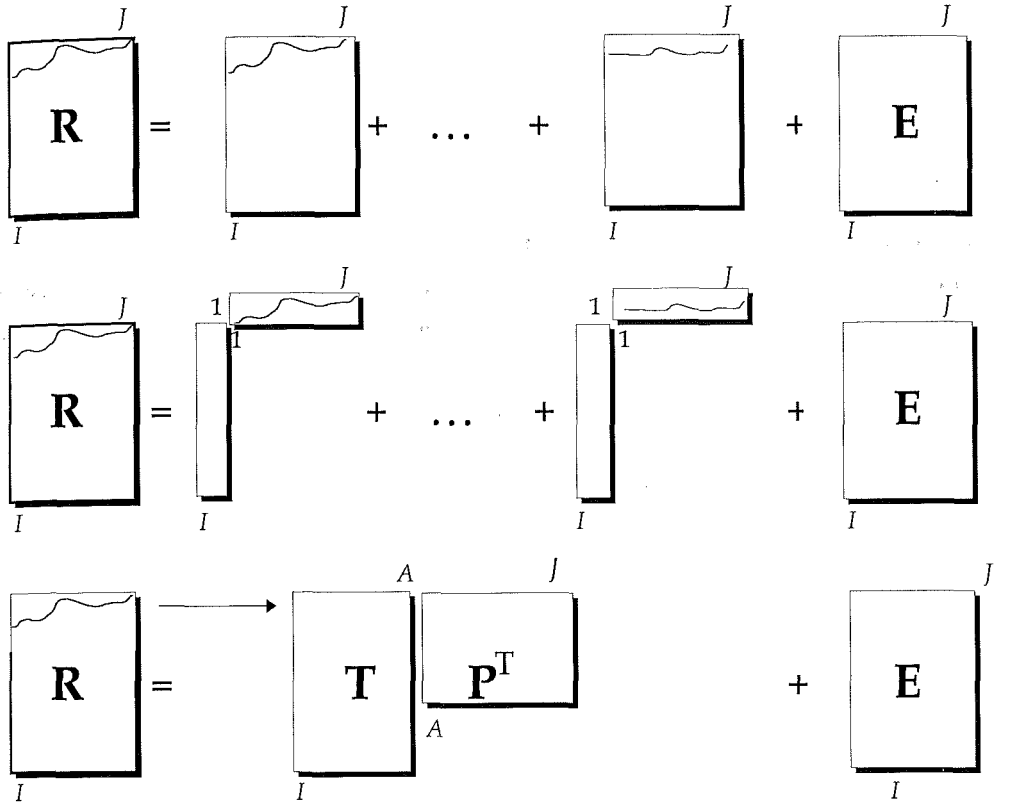


Figure 2.9. Example of the factor decomposition of a matrix of instrumental responses (here spectra).

$$t_{i,a} = w_{a,1}\, r_{i,1} + w_{a,2}\, r_{i,2} + \ldots + w_{a,J}\, r_{i,J} \qquad (2.62)$$

where the weighting coefficients $w_{a,j}$ are found during calibration. In PCR, $w_{a,j} = p_{a,j}$, the elements of **P**. The scores are unique to each calibration spectrum and are used instead of the original responses as variables that represent the sample in the regression with the concentration. In general, the assumed model for these methods is of the form[74]:

$$c_{i,k} = \theta_{0,k} + \theta_{1,k}\, t_{i,1} + \ldots + \theta_{a,k}\, t_{i,a} + \ldots + \theta_{A,k} t_{iA} + f_{i,k} = \theta_{0,k} + \mathbf{t}^{i\,T}\theta_k + f_{i,k} \qquad (2.63)$$

where $t_{i,a}$ is the ath score of the ith sample and $\theta_{k,a}$ are the model coefficients.

A (called the *pseudo-rank* of the calibration matrix) is the number of factors that are important for regression (and is usually much smaller than the number of measurements). It is supposed that $\mathbf{T}_{I\times A}\mathbf{P}_{J\times A}^{T}$ describes the part of **R** that comes from changes in the chemistry and is related to the constituents of interest, while $\mathbf{E}_{I\times J}$ contains information non important for predicting concentration as, for example, random noise. By retaining *A* factors, $\mathbf{R}_A=\mathbf{T}_{I\times A}\mathbf{P}_{J\times A}^{T}$ is an approximate reproduction of **R** with a lower mathematical rank that retains the non-random sources of variation and contains less noise. Thus the data compression step in these techniques is a way of filtering out noise, which is distributed throughout all loading vectors while the true spectral variation is generally concentrated in the early loading vectors. Discarding a part of the data introduces bias in the estimated coefficients but decreases their variance, thus improving the predictive ability of the model. For this reason PCR and PLS are called biased methods[43].

An infinite set of decompositions obeys eq 2.61 and different constrains distinguish the different regression methods that give a different new coordinate system[41]. In PCR (and usually in PLS) the score vectors $\mathbf{t}_a$ are orthogonal to each other. The additional constraint of orthonormal columns for **P** gives the principal component decomposition while **P** is not orthogonal in the other decompositions[63].

T, **P**, the model coefficients and the model size *A* are found during the calibration. If the score vectors are orthogonal, the resulting parameter estimates $q_{k,a}$ are stable. The prediction for a new sample is given by:

$$c_{un,k} = q_{0,k} + q_{1,k}\,t_{un,1} + \ \ldots\ + q_{a,k}\,t_{un,a} + \ldots + q_{A,k}\,t_{unA} = q_{0,k} + \mathbf{t}_{un}^{T}\mathbf{q}_k \qquad (2.64)$$

where $q_{k,a}$ are the estimations of $\theta_{k,a}$ and $t_{un,a}$ are the scores of the response of the unknown sample $\mathbf{r}_{un}$ calculated with eq 2.56. The prediction equation, expressed as a function of the scores, can be converted into eq 2.20, as a function of the measured responses. The PCR and PLS coefficients can also be found with eq 2.41, where $\mathbf{R}^{+}$ is calculated from $\mathbf{R}_A$, the matrix reproduced with only the significant factors which are supposed to be related to the chemical signal[52] and the factors associated mostly with random errors are not used.

Estimation of the optimal number of factors

Unlike CLS or ILS that calculate only one model, PCR and PLS models can be built with a different number of factors. However, an unnecessary large number leads to *overfitting* and bad prediction due to the inclusion of factors that model noise; and a too small number (*underfitting*) introduces systematic error since not enough terms are used to model all the spectral variations of the constituents of interest. Among others [75], the usual method of selecting the optimal number of factors A is to build models with an increasing number of factors and measure their predictive ability using samples of known concentration [43-45,47,73,76]. Then A corresponds to the first local minimum of the plot number of factors versus predictive ability. Statistical tests for determining if including an additional factor is significant have been also described [77,51]. The predictive ability is usually evaluated as $\text{PRESS}=\sum_i (c_i - c_{i,pred,a})^2$ (Prediction Residual Error Sum of Squares) where c_i and $c_{i,pred,a}$ are, respectively, the measured and predicted analyte concentration with a factors in the sample i). The way $c_{i,pred,a}$ is found defines two different validation methods[47, 78]:

- *External validation*. The $c_{i,pred,a}$ values are predicted from a set of I_P samples (*validation set*) not used in the model-building step and measured under the same conditions as the training set. The model with a factors used for prediction has been built using the training set (Figure 2.10) The predictive ability of the model is given by the *Root-Mean-Square Error of Prediction* (RMSEP) defined as:

$$\text{RMSEP} = \sqrt{\frac{\sum_{i=1}^{I_P} (c_i - c_{i,pred,a})^2}{I_P}} \qquad (2.65)$$

where c_i and $c_{i,pred,a}$ are, respectively, the measured and predicted analyte concentration with a factors in the sample i . RMSEP is equivalent to PRESS but it may be preferable since it has the same units as the concentration values. As I_P gets large, the RMSEP will approach the prediction error of the population of all future samples. The test set validation gives the best estimate of a model's performance since none of the samples in the validation set is used in the model building and

the final calibration equation is used for prediction. Its drawback is that, to reliably estimate the prediction ability, the data set must be representative for the future unknown samples and cover the expected range of concentration values. This may require a large number of test objects. Moreover it is rather wasteful and expensive due to the time and cost involved in generating the independent data set which is used for testing purposes only. Some alternative techniques overcome these problems and use all the available data both for calibration and for testing. Only cross-validation is considered here.

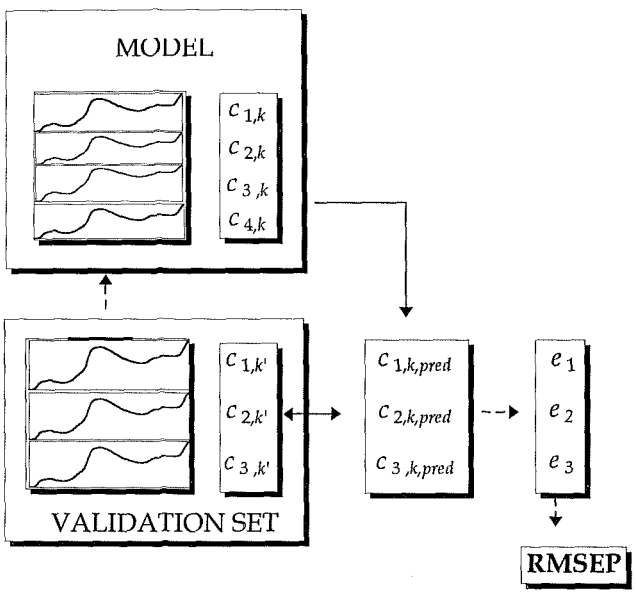


Figure 2.10 Graphical representation of test set validation.

In the *cross-validation* method, part of the data is left out, a model is constructed using the remaining data, and a prediction is made on the left-out data. This process is repeated until all the samples have been left out once. The most used is *leave-one-out cross-validation* where each sample is left out one at a time. An approximation to the prediction error is given by the *Root-Mean-Square Error of Cross-validation* (RMSECV):

UNIVERSITAT ROVIRA I VIRGILI
EXPERIMENTAL DESIGN APPLIED TO THE SELECTION OF SAMPLES AND SENSORS IN MULTIVARIATE CALIBRATION
Joan Ferré Baldrich
ISBN:978-84-691-1875-7/DL: T-337-2008

$$\text{RMSECV} = \sqrt{\frac{\sum_{i=1}^{I}(c_i - c_{i,pred,a})^2}{I}} \tag{2.66}$$

where I is the number of calibration samples and $c_{i,pred,a}$ is the predicted analyte concentration when a model with a factors is constructed without sample i. Since the predicted samples are not used to build the model, RMSECV is a good approximation to the prediction error for unknown samples and the model is validated without having to measure an entirely new set of data. Cross-validation is used also when it is difficult to obtain enough prediction samples to make RMSEP significant. In addition, since each sample is left out of the model during cross-validation, outliers can be detected by comparing the spectral reconstruction to the original training spectrum or the concentrations. A drawback of leave-one-out cross-validation is the large time required since the model is re-calculated for every sample left out. In such cases, leaving out groups of samples at a time can be preferable. This is also used in training sets that contain replicate spectra of the same sample.

It must be also said that the most desirable situation in multivariate calibration would be to have enough data to divide it into three sets: calibration set (to calculated the model), validation (or monitoring) set, to determine the optimal number of factors, and test (or prediction) set, an independent set not used to determine the model and that evaluates the final quality of the model[74]. However, when the number of samples is not so large to enable this division, the model is constructed with the calibration set, the factors selected with cross-validation (thus using the calibration set) and finally validated with the test set.

2.4.2.2.1 Advantages of factor-based regression methods (PCR and PLS)

Factor-based models combine the best features of both the CLS and ILS methods and are generally better in both accuracy and robustness:

1. Both PCR and PLS decompose the original matrix into factors and retain a subset of them to calculate the regression equation. This produces robust models for predicting concentrations of the desired constituents in very complex samples even that may contain contaminants not present in the original calibration mixtures.

2. The number of calibration samples, that in ILS must be at least equal to the number of wavelengths, in PCR and PLS is not necessarily large for mathematical requirements since only the coefficients for a reduced number of factors must be calculated.

3. Unlike in ILS, wavelength selection is not necessary in PCR and PLS. Usually the whole spectrum, or wide regions are used. This gives the signal averaging effect of full-spectral technique such as CLS and, together with the factors decomposition, makes models less susceptible to spectral noise.

4. The collinearity problem met in ILS is eliminated since the scores are usually orthogonal.

2.4.2.2.2 Limitations of factor-based regression methods (PCR and PLS)

1. Although PCR and PLS are full-spectrum methods and are able to accommodate some degree of non-linearities, the inclusion of non-informative wavelengths or that have non-linearities can degrade performance[44]. A careful wavelength selection is advisable to improve the prediction ability of these models.

2. Models are more complex to understand and interpret than the CLS and ILS methods and calculations are slower.

3. The method for selecting the important factors is a critical step in these methods and has been the subject of a large number of papers.

UNIVERSITAT ROVIRA I VIRGILI
EXPERIMENTAL DESIGN APPLIED TO THE SELECTION OF SAMPLES AND SENSORS IN MULTIVARIATE CALIBRATION
Joan Ferré Baldrich
ISBN:978-84-691-1875-7/DL: T-337-2008

2.4.2.3 Principal component analysis (PCA)

Principal component analysis (PCA) [41,79]constitutes one of the most important methods used in the chemometric literature and the factor-decomposition technique used in PCR. PCA transforms the correlated variables in the calibration data to a new set of A uncorrelated variables according eq 2.61 (without loss of generality $\mathbf{R}_{I\times J}$ is assumed to be mean-centered [62,63,79]:

$$\mathbf{R}_{I\times J} = \mathbf{t}_1\mathbf{p}_1^T + \ldots + \mathbf{t}_a\mathbf{p}_a^T + \ldots + \mathbf{t}_A\mathbf{p}_A^T + \mathbf{E} = \mathbf{T}_A\mathbf{P}_A^T + \mathbf{E} \qquad (2.61)$$

where the matrix of *scores* $\mathbf{T}_A=[\mathbf{t}_1, \ldots, \mathbf{t}_a, \ldots, \mathbf{t}_A]$ $(I\times A)$ has orthogonal columns ($\mathbf{t}_a^T\mathbf{t}_b=0$, $a\neq b$), the matrix of *loadings* $\mathbf{P}_A=[\mathbf{p}_1, \ldots, \mathbf{p}_a, \ldots, \mathbf{p}_A]$ $(J\times A)$ is orthonormal ($\mathbf{p}_a^T\mathbf{p}_b=0$, $a\neq b$ and $\|\mathbf{p}_a\|^2=\mathbf{p}_a^T\mathbf{p}_a=1$) and $\mathbf{E}_{I\times J}$ is the part of **R** not retained by the decomposition. The columns in **T** are ordered in decreasing order of explained variance so that the first factors (and thus a reduced number of variables) express the most important information in the data without a significant loss. By introducing normalized scores $\mathbf{u}_a=\mathbf{t}_a/\lambda_a^{1/2}$, eq 2.54 can be written as [63,80]:

$$\mathbf{R}_{I\times J}=\lambda_1^{1/2}\mathbf{u}_1\mathbf{p}_1^T + \ldots + \lambda_a^{1/2}\mathbf{u}_2\mathbf{p}_2^T + \ldots + \lambda_A^{1/2}\mathbf{u}_A\mathbf{p}_A^T + \mathbf{E} = \mathbf{U}_A\mathbf{D}_A\mathbf{P}_A^T + \mathbf{E} \qquad (2.67)$$

where $\lambda_a=\mathbf{t}_a^T\mathbf{t}_a$ is the square norm of the ath score vector prior to normalization and it is also the ath eigenvalue of $\mathbf{R}^T\mathbf{R}$, $\mathbf{D}_A=\text{diag}(\lambda_a^{1/2})$ $(A\times A)$ is diagonal and $\mathbf{U}_A=[\mathbf{u}_1,\ldots,\mathbf{u}_a,\ldots,\mathbf{u}_A]=\mathbf{T}_A\mathbf{D}_A^{-1}$ $(I\times A)$ is orthonormal and contains the vectors $\mathbf{t}_a$ normalized to length one. $\mathbf{R}_A=\mathbf{U}_A\mathbf{D}_A\mathbf{P}_A^T=\mathbf{T}_A\mathbf{P}_A^T$ is the compressed representation of **R** with some of the noise removed. The NIPALS algorithm[41,51] decomposes **R** according to eq 2.61. The singular-value decomposition (SVD)[1, 43,54] of **R** gives $\mathbf{R}=\mathbf{UDP}^T$ so that $\sigma_a=\lambda_a^{1/2}$ are the singular values of **R** and $\sigma_1\geq\sigma_2\geq\ldots\geq\sigma_A\geq 0$.The matrices $\mathbf{U}_A$, $\mathbf{D}_A=\text{diag}(\sigma_1,\ldots\ \sigma_A)$ and $\mathbf{P}_A$ are the retained part from partitioning the matrices $\mathbf{U}=[\mathbf{U}_A, \mathbf{U}_{-A}]$, $\mathbf{P}=[\mathbf{P}_A,\mathbf{P}_{-A}]$ and $\mathbf{D}=\begin{bmatrix}\mathbf{D}_A & \mathbf{0}\\ \mathbf{0} & \mathbf{D}_{-A}\end{bmatrix}$ (**0** is a matrix of zeros). $\mathbf{P}_A$ is the updated **P** matrix without the columns corresponding to irrelevant factors. It can also be seen that $\mathbf{D}_A^2=\mathbf{T}_A^T\mathbf{T}_A$

2.4.2.3.1Loadings

The loadings are the elements of the arbitrarily[79] normalized eigenvectors of $\mathbf{R}^T\mathbf{R}$, $\mathbf{p}_a$, with associated eigenvalue λ_a. Since $\mathbf{R}^T\mathbf{R}$ is a symmetric matrix, its eigenvectors are orthogonal. The equation of eigenvalues of $\mathbf{R}^T\mathbf{R}$ is:

$$\mathbf{R}^T\mathbf{R}\mathbf{p}_a=\mathbf{p}_a\lambda_a \tag{2.68}$$

and for the complete set of eigenvectors $\mathbf{P}=[\mathbf{p}_1,...,\mathbf{p}_A]$:

$$\mathbf{R}^T\mathbf{R}\mathbf{P} = \mathbf{P}\text{diag}(\lambda_a) \tag{2.69}$$

where $\text{diag}(\lambda_a)$ is the diagonal matrix of eigenvalues. By multiplying both sides of eq 2.69 by $\mathbf{P}^T$, the resulting equation

$$\mathbf{R}^T\mathbf{R} = \mathbf{P}\text{diag}(\lambda_a)\mathbf{P}^T \tag{2.70}$$

shows that $\mathbf{P}$ diagonalizes $\mathbf{R}^T\mathbf{R}$ associated with the eigenvalues $\text{diag}(\lambda_a)$, which agrees with $\mathbf{R}^T\mathbf{R}=(\mathbf{U}\mathbf{D}\mathbf{P}^T)^T(\mathbf{U}\mathbf{D}\mathbf{P}^T)=\mathbf{P}\mathbf{D}^T\mathbf{U}^T\mathbf{U}\mathbf{D}\mathbf{P}^T=\mathbf{P}\mathbf{D}^2\mathbf{P}^T$ where $\mathbf{D}^2=\text{diag}(\lambda_a)$.

2.4.2.3.2 Scores

Each vector of scores $\mathbf{t}_a$ is the projection of $\mathbf{R}$ on the basis vector $\mathbf{p}_a$ calculated as $\mathbf{t}_a = \mathbf{R}\mathbf{p}_a$. The scores on the first A principal components are $\mathbf{T}_A=\mathbf{R}\mathbf{P}_A=\mathbf{U}_A\mathbf{D}_A$. It is not necessary to use the inverse (or pseudo inverse) of $\mathbf{P}$ to solve the equation $\mathbf{R}=\mathbf{T}\mathbf{P}^T$ since the matrix $\mathbf{P}$ is orthonormal and it is enough to multiply both sides of this equation for $\mathbf{P}$. If $\mathbf{R}$ has been centered (thus $\mathbf{R}^T\mathbf{R}/(I\text{-}1)$ is a covariance matrix), the resulting scores are centered; otherwise, the scores are not centered. The scores for a new object of coordinates $\mathbf{r}_i^T$ are $\mathbf{t}^i{}_A{}^T=\mathbf{r}_i^T\mathbf{P}_A$.

UNIVERSITAT ROVIRA I VIRGILI
EXPERIMENTAL DESIGN APPLIED TO THE SELECTION OF SAMPLES AND SENSORS IN MULTIVARIATE CALIBRATION
Joan Ferré Baldrich
ISBN:978-84-691-1875-7/DL: T-337-2008

2.4.2.3.3 Eigenvalues

The sum of the eigenvalues is equal to the total variance of the data set and to the trace of the original matrix. Each eigenvalue λ_a divided by the trace represents the proportion of the total variance accounted for by the eigenvector $\mathbf{p}_a$.

2.4.2.3.4 Number of significant factors

Criteria such as the empirical indicator function by Malinowski[81] or a PRESS value using cross-validation [43,79,82], among others[83] have been proposed for selecting the number of factors that explain a significant variance of **R**.

2.4.2.3.5 Advantages of PCA

1. A large number of original variables can be reduced to few new variables that account for a significant portion of the information (variance) of the data. The reduced data can be interpreted as primary sources of variation of the original data. The eigenvectors which model statistically significant variation in the data are retained. This allows the graphical representation of the samples in the reduced space with a minimum loss of information, to identify natural associations of samples and/or variables and their relationship as well as outlier detection[79].

2. By deleting the principal components whose eigenvalues are nearly zero, linear dependencies are removed. If these PCs are associated to noise in the data, the noise in the reproduced data matrix has been reduced.

2.4.2.4 Principal component regression (PCR)

2.4.2.4.1 Calibration

PCR creates a quantitative model in a two-step process: the PCA scores $\mathbf{T}_A$ of I calibration samples are calculated for A factors and then the scores are regressed against the analyte concentration:

$$\mathbf{c}_k = \mathbf{1}\,\theta_{0,k} + \mathbf{T}_A\,\theta_k + \varepsilon \tag{2.71}$$

where θ_k is the vector of the regression coefficients and ε is a vector of independent and normally distributed errors. For column-centered data (i.e. the average calibration spectrum is substracted from each spectrum, and the average calibration concentration is substracted from each concentration) the resulting scores are centered and the intercept is eliminated from the fit (Figure 2.11):

$$\mathbf{c}_k = \mathbf{T}_A\theta_k + \varepsilon_k \tag{2.72}$$

$$\underset{I}{\begin{bmatrix} c_{1,k} \\ c_{2,k} \\ \cdot \\ c_{i,k} \\ \cdot \\ c_{I,k} \end{bmatrix}} = \underset{I}{\overset{A}{\begin{bmatrix} t_{1,1} & \cdot & t_{1,A} \\ t_{2,1} & \cdot & t_{2,A} \\ \cdot & \cdot & \cdot \\ t_{i,1} & \cdot & t_{i,A} \\ \cdot & \cdot & \cdot \\ t_{I,1} & \cdot & t_{I,A} \end{bmatrix}}} \underset{A}{\overset{1}{\begin{bmatrix} \theta_{1,k} \\ \cdot \\ \theta_{A,k} \end{bmatrix}}} + \underset{I}{\overset{1}{\begin{bmatrix} \varepsilon_{1,k} \\ \varepsilon_{2,k} \\ \cdot \\ \varepsilon_{i,k} \\ \cdot \\ \varepsilon_{I,k} \end{bmatrix}}}$$

Figure 2.11 Matrix representation of eq 2.72

Eq 2.72 can also be deduced by considering a model for centered data[73]:

$$\mathbf{c}_k = \mathbf{R}\,\beta_k + \varepsilon \tag{2.73}$$

UNIVERSITAT ROVIRA I VIRGILI
EXPERIMENTAL DESIGN APPLIED TO THE SELECTION OF SAMPLES AND SENSORS IN MULTIVARIATE CALIBRATION
Joan Ferré Baldrich
ISBN:978-84-691-1875-7/DL: T-337-2008

and multiplying by $\mathbf{P}_A\,\mathbf{P}_A^T=\mathbf{I}$:

$$\mathbf{c}_k = \mathbf{R}\,\mathbf{P}_A\,\mathbf{P}_A^T\,\beta_k + \varepsilon \qquad (2.74)$$

which gives eq 2.72 where $\mathbf{T}_A= \mathbf{R}\,\mathbf{P}_A$, $\theta_k= \mathbf{P}_A^T\,\beta_k$ and $\mathbf{P}_A$ only contains the loadings for the A significant factors. The least-squares solution for θ_k has the same form as the ILS solution but with $\mathbf{T}$ and $\mathbf{q}_k$ instead of $\mathbf{R}$ and $\mathbf{b}_k$:

$$\mathbf{q}_k = \mathbf{T}_A{}^{+}\mathbf{c}_k \qquad (2.75)$$

The coefficients $\mathbf{b}_k$ in eq 2.20 can be calculated from $\mathbf{q}_k$ as:

$$\mathbf{b}_{k,PCR}=\mathbf{P}_A\,\mathbf{q}_k \qquad (2.76)$$

$\mathbf{b}_{k,PCR}$ can also be estimated [47] from eqs 2.46 to 2.48 with $\mathbf{R}_A{}^{+}=\mathbf{P}_A\mathbf{D}_A{}^{-1}\mathbf{U}_A{}^T$ (proof: $\mathbf{b}_k= \mathbf{P}_A\mathbf{q}_k= \mathbf{P}_A\mathbf{T}^{+}\mathbf{c}_k= \mathbf{P}_A(\mathbf{T}_A{}^T\mathbf{T}_A)^{-1}\mathbf{T}_A{}^T\mathbf{c}_k =\mathbf{P}_A\mathbf{D}_A{}^{-2}\mathbf{D}_A\mathbf{U}_A\mathbf{c}_k =\mathbf{P}_A\mathbf{D}_A{}^{-1}\mathbf{U}_A{}^T\mathbf{c}_k =\mathbf{R}_A{}^{+}\mathbf{c}_k$ since $\mathbf{T}_A{}^T\mathbf{T}_A=\mathbf{D}_A{}^2$; $\mathbf{T}_A=\mathbf{U}_A\mathbf{D}_A$; $\mathbf{R}_A=\mathbf{U}_A\mathbf{D}_A\mathbf{P}_A{}^T$)

2.4.2.4.2 Prediction

The concentration of the analyte k in an unknown sample whose response is $\mathbf{r}_{un}$ can be predicted in two equivalent ways:

a) simplified prediction, using equations 2.49, 2.50 or 2.51 with the coefficients given by eq 2.76.

b) full prediction, using the loading vectors $\mathbf{P}_{J\times A}$ to transform $\mathbf{r}_{un}$ (centered) to its factor scores $\mathbf{t}_{un,A}{}^T= \mathbf{r}_{un}{}^T\mathbf{P}_A$ (Figure 2.12):

$$c_{un,k} = \bar{c} + \mathbf{t}_{un,A}{}^T\,\mathbf{q}_k \qquad (2.77)$$

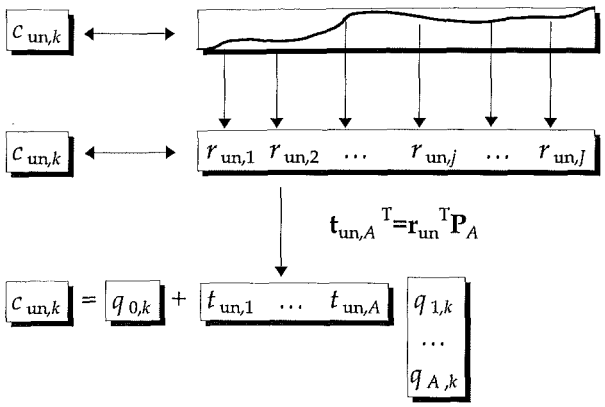


Figure 2.12. Matrix representation of the full prediction in PCR (eq 2.77).

2.4.2.4.3 Selection of factors in PCR

The usual procedure for selecting the optimal number of factors A in PCR is to calculate the predictive ability of several models built with a different number of factors (see §2.4.2.2). The factors can be included in the models in two different ways: (a) in decreasing order of explained variance of the response data matrix (i.e. factors ordered by magnitude of their eigenvalues, the *top-down* approach) or (b) ordered according to the importance for predicting each analyte. The most used strategy is (a). The PCs are calculated independently of any knowledge of the analyte concentrations and merely represent the largest common variations in the response data (e.g. spectra). Presumably, the major variations within the independent variables, accounted for in the PCs with the largest eigenvalues, are related to changes in the constituent concentrations. So, only these PCs are retained. The noise, which usually provides the smallest contribution to the data, is supposed associated to the factors with the smallest variances (that is, with the smallest eigenvalues). These factors are probably irrelevant for the prediction of chemical compositions and are rarely used in regression[84,85].

The approach (b) is justified by the fact that the underlying effects that are directly related to the concentration of the constituent of interest may be small in comparison with the irrelevant ones and may not appear among the PCs which explain a large percentage of the variance of the data. Then the PCR model built

UNIVERSITAT ROVIRA I VIRGILI
EXPERIMENTAL DESIGN APPLIED TO THE SELECTION OF SAMPLES AND SENSORS IN MULTIVARIATE CALIBRATION
Joan Ferré Baldrich
ISBN:978-84-691-1875-7/DL: T-337-2008

with the strategy (a) includes unnecessary factors not related to the constituent of interest that may degrade the prediction ability[86] of the model. It has been shown that PCR with selection of principal components instead of the usual top-down approach yields simpler and better models [86,87]. One way of determining the factors that are relevant for prediction is to build models with the factors included in all possible orders. However, this combinatorial problem can take a long computation time when the number of factors to be considered is large. The generalized simulated annealing (GSA) algorithm with the PRESS criterion [87] and a forward selection procedure where the PCs with the largest absolute correlation with the dependent variable enter one at a time in the model[86] have been used. Jouan-Rimbaud *et al.*[86] suggested that since the PCs are not correlated, simpler selection methods could probably be applied, but this required further investigations that were not made. Sun[84] presented the correlation principal component regression although no indications were given on how to perform the selection of the more relevant factors. Recently, Xie and Kalivas[88,89] proposed a forward selection procedure for PCR.

2.4.2.4.4 Advantages of PCR

Besides the advantages indicated in §2.4.2.2.1, the following can be noted:

1. The PCA data compression extracts the underlying effects in the **R** data and PCR uses these in the inverse regression to calculate the model coefficients and to predict the values of the dependent variable.

2. PCR combines the advantage of using all the spectral channels, excludes noise effects (which are relegated to the unused factors), and retains the ILS independence of uncalibrated components.

3. The problems present in collinear data such as NIR spectra have disappeared because the columns of **T** are orthogonal and the PCs of the smallest eigenvalues (the ones which would produce the largest variance in the estimated coefficients) have been deleted. This produces more reliable estimates of the model coefficients and hence a good predictive model useful for calibration of spectroscopic instruments.

ISBN:978-84-621-

2.4.2.4.5 Limitations of PCR.

The PCR calibration model is not completely free of problems:

1. The PCA factors are calculated independently of any knowledge of the concentration of the analyte of interest. The usual top-down method can introduce irrelevant factors that degrade the predictive ability of the model.

2.4.2.5 Partial least squares regression (PLS)

Partial least squares regression (PLS) is a factor-based calibration technique that uses both spectral and constituent concentration information in the decomposition process to find those factors with the greatest relevance for prediction. This is different from PCR, that first decomposes the spectral matrix into factors that represent the most common variations in the response data, completely ignoring their relation to the constituents of interest and then regresses the scores against the concentrations. The resulting PLS factors are more relevant for description of the concentration information than those calculated in PCR.

2.4.2.5.1 Calibration

The calibration equations for PLS are more complex than those of PCR and are not described here. Different versions of PLS algorithms can be found in a large number of publications [3,41,51,57,65,90-97]. Two main types of PLS algorithms exist: PLS-1, that calibrates for one constituent at a time, and PLS-2, that calibrates for more than one constituent simultaneously. The comparison of both methods can be found in Ref. 41.

UNIVERSITAT ROVIRA I VIRGILI
EXPERIMENTAL DESIGN APPLIED TO THE SELECTION OF SAMPLES AND SENSORS IN MULTIVARIATE CALIBRATION
Joan Ferré Baldrich
ISBN:978-84-691-1875-7/DL: T-337-2008

2.4.2.5.2 Prediction

The predicted concentration of the analyte k in an unknown sample whose response is r_{un} can be found using the eq 2.49 to 2.51, where the coefficients have been calculated using the PLS algorithm (see also Ref. 41 for more information of the prediction equations).

2.4.2.5.3 Advantages of PLS

1. Single step decomposition and regression; factors are directly related to constituents of interest rather than largest common spectral variations.

2. Calibrations are generally more robust provided that calibration set accurately reflects range of variability expected in unknown samples.

3. Enjoys the signal average advantages of other full-spectrum methods such as PCR and CLS [51].

2.4.2.5.4 Limitations of PLS

1. Models are more difficult to understand and interpret that CLS, ILS or PCR.

2. Calculations in PLS-1 are slower than most classical methods.

In the previous sections, four multivariate regression models have been presented. The selection of the samples and sensors used to build and validate the model is an important step that influences the quality of the predictions. In the following section, some ideas are given about the methodological selection of the best samples and sensors for calibration.

2.5 Optimal design in multivariate calibration

Calibration relates analyte concentrations and instrumental responses with the aim of achieving an acceptable predictive quality (in terms of trueness and precision) over all the experimental domain and applicable to the largest number of unknown samples possible. The mathematical expressions of the multiple linear regression (MLR) model, CLS, ILS and PCR are compared in the Table 2.1 along with their least-squares solution.

Table 2.1. Comparison of different regression models

Model	*Model expression*	*Least-squares solution*	*Vector of dependent variables*	*Matrix of the model*		
					Rows are:	*Columns are:*
MLR	$\mathbf{y} = \mathbf{X}\beta + \varepsilon$	$\mathbf{b} = \mathbf{X}^{+}\mathbf{y}$	$\mathbf{y}$	$\mathbf{X}$	experiments	variables
CLS	$\mathbf{r}_{un} = \mathbf{S}\,\mathbf{c}_{un,true} + \varepsilon$	$\mathbf{c}_{un} = \mathbf{S}^{+}\,\mathbf{r}_{un}$	$\mathbf{r}_{un}$	$\mathbf{S}$	sensors	pure spectra
ILS	$\mathbf{c}_k = \mathbf{R}\beta_k + \varepsilon$	$\mathbf{b}_k = \mathbf{R}^{+}\mathbf{c}_k$	$\mathbf{c}_k$	$\mathbf{R}$	samples	sensors
PCR	$\mathbf{c}_k = \mathbf{T}\theta_k + \varepsilon$	$\mathbf{q}_k = \mathbf{T}^{+}\mathbf{c}_k$	$\mathbf{c}_k$	$\mathbf{T}$	samples	scores

It can be seen that the dependent variables **y** in MLR are either the spectrum of the unknown sample $\mathbf{r}_{un}$ (CLS) or the concentration of the analyte under study in the calibration samples $\mathbf{c}_k$ (ILS, PCR). A row of the calibration matrix **X** corresponds either to the absorbances of the *K* analytes at a given wavelength (CLS) or to one calibration sample represented by the absorbances at *J* wavelengths (ILS) or its scores (PCR). A column of the calibration matrix (the settings of one variable in all the experimental points) can be either the spectrum of one pure analyte (CLS), the absorbance at a given wavelength in all the calibration samples (ILS) or the score of these samples in a given factor (PCR). The coefficients of the MLR model correspond to the concentration of the *K* analytes in the unknown sample in CLS.

The variance of the predicted concentration in these models depends on:

a) three sources of error: the measured responses from the unknown sample, measured responses from the calibration samples and the analyte concentrations in the calibration samples.
b) the mathematical expression of the model.
c) the points in the calibration matrices (the calibration space)
d) the position in the calibration space of the sample to be predicted (if the point is close or away from the points used for calibration).

The degree of complexity of the available expressions to calculate this variance in the different calibration models[1,41,47,58,62,98] depends on the assumptions made referent to the points a) to d). An usual simplification is to assume that the errors in the independent variables are neglected and to use the MLR expressions to calculated the variances. In this case, $var(c_{un,k})$ in ILS is given by eq 2.17 and only depends on the values in **R** (the absorbances of each calibration sample) and not on the values of the concentrations. In the same way, $var(\mathbf{c}_{un})$ in CLS depends only on the variance of the measured responses in $\mathbf{r}_{un}$ and on the matrix **S** (eq 2.12). With these assumptions, the DOE can be applied for the optimal building of the multivariate calibration models so that a reliable estimation of the searched relationship is at the minimum cost found[99].

As already indicated in §2.3.4, the values in the calibration matrices in MLR influence in the way that the measurements errors propagate to the predicted concentration. The DOE [16,21] indicates the most appropriate settings for the variables in each row of **X** to find a reliable estimation of the coefficients. However, classical designs (e.g. factorial-type designs) require the independent variables be manipulated according to the specified design strategy. This, for example, would require preparing calibration samples with specific values of absorbance at each of the measured wavelengths in ILS or scores in PCR and in CLS having pure analytes with the necessary absorbance values at each wavelength. This is not so readily used in multivariate calibration problems where it is impossible to make calibration samples of a determined composition and with prearranged values of the independent variables (specially when they are spectral measurements which are function of the chemical values). These situations are frequent in the quantitative analysis of natural products (i.e. water, flour, meat,...), whose chemical composition

cannot be controlled. Then the design variables in **S**, **R** or **T** are interrelated and cannot be manipulated independently of one another since they are influenced by many factors outside the experimenter's control, relating to the nature of the substance. These cases also arise in structure-activity relationship studies, where the values of the independent variables are fixed for the chemical configuration.

Therefore, the classical designs are difficult to use due to the impossibility of preparing samples with complex matrices and well determined instrumental responses. The alternative consists of obtaining a large list of possible points (rows in **S**, **R** or **T**) and select among them the most appropriate for calibration (generally the more economical subset that has the sufficient information for the model). This means that the best wavelengths(rows of **S**) are selected in CLS from the full spectra of the pure components (the matrix **S**). This is an advantage in ILS and PCR since the best calibration samples can be selected from a series of all the available candidate samples characterized by "inexpensive" multivariate measurements (e.g. spectroscopic data which can be collected with little labor and cost, **R**) or scores (**T**). For calibration, the analyte concentration must only be determined in the few selected samples using the more-time consuming referee or well-established method. Compared to analyzing a full set of samples, the cost of the model is being reduced.

The requirements for this selection step have been indicated in §2.3.4. The selection algorithms select the subset of points among all the points available for multiple linear regression (MLR) optimizing the desired criterion. An adequate criterion is the optimality of the selected subset for estimating the parameters of the model, since a low variance in the parameters should result in a good predictive power over the calibration range (e.g. the D-criterion and the A-criterion). Other criterion to consider is the quality of the predictions furnished by the model built (the G-criterion). Also measures of collinearity must be considered, specially for the selection of samples in ILS. The collinearity in the columns of **R** produces large uncertainty in the coefficients in the model and thus a large uncertainty in the estimated concentrations. In this case, wavelength selection is important to reduce the collinearity before (or simultaneously) to the selection of the calibration samples. PCR does not present these collinearity problems since the scores are orthogonal.

UNIVERSITAT ROVIRA I VIRGILI
EXPERIMENTAL DESIGN APPLIED TO THE SELECTION OF SAMPLES AND SENSORS IN MULTIVARIATE CALIBRATION
Joan Ferré Baldrich
ISBN:978-84-691-1875-7/DL: T-337-2008

2.6 Collinearity in multivariate calibration

2.6.1 Definition of collinearity and singularity

Multiple linear regression regresses several independent variables x on a dependent variable y (eq 2.6). Collinearity (also called multicollinearity) is defined as approximate linear dependence of at least one of the columns $\mathbf{x}_j$ of $\mathbf{X}$ with other/s column/s[MLII]. Singularity accurs when the variables are perfectly correlated.

2.6.2 Problems caused by collinearity

Collinearity and singularity concern an ill-conditioning of the matrix $\mathbf{X}^T\mathbf{X}$ that causes numerical and statistical problems in MLR related to stability and ability for matrix inversion:

1. At least one diagonal element of $(\mathbf{X}^T\mathbf{X})^{-1}$ is large and the associated least-squares estimated coefficient has large variance. The more collinear the x variables are, the more unstable becomes the linear system i.e. more susceptible to large changes in $\mathbf{b}$ produced by small changes in $\mathbf{X}$ or $\mathbf{y}$ due to noise. Unacceptable signs or too large values of the coefficients can be obtained [100] which affects their chemometric interpretation.

2. Numerical difficulties in $(\mathbf{X}^T\mathbf{X})^{-1}$, that cannot be calculated in case of singularity or may have large values in case of collinearity. For each exact linear dependence in the columns of $\mathbf{X}$ there is one zero eigenvalue of $\mathbf{X}^T\mathbf{X}$. Near-linear dependencies result in small eigenvalues.

3. Collinearity makes it more difficult to interpret the impact of each regressor on the response: correlated estimates cannot be interpreted separately and are unstable and unreliable. A regression coefficient is the partial derivative of the response with respect to a regressor variable[100]. It is desirable to have independent estimations of the coefficients.

4. The *t*-test can indicate statistical insignificance of the coefficients owing to large variances of regression coefficient estimates[76].

5. The prediction for new x measurements may be good at points with combinations of x's similar to those in the calibration data. Prediction at combinations different from these or extrapolation outside the range of the data can be adversely affected and have large errors.

6. Collinearity can exist in models with a good fit (a high multiple correlation coefficient). Since the residuals in the regression may be very small but the coefficients are estimated poorly, the traditional analysis of lack of fit does not signal potential collinearity problems[100].

2.6.3 A graphical representation of collinearity

The collinearity problem is illustrated in Figure 2.13, with a training set with two measured variables, x_1 and x_2. The *sample domain*[101] , rectangle ABCD, is delimited by the highest and lowest values of these variables. If the fitted model is the plane given by $y=b_0 + b_1x_1 + b_2x_2$, this can be interpreted as "a table whose legs are situated in the coordinates of each point and the lengths of the legs are the measured ys". Due to a different measurement error in each point, not all the "legs" fit the table so the inclination of the table (given by the model coefficients) has some uncertainty. This uncertainty is smaller in the direction AC since the table has legs at the extremes and an error in the length of a leg has less effect on the inclination of the table. The uncertainty is higher in the direction BD. A new prediction in a point in the direction AC has a small uncertainty since the model is stable in this direction, but a predicted point near the vertex B (the point in black) has a high uncertainty due to the uncertainty of the table in that direction.

The PCA decomposition used in the PCR model defines a new variable along the direction AC and another in the direction BD and the new experimental domain as the largest and smallest values of the scores along these two PCs. A sample in the vertex B would be detected as an outlier. In addition not considering the direction BC makes more stable predictions.

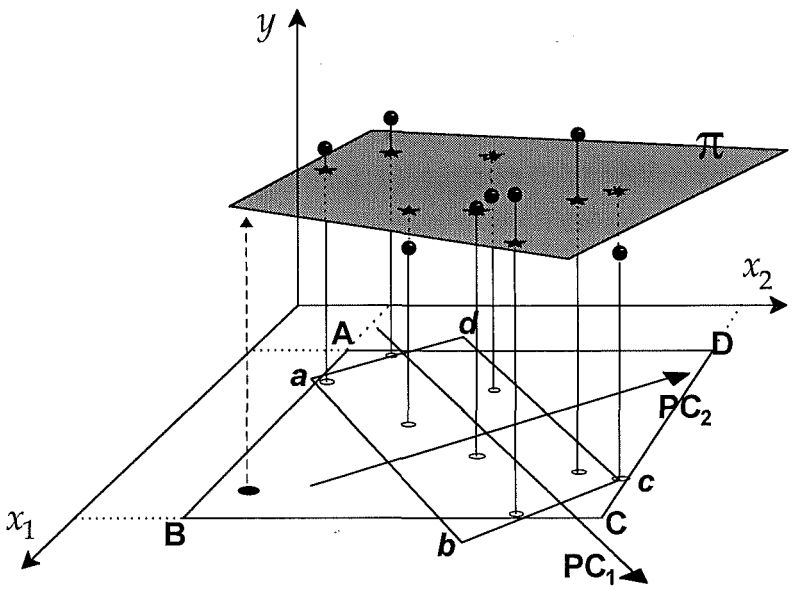


Figure 2.13. The training set has values that are very collinear: the x_1 values increase as the corresponding values for x_2 increase.

2.6.4 Detection and measures of collinearity

Several measures enable the extent of the collinearity problem to be evaluated: the correlation matrix of the regressor variables, the eigenvalues of the calibration matrix (and related measures such as the condition indices and the condition number), the tolerance and variance inflation factors of the estimated coefficients and the variance-decomposition proportions. A discussion about these measures can be found in references 34,100-106. The variance inflation factors (explained below) and the variance-proportion decompositions (explained in the section §4.3) are the diagnostic tools used in this thesis .

Variance inflation factors

The variance inflation factor (VIF) of the regression coefficient b_j is the jth diagonal element of the inverse of the correlation matrix of the variables. It can also be calculated as [106]

$$VIF_j = (1-R_j^2)^{-1} \quad (2.78)$$

where R^2_j is the multiple correlation coefficient of x_j regressed on all the other terms in the model. An equivalent definition is [10]:

$$VIF_j = UVIF_j \sum_{i=1}^{I} \left[x_{ij} - \bar{x}_j \right]^2 \quad (2.79)$$

The VIF_j measures the increase in variance in the fitted model due to the collinearity compared to a design of uncorrelated x-variables. The VIFj has a range 1 (non-correlated coefficients) to infinity (prefect correlation). Values larger than 1 indicate that the variable is affected by collinearity and larger than 10 that the correlation among the variables is so high that the coefficient is likely to be poorly estimated[107,108]. Since the maximum VIF is a lower bound on the condition number, a large VIF implies also a large condition number. The VIFs have been recommended as a general diagnostic measure of collinearity[104] and are used to measure if the selected points contain the sufficient information (if they are orthogonal enough) to estimate the model correctly [10,31]. Every wavelength or sample selection methodology can use the VIFs to measure the quality of the selected subset. In §4.6, the VIFs is shown to be equivalent to the Lorber's definition of selectivity. However, since the VIFs are a global measure, they do not indicate which regressors are involved and are unable to distinguish among coexisting spectral overlap situations of three or more components.

2.6.5 Influence of collinearity in multivariate calibration

Collinearity largely affects the prediction error in multivariate calibration models. Elimination of collinearity is important in the instrumental methods of analysis. Different sources of collinearity and their solutions are:

1. A bad experimental design of the analyte concentrations in the calibration samples. An example is the artificial calibration samples made by dilutions of a

single mixture with of all constituents of interest. The concentrations of all constituents vary together and their spectra all increase and decrease in sympathy. Multivariate models, which correlate changes in the concentrations to changes in the spectra, will fail since they will detect only one cause of variation regardless of how many constituents were mixed together in the original mixture. To an eigenvector-based model, one only factor will contain nearly all the variance in the data. A sample that does not have exactly the same ratio of constituent concentrations as the calibration samples will be predicted wrong. Collinearity can be reduced here with properly designed calibration mixtures having different ratios of the concentration on the components of interest.

2. Physical constraints in model or in data so that only certain combinations of the independent variables can be evaluated. This could be solved by adequately selecting the regression method so that it can handle collinear data, such as factor-based regression or ridge regression.

3. In CLS (eq 2.22), the overlap of the pure component spectra produces collinear columns in **S** and large variances and covariances of the concentration estimates. This gives an unstable system of equations and small relative changes in **r** due to measurement error can produce large relative changes in **c** so that misleading results can be obtained (e.g. negative concentration values for analytes that are present). The effects of collinearity can be reduced (but not eliminated unless completely selective sensors are available) with a correct choice of the wavelengths, which affects both the trueness and precision of the predicted concentrations. Criteria for wavelength selection are usually based on some measure of orthogonality in the **S** matrix, such as the selectivity by Lorber[17]. This and other criteria are discussed in the chapter §4.

4. In ILS, singularity is produced by a number of calibration samples inferior to the number of measured variables (over-estimated regression).This can be solved by using more calibration samples, removing redundant variables (e.g. with genetic algorithms or stepwise multiple linear regression, although this last method is not sensitive to the collinearity of the independent variables) or using factor-based models such as PCR and PLS, which reduce the number of regressor variables. Another source of collinearity in ILS is the correlation between the

columns of **R** produced, specially in NIR or UV-vis spectra where the number of responses can easily approach 1000, by the similar absorbances at adjacent wavelengths that tend to increase and decrease together in the calibration samples. This causes a large sensitivity of the estimated $\mathbf{b}_k$ to small changes in $\mathbf{c}_k$ (see §2.6.2). The large variance (low precision) of the coefficients produces large variances for the predicted concentration in unknown samples. For these reasons ILS is always accompanied of wavelength selection to reduce the collinearity. PCR or PLS are usually employed instead of ILS. These methods reduce the variance of the coefficients compared with ILS by discarding part of the information of the data to estimate the regression coefficients. The resulting estimators are biased, but they may be preferable to ILS.

ISBN:978-84-691-1875-7/DL. T-337-2008

2.7 References

1. Lorber A., Kowalski B.R. *J. Chemom.* 2 (1988) 93-109.
2. Searle S.R. *Matrix Algebra Useful for Statistics.* John Wiley & Sons. New York. 1982.
3. Manne R. *Chem. Intell. Lab. Syst.* 2 (1987) 187-197.
4. De Jong S. *J. Chemom.* 9 (1995) 323-326.
5. Deming S.N., Morgan S.L. *Experimental Design: a chemometric approach.* Elsevier. Amsterdam. 1987
6. Wold S., Sjöström M., Carlson R., Torbjörn L., Hellberg S., SkagerbergB., Wikström C., Öhman J. *Anal. Chim. Acta* 191 (1986) 17-32.
7. Hellberg S., Sjöström M., Skagerberg B., Wold S. *J. Med. Chem.* 30 (1987) 1126-1135.
8. Hellberg S., Sjöström M., Skagerberg B., Wikström, C., Wold S. *Acta Pharm. Jugosl.* 37 (1987) 53-65.
9. Peissik A. *Methodologie de la recherche experimentale: properties et caracteristiques des matrices d'experiences pour les modeles polynomiaux du second degre.* Thesis. 1995. Université d'Aix Marseille III.
10. Sergent M., Mathieu D., Phan-Tan-Luu R., Drava G., *Chem. Intell. Lab. Syst.* 27 (1995) 153-162.
11. Meloun M., Militký J., Forina M. *Chemometrics for analytical chemistry. Vol 2. PC-aided Regression and Related Methods.* Ellis Horwood: Great Britain 1994
12. Lebart L., Morineau A., Fénelon J.P. *Traitement des données statistiques. Méthodes et programmes.* Dunod. 2 Ed. Bordas. Paris 1982.
13. Weisberg S. *Applied Linear Regression* 2nd Edition, Wiley, New York 1985.
14. Atkinson A.C., Humter W.G. *Technometrics* 10 (1968) 271-289.
15. Atkinson A.C., Donev.A.N. *Optimum Experimental Designs.* Oxford Statistical Science Publications: Oxford 1992
16. Carlson R. *Design and optimization in organic synthesis.* Elsevier: The Netherlands 1992
17. Lorber A. *Anal. Chem.* 58 (1986) 1167-1172.
18. Goupy J. *Chem. Intell. Lab. Syst.* 33 (1996) 3-16.
19. Araujo P.W., Brereton R.G. *Trends Anal. Chem.* 15 (1996) 26-31.
20. Araujo P.W., Brereton R.G. *Trends Anal. Chem.* 15 (1996) 63-70.
21. Box G.E.P., Hunter W.G., Hunter J.S. *Statistics for Experimenters* Wiley: New York. 1978

22. Broudiscou A., Leardi R., Phan-Tan-Luu R., *Chem. Intell. Lab. Syst.*, 35 (1996) 105-116.
23. Pukelsheim F. *Optimal Design of Experiments*. Wiley. New York. 1993
24. Fedorov V.V. *Theory of Optimal Experiments*.(translated and edited by W.J. Studden and E.M. Klimko) Academic Press: New York, 1972
25. Kiefer J. *J. Royal. Statist. Soc. B.* 21 (1959) 272-319.
26. Pázman A. *Foundations of Optimum Experimental Design*. D.Riedel Publishing Company. Dordrecht. 1986.
27. Nishii R. *Discrete Mathematics*, 116 (1993) 209-225.
28. Pukelsheim F., Rosendberg J.L. *J. Amer. Stat. Assoc.* 88 (1993) 642-649.
29. Atkinson A.C. *Chem. Intell. Lab. Syst.*, 28 (1995) 35-47.
30. Khuri A.I., Cornell J.A *Response Surfaces. Designs and Analyses*. Marcel Dekker: USA 1987
31. Mathieu D. *Contribution de la Méthodologie de la Recherche Expérimentale à l'étude des relations Structure-Activité*. Thesis. Marseille, 1981.
32. Steinberg D.V., Hunter W.G. *Technometrics*, 26 (1984) 71-130.
33. Papakyriazis P.A. *J. Econometrics* 7 (1978) 351-372.
34. Willan A.R., Watts D.G.. *Technometrics* 20 (1978) 407-412.
35. De Aguiar P.F., Bourguignon B., Khots M.S., Massart D.L., Phan-Tan-Luu R. *Chem. Intell. Lab. Syst.*, 30 (1995) 199-210.
36. Franquart P. *Optimisations Multi-criteres et Methodologie de la Recherche Experimentale*. Thesis. Université d'Aix-Marseille III. 1992.
37. Naes T. *J. Chemom.* 1 (1987)121-134.
38. Kennard R.W., Stone L.A. *Technometrics* 11 (1969) 137- 148.
39. Jouan-Rimbaud D., Khots M.S., Massart D.L., Last I.R., Prebble K.A. *Anal. Chim. Acta* 315 (1995) 257-266.
40. Zemroch P. J. *Technometrics* 28 (1986) 39-49.
41. Martens H.; Naes T. *Multivariate Calibration*, Wiley: New York, 1989.
42. Sanchez E., Kowalski B.R. *J. Chemom.* 2 (1988) 247-263.
43. Kowalski B.R., Seasholtz M.B. *J. Chemom.* 5 (1991)129-145.
44. Thomas E.V. *Anal. Chem.* 66 (1994) 795A-804A.
45. Geladi P., Kowalski B.R. *Anal. Chim. Acta* 185 (1986) 1-17.
46. Martens H., Karstang T., Naes T. *J. Chemom.* 1 (1987) 201-219.
47. Marbach R., Heise H.M. *Chem. Intell. Lab. Syst.* 9 (1990) 45-63.
48. Frank I.E., Friedman J.H. *Technometrics* 35 (1993) 109-139.

49. Marbach R., Heise H.M. *Trends Anal. Chem.* 11 (1992) 270-275.
50. Brown C.W., Lynch P.F., Obremski R.J , Lavery D.S. *Anal. Chem.* 54 (1982) 1472-1479.
51. Haaland D.M., Thomas E.V. *Anal. Chem.* 60 (1988) 1193-1202.
52. Booksh K.S., Kowalski B.R *Anal. Chem. 66* (1994) 782A-804A.
53. Beebe K.R., Kowalski B.R. *Anal. Chem.* 59 (1987) 1007A-1017A.
54. Lang P.M., Kalivas J.H. *J. Chemom.* 7 (1993) 153-164.
55. Kalivas J.H., Lang P. M. *Chem. Intell. Lab. Syst.* 32 (1996) 135-149.
56. Maris M.A., Brown C.W., Lavery D.S. *Anal. Chem.* 55 (1983) 1694-1703.
57. Lorber A., Wagen L.E., Kowalski B.R., *J. Chemom.* 1 (1987) 19-31.
58. Faber K., Kowalski B.R. *J. Chemom.* 11 (1997) 181-238.
59. Bauer G., Wegscheider W., Ortner H.M. *Spectrochimica Acta 46B* (1991) 1185-1196
60. Otto M., Wegscheider W. *Anal. Chem.* 57 (1985) 63-69.
61. Schmidt P.C., Glombitza B.W. *Trends Anal. Chem.* 14 (1995) 45-48.
62. Naes T., Martens H. *J. Chemom.* 2 (1988) 155-167.
63. Kvalheim O.M. *Chem. Intell. Lab. Syst.* 8 (1990) 59-67.
64. Sekulic S., Seasholtz M.B., Wang Z., Kowalski B.R. *Anal. Chem.* 65 (1993) 835A-845A.
65. Lorber A., Kowalski B.R. *Appl. Spectrosc.* 42 (1988) 1572-1574.
66. Naes T. *Chem. Intell. Lab. Syst.* 5 (1989) 155-168.
67. Seasholtz M.B., Kowalski B.R. *J. Chemom.* 6 (1992) 103-111.
68. Lorber A., Harel A., Goldbart Z., Brenner. B. *Anal. Chem.* 59 (1987) 1260-1266.
69. Lorber A., Faber K., Kowalski B.R. . *Anal. Chem.* 69 (1997) 1620-1626.
70. Lorber A. , Kowalski B.R. *J. Chemom.* 2 (1988) 67-79.
71. Jouan-Rimbaud D., Massart D.L., Leardi R., De Noord O.E. *Anal. Chem.* 67 (1995) 4295-4301.
72. Leardi R. *J. Chemom.* 8 (1994) 65-79.
73. Vigneau E., Devaux M.F., Qannari E.M., Robert P. *J. Chemom.* 11 (1997) 239-249.
74. Shaffer R.E. Small, G.W. Arnold M.A. *Anal. Chem.* 68 (1996) 2663-2675.
75. Lorber A., Kowalski B.R. *Appl. Spectrosc.* 44 (1990) 1464-1470.
76. Militký, J., Meloun M. *Anal. Chim. Acta* 277 (1993) 267-271.
77. Osten D.W. *J. Chemom.* 2 (1988) 39-48.
78. Wold S., *Technometrics* 20 (1978) 397-405.
79. Wold S., Esbensen K., Geladi P. *Chem. Intell. Lab. Syst.* 2 (1987) 37-52.

UNIVERSITAT ROVIRA I VIRGILI
EXPERIMENTAL DESIGN APPLIED TO THE SELECTION OF SAMPLES AND SENSORS IN MULTIVARIATE CALIBRATION
Joan Ferré Baldrich
ISBN:978-84-691-1875-7/DL: T-337-2008

80. Kvalheim O.M., Karstang T.V. *Chem. Intell. Lab. Syst.* 7 (1989) 39-51.
81. Malinowski, E. R. *Factor Analysis in Chemistry*, 2nd Ed. John Wiley & Sons Inc: New York 1991.
82. Eastment H.T., Krzanowski W.J. *Technometrics* 24 (1982) 73-77.
83. Todeschini R. *Anal. Chim. Acta* 348 (1997) 419-430.
84. Sun J. *J. Chemom.* 9 (1995) 21-29.
85. Sun J. *J. Chemom.* 10 (1996) 1-9.
86. Jouan-Rimbaud D., Walczak B., Massart D.L., Last I.R., Prebble K.A. *Anal. Chim. Acta* 304 (1995) 285-295.
87. Sutter J.M., Kalivas J.H., Lang P.M. *J. Chemom.* 6 (1992) 217-225.
88. Xie Y., Kalivas J.H. *Anal. Chim. Acta* 348 (1997) 19-27.
89. Xie Y., Kalivas J.H. *Anal. Chim. Acta* 348 (1997) 29-38.
90. Höskuldsson A. *J. Chemom.* 6 (1995) 91-123.
91. Höskuldsson A. *J. Chemom.* 2 (1988) 211-228.
92. De Jong S., Ter Braak C.J.F. *J. Chemom.* 8 (1994) 169-174.
93. Geladi P. *J. Chemom.* 2 (1988) 231-246.
94. Lindgren F., Geladi P., Wold S. *J. Chemom.* 8 (1994) 377-389.
95. Lindgren F., Geladi P., Wold S. *J. Chemom.* 7 (1993) 45-59.
96. Marengo E., Todeschini R. *Chem. Intel. Lab. Syst.* 12 (1991) 117-120.
97. Rännar S., Lindgren F., Geladi P., Wold S. *J. Chemom.* 8 (1994) 111-125.
98. Karstang T.V., Toft J., Kvalheim O.M. *J. Chemom.* 6 (1992) 177-188.
99. Araujo P.W., Brereton R.G. *Trends Anal. Chem.* 15 (1996) 156-163.
100. Myers R.H *Classical and modern regression with applications 2nd edition.* Duxbury Press: Belmont 1990.
101. Mandel J. *J. Res. of the National Bureau of Standards* 90 (1985) 465-476.
102. Belsley D.A., Kuh E., Welsch R.E.. *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity.* Wiley: New York 1980 Chapter 3.
103. Berk K.N *J. Amer. Statist. Assoc.* 72 (1977) 863-866.
104. H. M. Wadswoth. *Handbook of Statistical Methods for Engineers and Scientist.* McGraw Hill: USA 1990.
105. Snee R.D. *J. Qual. Tech.* 5 (1973) 67-79.
106. Kalivas J.H. *J. Chemom.* 3 (1989) 409-418.
107. Snee R.D. *Technometrics* 19 (1977) 415-428.
108. Marquardt D.W. *Technometrics.* 12 (1970) 591-612.

UNIVERSITAT ROVIRA I VIRGILI
EXPERIMENTAL DESIGN APPLIED TO THE SELECTION OF SAMPLES AND SENSORS IN MULTIVARIATE CALIBRATION
Joan Ferré Baldrich
ISBN:978-84-691-1875-7/DL: T-337-2008

Chapter 3

Selection of Calibration Samples and Factors in Principal Component Regression

ISBN:978-84-691-1875-7/DL: T-337-2008

3.1 Introduction

3.1.1 Aim of the chapter

The aim of this chapter is to propose a new methodology for selecting the best subset of calibration samples for PCR from the instrumental responses of a large set of samples. Only the selected samples are submitted to chemical analysis and calibration, thus reducing the time and cost of the calibration step. As a part of the methodology, a fast method for selecting the most important principal components for regression is developed.

3.1.2 Structure of the chapter

After the introduction, containing the aim of the chapter, its structure and a bibliographic revision, the sections §3.2 to §3.7 contain the main contents structured in different papers:

§3.2 is the paper *Selection of best calibration sample subset for multivariate regression.* Joan Ferré, F. Xavier Rius *Anal. Chem.* 68, (1996) 1565-1571. Here a sample selection methodology in PCR, based on the D-optimality criterion and the Fedorov's exchange algorithm, is described. The algorithm is also reviewed in the last part of this paper.

§3.3 is the paper *Determination of ethylene content in poly(propylene-ethylene) copolymers using near-infrared spectra (NIR) and multivariate calibration* Villagrasa C., Ferré J., Larrechi M.S., Rius F.X., García C. (*in preparation*). Here near-infrared (NIR) data of copolymers is used to compare the predictive ability of PLS and of PCR with the factors selected according to the methodology described in §3.2.

§3.4 is the paper *Constructing D-optimal designs from a list of candidate samples.* Joan Ferré, F. Xavier Rius *Trends Anal. Chem.* 16, (1997) 70-73. Here, the Fedorov's algorithm, that is seldom used in the analytical literature, is presented and compared

UNIVERSITAT ROVIRA I VIRGILI
EXPERIMENTAL DESIGN APPLIED TO THE SELECTION OF SAMPLES AND SENSORS IN MULTIVARIATE CALIBRATION
Joan Ferré Baldrich
ISBN:978-84-691-1875-7/DL: T-337-2008

with the popular Kennard-Stone's algorithm[1] and the random division of samples into calibration and validation sets. Kennard-Stone's algorithm selects samples spread over the experimental domain. In contrast, the samples selected with the Fedorov's algorithm tend to lie at the extremes of the experimental domain for a linear model of first degree. These methods are applied in a MLR model to predict the octane index in fuel samples

§3.5 is the paper *Selection of calibration points for PCR in QSAR studies* . Joan Ferré, F. Xavier (*in preparation*) where the Fedorov's algorithm is applied in quantitative structure-activity relationship (QSAR) studies to select calibration points characterized by the scores on some PCs of a series of properties. The algorithm is an alternative to selecting the samples for its similarity to the points of a given experimental design (usually a factorial design).

§3.6 contains the paper *Assessing the validity of principal component regression models in different analytical conditions*. Rius A.; Callao M.P., Ferré J.; Rius F.X., *Anal. Chim. Acta* 337 (1997) 287-296. This paper proposes a methodology for assessing, before using the piecewise direct standardization (PDS) technique, if a PCR model is valid when the actual working conditions are different from those used for modeling. The contribution to this work consists on applying the D-optimality criterion for selecting, from a large set, the minimum number of samples that must be measured. These already analyzed samples can be used in the standardization process in case that it is necessary.

3.1.3 Bibliographic revision and comments

To estimate the coefficients in a multivariate calibration model a set of calibration samples with known instrumental responses (e.g. absorbance at different wavelengths) and analyte concentrations is required.

The design or selection of the calibration samples must consider and satisfy statistical and economical requirements. The economical aspect mainly comprises the cost and time involved to obtain the calibration samples. Measuring the instrumental

responses is supposed to be cheaper and less time-consuming than determining the analyte concentration. This last step may require cumbersome analyses with a well established or reference method of analysis, which might be expensive, slow or undesirable. The statistical performance characteristics are related to the quality of the values predicted with the model. The spectra and composition of the calibration samples should emulate the unknown samples as closely as possible. The composition should span the expected range of concentration values of the future unknown samples and the spectra should be representative of all the constituents that contribute to the instrumental response in the unknown sample to enable the model to recognize the information for the constituents of interest. All phenomena (with chemical, physical or other basis) that vary in the unknown samples and influence the instrumental measurements must also vary in the calibration set over the same ranges. Martens and Naes[2] and Gemperline[3] also commented these ideas.

Commonly, calibration sets have a relatively large number of samples (maybe hundreds) to achieve a statistical representation of all sample properties. Lorber and Kowalski[4] proved that, to improve the prediction quality, it is always advantageous to add samples to the calibration. Although the mathematical expression of the Sherman-Morrison-Woodbury theorem used in their proof contains a small erratum ($\mathbf{X}^T\mathbf{X}$ should be written instead of $\mathbf{X}$; the correct expression can be found in Meyers[5] page 459 or in Weisberg[6] page 293), the conclusions are not affected. However, due to the effort of analyzing each calibration sample with the well-established technique, the considerable cost of obtaining a large calibration set may not always be compensated by an equal increase in the quality of the model. Faber and Kowalski[7] commented that increasing the number of calibration samples above some limit has only a marginal effect on the prediction error. In addition, Honigs *et al.*[8] mentioned some specific drawbacks to the use of a large sample set, such as the possibility that a property that is only present in a few calibration samples be ignored by the model if many other samples do not present this property. A small training set selected to contain a high degree of variability could avoid that problem.

The experimenter usually employs his/her subjective criterion to decide when the number of the calibration samples is 'sufficient' and when their composition spans correctly the experimental domain. The number considered as 'sufficient' depends on the nature, cost and difficulty in obtaining the samples. Moreover, the available

samples are usually randomly divided into calibration and validation sets. Several authors[9,10] found that calibrations based on random choice may perform well but there is also a chance to obtain much worse results than with a careful selection using a methodological approach. The validation of the models also depends on how well the calibration set represents the validation set.

Owing to the large importance of the calibration set on the predictive ability of the model, the selection of the calibration samples should not depend upon scarcely rigorous criteria. The idea here is to employ a mathematical criterion to reduce the cost of the calibration by selecting an adequate number of calibration samples that gives a compromise between the performance criteria and the cost of the model. Since it is easy to perform measurements on a large number of samples, there is a large probability that the calibration set contains the most important variations in the data.

3.1.3.1 Bibliographic revision of calibration sample selection

Table 3.1 resumes different approaches found in the literature for selecting calibration samples in PCR from a large list when they cannot be synthesized. These procedures often use spectral data, which can be collected with little labor and cost, and characterize the samples by their scores on a certain number of principal components (PCs). Other approaches found, although not based on PCR, are also indicated but not commented.

3.1.3.2 Comments to the existing approaches

The idea of all these methods is to select a representative sample among other similar and avoid discarding useful information. The approaches based on the Kennard-Stone-like algorithms[1,11] and clustering[10] have the interesting advantage of collecting points evenly spread over the whole experimental domain and span the variation as uniformly as possible. In this way, the selected samples are not too close to any of the others and can be used for check the fit of the model or add new terms if necessary. However, some general objections can be made to the methods indicated in the Table 3.1.

UNIVERSITAT ROVIRA I VIRGILI
EXPERIMENTAL DESIGN APPLIED TO THE SELECTION OF SAMPLES AND SENSORS IN MULTIVARIATE CALIBRATION
Joan Ferré Baldrich
ISBN:978-84-691-1875-7/DL: T-337-2008

Table 3.1. Sample selection approaches proposed by several authors.

Authors	*Proposed approach. Comments*
Hruschka and Norris[12]	Selection using concentration and instrumental responses of all the samples in ILS.
Honigs *et al.*[8]	Subtractions from the spectral data to choose the spectrally unique samples for calibration from a large set. The algorithm spans the spectral variation as much as possible.
Zemroch[13]	Clustering of the candidate points and selection of one point from each cluster in MLR.
Naes[10]	Clustering of the samples using their scores on a certain number of PCs of NIR spectra. The sample farthest away from the center of every cluster is selected as representative.
Puchwein[9]	Iterative elimination of similar samples from a large data set using sample scores and distances between data points. The samples retained for calibration have the largest Mahalanobis distance from the origin and are representative of the complete original data set.
Lorber and Kowalski[4]	Algorithm for sensor selection that can be modified for sample selection. The calibration samples, that are optimal for all analytes, are selected after measuring the response of the unknown sample. The optimal calibration set may change (thus different samples must be analyzed) for each unknown sample., which does not reduce the cost of the calibration.
Schostack and Malinowski[14]	Iterative key set factor analysis (IKSFA) to select the key set, the preferred set of analytical wavelengths or calibration samples that best characterizes a multicomponent system.
Isaksoon and Naes[15]	Compared Naes[10] and Honigs *et al.*[8] approaches in PCR. The Naes[10] approach performed better in terms of prediction error.
Kalivas[16]	Generalized simulated annealing (GSA) to select calibration samples from a set of NIR spectra by minimizing the Mahalanobis distance between the unknown sample and the average spectrum. Not applied to PCR but to PLS with one latent variable. The reason for this one latent variable was not indicated, nor the criteria to estimate the appropriate number of calibration samples.
Hitchcock *et al.*[17]	Design of optimal calibration concentration matrices for spectroscopic data and PLS. They assume that the analyst knows the components present in the unknown sample, the proper number of calibration samples and is able to artificially generate them.
Tong-Hua *et al.* [18]	A genetic algorithm is used to select calibration samples. Its number is decided a priori.
Marengo and Todeschini[11]	Algorithm for selecting experiments uniformly distributed from the set of candidates using the original variables, not PCA scores. It does not require any preliminary hypothesis about a regression model. Similar to Kennard-Stone's algorithm[1].
Aastveit and Marum[19]	They compared different strategies for sample selection in PCR, including the Naes's[10] clustering method. Local calibration methods performed the best.
Naes and Isaksoon[20]; Araujo and Brereton[21]	General rules for selection of samples for calibration, specially for ILS models
Jouan-Rimbaud *et al.* [22]	Kennard-Stone algorithm to select calibration samples in ILS.

One objection is the lack of a clear mathematical criterion to decide the sufficient (or optimal) number of calibration samples to correctly estimate the model coefficients. The authors either did not give any criterion[4,11,16] or decided this number beforehand depending on how many samples one wants to or can afford to use in the calibration [8,10,13,]. For example, Naes[10] divided randomly the available samples into calibration and validation sets, and decided arbitrarily a number of samples from the calibration set to be selected using clustering. Honigs *et al.*[8] continued the iterative procedure until the desired number of spectra was selected. Puchwein[9] submitted to factor analysis the raw data of the reduced number of samples and the transformation matrix derived was used to recalculate the factor scores of the whole data set. The subset was assumed to still represent the original set if the redefined subset region contained all or at least most of the original samples. However, he was not able to formulate a rule to stop the sample reduction automatically when the minimum subset was reached. When the experimenter is not able to decide the exact number of samples that are adequate to assure the quality of the estimated model coefficients, these approaches may fail.

Another limitation is that the selections based on distances require the PCs used in the model (which should be the ones with the best predictive ability) to be specified beforehand (e.g. the scores on the important PCs are used to compute the Mahalanobis distance in the clustering method[10]). However, this cannot be known only from the instrumental responses matrix (see §2.4.2.4.3); they must be selected considering the predictive ability of PCR models made with an increasing number of factors. Moreover, the factors should be included in these models in order of their correlation with the concentration, not in order of the percentage of explained variance of the data matrix. To make sure that the optimal number of components is incorporated it was suggested using experience with similar systems on how many PCs had main predictive relevance or to randomly select a few calibration samples and estimate the optimal number of factors with cross-validation or leverage correction [10,15]. This number could then be used in a search for additional calibration samples according to the selection procedure. Naes[10] selected the optimal factors for clustering using all the available samples and their analyte concentrations. So did Xie and Kalivas[23] and Sutter *et al.*[24] and in their optimization procedure to find the optimal set of factors. These approaches cannot be used here since the aim of the methodology is avoid analyzing all the samples. Puchwein[9] first considered the factors required to regenerate the raw data matrix within the measurement error and

the factors were selected by regressing the concentration against the scores of each factor individually. The factors were introduced into the final model by the absolute value of their regression coefficients in decreasing order. Isaksoon and Naes[15] selected the PCs with the largest variance but did not give any criterion to justify the number of factors used. They indicated that further studies on how to select an optimal number of PCs should be performed.

Considering the mentioned comments it can be stated that a strategy of selection of the calibration sample subset for PCR requires:

1. A criterion for judging the quality of each subset of N candidates. The selected subset is the one that optimizes this criterion over all the other possible subsets. Faber and Kowalski[7] also indicated that the a suitable selection criterion should provide a design of the calibration samples good enough to avoid extrapolations.

2. An optimization algorithm to find the optimal subset avoiding the examination of all possible combinations of subsets of samples.

3. A criterion for comparing the subsets with a different number of samples, so that the optimal N can be decided.

4. A method for selection of the best predictive factors for PCR in case that the optimization criterion needs them to be specified beforehand

A selection methodology should consider the above indicated steps using the instrumental responses of a large set of samples but analyzing only the minimum number required. This has been solved in the paper presented in §3.2.

3.1.4 References

1. Kennard, R.W. ; Stone, L.A. *Technometrics* 11 (1969) 137- 148.
2. Martens H.; Naes T. *Multivariate Calibration*, Wiley: New York, 1989.
3. Gemperline P.J. *J. Chemom.* 3 (1989) 549-568.

UNIVERSITAT ROVIRA I VIRGILI
EXPERIMENTAL DESIGN APPLIED TO THE SELECTION OF SAMPLES AND SENSORS IN MULTIVARIATE CALIBRATION
Joan Ferré Baldrich
ISBN:978-84-691-1875-7/DL: T-337-2008

4. Lorber A. , Kowalski B.R. *J. Chemom.* 2 (1988) 67-79.
5. Myers R.H *Classical and modern regression with applications 2nd edition.* Duxbury Press: Belmont 1990.
6. Weisberg S. *Applied Linear Regression* 2nd Edition, Wiley, New York 1985.
7. Faber K., Kowalski B.R. *J. Chemom.* 11 (1997) 181-238.
8. Honigs D.E.; Hietfje G.M.; Mark H.L.; Hirschfeld T.B. *Anal. Chem.* 57 (1985) 2299-2303.
9. Puchwein G. *Anal. Chem.* 60 (1988) 569-573.
10. Naes, T. *J. Chemom.* 1 (1987)121-134..
11. Marengo E. ; Todeschini R. *Chem. Intel. Lab. Syst.* 16 (1992) 37-44.
12. Hruschka W.R., Norris K.H. *Applied Spectrosc.* 36 (1982) 261-265.
13. Zemroch, Peter J. *Technometrics* 28 (1986) 39-49.
14. Schostack, K. J.; Malinowski, E. R. *Chem. Intel. Lab. Syst.* 6 (1989) 21-29.
15. Isaksoon T.; Naes T. *Appl. Spectrosc.* 44 (1990) 1152-1158.
16. Kalivas, J.H. *J. Chemom.* 5 (1991) 37-48.
17. Hitchcock K., Kalivas, J.H., Sutter J.M. . *J. Chemom.* 6 (1992) 85-96.
18. Tong-Hua Li, Lucasius C.B., Kateman G. *Anal. Chim. Acta* 268 (1992) 123-134.
19. Aastveit A.H.; Marum P., *Appl. Spec.* 47 (1993) 463-469.
20. Naes T., Isaksoon T. *NIR news* 5 (1994) 16-17.
21. Araujo P.W., Brereton R.G. *Trends Anal. Chem.* 15 (1996) 156-163.
22. Jouan-Rimbaud D., Khots M.S., Massart D.L., Last I.R., Prebble K.A. *Anal. Chim. Acta* 315 (1995) 257-266.
23. Xie Y., Kalivas J.H. *Anal. Chim. Acta* 348 (1997) 19-27.
24. Sutter, J.M.; Kalivas J.H. ; Lang P.M. *J. Chemom.* 6 (1992) 217-225.

3.2 Selection of the Best Calibration Sample Subset for Multivariate Regression

Anal. Chem. 68 (1996) 1565-1571

Joan Ferré and F. Xavier Rius.

Departament de Química. Universitat Rovira i Virgili.
Pl. Imperial Tarraco, 1, 43005-Tarragona. SPAIN

This paper discusses a methodology for selecting the minimum number of calibration samples in principal component regression (PCR) analysis. The method uses only the instrumental responses of a large set of samples to select the optimal subset, which is then submitted to chemical analysis and calibration. The subset is selected to provide a low variance of the regression coefficients. The methodology has been applied to UV-visible spectroscopy data to determine Ca^{2+} in water and near-IR spectroscopy data to determine moisture in corn. In both cases, the regression models developed with a reduced number of samples provided accurate results. As far as precision is concerned, a similar root-mean-squared error of cross-validation (RMSECV) is found when comparing the new methodology with the results of the regression models that use the complete set of calibration samples and PCR. The number of analyzed samples in the calibration set can be reduced by up to 50%, which represents a considerable reduction in costs.

Received for review May 22, 1995. Accepted January 18, 1996.

Multivariate regression methods [1-4] such as classical least-squares (CLS), inverse least-squares (ILS), principal component regression (PCR), or partial least-squares regression (PLSR) enable mathematical models to be developed that relate multivariate instrumental responses r_j (e.g.: spectral intensities) from many calibration samples to the known analyte concentrations in these samples (c_i) according to eq 1:

$$c_i = f(r_1, r_2, \ldots, r_j) + e_i \qquad (1)$$

where e_i is the residual associated with the ith concentration. This relationship can then be used to predict analyte concentrations for unknown samples from their instrumental responses. When calibration standards are not easily synthesized (e.g., natural samples), they are selected from among all the available samples. For this calibration set to be obtained, the instrumental responses must be measured (which is usually relatively quick and easy, as in spectroscopic analysis) and the analyte concentrations determined. These concentrations are usually determined with a reference or well-established method that may be slow, expensive, or cumbersome. Lorber and Kowalski[5] showed that, to establish suitable prediction models, the higher the number of calibration samples providing supplementary information the better. However, this might mean that the cost and time spent in obtaining a high number of calibration standards is not affordable in cost-effectiveness terms. The analyst might be interested in using not all but only the minimum (or a reduced) number of calibration samples, provided that the developed model is able to furnish prediction values of the desired quality.

In this paper we report a procedure for selecting an adequate subset of calibration samples for PCR from the instrumental responses of a large number of samples. Only the selected samples are submitted to the more time-consuming chemical analysis and to principal component modeling. The selected samples give the lowest variance for the estimated regression coefficients and enable the principal components (PCs) that provide the best predictive PCR model to be selected. The results compare well to the ones obtained using the complete set of calibration samples.

Several procedures for selecting a subset of calibration samples for PCR from a large data set have been proposed.[6,7] Clustering,[8] iterative elimination of similar samples by using the Mahalanobis distance,[9] and iterative key set factor analysis

(IKSFA)[10] have been applied to spectral data and make use of the scores on a certain number of PCs. In quantitative structure-activity relationship studies (QSAR), principal component analysis followed by sample selection to fit factorial and fractional factorial designs have been reported.[11-15] Other approaches for selecting calibration samples, although they do not focus on PCR, have been used and compared.[16-21] Recently, Naes and Isaksoon[22] gave some general principles for selecting calibration samples.

Although promising results have been obtained, some general objections can be made to these approaches. While they aim to span as much of the experimental domain as possible, the mathematical expression of the regression model is seldom considered. Moreover, no unambiguous mathematical criterion is reported (except for Puchwein's approach[9]) to decide how many calibration samples are required. The experimenter must usually decide a priori what a "sufficient" number of calibration standards is, using subjective criterion. In such cases, there is no guarantee that the selected samples will contain the necessary information to build a model which should provide accurate and precise predictions throughout the experimental domain. Moreover, as is shown below, the random separation of samples into calibration and test sets can give models with poor prediction ability if the calibration set does not contain enough information to allow correct estimation of the regression coefficients.

In the field of experimental design theory,[23-29] a variety of algorithms[25-29] have been used to select subsets of calibration samples for multiple linear regression (MLR) models. A usual selection criterion is to minimize the variance of the estimated regression coefficients, but this is very time-consuming when the number of predictor variables is high and is not suitable for highly collinear data. These two characteristics are common, for example, in spectroscopic data. PCR can overcome these problems since it can deal with collinear data and a large number of variables.

The methodology presented here uses Fedorov's[28-29] exchange algorithm to select an appropriate set of calibration samples for PCR models after scaling the scores. This algorithm makes use of the mathematical expression of the model, so the experimenter must know which PCs are relevant for regression in order to include them in the model. As this is a rather strict condition when the experimenter faces the regression problem for the first time, the PCs that are important for regression are

found by building a preliminary screening model that uses the minimum number of necessary samples. The important factors are selected from the absolute values of this model's coefficients. Subsequently, a definitive model containing the selected factors can be postulated, and the definitive sample subset that will be used to build it is selected. Finally, the model is validated by using the cross-validation technique.

Background and Theory

Notation. Matrices are represented by bold capital letters, column vectors by bold lowercase letters, and scalars by italic characters. The superscript T means transposed. The subindices in a matrix indicate its dimensions. Let $\mathbf{R}_{I\times J}$ be the column mean-centered matrix of instrumental response data for I samples and J sensors and $\mathbf{c}_{I\times 1}$ the vector of the kth analyte concentration in the I calibration samples.

Principal component regression formulation. In PCR, $\mathbf{R}_{I\times J}$ is decomposed according to

$$\mathbf{R}_{I\times J} = \mathbf{T}_{I\times P}\mathbf{P}_{J\times P}^{T} + \mathbf{E}_{I\times J} \qquad (2)$$

where the columns in $\mathbf{T}_{I\times P}$ are $P \leq min(I,J)$ uncorrelated underlying factors that might be important for prediction, $\mathbf{P}_{J\times P}$ is the loading matrix, and $\mathbf{E}_{I\times J}$ is a matrix of residuals. After determining which $Q \leq P$ principal components are important for regression, $\mathbf{c}_{I\times 1}$ is regressed versus $\mathbf{T}_{I\times Q}$ according to

$$\mathbf{c}_{I\times 1} = \mathbf{T}_{I\times(Q+1)}\beta_{(Q+1)\times 1} + \mathbf{f}_{I\times 1} \qquad (3)$$

where $\beta_{(Q+1)\times 1}$ is the vector containing the regression coefficients, a column vector of ones has been appended to $\mathbf{T}_{I\times Q}$ to account for a constant term, and the elements of $\mathbf{f}_{I\times 1}$ are the calibration error terms. An estimate of β can be found using the least-squares method.[2]

The Q factors that provide the best predictive model are usually found by cross-validation[2] of several regression models built with a different number of factors. Since the factors associated with the largest eigenvalues do not necessarily give the best

ISBN:978-84-691-1875-7/DL: T-337-2008

predictive model,[30,31] all possible combinations of factors should be checked. Optimization methods such as generalized simulated annealing (GSA)[30] enable the best subset to be found and make it unnecessary for all possible combinations to be checked. Although GSA has provided promising results, it uses the whole set of samples and may often be quite time-consuming. We propose a faster approach based on screening the factors which are important for modeling the concentration using instrumental response and analyte concentration values for a reduced number of samples.

Sample Selection and Model Building. The devised procedure consists of the following steps:

(1) $\mathbf{R}_{I\times J}$ is first decomposed according to eq 2. The P factors in $\mathbf{T}_{I\times P}$ that contain important information are selected on the basis of previously acquired experience about similar systems, criteria about significant eigenvalues (e.g., Malinowski's IND function[32]), or the commonly used cross-validation technique.[33,34] It is preferable to overdetermine P to ensure that all factors containing information are taken into account.

(2) The range-midrange transformation[35] is performed on $\mathbf{T}_{I\times P}=\{t_{ip}\}$ for every sample $i=1, ..., I$ and every factor $p=1, ..., P$ so as to obtain $\mathbf{S}_{I\times P}=\{s_{ip}\}$:

$$s_{ip}=(t_{ip}-C_p)/R_p \qquad (4)$$

where:

$$C_p= [\max (t_{ip}) + \min (t_{ip})] / 2$$
$$R_p= [\max(t_{ip}) - \min (t_{ip})] / 2$$

with $\max(t_{ip})$ and $\min(t_{ip})$ being the highest and lowest score values for the pth factor. This scaling technique makes the scaled scores s_{ip} for every factor span the range [−1,+1]. Although this scaling technique can be sensitive to outliers, it is essential since it enables the regression coefficients calculated in step 4 to be compared.

(3) A subset of N samples ($P+1\leq N< I$) is selected according to the procedure developed in the next section and their analyte concentrations ($\mathbf{c}_{N\times 1}$) are chemically determined.

(4) The coefficients in model eq 5 are determined:

$$\mathbf{c}_{N\times 1} = \mathbf{S}_{N\times(P+1)}\beta_{(P+1)\times 1} + \mathbf{f}_{N\times 1} \quad (5)$$

where $\mathbf{S}_{N\times(P+1)}$ is a submatrix of $\mathbf{S}_{I\times P}$, to which the column vector of ones has been appended to account for the constant term and whose rows are the N samples selected in step 3. The magnitude of each estimated coefficient in eq 5 indicates the ability of its corresponding factor to model the concentration values. As the regressor variables are equally scaled between -1 and +1, the absolute values of the coefficients can be compared among one another. A high value indicates that the corresponding factor explains a considerable variance of the concentration. The Q factors with largest coefficients are selected for the final regression model since they provide the best modelling ability. The factors with coefficient values near to zero only model random error and are discarded. Should irrelevant factors be included in the model, the quality of the prediction will decrease.

To classify a factor as non-significant, two alternative approaches are presented here: a t-test for every coefficient[36] and the leave-one-out cross-validation error for a series of models made with the selected samples and by adding the factors in decreasing order according to the absolute value of their coefficients. The factors that give the model with the minimum error of prediction are selected.

(5) The final PCR model is built with the Q important factors and a new selected subset of M samples according to eq 6:

$$\mathbf{c}_{M\times 1} = \mathbf{S}_{M\times(Q+1)}\beta_{(Q+1)\times 1} + \mathbf{f}_{M\times 1} \quad (6)$$

where $\mathbf{S}_{M\times(Q+1)}$ is a submatrix of $\mathbf{S}_{I\times P}$ in which each row corresponds to a selected sample, each column corresponds to a selected factor, and a column of ones has been appended. The analyte concentrations of these new selected samples ($\mathbf{c}_{M\times 1}$) are determined according to the chemical procedure used. The final model can be validated using the cross-validation technique. Alternatively, the samples that have not been used in the model building step can be used as a test set to validate the model, but this requires analyzing these latter samples. The selection of a subset of test samples for model validation could be the subject for future research.

ISBN:978-84-691-1875-7/DL: T-337-2008

Selection of the Optimal Set of Calibration Samples. The criterion for sample selection is that the N calibration samples should provide regression coefficients with the lowest variance of all of the subsets of N samples. The global precision of the estimated coefficients in eq 5 is given by the volume of their 100(1-α)% confidence region, which is proportional to $[\mathrm{Det}(\mathbf{S}_{N\times(P+1)}{}^{\mathrm{T}}\mathbf{S}_{N\times(P+1)})]^{-1/2}$, where Det denotes determinant.[24,27-29] Maximizing $\mathrm{Det}(\mathbf{S}_{N\times(P+1)}{}^{\mathrm{T}}\mathbf{S}_{N\times(P+1)})$ by selecting which N samples are included in $\mathbf{S}_{N\times(P+1)}$ minimizes the volume of the confidence region and helps to achieve minimum variance in the coefficients[25]. This criterion is known as the D-optimality criterion. Every definite model, implicit in the matrix $\mathbf{S}_{N\times(P+1)}$ (and later in $\mathbf{S}_{M\times(Q+1)}$) in which each column is related to a coefficient, has a different optimal calibration set.

Although a search for all combinations of N samples ensures that the subset that maximizes $\mathrm{Det}(\mathbf{S}_{N\times(P+1)}{}^{\mathrm{T}}\,\mathbf{S}_{N\times(P+1)})$ is found, the time required for such a search makes it impractical when the number of available samples, I, is high. Several algorithms make examining all possible combinations unnecessary.[25,27] We used Fedorov's exchange algorithm[27-29] found in the NEMROD 3.0 software package,[37] since it is specially designed to search for D-optimal subsets from a list of I candidate samples. D-optimal subsets are found for different N: from N equal to the number of coefficients in the PCR model (i.e, the minimum required to solve the linear system of equations) to a number $N<I$ defined according to the user's needs. Of the D-optimal subsets, the one that is selected for calibration is the one that contains the maximum information *per sample* to estimate the coefficients, which is given by $\log(\mathrm{Det}(\mathbf{M}_N))$ with $\mathbf{M}_N=(\mathbf{S}_{N\times(P+1)}{}^{\mathrm{T}}\,\mathbf{S}_{N\times(P+1)})/N$.[29]

A similar procedure is used to select the final subset of M samples in eq 6, and the matrix $\mathbf{S}_{M\times(Q+1)}$ is used. As the optimal calibration subset is model-dependent, the N samples used to build eq 5 are not necessarily the most suitable for eq 6. Since the aim is to use the minimum number of calibration samples, rather than discard the previously analyzed N samples, the selection algorithm will add (if necessary) M-N samples to the already analyzed N samples so that the D-optimal subset contains the necessary information for a good estimation of the coefficients of the final model.

It should be noticed that the selection algorithm uses only the instrumental responses. Hence, a selected calibration set used in the screening step, $\mathbf{S}_{N\times(P+1)}$ is

UNIVERSITAT ROVIRA I VIRGILI
EXPERIMENTAL DESIGN APPLIED TO THE SELECTION OF SAMPLES AND SENSORS IN MULTIVARIATE CALIBRATION
Joan Ferré Baldrich
ISBN:978-84-691-1875-7/DL: T-337-2008

optimal for all the analytes present in a sample, provided that the same model is postulated. This is no longer true for $\mathbf{S}_{M\times(Q+1)}$, which includes only the factors that are important for predicting each specific analyte.

Validation of the Methodology Developed. To assess the accuracy and precision of the PCR model built with a selected subset, the selected M samples are used as the calibration set while the remaining (I-M) samples are used as the test set. The accuracy is checked by a joint statistical test for the slope and the intercept of the linear regression between the measured versus predicted concentration values in the test set.[38] The multivariate model is regarded as being accurate if the theoretical values of intercept zero and slope unity are included within the ellipse which describes the joint confidence interval of the calculated straight line. The precision is measured by the root-mean-square error of prediction[2] (RMSEP) for the test set. In addition, the analyst can calculate the root-mean-square error of cross-validation (RMSECV) for the model built with the M selected samples, given in eq 7:

$$\text{RMSECV} = [\sum_{i=1}^{M} (c_{i_{\text{CV}}} - c_i)^2 / M]^{1/2} \qquad (7)$$

where $c_{i\text{CV}}$ is the predicted concentration for the ith sample in a model developed without the ith sample. The RMSEP of these models was also compared with the root mean square error of cross-validation (RMSECVT) for the model developed using the I available samples for calibration.

Experimental section

Samples and software. The following sample sets were used to check the validity of the method proposed for sample selection.

(1) Data set I consists of 24 UV–visible spectra. Rius et al.[39] determined Ca^{2+} by using the absorbance of their complexes with 2,2'-(1,8-dihydroxy-3,6-disulfonaphthylene-2,7-bisazo)-bis(benzenearsonic acid) (arsenazo III) in 24 natural water samples. Each spectrum consists of 101 variables, corresponding to the

absorbance values at wavelengths from 450nm to 650 nm. The actual content of calcium in water samples was determined by AAS and ranges between 3.1 and 40.6 ppm.

(2) Data set II consists of near-IR spectra of 46 corn samples at 19 fixed wavelengths reported by Puchwein.[9] The moisture of corn is the constituent of interest, and its content was determined by oven-drying. It ranges between 3.63% and 19.39%.

All computations were performed with home-made Matlab[40] subroutines. The instrumental response matrices were first mean-centered and Matlab SVD was used to evaluate the factors. The Matlab source codes are available on request.

Results and discussion

Data set I: Calcium in water samples. *Selection of the Important Factors To Be Used in the Screening Model.* P=10 factors were regarded as possibly containing important information according to PCA cross-validation. Hence, the minimum number of samples in the calibration set was 11 to enable the constant term and the regression coefficients associated to each factor in eq 5 to be estimated. Fedorov's exchange algorithm searched for the subset of N=11-24 samples that maximize $\mathrm{Det}(\mathbf{S}_{N\times 11}{}^{T}\mathbf{S}_{N\times 11})$. The plot $\log(\mathrm{Det}(\mathbf{M}_N))$ versus number of selected samples N (Figure 1) shows that the subset containing 15 samples ($\mathbf{S}_{15\times 11}$) has the maximum information per sample. Therefore, $\mathbf{S}_{15\times 11}$ and the calcium concentration for these 15 samples were used to build the screening regression model. The regression results are summarized in Table 1. Comparing the values in column 6 with the tabulated t-value for α =0.05 and 4 degrees of freedom (i.e, 15 samples minus 11 coefficients), $t_{0.05,4}$= 2.13, the factors numbered 7, 8 and 10 are discarded for regression. The same result is found by looking at the minimum cross-validation error.

Final regression model. The exchange algorithm was run again to maximize $\mathrm{Det}(\mathbf{S}_{M\times 8}{}^{T}\mathbf{S}_{M\times 8})$, where each column in $\mathbf{S}_{M\times 8}$ corresponds to the sample scores on the selected factors 1–6 and 9, and there is a column of ones to account for the constant

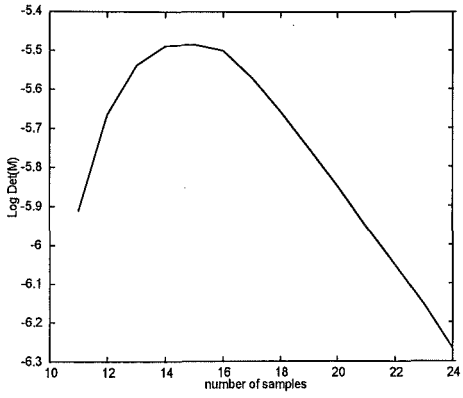


Figure 1. Number of selected samples (N) versus log(Det($\mathbf{M}_N$)). Calcium, 10 factor model.

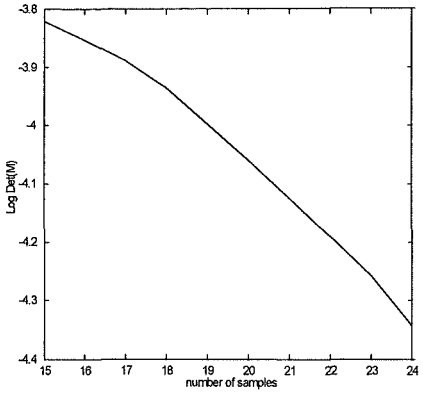


Figure 2. Number of selected samples (M) versus log(Det($\mathbf{M}_M$)). Calcium in water, seven selected factor model.

Table 1. Calcium in water samples. Regression Results for the Screening Model of 10 Factors and 15 Selected Samples [a]

PC	eigenvalue ($\times 10^{-3}$)	% variance	cumulative % variance	coeff	t-test	CV error
1	67.810	81.87	81.87	16.79	9.77	7.51
2	14.013	16.92	98.79	8.20	5.25	6.71
9	0.001	0.00	98.79	4.44	2.78	6.07
3	0.777	0.94	99.73	4.35	3.09	5.85
6	0.003	0.00	99.73	3.83	2.26	5.86
4	0.210	0.26	99.99	3.73	2.33	6.08
5	0.003	0.00	99.99	3.55	2.47	4.45
8	0.001	0.00	99.99	1.48	0.85	5.29
7	0.001	0.00	99.99	1.17	0.79	6.30
10	0.001	0.00	99.99	0.66	0.46	8.29

[a] Column 1 lists the PCs numbered according to decreasing eigenvalues. Columns 2-4 list the eigenvalue associated with each PC, the percentage of explained variance and the cumulative percentage of explained variance, respectively. Column 5 lists, in decreasing order of magnitude, the absolute value of the regression coefficient corresponding to each factor in column 1. Column 6 contains the calculated t-values (for comparison, tabulated t-value: $t_{0.05,4}$= 2.13). The last column shows the prediction error according to the leave-one-out cross-validation procedure for the selected samples using the factors cumulatively according to the ordered list in column 1.

term. To use the already-analyzed samples, the algorithm added samples to the 15 previously selected ones. The subset with maximum log(Det($\mathbf{M}_M$)) did correspond to the same 15 samples since any addition of new samples gave rise to a decrease in the log(Det($\mathbf{M}_M$)) (Figure 2). Thus, the 15 previously selected samples already contained enough information for the final model and no additional samples were needed, making any further analyte determination unnecessary.

Model validation. The results of the PCR model built with 15 samples and factors 1-6 and 9 are shown in Table 2. All the factors used in the final model are important, as shown by the *t*-test (tabulated $t_{0.05,7}$= 1.89) and the cross-validation error that reaches a minimum when the model is made with all the selected factors. The nine samples not used for model building were used as a test set. From the *F*-test for the joint confidence interval for the slope and intercept of the linear regression between the measured versus predicted concentration values, the model was considered to be accurate at an α = 0.425 level of significance. As far as precision was concerned, a very acceptable value of RMSEP = 1.86 was obtained. On the other hand, the value of RMSECV = 4.45 is higher than the RMSECVT = 3.00 obtained using the initial 24 samples. This could be explained if all 15 samples are important for building the PCR model. Deleting only one sample to calculate RMSECV using the leave-one-out procedure can considerably change the model, giving rise to a loss in precision.

Table 2. Calcium in Water Samples. Regression results for the Final Model with 15 Selected Samples and Factors 1-6 and 9[a]

PC	coeff	*t*-test	CV error
1	17.13	11.85	7.51
2	7.73	6.05	6.71
3	4.44	3.66	6.01
9	4.36	3.17	5.85
6	3.64	2.51	5.86
5	3.34	2.73	5.03
4	3.33	2.49	4.45

[a] The columns have the same meaning as columns 5-7 in Table 1. The first column is the number of each PC assigned in table 1, listed according to the absolute values of the coefficients for the final regression model.

UNIVERSITAT ROVIRA I VIRGILI
EXPERIMENTAL DESIGN APPLIED TO THE SELECTION OF SAMPLES AND SENSORS IN MULTIVARIATE CALIBRATION
Joan Ferré Baldrich
ISBN:978-84-691-1875-7/DL: T-337-2008

Validation of the methodology. The performance of the developed methodology was also compared with the following methods for PCR modeling by evaluating their prediction ability. (a) The commonly used PCR method, where the complete set of samples is used for calibration and the factors are introduced into the model in decreasing order of eigenvalues. The results are given in Table 3, column A. (b) All possible models made with all possible combinations of PCs, where the complete set of samples is used for calibration and the model with minimum RMSECVT is selected for each number of factors. The results are given in Table 3, column B. (c) The same procedure as (b) but using only the 15 selected samples. The RMSECV results are listed in column C in Table 3.

Table 3, column B shows that models built with a subset of selected factors provide smaller RMSECVT values than models in which the factors are introduced in decreasing order of their eigenvalues (Table 3, column A). In this case, factor 7 probably models information not related to the calcium concentration, thus

Table 3. RMSECV Values of Regression Models for Calcium with a Different Number of Factors[a]

A) Complete set [a]	B) Complete set [b]	C) 15 selected samples [c]
15.39 (zero factors)	15.39 (zero factors)	18.51 (zero factors)
6.78 1	6.78 1	7.51 1
5.61 1 2	5.61 1 2	6.71 1 2
4.99 1 2 3	4.99 1 2 3	6.01 1 2 3
4.75 1 2 3 4	4.61 1 2 3 5	5.70 1 2 3 5
4.25 1 2 3 4 5	4.19 1 2 3 5 9	5.33 1 2 3 4 5
3.96 1 2 3 4 5 6	3.55 1 2 3 5 6 9	5.03 1 2 3 5 6 9
4.24 1 2 3 4 5 6 7	3.00 1 2 3 4 5 6 9	4.45 1 2 3 4 5 6 9
4.81 1 2 3 4 5 6 7 8	3.02 1 2 3 4 5 6 9 10	4.83 1 2 3 4 5 6 9 10
3.47 1 2 3 4 5 6 7 8 9	3.11 1 2 3 4 5 6 7 9 10	6.23 1 2 3 4 5 6 7 9 10
3.57 1 2 3 4 5 6 7 8 9 10	3.57 1 2 3 4 5 6 7 9 10 8	8.29 1 2 3 4 5 6 7 8 9 10

[a] The complete set of samples and models made with increasing number of PCs introduced in order of decreasing eigenvalues (usual PCR regression). [b] The complete set of samples for models made with an increasing number of PCs selected according to their best performance. [c] All possible combinations of models for the 15 selected samples only; the best subset of factors is indicated .

scaled scores of few selected samples. The higher cross-validation errors obtained when using only the 15 selected samples (Table 3, column C) can be explained if the 15 selected samples are important for modelling, and when one is deleted, the model decreasing the prediction properties when it is included in the model. It should be noticed that factors 1–6 and 9, which give the model with the lowest RMSECVT, have also been selected as important for the methodology developed here using the is considerably altered. Another explanation could be that the remaining nine samples are not important for modelling, since they are similar to other selected samples. Thus, their deletion does not considerably change the model, which makes the prediction errors smaller.

To show that the proposed methodology, in the great majority of cases, can perform better than the random division of samples into a calibration and an evaluation set, 10,000 models were built by randomly dividing the 24 samples into a calibration set of 15 samples and a test set of nine samples and their RMSEP was evaluated. Only 78 of them gave lower RMSEP values than the model built with the selected samples (Figure 3). This shows that the random division of samples into training and test sets is not a guarantee for building a good quality model, specially if the randomly selected calibration set does not adequately span the experimental domain.

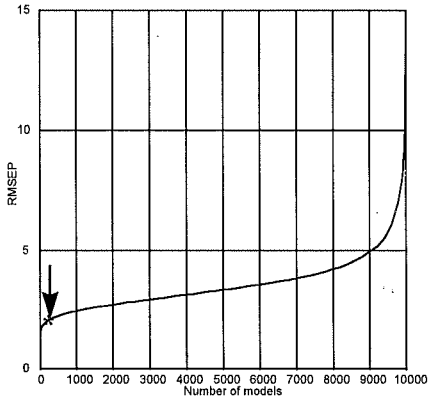


Figure 3. RMSEP of 10 000 models for calcium in water made by randomly dividing the samples into calibration and evaluation sets. The arrow points to the RMSEP of the model built using the methodology proposed.

Data set II. NIR spectra of 46 corn samples. *Selection of the Important Factors To Be Used in the Screening Model.* Initially evaluating the factors according to Malinowski's IND function resulted in selecting 15 factors which were considered to be important. The exchange algorithm searched for the subsets containing N=16-46 samples which maximized the Det($S_{N\times16}{}^T$ $S_{N\times16}$) function. The plot log(Det(M_N)).

UNIVERSITAT ROVIRA I VIRGILI
EXPERIMENTAL DESIGN APPLIED TO THE SELECTION OF SAMPLES AND SENSORS IN MULTIVARIATE CALIBRATION
Joan Ferré Baldrich
ISBN:978-84-691-1875-7/DL: T-337-2008

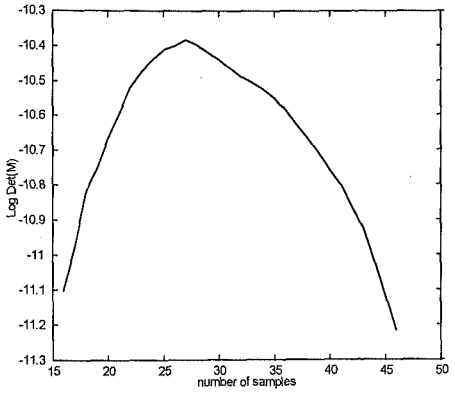


Figure 4. Number of selected samples (N) versus log (Det($\mathbf{M}_N$)). Corn samples, 15 factor model.

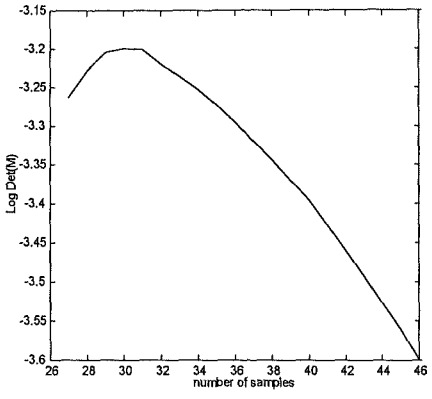


Figure 5. Number of selected samples (M) versus log (Det($\mathbf{M}_M$)). Corn samples, five selected factor model.

Table 4. Moisture of Corn. Regression Results for the Model Made with 15 Factors and the 27 Selected Samples [a]

PC	eigenvalue ($\times 10^{-3}$)	% variance	cumulative % variance	coeff	*t*-test	CV error
1	20.065	92.28	92.28	8.26	48.17	2.91
2	1.617	7.44	99.72	4.85	34.14	0.98
4	0.016	0.07	99.79	1.26	7.31	0.68
3	0.033	0.15	99.94	0.78	5.68	0.46
7	0.000	0.00	99.94	0.54	2.78	0.41
6	0.002	0.01	99.95	0.24	1.17	0.42
9	0.000	0.00	99.95	0.24	1.59	0.42
13	0.000	0.00	99.95	0.23	1.21	0.42
15	0.000	0.00	99.95	0.16	1.10	0.44
10	0.000	0.00	99.95	0.14	0.94	0.42
14	0.000	0.00	99.95	0.10	0.71	0.44
12	0.000	0.00	99.95	0.10	0.57	0.46
11	0.000	0.00	99.95	0.04	0.19	0.48
8	0.000	0.00	99.95	0.03	0.16	0.52
5	0.010	0.05	100.00	0.02	0.11	0.59

[a] The meanings of the columns are the same as for Table 1.

versus number of selected samples N (Figure 4) shows that the subset of 27 samples is the one having the most information per sample. From the experimentally determined moisture content in these selected samples[9], the screening regression model was built. The regression results are listed in Table 4. Only factors 1–4 and 7 are statistically significant according to the t-test (calculated $t_{0.05,11}$=1.80). This agrees with the global minimum error obtained by cross-validation.

Final regression model. Using factors 1- 4 and 7 and the exchange algorithm, the subset of 30 samples had maximum log(Det($\mathbf{M}_M$)) (Figure 5), so only three new samples had to be analyzed for their moisture content to be determined. The regression results for the final model are listed in Table 5. All the factors used are important for prediction as indicated by the t-test ($t_{0.05,24}$= 1.711) the results of which agree with the ones obtained using the minimum RMSECV when the model is built with all the selected factors.

Model validation. The model is accurate according to the F-test, with α=0.398. It should be pointed out that a very good precision value of RMSEP = 0.44 is obtained. Moreover, RMSECV = 0.39 is comparable to RMSECVT = 0.41, indicating that the selected samples cover the experimental domain quite well.

Table 5. Regression Results for the Final PCR Model Made with Factor Numbers 1, 2, 3, 4, and 7 and the 30 Selected Samples [a]

PC	coeff	t-test	CV error
1	8.32	57.27	2.90
2	4.82	41.98	0.92
4	1.31	8.80	0.66
3	0.78	6.85	0.44
7	0.38	2.74	0.39

[a]The meanings of the columns are the same as for Table 1.

Conclusions

A new procedure for selecting the best calibration sample subset for PCR has been developed. It is based on D-optimal design theory and makes use of only the easily obtainable multivariate instrumental responses. The cost in time and effort of the calibration process is substantially reduced because only the selected samples are submitted to chemical analysis by using a reference method. In addition, an approach for quickly selecting the factors with high modeling ability in PCR has been devised. The overall methodology has proven to provide accurate and precise results. The method is of a general nature, and it can be applied to data sets obtained using very different instrumental techniques.

Although only the selected samples are used for calibration, the eigenvector structure is computed for all the instrumental responses available. In this way, all possible causes of variability are taken into account, so ensuring that a representative subset which covers the experimental domain of the chemical constituents is found. However, it should be pointed out that, when few calibration samples are to be used, the quality parameters of the methodology depend very much on the reliability of the analytical results carried out using the reference method.

To sum up, the whole procedure tries to reach a compromise between the quality, in terms of accuracy and precision, that the experimenter demands of the model and the effort, in terms of time and cost, that he or she is ready to put in to build it.

Several research areas related to the present methodology can be developed in the future. The approach developed is being tested by our group to select the minimum number of appropriate samples for multivariate instrument standardization. In addition, since prediction is the main function of the multivariate model, the quality of the predictions furnished by the model, measured by var(c), (G-optimal designs) instead of by minimizing the error of the coefficient estimates, could be a more suitable criterion for selecting the samples to build the final PCR model. This new approach would overcome the distressing problem of not taking into account the errors in the regressor variables to build the multivariate model. However, this would require the development of new sample selection algorithms. Furthermore, a new method to simultaneously designate the calibration and validation samples could be

developed. This methodology would select, from the overall data set, the sample subset that would provide the best estimation of the model coefficients and the samples that give the best indication of the quality of the model.

Acknowledgment

The authors express their gratitude to A. Rius and G. Puchwein for providing the UV-vis and near-IR spectral data, respectively. J. F. thanks the Comissionat per a Universitats i Recerca of Generalitat de Catalunya, for providing a doctoral fellowship (FI/94-7001). Financial support from the Spanish Ministry of Education and Science (DGICyT project BP93-0366) is gratefully acknowledged.

Supporting information available

Mathematical expressions of Fedorov's exchange algorithm. (2 pages). Ordering information is given on any current masthead page.

SUPPLEMENTARY MATERIAL

Fedorov's exchange algorithm[28-29] for selecting D-optimal matrices has three main steps:

1) Initiation: the initial matrix $\mathbf{S}_{N\times(P+1)}^{(0)}$ is built with N randomly selected samples with the condition that the matrix $\mathbf{S}^{(0)T}\mathbf{S}^{(0)}$ must be non-singular. Alternatively, the experimenter can choose the samples it is made up of. In the case of $\mathbf{S}_{M\times(Q+1)}^{(0)}$, only M-N samples are randomly added to the already selected N samples. For simplicity the dimensions are not indicated: $\mathbf{S}^{(0)}$

2) Iteration number *j*:

1. The pair of samples (s^o, s^i) that gives the maximum increase in $Det(S^{(j+1)T}S^{(j+1)})$ is selected. s^o ("out")=$(1, s^o_1, s^o_2,..., s^o_p)^T$ is the sample that leaves the matrix $S^{(j)}$, and s^i("in")=$(1, s_1^i, s_2^i,..., s_p^i)^T$ is one of the candidate samples that enters the matrix.

2. s^o is replaced by s^i in the matrix $S^{(j)}$ so a new matrix $S^{(j+1)}$ is obtained .

3) Stop criterion: the algorithm stops when the increase in $Det(S^{(j+1)T}S^{(j+1)})$ is zero or less than a critical value.

Mathematical expressions:

When a sample s^i enters the initial matrix of experiments $S^{(j)}$ and a sample s^o leaves at the same time, it can be shown [41] that

$$Det(S^{(j+1)T}S^{(j+1)}) = Det\,(S^{(j)T}S^{(j)})(1 + \Delta(s^o,s^i))$$

where: $\Delta(s^o,s^i) = d(s^i) - d(s^o) - d(s^i)\times d(s^o) + [d(s^o,s^i)]^2$ with :

$d(s^o,s^i) = s^{oT}(S^{(j)T}S^{(j)})^{-1}s^i = s^{iT}(S^{(j)T}S^{(j)})^{-1}s^o$

$d(s^i) = d(s^i,s^i)$

$d(s^o) = d(s^o,s^o)$

The maximum increase in $Det(S^{(j)T}S^{(j)})$ is achieved by the pair of samples with largest $\Delta(s^o,s^i)>0$, and this is what the algorithm looks for. The exchanges are faster when only considering the samples with $d(s^i) - d(s^o)>0$ which is a necessary condition for $\Delta(s^o,s^i)>0$. A disadvantage of this algorithm is that the final solution depends on the choice of the initial matrix since it can reach local maxima instead of global maxima. This problem can be solved by repeating the procedure several times with different initial matrices. In our experience, five to ten restarts are enough to find the best solution several times.

Literature cited

1. Beebe, K.R.; Kowalski B.R. *Anal. Chem.* **1987**, *59*, 1007A-1017A.
2. Martens H.; Naes T. *Multivariate Calibration*, Wiley: New York, 1987
3. Kowalski, BR.; Seasholtz M.B. *J. Chemom.* **1991**, *5*, 129-145.
4. Geladi P.; Kowalski B.R. *Anal. Chim. Acta* **1986**, *185*, 1-17.
5. Lorber A; Kowalski B.R. *J. Chemom.* **1988**, *2*, 67-79.
6. Gemperline P.J. *J. Chemom.* **1989**, *3*, 549-568.
7. Aastveit A.H.; Marum P., *Appl. Spectrosc.* **1993**, *47*, 463-469.
8. Naes, T. *J. Chemom.* **1987**, *1*, 121-134.
9. Puchwein G. *Anal. Chem.* **1988**, *60*, 569-573.
10. Schostack, K. J.; Malinowski, E. R. *Chem. Intel. Lab. Syst.* **1989**, 6, 21-29.
11. Skagerberg B.; Bonelli D.; Clementi S.; Cruciani G.; Ebert C. *Quant. Struct.-Act. Relat.* **1989**, *8*, 32-38.
12. Norinder U.; Högberg T. *Acta Chemica Scand.* **1992**, 24, 363-366.
13. Hellberg S.; Sjöström M.; Skagerberg B.; Wold S. *J. Med. Chem.* **1987**, *30*, 1126-113536.
14. Wold S.; Sjöström M.; Carlson R.; Torbjörn L.; Hellberg S.; SkagerbergB.; Wikström C.; Öhman J. *Anal. Chim. Acta* **1986** *191* 17-32.
15. Carlson R. *Design and optimization in organic synthesis.* Elsevier: The Netherlands 1992
16. Zemroch, Peter J. *Technometrics* **1986**, *28*, 39-49.
17. Honigs D.E.; Hietfje G.M.; Mark H.L. ; Hirschfeld T.B. *Anal. Chem.* **1985** 57, 2299-2303.
18. Marengo E. ; Todeschini R. *Chem. Intel. Lab. Syst.* **1992**, *16*, 37-44.
19. Kalivas, J. *J. Chemom.* **1991**, *5*, 37-48.
20. Hruscha W.; Norris, K. *Appl. Spectrosc.* **1982**, *36*, 261-265.
21. Isaksoon T.; Naes T. *Appl. Spectrosc.* **1990**, *44*, 1152-1158.
22. Naes T.; Isaksoon T. *NIR news* **1994** ,*5*, 16-17
23. Box, G.E.P. ; Hunter, W.G.; Hunter, J.S. *Statistics for Experimenters* Wiley: N.Y. 1978
24. Nishii, Ryuei. *Discrete Mathematics* **1993**, *116*, 209-225.
25. Steinberg D.V.; Hunter W.G. *Technometrics* **1984**, *26*, 71-130
26. Kennard, R.W. ; Stone, L.A. *Technometrics* **1969** ,*11*, 137- 148
27. Atkinson, A.C. ; Donev.A.N. *Optimum Experimental Designs.* Oxford Statistical

Science Publications: Oxford 1992

28. Fedorov, V.V. *Theory of optimal experiments.*(translated and edited by W.J. Studden and E.M. Klimko) Academic Press: New York, 1972
29. Mathieu D. *Contribution de la Méthodologie de la Recherche Experimentale à l'étude des relations Structure-Activité.* Thèse Sciences. Marseille, 1981
30. Sutter, J.M.; Kalivas J.H. ; Lang P.M. *J. Chemom.* **1992**, *6*, 217-225.
31. Sun, J. *J. Chemom* **1995**, *9*, 21-29.
32. Malinowski, E. R. *Factor Analysis in Chemistry*, 2nd de; John Wiley & Sons Inc: New York 1991
33. Wold S., *Technometrics* **1978**, *20*, 397-405.
34. Eastment H.T.; Krzanowski W.J. *Technometrics* **1982**, 24, 73-77
35. Mandel, J. *J. Res. of the National Bureau of Standards* **1985**, *90*, 465-476.
36. Khuri, A.I.; Cornell, J.A *Response Surfaces. Designs and Analyses*; Marcel Dekker: New York, 1987; pp 36-38.
37. Mathieu, D.; Phan-Tan-Luu, R. NEMROD ver. 3.0. L.P.R.A.I. - Université d'Aix-Marseille 1995
38. Mandel, J.; Linning, F. J. *Anal. Chem.* **1957**, *29*, 743-49
39. Rius A.; Callao M.P.; Rius F.X. *Anal. Chim. Acta*, in press.
40. Matlab. The Mathworks, South Natick, MA, 1994.

3.3 Determination of ethylene content in poly(propylene-ethylene) copolymers using near-infrared spectra (NIR) and multivariate calibration

(submitted)

Villagrasa C., Ferré J., Larrechi M.S., Rius F.X., García C* [1].

Department of Chemistry, Universitat Rovira i Virgili. Pl. Tarraco, 1. 43005 Tarragona.
[1] *Transformadora de Polipropileno TDP. PO Box 1175. Tarragona. Spain*

A new method for determining the ethylene content in poly(propylene-ethylene) copolymers using near-infrared spectra (NIR) in the 1666-1767 nm range and multivariate calibration is discussed. Three multivariate calibration methods were studied; principal component regression (PCR), principal component regression selecting the factors according to its predictive ability (PCRSF), and partial least-squares regression (PLS). PLS was found to have the best precision. The absence of bias in this model was assessed by performing the joint statistical test of the slope and the ordinate in the regression of $c_{pred\text{-}cv}$ versus c_{known} for the calibration samples taking, into account the errors in both axes.

1.Introduction

Heterophasic poly(propylene-ethylene) copolymers are widely used for industrial purposes because of their extreme toughness. This property is usually measured by the impact strength,[1,2] which is related, among other variables, to the ethylene content. Determining the percentage of ethylene is an important analysis in the quality control process of these plastics.

In industry, ethylene concentration is usually determined by using the infrared (IR) spectra of the pressed films of these copolymers[3,4]. The area of the bands between 750 and 690 cm^{-1}, which corresponds to the absorption of the methylenic sequences, is used to calculate the ethylene concentration. This area, however, is previously divided by the area of the bands between 4361 and 3950 cm^{-1} to correct for the thickness of the film. Univariate linear calibration of these data is then carried out using internal standards as a reference. The variability coefficient of this analysis is between 5-10%. It is sometimes difficult to accurately determine ethylene concentration because of the presence of talc which produces a band in the IR spectra that overlaps the ethylene band (Figure 1), thus making univariate calibration unsuitable in this case.

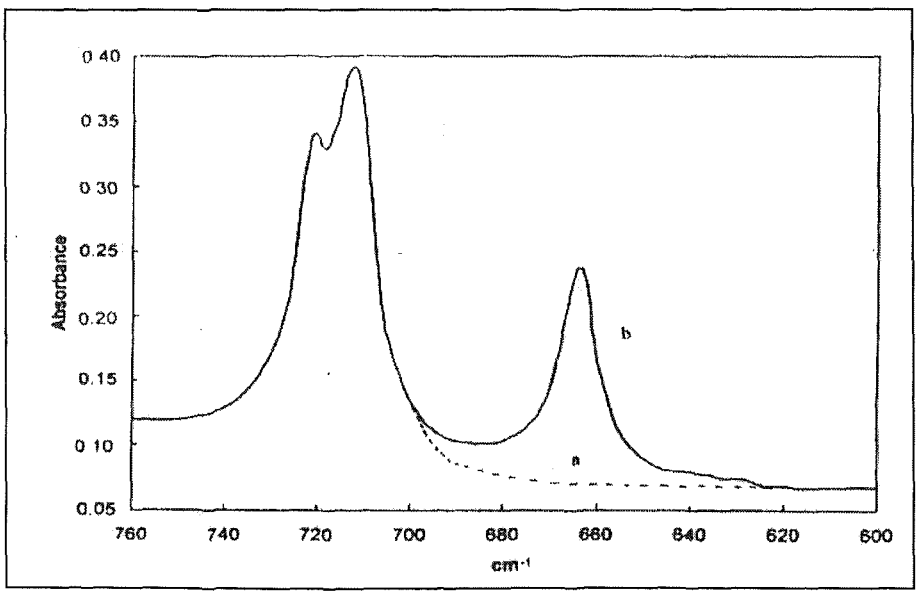


Figure 1. A typical IR absorption spectrum of poly(propylene-ethylene) copolymer a)without talc b)with talc

In this study, an alternative method for determining ethylene in poly(propylene-ethylene) copolymers is described. It uses the near-infrared (NIR) spectra of the samples between 1666 and 1767 nm which correspond to the first overtone of CH stretching bands[5]. Although the use of NIR spectroscopy to determine structural properties in different types of plastics such as polyethylene and polyurethane has been reported[6,7] to our knowledge, neither quantitative nor qualitative application of poly(propylene-ethylene) copolymers have been described. Multivariate calibration must be applied to quantify the ethylene concentration since there are no selective wavelengths. Three multivariate calibration methods were studied and compared: principal component regression (PCR), PCR selecting factors for their prediction ability (PCRSF) and partial least-squares regression (PLS).

2. Theoretical background

Notation. Matrices are represented by bold capital letters, column vectors by bold lower-case letters and scalars by italic characters. The superscript ^{T} means transposed. The subindices in a matrix indicate its dimensions. Let $\mathbf{R}_{I\times J}$ be the column mean-centered matrix of instrumental response data for I samples and J sensors and $\mathbf{c}_{I\times 1}$ the vector of the ethylene concentration in the I calibration samples.

Principal component regression. In PCR, $\mathbf{R}_{I\times J}$ is decomposed according to:

$$\mathbf{R}_{I\times J} = \mathbf{T}_{I\times P}\mathbf{P}_{J\times P}^{T} + \mathbf{E}_{I\times J} \qquad (1)$$

where the columns in $\mathbf{T}_{I\times P}$ are $P \leq min(I,J)$ uncorrelated underlying factors, $\mathbf{P}_{J\times P}$ is the loading matrix and $\mathbf{E}_{I\times J}$ is a matrix of residuals. After determining which Q principal components are important for regression, $\mathbf{c}_{I\times 1}$ is regressed versus $\mathbf{T}_{I\times Q}$ according to:

$$\mathbf{c}_{I\times 1} = \mathbf{T}_{I\times(Q+1)}\,\beta_{(Q+1)\times 1} + \mathbf{f}_{I\times 1} \qquad (2)$$

where $\beta_{(Q+1)\times 1}$ is the vector of regression coefficients, a column vector of 1's has been appended to $\mathbf{T}_{I\times Q}$ to account for a constant term and the elements of $\mathbf{f}_{I\times 1}$ are the calibration error terms. An estimate of β can be found using the least-squares method[8]. The Q factors that provide the best predictive model can be found by cross-validation[9] of several regression models built with a different number of factors. These factors are put into the model depending on the value of their corresponding eigenvalue.

Principal component regression with selection of factors (PCRSF). It has been shown that the factors with the largest eigenvalues do not necessarily give the best predictive PCR model[10]. An alternative PCR model can be built by considering only the most-predictive principal components, which can be selected with a screening method. The procedure used, which has the advantage of being very fast, is as follows:

(1) $\mathbf{R}_{I\times J}$ is decomposed according to equation (1). The number of factors in $\mathbf{T}_{I\times P}$ is selected on the basis of the optimal number of factors used to build the 'usual' PCR model and it is overdetermined to ensure that all factors containing information are taken into account.

(2) The Range-Midrange Transformation[11] is performed on $\mathbf{T}_{I\times P}=\{t_{ip}\}$ for every sample $i=1,\ldots,I$ and every factor $p=1,\ldots,P$ so as to obtain $\mathbf{S}_{I\times P}=\{s_{ip}\}$:

$$s_{ip}=(t_{ip}-C_p)/R_p \qquad (3)$$

where:

$$C_p = [\max(t_{ip}) + \min(t_{ip})] / 2$$
$$R_p = [\max(t_{ip}) - \min(t_{ip})] / 2$$

with $\max(t_{ip})$ and $\min(t_{ip})$ being the highest and lowest score values for the pth factor. This scaling technique makes the scaled scores s_{ip} for every factor span the range [–1,+1].

(3) The coefficients in model equation (3) are determined:

$$\mathbf{c}_{I\times 1} = \mathbf{S}_{I\times(P+1)}\,\beta_{(P+1)\times 1} + \mathbf{f}_{I\times 1} \qquad (4)$$

where $\mathbf{S}_{I\times(P+1)}$ is $\mathbf{S}_{I\times P}$ to which the column vector of 1's has been appended to account for the constant term. The magnitude of each estimated coefficient in equation (4) indicates the ability of its corresponding factor to model the concentration values. As the regressor variables are equally scaled between -1 and +1, the absolute values of the coefficients can be compared among one another. A high value indicates that the corresponding factor explains a considerable variance of the concentration. The factors are selected in the order indicated by the absolute value of the coefficients.

(4) The cross-validation technique is used to find out the optimal number of factors, by building the model including the factors in the order found. The factors that give the model with the minimum error of prediction are selected.

Partial least-squares (PLS). The PLS model was also applied[8]. Different from PCR, whose decomposition is based entirely on spectral variations, PLS takes the concentrations into account when decomposing the spectral matrix. Thus, PLS sacrifices some of the fit of the spectral data in order to achieve better correlation with the predicted concentrations.

Model Validation. The prediction ability was assessed by the root mean squared error of cross-validation (RMSECV). The absence of bias in the method was assessed by performing a joint statistical test of the slope and the intercept of the regression line, calculated by regressing the predicted concentration using cross-validation, $c_{i,pred-cv}$, versus the known concentration ($c_{i,known}$), taking into account the errors in both axes[12]. The variance of known concentration and the variance of predicted concentration were required in this test. The variance of the predicted concentration was given by the Unscrambler program[13] and the variance of the known concentrations was considered constant and equal to 10% of the $c_{i,known}$, since this is the error accepted in this kind of analysis.

3. Experimental section

Samples. 86 samples of pressed films with a thickness of 150 μm were prepared in a thermostatized press at 210⁰ C. Their ethylene content had previously been determined by IR and univariate calibration[3] and had been between 4 and 12%. The variability of spectral measurements due to the thickness and the non-homogeneity of the pressed films was taken into account by using several pressed films of the same ethylene concentration.

Spectroscopy. A Galaxy 3040 (Unicam) FTIR spectrophotometer with a tungsten lamp and PbSe detector connected to a PC 286 was used. The software FIRST version enhanced V1.52 selected the data acquisition conditions and provides an ASCII file. The spectral data consisted of 178 absorbance values between 1666 and 1767 nm. Five typical spectra are shown in Figure 2.

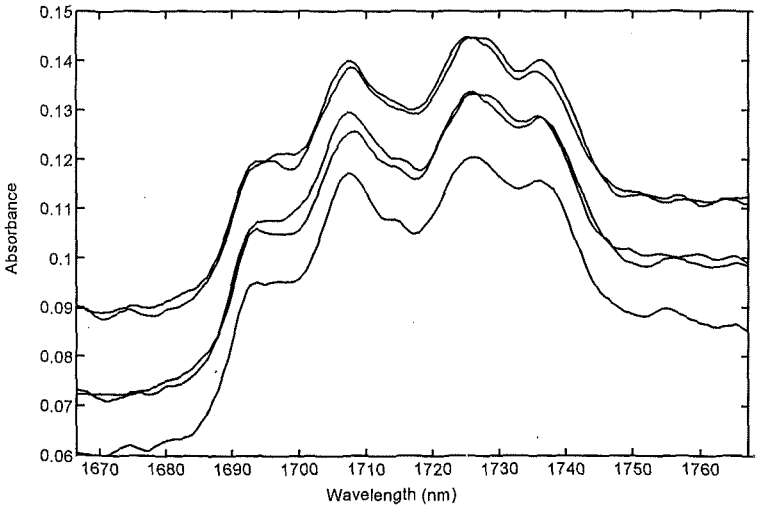


Figure 2. Examples of five digitalized near-infrared (NIR) absorption spectrum of poly(propylene-ethylene) copolymer.

Statistical methodology and software. The UNSCRAMBLER[13] program was used for PCR and PLS, and MATLAB[14] home-made subroutines were used for PCR selecting factors and the joint statistical test. Mathematical treatments of the data such as base line drift correction and first and second derivatives were explored to determine the linear-regression equation with the best fit.

4. Results and discussion

The application of the analytical procedure to the 86 analyzed samples provided the data matrix $\mathbf{R}_{86x178}$ where each row represents the NIR absorption spectrum of a sample. The columns in **R** were mean centered before calibration and were not subjected to a standardization procedure since, as they are spectroscopic data, they are expressed in the same units and there are no variables that incorporate special conditions of noise.

The initial representation of the sample scores in the space of the first two principal components, Figure 3, (99.9% of the x-variance explained) shows no tendencies suggesting different groups of samples. Nor does it allow the presence of any outliers to be observed at this point.

PCR, PCR selecting factors (PCRSF) and PLS were used to establish the relationship between the ethylene content and spectral data. The number of significant factors in the calibration model was assessed by the leave-one-out cross-validation procedure. Table 1 shows the root mean squared error of cross-validation (RMSECV) and the percentage of variance explained for the y variable for each factor in the three models.

The PLS model with four factors and a prediction error RMSECV= 0.719 (86% of total y variance explained) had the best prediction ability although it is only slightly higher than the prediction ability of PCR and PCRSF. PLS seems to predict better than PCR when there are random linear base lines or independently varying major spectral components which overlap with the spectral features of the analysis[15].

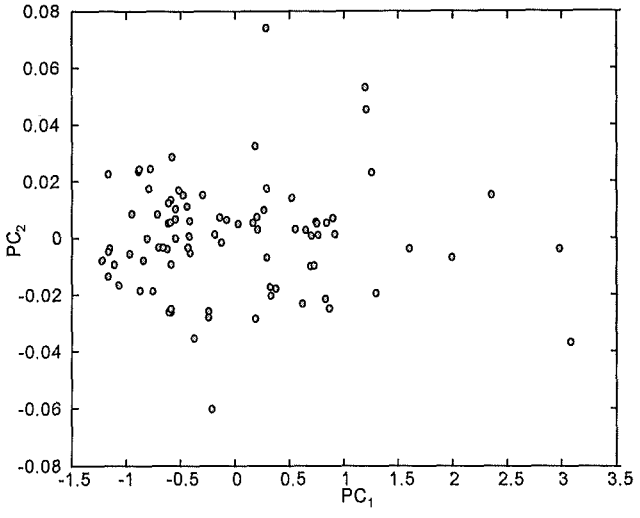


Figure 3. Score plot for the two principal components of PCR model (explained variance 99.9%).

Table 1. Root mean square error of cross-validation and percentage of Y variance in PCR, PCRFS and PLS models.

PCR			PCRSF			PLS		
Factor	% Var-*y*	RMSECV	Factor	% Var-*y*	RMSECV	Factor	% Var-*y*	RMSECV
0	0.0	1.77	0	0.0	1.77	0	0.0	1.77
1	24.5	1.539	4	51.9	1.258	1	24.5	1.390
2	24.7	1.557	1	76.4	0.913	2	41.2	0.843
3	25.8	1.554	2	76.6	0.892	3	81.7	0.770
4	77.7	0.864	7	78.3	0.868	4	86.0	0.719
5	77.5	0.872	3	79.4	0.836	5	92.0	0.719
6	79.1	0.843	6	81.0	0.807	6	94.8	0.730
7	80.8	0.813	5	81.2	0.813	7	96.7	0.730

In order to improve the prediction ability of the PLS model, the variance in base line offset by offset subtraction and derivative transformation[16] was removed. Offset subtraction was performed by subtracting the absorbance at 1680 nm from the absorbance values at all other wavelengths. Both offset substraction and derivative transformation gave similar or even worse correlations with the ethylene content for the PLS model, so they were no longer considered.

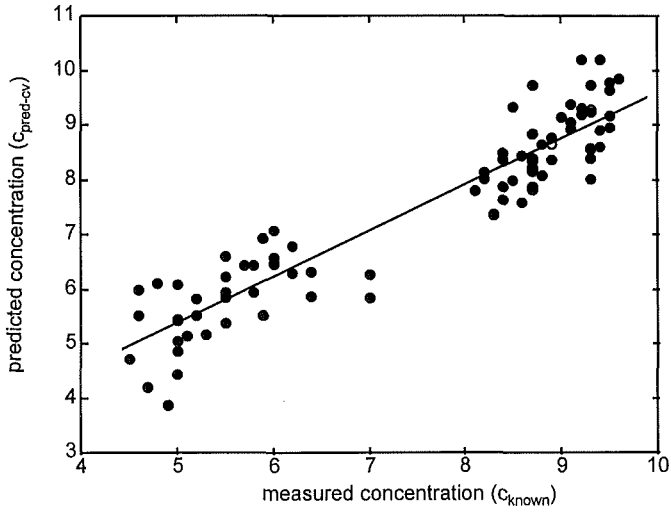


Figure 4. Measured concentration ($c_{i,known}$) of ethylene versus predicted concentration ($c_{i,pred-cv}$) from NIR spectra using PLS model.

As far as the PCR models are concerned, the optimal number of factors is 4 in the PCR model (RMSCV=0.864) and 6 for the PCRSF model, which has a slightly smaller prediction error (RMSECV=0.807). Although the prediction error in the two models is similar when a larger number of factors is used, some important considerations must be taken into account: the first three factors in PCRSF have a greater prediction ability than PCR, which agrees with the consideration that the factors are selected according to their capacity for explaining the variance of the response y. Thus, it can be thought that three factors can lead to the most robust model. Moreover, factor number 5 does not provide information correlated with the ethylene content, so this factor was not included in the PCRSF model.

The absence of bias was assessed only for the PLS model because it had the best prediction ability. Figure 4 shows the plot of predicted concentrations ($c_{pred-cv}$) versus measured concentrations (c_{known}) for the PLS model. Although two groups of samples are observed, the objective was to develop a unique model for predicting any sample for its easy applicability in an industrial process. The joint confidence interval test for the slope and the offset indicated that the methodology gives accurate results for the PLS at a 99% level of significance. Figure 5 shows the confidence ellipse the centre of which corresponds to the intercept=0.56 and slope=0.93. It can be seen that the theoretical point (0,1) is included in the confidence interval of the ellipse

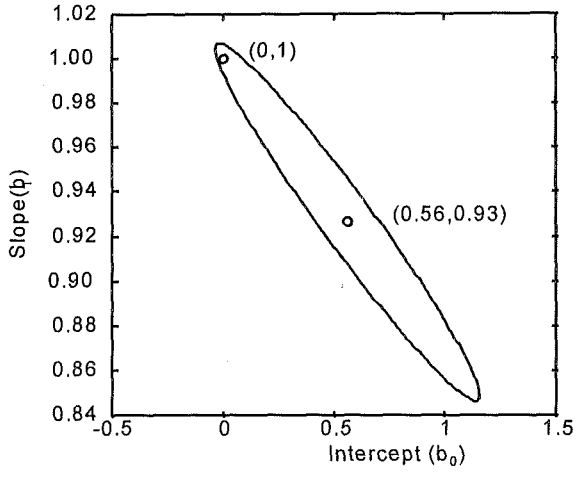


Figure 5. Confidence ellipse for the joint test of the slope and the intercept of the straight line obtained regressing $c_{i,\text{known}}$ on $c_{i,\text{pred-cv}}$ taking into account the uncertainties in both axes (level of significance=99%).

5. Conclusions

A new method for determining the ethylene content in poly(propylene-ethylene) copolymers has been developed. The resulting model gives accurate and precise predictions of the ethylene concentration, and the best prediction ability is obtained with the PLS model. These results are within the limits permitted in this kind of analysis. The new method is a very interesting alternative to the usual method, because of its swiftness and precision. NIR spectroscopy makes a multicomponent analysis possible, so it to be expected that other important parameters in quality control can also be determined by the same technique.

Acknowledgements

Financial support from the Spanish Ministry of Education an Science (DGICyT) project PB93-0.366, is gratefully acknowledged. J. F. thanks the Comissionat per a Universitats i Recerca of Generalitat de Catalunya, for providing a doctoral fellowship (FI/94-7001).

References

1. Lacosta Berna, J. M., *Revista de Plasticos Modernos*, 1993, **448**, 381-389
2. Instrumented Impact testing of Plastics and Composites ASTM STP 936.1987
3. Painter, P. C., Coleman,M. M., Koeing, J. L., *The theory of Vibrational Spectroscopy and Its Application to Polymeric Materials*, Wiley, New York,1964
4. Bower, D. I. , Maddams, W. F., *The Vibrational Spectroscopy of Polymers*, Cambridge University Press, Cambridge, UK, 1989
5. Miller, C. E., *Applied Spectroscopy Reviews*, 1991, **26(4)**, 277-339.
6. Hildrum, K. I., Isaksson, T., Naes T., Tandberg, A., *Near Infra-Red Spectroscopy*, Chichester,1992
7. Miller, C. E., Eichinger, B. E., *J. Appl. Poly. Sci.*, 1991, **42**, 2169-2190.
8. Martens H., Naes T., *Multivariate Calibration*, Wiley, New York, 1987
9. Wold S., *Technometrics* 1978, **20**, 397-405.
10. Ferré, J.; Rius F.X. *Anal. Chem.* 1996, **68**,1565-1571.
11. Mandel, J. *J. Res. of the National Bureau of Standards* 1985, **90**, 465-476.
12. Riu J., Rius F.X. *Anal. Chem.*, 1996, **68**, 1851-1857.
13. UNSCRAMBLER II, version 4.0. CAMO A/S. Norway.
14. MATLAB. The Mathwoks, South Natick, MA, USA.
15. E. V. Thomas, D. M. Haaland, *Anal. Chem.*, 1990, **62**, 1091-1099.
16. Kelly,J.J., Barlow, C. H., Jinguji,T. M., Callis, J. B., *Anal. Chem.*, 1989, **61**, 313-320.

3.4 Constructing D-optimal designs from a list of candidate samples

Trends Anal. Chem. 16 (1997) 70-73

Joan Ferré, F.Xavier Rius

Departament de Química. Universitat Rovira i Virgili.
Pl. Imperial Tarraco, 1, 43005-Tarragona. SPAIN

The main characteristics of a Matlab program to select D-optimal subsets of calibration samples for multiple linear regression are described. The performance of Fedorov's exchange algorithm to select samples is compared with the Kennard-Stone algorithm and the random selection of samples into training and test sets.

1. Introduction

Multiple linear regression is based on the regression of J independent variables measured in N calibration samples against a vector of observed values described by Eq. (1):

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{e} \quad (1)$$

where $\mathbf{y}$ is the $N\times1$ vector of the dependent variable, $\mathbf{X}$ is the $N\times P$ model matrix where each row corresponds to a calibration sample and each column is a predictor variable corresponding to a coefficient in the model (usually $J \leq P$ since higher order and/or interaction terms may be included in the model), β are the coefficients to be estimated and $\mathbf{e}$ is an $N\times1$ vector of error terms. The least-squares estimate of β is given by Eq. (2):

$$\mathbf{b} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \quad (2)$$

where ^{T} means transposed. It is desirable to use the smallest set of N calibration samples that reliably estimates the coefficients in order to reduce the cost of building the model. Although classical experimental designs [1] might be used, they sometimes involve a large number of calibration samples and/or may not be described for models other than the usual first or second order ones. What is more, these designs cannot be used if the experimental domain is irregularly shaped or if only a predefined list of samples is available as in the spectroscopic calibration of natural samples or quantitative structure-activity relationship (QSAR) studies. In such cases, the calibration samples are usually selected from a list of possible candidates using the Kennard-Stone algorithm [2], artificial neural networks [3] or random selection, among other methods.

In this paper we present a program to select calibration samples from a list of I candidates based on the D-optimality criterion. Of all the sets of $N \leq I$ samples that can constitute matrix $\mathbf{X}$, the N samples that maximise $\mathrm{Det}(\mathbf{X}^T\mathbf{X})$, where Det denotes determinant, are selected avoiding the need to search for all possible N sample combinations [4-6]. The selected samples minimise the volume of the $100(1-\alpha)\%$ confidence region of the coefficients, thus producing reliable estimations [7]. The selection is made without considering the y responses, which are only measured for

ISBN:978-84-691-1875-7/DL: T-337-2008

the selected samples. The performance of this algorithm is compared with the Kennard-Stone algorithm and the random selection of calibration samples in the prediction of the octane index in fuels from their proton NMR spectra.

2. The capabilities of the program

2.1 The algorithm

The algorithm [4-6] starts with a group of samples randomly chosen from the I candidate samples, iteratively changes one of the selected samples for a sample from the list of candidates that leads to a maximum increase in $\mathrm{Det}(\mathbf{X}^T\mathbf{X})$ and stops when a change can no longer increase $\mathrm{Det}(\mathbf{X}^T\mathbf{X})$. To avoid finding local optimal subsets, the algorithm is re-started several times with a different random set of samples. In our experience, five to ten times is enough to find the best solution. Scheme 1 shows the program listing corresponding to the selection subroutine in Matlab code [8].

As an additional option in the algorithm, the user can select samples that have to be included in all the selected subsets. This is useful when the y values of certain samples have already been measured.

To help choose the most suitable subset for calibration, the program also calculates, for each selected subset $\mathbf{X}$ of N samples, the values of $\log(\mathrm{Det}(\mathbf{X}^T\mathbf{X}))$, $\log(\mathrm{Det}(\mathbf{M}))$, where $\mathbf{M}=\mathbf{X}^T\mathbf{X}/N$, $\mathrm{Trace}(\mathbf{X}^T\mathbf{X})^{-1}$ and the variance inflation factors (VIF). These values are related to the expected quality of the predictions which have been made by the model built with the selected samples[4]. The subset with maximum $\log(\mathrm{Det}(\mathbf{M}))$ (which measures the information content *per sample*) is usually used for calibration provided that the other quality measures (VIF, $\mathrm{Trace}(\mathbf{X}^T\mathbf{X})^{-1}$, etc) are acceptable.

2.2 The computer program

All algorithms were implemented in home-made Matlab subroutines which can easily be edited by the user to adapt the program to his/her needs. The user

```
[n,p]=size(XC);
for ne=ni:nf
    for rep=1:nrepet
        a=randperm(n);a=(sort(a(:,1:ne)))'; X=XC(a,:);canvi=1;niter=0;
        while canvi==1 & niter<2*n
            niter=niter+1;canvi=0;deltabo=0;fora=[1:n]';fora(a,:)=[];
            A=XC*inv(X'*X)*XC';
            d=diag(A);Dfora=d;Dfora(a,:)=[];
            for s=1:ne
                dsurt=d(a(s));delta=Dfora-dsurt;
                for i=1: (n-ne)
                    if delta(i)>0
                        y=delta(i)-dsurt*Dfora(i)+A(a(s), fora(i)) ^2;
                        if y>deltabo;deltabo=y;surt=s;entra=i; canvi=1;end
                    end
                end
            end
            if canvi==1;a(surt)=fora(entra);a=sort(a);X=XC(a,:);end
        end
        disp(a');
    end
end
```

Scheme 1. Main body of the selection algorithm, which selects the D-optimal subset of samples from the list of candidate samples (**XC** matrix)

communicates with the algorithms through graphical menus such as editing dialogue boxes or multi-option menus. Since Matlab can run under Windows, the results displayed can be easily transferred to any word processor by doing copy/paste .

3. Comparative study

3.1 Samples and software

Proton NMR spectra were recorded for 75 gasoline samples and the areas of the three zones corresponding to the aromatic proton signals (chemical shift, δ=6.6-8.0), olefinic protons (δ=4.5-6.0) and CH_3 alpha aromatic and paraffinic protons (δ=3.0-0.6) were integrated. Samples were prepared by diluting 50% in CCl_4 and tetramethylsilane (TMS) was the internal reference. The objective was to predict the octane index from proton NMR spectra with a model built from a reduced number of calibration samples, which had been previously analysed using the ASTM reference method.

3.2 Methodology

Calibration sets containing from 4 to 60 samples were compiled using the Fedorov and Kennard-Stone algorithms and random selection . The samples not used for calibration were used for prediction. In the case of random selection, 50 random models were built and the mean value of the prediction ability was evaluated for each number of selected samples.

3.3 Results

The Fedorov selection took about 2 min using a Hewlett-Packard computer with Pentium 90Mhz CPU and 16 Mbyte RAM. The prediction error, in terms of root-mean-squared error of prediction for the test samples, RMSEP, is plotted as a function of the number of selected samples in Fig. 1. It can be observed that Fedorov's algorithm performs better in almost the whole range of samples while the models built with randomly selected calibration samples always give worse predictions except for subsets containing a rather reduced number of samples.

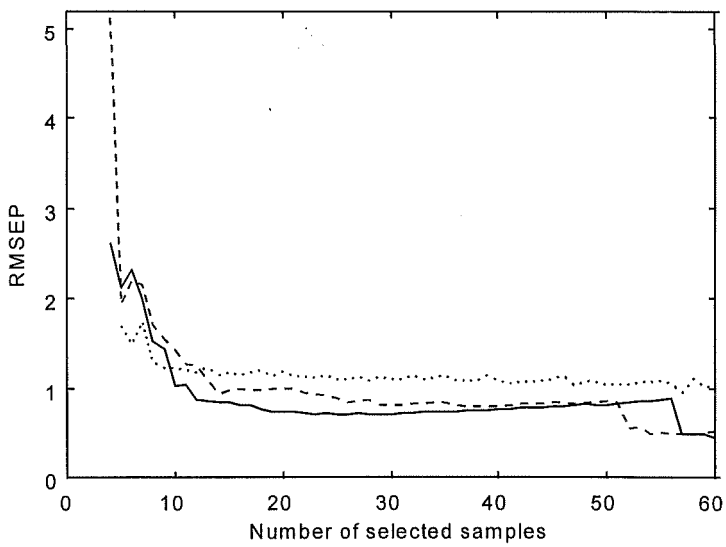


Fig. 1. Prediction error, in terms of root-mean- squared error of prediction for the test samples (RMSEP) as a function of the number of selected samples for the three selection methods. (—) Fedorov, (- -) Kennard-Stone , (···) random selection.

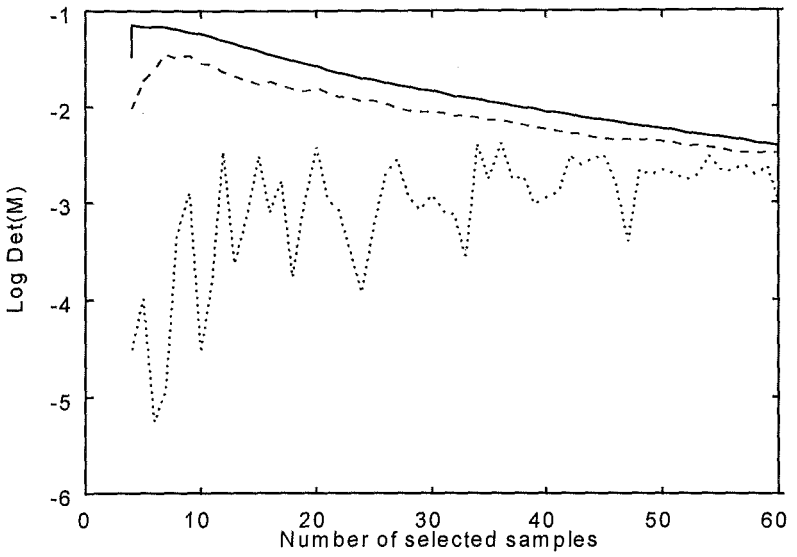


Fig. 2. Evolution of log(Det(**M**)) versus the number of selected samples. (—) Fedorov, (- -) Kennard-Stone , (···) random selection.

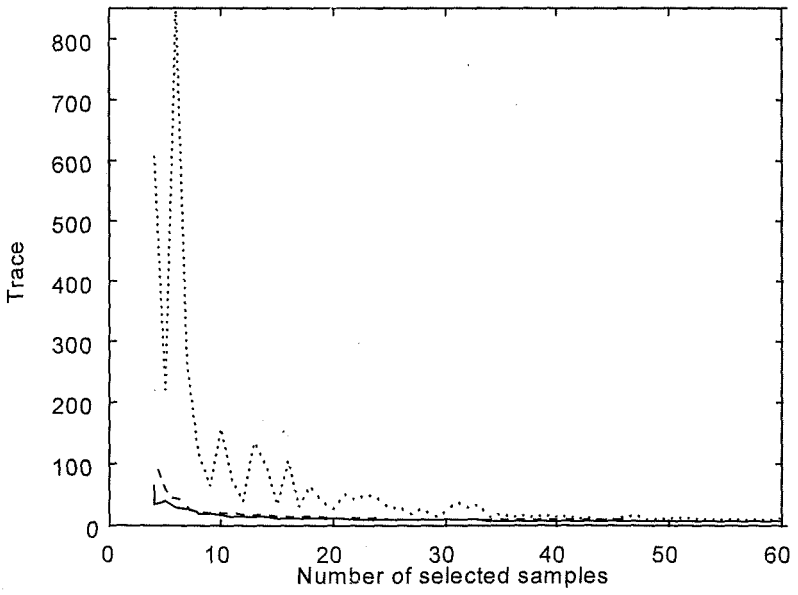


Fig. 3. Evolution of Trace($\mathbf{X}^T\mathbf{X}$)$^{-1}$ for each subset of calibration samples. (—) Fedorov, (- -) Kennard-Stone , (···) random selection.

The evolution of log(Det(M)) versus the number of selected samples is shown in Fig.2. Taking into account that higher values for Det(M) are desired, Fedorov's algorithm provides the best values for the whole range of selected subsets. It can also be observed that when selecting the calibration samples on a random basis there is a considerable instability. The set of samples recommended for building the regression model is the one that maximises log(Det(M)), which corresponds to four to seven samples.

Fig. 3 shows the evolution of the Trace$(\mathbf{X}^T\mathbf{X})^{-1}$ for each selected subset. In this case, small values are desired. It can seen that the random selection of calibration samples can produce higher values for the trace giving unreliable estimates of the coefficients in the model. Note that although D-optimality does not necessarily produce minimum trace designs they are always slightly smaller than the ones produced by the Kennard-Stone algorithm.

4.Conclusions

The advantages of selecting D-optimal calibration subsets with Fedorov's algorithm have been shown. The samples are specially selected for the model that must be built, which can have a degree of complexity of any order and is not restricted to the first- or second-order models used in classical experimental designs. This is in contrast with Kennard-Stone or the random selection of calibration samples, which select samples independently of the model equation. Additionally, in contrast to the classical designs, where the number of samples is fixed, the algorithm selects D-optimal subsets for any number of samples. In such a case, the experimenter can reach a compromise between the information content that is required of the calibration set and the cost he/she is willing to pay for it.

Acknowledgements

J. Ferré thanks the Comissionat per a Universitats i Recerca of Generalitat de Catalunya, for providing a doctoral fellowship (FI/94-7001). Financial support from the Spanish Ministry of Education and Science (DGICyT project BP93-0366) is acknowledged.

References

[1] G.E.P. Box, W.G. Hunter and J.S. Hunter, Statistics for Experimenters , Wiley, New York , 1978

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

[3] I. Ruisánchez, J. Lozano, M. S. Larrechi, F.X. Rius and J. Zupan, Anal. Chim. Acta, in press.

[4] P.F. De Aguiar, B. Bourguignon, M.S. Khots, D.L. Massart and R. Phan-Than-Luu, Chemom. Intell. Lab. Syst., 30 (1995) 199.

[5] D. Mathieu PhD Thesis, Marseille, 1981

[6] V.V. Fedorov, Theory of Optimal Experiments, translated and edited by W.J. Studden and E.M. Klimko Academic Press, New York, 1972.

[7] J. Ferré and F.X. Rius, Trends Anal. Chem., submitted for publication.

[8] Matlab, The Mathworks, South Natick, MA.

UNIVERSITAT ROVIRA I VIRGILI
EXPERIMENTAL DESIGN APPLIED TO THE SELECTION OF SAMPLES AND SENSORS IN MULTIVARIATE CALIBRATION
Joan Ferré Baldrich
ISBN:978-84-691-1875-7/DL: T-337-2008

3.5 Selection of calibration points for principal component regression in quantitative structure-activity relationship studies

(in preparation)

Joan Ferré and F. Xavier Rius

Departament de Química. Universitat Rovira i Virgili.
Pl. Imperial Tarraco, 1, 43005-Tarragona. SPAIN

D-optimality is used as a criterion for selecting calibration points for quantitative structure-activity relationship (QSAR) studies with the principal components as design variables. Fedorov's exchange algorithm and genetic algorithms (GA) are used to find the D-optimal subsets.

UNIVERSITAT ROVIRA I VIRGILI
EXPERIMENTAL DESIGN APPLIED TO THE SELECTION OF SAMPLES AND SENSORS IN MULTIVARIATE CALIBRATION
Joan Ferré Baldrich
ISBN:978-84-691-1875-7/DL: T-337-2008

1. Introduction

In quantitative structure-activity relationship (QSAR) studies, physicochemical variables of test compounds are regressed versus a property of interest. To reduce the number of experimental points (e.g. chemical reactions to measure the property of interest) and thus the cost of the calibration step, classical statistical designs for the regressor variables can be used[1]. However, the entities manipulated by the chemist (e.g. a structural fragment such as $-NO_2$ or –Br in a molecule) rarely enter directly into the model and they are represented by values (e.g. electronic properties) which are used as regressor variables. In such cases, the values of these properties cannot be independently manipulated, only measured, and may vary collinearly to each other. An experimental design with predefined levels of these measured variables is impossible. The procedure is then to select the most representative points from a list of candidates which are represented either by the values of the measured variables or by the factor scores on several principal components. The scores are useful in complicated systems with a large number of variables and whose designs would require a large number of experimental runs. To span the pertinent experimental region properly, the points whose coordinates are the most similar to the points of a classical design (e.g. a fractional factorial design) are selected[2-6]. Another procedure for selecting points is to use algorithms such as Fedorov's[7-9] (which optimizes the D-criterion) or Kennard-Stone's[10]. However, the studies available[11-14] use the original variables, not scores.

The case study in this section is based on the QSAR studies by Skagerberg *et al.*[15]. They evaluated four principal components for a set of one hundred aromatic substituents characterized by nine descriptor variables. The substituents whose scores on the first three principal components were placed approximately at the vertices of a cube representing a 2^3 factorial design were selected for modeling. They used only three factors because of the weak chemical significance of the fourth component and because a a four variable factorial design might involve a large number of points. However, since the scores do not cover all the possible values within the experimental domain it was impossible to construct a design with substituents that had coordinates which were exactly on the corners of the cube. They then divided the candidate points into eight groups that corresponded to a corner of the cube and subjectively selected one point from each group.

Since a factorial design is D-optimal, a D-optimal set of points selected using optimization algorithms will be the closest possible to the factorial design points. In this section a method is presented for selecting the aromatic substituents in QSAR studies using Fedorov's exchange algorithm and the scores as design variables. In addition, genetic algorithms are used to produce a list of quasi-D-optimal solutions so that a subset which is not D-optimal can be selected if additional considerations, such as the cost of evaluating the points, makes that solution more suitable.

2. Theoretical background

Principal component regression regresses the scores on A factors of I calibration points against the $I\times1$ vector of observed values of the dependent variable ($\mathbf{y}$) according to

$$\mathbf{y} = \mathbf{T}\beta + \mathbf{e} \qquad (1)$$

where $\mathbf{T}$ is the $I\times P$ model matrix in which each row corresponds to a calibration point and each column corresponds to a coefficient in the model, β are the coefficients to be estimated and $\mathbf{e}$ is an $I\times1$ vector of error terms. The least-squares estimate of β is given by :

$$\mathbf{b} = (\mathbf{T}^T\mathbf{T})^{-1}\mathbf{T}^T\mathbf{y} \qquad (2)$$

where ^{T} means transposed. Several criteria based only on the model matrix can be used to characterize the possible performance of the model [7,8]:

1. logDet($\mathbf{T}^T\mathbf{T}$) where Det denotes determinant. It measures the information content in the model matrix.

2. logDet($\mathbf{M}$), where $\mathbf{M}=\mathbf{T}^T\mathbf{T}/I$. This is a measure of the information content *per point* and enables matrices with a different number of points to be compared.

3. Trace$(\mathbf{T}^T\mathbf{T})^{-1}$. It measures the global variance of the estimated coefficients.

4. The maximal variance function, which is given by $d_{max}=\max(\mathbf{t}^T(\mathbf{T}^T\mathbf{T})^{-1}\mathbf{t})$, where $\mathbf{t}^T$ is a row in the matrix of candidate points. It is related to the variance of the predicted response at point **t**.

5. G-efficiency = $P/(d_{max}\times I)$. This takes into account the number of coefficients, the number of points and the largest variance function in the design.

6. Variance coefficients (UVIF), which are the diagonal values of $(\mathbf{T}^T\mathbf{T})^{-1}$ and are proportional to the variance of the model coefficients.

7. Variance inflation factors (VIF's) for each coefficient, which measure the collinearity in the columns in **T** . VIF's indicate whether the information content in the selected points is sufficient to obtain precise estimations of the coefficients in the model. For values larger than 10 the coefficient might be unreliably estimated due to the collinearity[16].

The model matrix should have optimal values of these properties: maximum Det($\mathbf{T}^T\mathbf{T}$), G-efficiency and Det(**M**), and minimum Trace$(\mathbf{T}^T\mathbf{T})^{-1}$, coefficient variance function and variance inflation factors (the value of which should never be higher than 10) for all the coefficients. Hence, these criteria can be used to select the smallest set of I points in **T** that reliably estimates the coefficients, before measuring the corresponding responses **y**. Either classical experimental designs[1] (with the drawbacks already commented) or optimization algorithms can be used for selecting the most suitable points. The most frequently used algorithms are the ones that maximize Det($\mathbf{T}^T\mathbf{T}$) (called the *D-criterion*) and avoid searching for all possible N point combinations. Here, the Fedorov exchange algorithm (FA) [6,7,9] and genetic algorithms (GA) are used to carry out the selection in a reasonable time. In addition to the optimal solution, GAs can find sub-optimal sets with a given number of points. This enables subsets of a quality similar to the D-optimal to be selected, in case that the D-optimal have undesirable properties (such as undesirable reactants). For each subset of N points, the performance characteristics are evaluated. This helps to choose the most suitable matrix to perform the calibration, taking into account the number of points that it involves. The subset with maximum information per point (maximum logDet(**M**)) is usually selected for calibration provided that the other quality measures (VIF, Trace$(\mathbf{T}^T\mathbf{T})^{-1}$, etc.) are acceptable.

3. Experimental section

Data Set. We used the scores on three PCs for one hundred substituents from the data published in Table 4 of the paper by Skagerberg *et al.*[15].

Procedure. After adding a column of ones to **T** (to account for the constant term in a factorial design), the FA found the D-optimal designs with 4 to 15. The analysis was re-started 10 times to avoid finding only local optima. After deciding the optimal number of points for regression, the GA selected the best 10 sets with this number of points according to the D-optimality criterion, thus finding a list of sets with similar qualities. The quality parameters of the matrices generated by the algorithm were plotted against the number of points to make it easier find the optimal subset.

4. Results and discussion

Figure 1 shows the values of logDet(**M**), Trace$(\mathbf{T}^T\mathbf{T})^{-1}$, d_{max} and *G*-efficiency for the D-optimal subsets as a function of the number of points. In addition to the global optimum, local optima were found for the subsets of 5, 6, 7, 8, 9 and 10 points. So, in the graph, the different solutions found con be seen by vertical lines.

It can be seen that logDet(**M**) has the highest values at 5 and 8 points, so these sets contain the most information per point of all the other subsets. At more than 8 points, logDet(**M**) starts decreasing. The trace decreases sharply for the sets with fewer than 8 or 10 points. Of the two solutions found for 5 points, one produces rather poor values of logDet(**M**) and Trace$(\mathbf{T}^T\mathbf{T})^{-1}$. The largest variance function always decreases as more points are added to the design and for more than 10 points the change gets smaller and smaller. The efficiency shows that the set of 5 points, despite being M-optimal, has such a large d_{max} that its efficiency is smaller than that of the designs with 8, 9 or 10 points. The 8-point design is G-efficiency optimal. Taking all this into account, the 8-point set is most suitable for modeling (unless other constraints, such as the economic aspect of the selected points, make the 5-point set preferable). In our case, FA found two different 8-point solutions. The values of the performance characteristics of these solutions and of the solution used by Skagerberg *et al.*[15] are

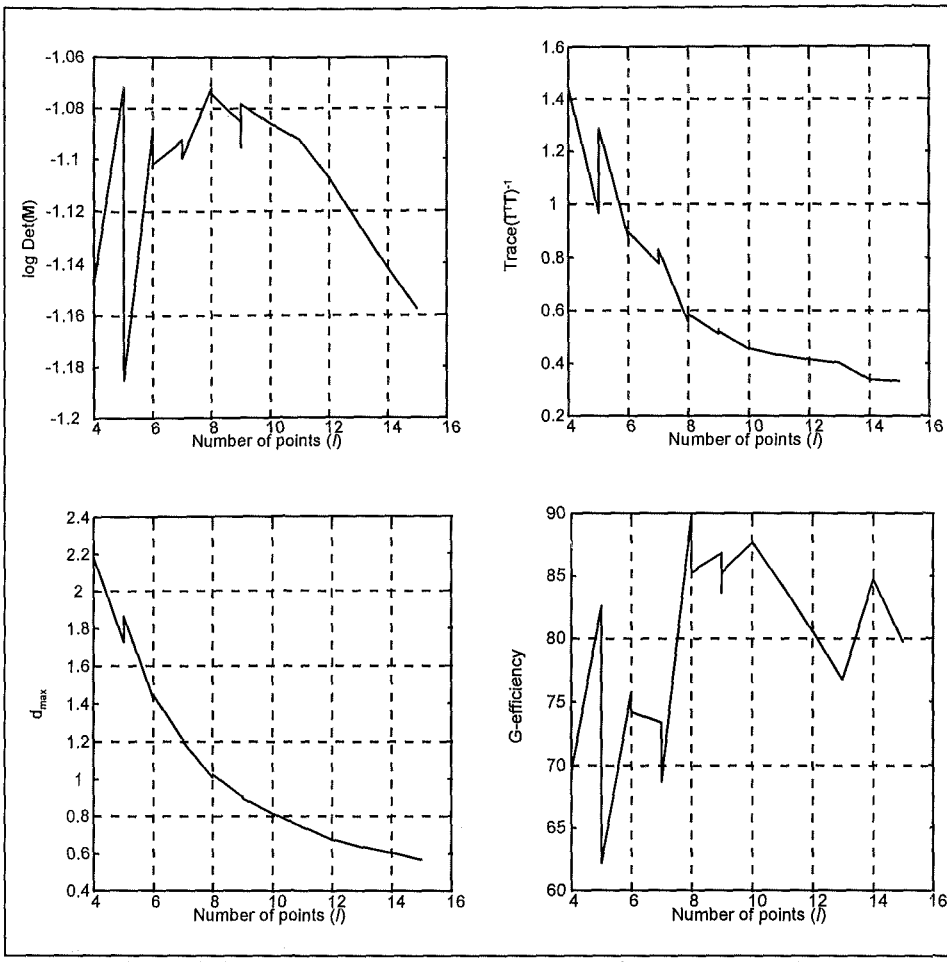


Figure 1. logDet(**M**), Trace($\mathbf{T}^T\mathbf{T}$)$^{-1}$, d_{max} and *G*-efficiency for the D-optimal subsets of 4 t
points selected with the FA.

compared in Table 1. It should be noted that each of the selected points is include
one of the 8 subregions defined by the 2^3 factorial design used by Skagerberg *et*
However, the points selected using FA give lower VIFs and variance coefficient
well as better values of logDet(**M**), Trace($\mathbf{T}^T\mathbf{T}$)$^{-1}$, d_{max} and *G*-efficiency . So, these
will estimate the coefficients in the model more precisely. Figure 2 compares
points selected by Fedorov and the ones used by Skagerberg *et al.*[15] in the space o
three principal components (PCs) used for modeling. Observe how the p
selected by the algorithm are more external than the ones proposed by Skagerbe
al.[15].

Table 1. Performance characteristics of the selected 8-point matrices: the sets of points selected by FA and Skagerberg *et al.*[15] The numbers that identify the points are the same as in Table 4 in Skagerberg *et al.*[15].

		Fedorov	Fedorov	Skagerberg
Number of points		8	8	8
Selected points		4,7,10,42,93 95,96,97	4,10,13,23,92 95,96, 97	1,8,10,33,65 71,85,93
Number of times selected		4	6	
logDet ($\mathbf{T}^T\mathbf{T}$)		2.538	2.540	1.350
logDet ($\mathbf{M}$)		-1.074	-1.072	-2.263
Trace $(\mathbf{T}^T\mathbf{T})^{-1}$		1.03	1.01	2.56
Maximal variance function		0.59	0.56	0.75
G-efficiency (%)		85.2	89.8	66.8
variance coefficients	b_0	0.15	0.13	0.18
	b_1	0.34	0.31	0.70
	b_2	0.28	0.28	0.65
	b_3	0.26	0.28	1.02
VIF	b_0	0	0	0
	b_1	1.07	1.06	1.18
	b_2	1.01	1.02	1.21
	b_3	1.06	1.06	1.27

Table 2 shows the ten best solutions for the Det($\mathbf{T}^T\mathbf{T}$) criterion for 8 points found using the genetic algorithm in decreasing order of Det($\mathbf{T}^T\mathbf{T}$). The two best solutions are the ones found by FA but others could also be acceptable (e.g. the third solution has a Trace$(\mathbf{T}^T\mathbf{T})^{-1}$ value of 1.011, smaller than that of the D-optimal solution, and this is also a desirable property for the calibration matrix).

Table 2. Performance characteristics of the 8-point matrices with the largest Det($\mathbf{T}^T\mathbf{T}$). The numbers that identify the points are the same as in Table 4 in Skagerberg *et al.*[15].

Selected points								log Det($\mathbf{T}^T\mathbf{T}$)	Trace$(\mathbf{T}^T\mathbf{T})^{-1}$
4	10	13	23	92	95	96	97	2.540	1.019
4	7	10	42	93	95	96	97	2.538	1.027
7	10	13	23	92	95	96	97	2.533	1.011
7	10	23	42	93	95	96	97	2.531	1.040
7	10	23	42	92	95	96	97	2.530	1.010
7	10	13	23	81	95	96	97	2.527	1.042
7	10	23	42	81	95	96	97	2.524	1.044
4	10	23	42	92	95	96	97	2.522	1.030
4	7	10	13	92	95	96	97	2.520	1.028
7	10	13	23	93	95	96	97	2.519	1.062

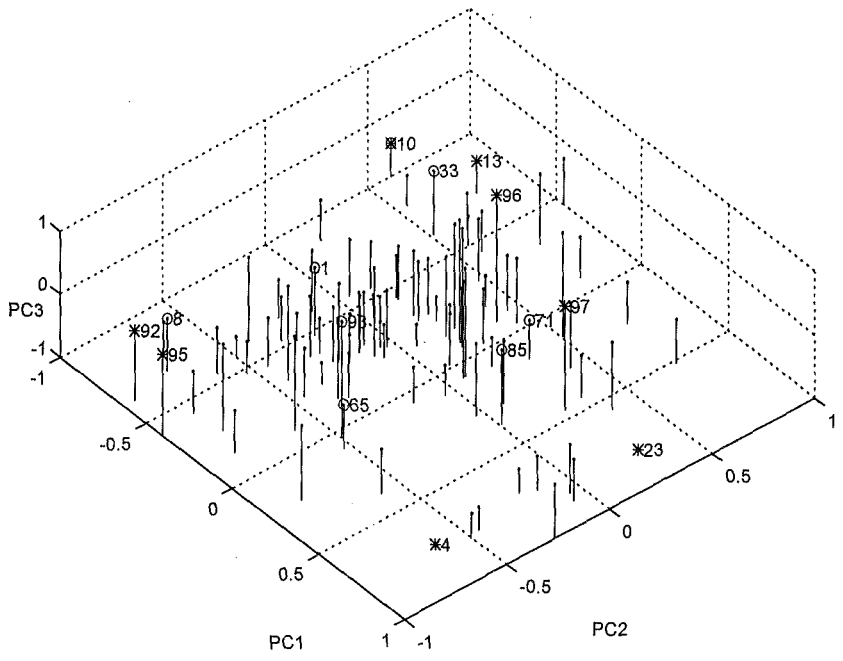


Figure 2. The 100 candidate points in the space of the three PCs. (*) Points selected by FA. (o) Points by Skagerberg *et al.*[15]. The numbers that identify the points are the same as in Table 4 in Skagerberg *et al.*[15]. The vertical lines show the points in the space better.

Conclusions

There are several advantages of using optimization algorithms to build D-optimal designs for QSAR studies:

1. There is no need to represent the selected points graphically. This enables points which fit designs of more than three variables, containing interaction terms, etc. ... to be selected. No subjective approach is needed to decide which points are most similar to the levels of a factorial design.

2. The algorithm can select optimal subsets for any number of points equal to or larger than the number of coefficients. This overcomes the difficulties of using the specific number of a classical design that could be too large. Thus the

experimenter can choose the number which is a compromise between the information content required and the cost. In the case studied, a D-optimal solution of five points was found for three variables.

3. These algorithms also enable some candidate points to be always included in all the selected subsets. This is useful when the observed values y of some points have already been measured and therefore it is information that the user already knows. In addition, if some points are preferable to others for practical or economic reasons, they can be weighted according to the experimenter's criterion. By so doing, for instance, a point corresponding to an expensive experiment would only be selected when its information content is important enough to compensate for the additional cost of using it.

4. The variance inflation factors indicate whether the information content in the selected points is sufficient to obtain precise estimations of the coefficients in the model. Additional performance statistical parameters such as the Det($\mathbf{T}^T\mathbf{T}$), the Trace($\mathbf{T}^T\mathbf{T}$)$^{-1}$ and the G-efficiency for the matrix of selected points give extra information on the suitability of the selected subset. In our case, a subset of five points and eight points are very similar according to the M-criterion. Therefore, if other considerations such as cost do not influence the final decision, it is advisable to perform the eight selected experiments instead of five since the trace and VIF values are smaller.

5. The genetic algorithm, used to select D-optimal sets, found the same solutions as Fedorov's algorithm and, in addition, generated a list of sub-optimal solutions ordered according to the optimality criterion. Hence, an almost D-optimal solution can be selected taking into account additional qualities such as the trace of the dispersion matrix being smaller than the D-optimal solution or the lower cost of the experimentation.

6. The statistical properties of the D-optimal calibration matrices are superior to over the set used by Skagerberg *et al.*[15] and supposedly to lead to models with a better prediction ability. Unfortunately, this could not be assessed since the values of the dependent variable were not available. However, the advantages of these matrices, suggest that it is advisable to use the proposed method in future problems.

References

1. Box G.E.P., Hunter W.G., Hunter J.S. *Statistics for Experimenters* Wiley: New York. 1978.
2. Hellberg S., Sjöström M., Skagerberg B., Wold S. *J. Med. Chem.* 30 (1987) 1126-1135.
3. Wold S., Sjöström M., Carlson R., Torbjörn L., Hellberg S., SkagerbergB., Wikström C., Öhman J. *Anal. Chim. Acta* 191 (1986) 17-32.
4. Norinder U., Högberg T. *Acta Chemica Scand.* 24 (1992) 363-366.
5. Hellberg S., Sjöström M., Skagerberg B., Wikström C., Wold S. *Acta Pharm. Jugosl.* 37 (1987) 53-65.
6. Carlson R. *Design and optimization in organic synthesis.* Elsevier: The Netherlands 1992.
7. De Aguiar P.F., Bourguignon B., Khots M.S., Massart D.L., Phan-Tan-Luu R. *Chem. Intell. Lab. Syst.*, 30 (1995) 199-210.
8. Mathieu D. *Contribution de la Méthodologie de la Recherche Expérimentale à l'étude des relations Structure-Activité.* Thèse Sciences. Marseille, 1981.
9. Fedorov V.V. *Theory of Optimal Experiments*(translated and edited by W.J. Studden and E.M. Klimko) Academic Press: New York, 1972.
10. Kennard R.W. , Stone L.A. *Technometrics* 11 (1969) 137- 148.
11. Mathieu D., Phan-Tan-Luu R., Elguero J. *Bull. Soc. Chim. Belg.* 89(1980) 267-279.
12. Claramunt R.M., Gallo R., Elguero J., Mathieu D., Phan Tan Luu R., *J. Chimie Physique* 78 (1981) 805-814.
13. Cativela C., Melendez E., Calvete H., Elguero J., Mathieu D., Phan Tan Luu R. *An. Quím,* 80 (1984) 91-97.
14. Claramunt R.M., Elguero J., Mathieu D., Phan Tan Luu R., *An. Quím.* 80 (1984) 30-38.
15. Skagerberg B., Bonelli D., Clementi S., Cruciani G., Ebert C. *Quant. Struct.-Act. Relat.* 8 (1989) 32-38.
16. Harrison M. Wadswoth. *Handbook of Statistical Methods for Engineers and Scientist.* McGraw Hill: USA 1990.

3.6 Assessing the validity of principal component regression models in different analytical conditions

Anal. Chim. Acta 337 (1997) 287-296

A. Rius, M.P. Callao, J. Ferré and F.X. Rius*

Departament de Química, Universitat Rovira i Virgili, Pl. Imperial Tàrraco 1, 43005 Tarragona, Spain.

This study proposes a methodology for assessing the validity of principal component regression models when the experimental conditions which have been used in the process of modeling may have changed. The methodology proposed is based on the procedure for selecting the validation sample subset which includes the D-optimal criterion and application of Fedorov's exchange algorithm. Two basic performance characteristics define the validity of the models: trueness is assessed by linear regression using the joint confidence test for the slope and the intercept and precision is estimated by bias corrected MSEP and RRMSEP. The methodology is validated with a simulated data set and three real data sets corresponding to models constructed for spectrophotometric data from determinations of various analytes in waters using sequential injection analysis (SIA). Using a reduced number of samples can be very useful in several applications, such as in process analytical control, and is specially useful as an initial step to check the need for standardization.

Keywords: Principal component regression; Sample selection; Chemometrics

Received 5 February 1996; accepted 7 August 1996

1. Introduction

Multivariate calibration has been a well established technique in the field of chemical analysis for a considerable time, and recently, it has been successfully applied to the field of process analytical chemistry [1]. For these applications in particular, multivariate calibration models need to be used for some time with no need for recalibration, since they are usually constructed from a great deal of data. Unfortunately, however, there may be some situations in which the model is not valid, for example an instrument change, a change in the environmental conditions, alterations or drifts in some of the parts of the instrument, or other changes in the working or measuring conditions [2,3].

In any of these circumstance, the constructed model cannot be used for prediction and either the instrument has to be recalibrated, with the work and cost it requires, or some sort of standardization procedure has to be applied [2]. Standardization techniques are arousing great interest at present. Different strategies have been described, particularly for spectroscopic instruments in near infrared (NIR) [4-9], multivariate calibration models in general [3, 10, 11] and second order instruments [12]. Also, several approaches have been proposed for selecting the most suitable samples to carry out the standardization [3,13,14]

All the studies published so far basically deal with the methodology of standardization itself, but there are other aspects which are no less important and are yet to be studied, such as establishing a methodology to determine whether the multivariate model is applicable in the new conditions.

The present study discusses a methodology which enables the validity of a multivariate model in different instrumental conditions to be checked. The validity of the model is established by maintaining the two basic performance characteristics, trueness and precision. This method is, then, a step prior to the possible standardization of multivariate regression models, but it may also be applied to validate the standardization methodology after it has been applied.

The quality of the model in the new conditions is validated with a subset of the calibration samples because, if many are used, there is no gain in time or cost in

comparison to recalibrating the instrument. The proposed methodology consists of selecting this subset of samples belonging to the set used in the construction of the model. The values of the instrumental responses in the new conditions are measured from this reduced set, for which the concentration values of the analyte of interest are kept constant. When the regression model constructed is applied to the responses obtained in the new conditions, new predicted values are obtained; studying the trueness and precision for these new values of predicted concentrations gives information about the further usefulness of the model.

2. Theory

2.1. Steps for multivariate model validation

To validate multivariate models in different instrumental conditions the following steps must be followed:

(1) For the set of I calibration samples: Multivariate response measurement in initial conditions and determination of analyte concentration using a reference method.
(2) Construct the PCR model with the I calibration samples. Validation of this model: Evaluation of the trueness by linear regression and of precision by calculating bias corrected MSEP and RRMSEP.
(3) Select a subset of N samples from the initial calibration set.
(4) Check that the N samples selected adequately represent the total set of samples in the initial conditions.
(5) Measure the multivariate responses for the N samples in the new instrumental conditions.
(6) Using the multivariate model constructed in stage 2, calculate the new concentrations predicted for the N samples considered.
(7) Check the goodness of the initial model used in the new conditions taking into account only the set of N samples chosen: Study the precision and assess the trueness.

2.2. *Trueness validation*

The possible presence of bias is detected using linear regression. The plot of the predicted value of concentration versus the known concentration for each sample ideally should be a straight line with a slope of 1 and an intercept of 0. The joint interval test for the slope and the intercept [15,16] assesses whether the theoretical value (1,0) is really within the confidence ellipse. The *F*-value is calculated using the equation:

$$N(b - b)^2 + 2\,\Sigma x\,(b - b)(m - m) + \Sigma x^2\,(m - m)^2 = 2Fs^2 \qquad (1)$$

where N is the number of samples, b = 0, m = 1, b is the intercept, m is the slope, x is the known concentration, and s is the residual standard deviation. The calculated *F*-value is compared to the value of the tabulated *F*-value for a fixed probability of a type I error (α) considering 2 and N-2 degrees of freedom. But in practice, it is better to calculate the value of α necessary for the point (0,1) to be inside the confidence region than check if the joint confidence test provides a significant difference for a given α. The values which are given in the Section 4 correspond to the significance level of the tests, 1-α. (For example, 1-α=0.92 is equivalent to a significant difference for a given α=0.08, but there is no significant difference for a given α=0.1).

Prior to applying the test, the regression line is validated by means of the residuals analysis, and the absence of influential points is proved with the Cook test [15].

2.3 *Precision*

Precision is estimated according to two criteria: the bias corrected mean square error of prediction is estimated according to the equation:

$$\mathrm{MSEP}_{bc} = \frac{\sum (c_{i,\mathrm{pred}} - c_i)^2 - \frac{\left[\sum (c_{i,\mathrm{pred}} - c_i)\right]^2}{I}}{I - 1} \qquad (2)$$

and the bias corrected relative root mean square error of prediction by using the equation:

$$RRMSEP_{bc} = \frac{100}{c_m}\sqrt{\frac{\sum (c_{i,pred} - c_i)^2 - \frac{\left[\sum (c_{i,pred} - c_i)\right]^2}{I}}{I-1}} = \frac{100}{c_m}\sqrt{MSEP_{bc}} \quad (3)$$

where I is the number of samples, c_i is the i sample concentration, $c_{i,pred}$ is the predicted concentration for the ith sample (cross-validated prediction for the initial conditions), and c_m is the mean of the c_i concentrations.

If the precision obtained experimentally satisfies the needs of the user and bias is not detected at the chosen significance level, the model is considered to be valid in the new conditions tested.

2.4 Sample selection

2.4.1 Criterion for sample selection

The aim is to select, from the I samples used to build the model in conditions 1, a subset with the minimum number of samples (N) which can give a "good estimate" of the accuracy and precision of the model in conditions 1 and, after the responses have been measured again for these samples, in the new conditions. We select the N samples which are the best representative of the experimental, which in turn, will enable us to determine whether there has been any change which makes the model unusable for prediction better. We assume that the best representative samples of the variable space selected are the ones which give a good estimate of the model coefficients (minimum variance of the regression coefficients in the model). Let $\mathbf{T}_{Nx(P+1)}$ be the score matrix for these N samples on the P principal components of the model with a column of 1s added to take into account the constant term. The 100(1-α)% confidence region for the coefficients in the model is given by Eq (4) [17]:

$$(\beta - \mathbf{b})^T \mathbf{T}^T_{N\times(P+1)} \mathbf{T}_{N\times(P+1)} (\beta - \mathbf{b}) \leq (P+1)\, s^2\, F_{P+1,df,\alpha} \qquad (4)$$

where s^2 is the estimated variance for the measured concentration of analyte with *df* degrees of freedom and $F_{P+1,df,\alpha}$ is the α per cent point of the *F* distribution on *P*+1 and *df* degrees of freedom. The precision of the estimated coefficients is given by the volume of this confidence region which is proportional to $[\mathrm{Det}(\mathbf{T}^T_{N\times(P+1)} \mathbf{T}_{N\times(P+1)})]^{-1/2}$, where Det denotes determinant [17-20]. The minimum variance in the coefficients of the model can be achieved by selecting the *N* samples included in $\mathbf{T}_{N\times(P+1)}$ that maximize $\mathrm{Det}(\mathbf{T}^T_{N\times(P+1)} \mathbf{T}_{N\times(P+1)})$ [21]. This criterion is known as the D-optimal criterion.

For a given number of samples, *N*, those which give higher values for $\mathrm{Det}(\mathbf{T}^T_{N\times(P+1)} \mathbf{T}_{N\times(P+1)})$ will be the ones which give the coefficients with least variance.

2.4.2 Algorithm for sample selection

The complete search for all possible combinations of N samples in order to find the subset that maximizes $\mathrm{Det}\,(\mathbf{T}^T_{N\times(P+1)} \mathbf{T}_{N\times(P+1)})$ is very time consuming when the number of available samples, *I*, is high. So, we used Fedorov's [17-19] exchange algorithm found in Mathieu [19] which is designed to search for D-optimal designs when there is a list of candidate samples. For a given model of *P*+1 coefficients and a given *N*>*P*+1, the algorithm selects the subset that makes $\mathrm{Det}\,(\mathbf{T}^T_{N\times(P+1)} \mathbf{T}_{N\times(P+1)})$ maximum from among the *I* available samples.

2.4.3. Criterion for selecting optimal N

$\mathrm{Det}(\mathbf{T}^T_{N\times(P+1)} \mathbf{T}_{N\times(P+1)})$ always increases when a new sample is added to matrix **T**[19]. Therefore, the best sample set would be one that contains them all. What we are attempting to do is avoid analysing them all again, so this is not useful. The one we will choose will be the subset with the most information per sample which is

given by log (Det($\mathbf{M}_N$)) with $\mathbf{M}_N=(\mathbf{T}^T_{N\times(P+1)}\mathbf{T}_{N\times(P+1)})/N$ [21].

Consequently, of all the D-optimal subsets for different N, the subset with maximum log(Det($\mathbf{M}_N$)) is selected since it gives the best precision per sample for estimating the coefficients of the regression model, and so for seeing if the model has to be modified or not.

3. Experimental

3.1 Sequential Injection Analysis data set

The sample sets worked with are for determining the concentration of Ca, Mg and sulphates in natural waters using a Sequential injection analysis (SIA) method with multivariate spectrophotometric detection. The conditions in which the spectra were obtained have been reported elsewhere [22, 23].

Data set 1: 26 samples of natural waters, of which we have the spectra and the Ca and Mg concentrations (between 40 and 120 ppm). We shall call the models Caa and Mga. The spectra were obtained for all the samples in different experimental conditions.

Data set 2: 25 samples of natural waters of which we have the spectra and the Ca and Mg concentrations (between 0 and 40ppm). We shall call the models Cab and Mgb. The spectra were recorded for all the samples in two new experimental conditions.

Data set 3: 38 samples of natural waters of which we have the spectra and the sulphate concentrations (between 0 and 500 ppm). We shall call the model SO_4. For all the samples, the spectra were recorded in new conditions.

In all cases, the new experimental conditions mean analysing the samples after a short time and with a newly prepared reagent.

Simulated data

The spectra of 100 samples (101 variables corresponding to 101 sensors) were simulated by adding the following (see Fig. 1):

- the spectra of the reagent and/or the sample matrix (Fig. 1a);
- the spectra of two randomly generated interferent components (Fig. 1b and 1c); and
- the spectra of the mixture of two pure components. The mixtures were generated by a 10^2 complete factorial design (two components at ten levels) (Fig. 1d and 1e).

Condition 1: The response matrix $\mathbf{R_0}$ is calculated by Eq. (5):

$$\mathbf{R_0} = \mathbf{CS} + \mathbf{1B} + \mathbf{r_1 I_1} + \mathbf{r_2 I_2} \qquad (5)$$

where **C** is the matrix of concentrations (100×2) for the two components considered, **S** is the matrix with the two spectra of these components (2×101), **1** is a column vector of 1s (100×1), **B** is a row vector with the spectrum of the reagent and/or the sample matrix (1×101), $\mathbf{I_1}$ and $\mathbf{I_2}$ are the spectra of the two interferences (1×101), and $\mathbf{r_1}$ and $\mathbf{r_2}$ are two column vectors (100×1) which represent the intensity of the interferences (random values between 0 and 40).

Finally, 1% of noise is added, obtaining the response matrix $\mathbf{R_1}$ from which the simulated model, called sim, will be built:

$$\mathbf{R_1} = \mathbf{R_0} + 1\% \text{ random noise} \qquad (6)$$

A spectrum example is shown in Fig. 1(f).

To simulate new experimental conditions, the $\mathbf{R_1}$ matrix is modified in the ways described below:

Condition 2: The same matrix but with different noise, Eq. (7):

$$\mathbf{R_2} = \mathbf{R_0} + 1\% \text{ random noise} \qquad (7)$$

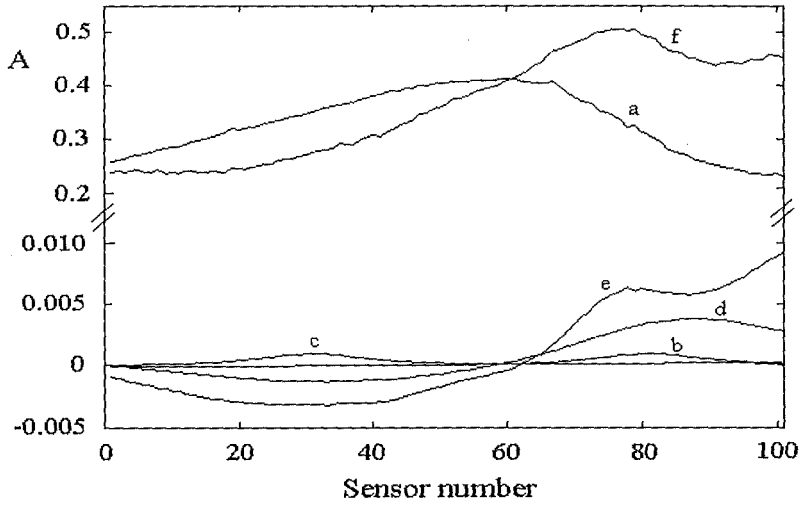


Fig. 1. Spectra of the components used to generate the simulated samples spectra: (a) reagent and/or sample matrix; (b) and (c) interferents; (d) and (e) pure components; (f) example of a sample spectrum.

This would be the case that the spectra of all the samples be repeated in the same instrument with no other changes.

Condition 3: Spectra shifted between +0.005 and +0.006 units of absorbance, Eq. (8):

$$\mathbf{R}_3 = \mathbf{R}_0 + \mathbf{r}_d \mathbf{1} + 1\% \text{ random noise} \qquad (8)$$

where $\mathbf{r}_d$ is a column vector (100×1) having has random values between 0.005 and 0.006, and $\mathbf{1}$ is a row vector (1×101) which contains 1s. This would be the case for a positive shift of all the spectra but where each spectrum has undergone a shift that is different from the others.

3.3 *Software*

All the algorithms for multivariate calibration, sample selection, statistical tests, simulation and other calculations were programmed with Matlab 4.0 [24]

4. Results and discussion

4.1. Validation of multivariate models

For each of the data sets described in the experimental section (Caa, Mga, Mgb, SO_4 and sim), a PCR model is constructed from the spectra recorded in condition 1. The slope and the intercept of the regression line (concentrations predicted by cross-validation vs. the actual concentrations), joint confidence test for slope and intercept, the $MSEP_{bc}$ and the $RRMSEP_{bc}$ are calculated. The results obtained are shown in Table 1(a). It can be seen that there are no significant differences at the significance level $\alpha = 0.05$, for any model, and the relative prediction error goes from 3% for the simulated data set to 27% for the Mga data set. These uncertainties are regarded as being acceptable for this kind of automated, quick analysis.

Table 1. Performance characteristics of the models in the initial conditions

PCR model	Cab	Mgb	Caa	Mga	SO_4	sim
Number of samples	25	25	26	26	38	100
Number of factors	4	4	4	4	4	4
(a) All samples, initial conditions (cross-validated predictions)						
Joint conf. test (1–α)	0.54	0.30	0.45	0.95	0.22	0.29
Slope	0.91	0.96	0.95	0.93	0.98	0.99
Intercept	1.13	0.37	3.32	3.15	4.29	0.04
$MSEP_{bc}$	10.33	4.49	34.9	36.38	528.6	0.08
$RRMSEP_{bc}$ (%)	24.3	18.9	10.3	27.0	8.7	2.8
(b) Selected samples, initial conditions						
Number of samples selected by D-optimal	7	7	7	7	8	7
Joint conf. test (1–α)	0.53	0.42	0.81	0.92	0.05	0.20
Slope	0.92	0.98	0.93	0.87	0.99	1.00
Intercept	0.91	0.00	4.17	2.66	1.60	0.02
$MSEP_{bc}$	9.93	1.82	34.79	17.01	277.0	0.03
$RRMSEP_{bc}$ (%)	25.5	13.0	9.7	19.5	8.1	1.9
(c) *F*-test for $MSEP_{bc}$ (Eq.(9))						
1–α	0.48	0.89	0.56	0.85	0.83	0.90

4.2 Selection of the best subset of samples

The sample selection algorithm provides us with the samples chosen for each subset of N calibration samples. As far as the precision of the results is concerned, the plot of $MSEP_{bc}$ for the selected samples versus the number of samples of the chosen subset (Fig. 2) shows that for the different models tested, $MSEP_{bc}$ is maintained practically constant in relation to N. This means that if the samples are chosen according to the D-optimal criterion a good estimate is obtained for the $MSEP_{bc}$ of the model. To check whether the prediction errors calculated from all the samples are statistically comparable to the ones calculated from the selected samples, an F-test was carried out [25] using Eq. (9):

$$F = \frac{(MSEP_{bc})_1}{(MSEP_{bc})_2} \qquad (9)$$

The results are shown in Table 1c, and it can be seen that in all cases there are no statistical differences between the $MSEP_{bc}$ compared. So, it can be considered that the samples chosen adequately represent the total set of samples as far as precision is concerned.

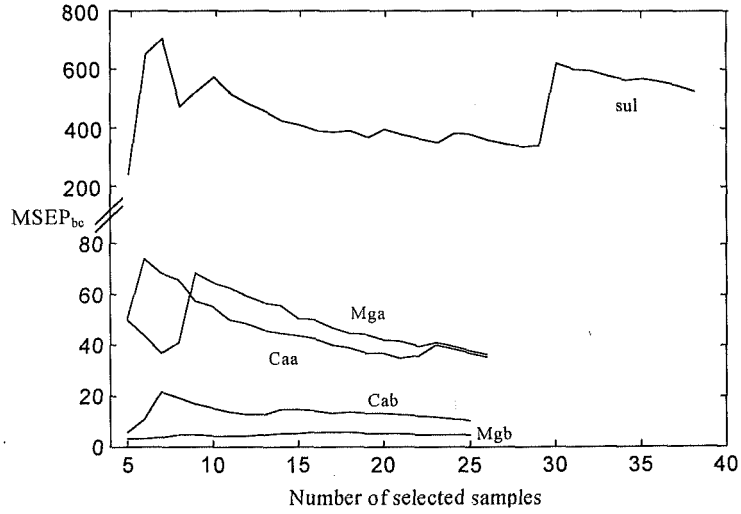


Fig. 2. Plot of the $MSEP_{bc}$ vs. the number of selected samples, for several models.

With respect to checking trueness using the joint confidence interval test, the plot of 1-α versus the number of samples selected is shown in Fig 3(a). Considerable oscillations in the values of 1-α can be seen in the graph (for example, for the Cab model, 1-α=0.7 if the number of samples selected is 6, 0.04 if 7 are selected, or 0.54 for all of the samples). This is due to the fact that the number of samples is a very important parameter in the result of this test (see Eq. (1)) and when the test is applied to a regression line obtained from a reduced set of samples, such considerable variations are often produced. One way of preventing these oscillations is by applying the test to the regression line which includes the predictions of selected samples in the new conditions added to the cross-validation predictions of all samples in condition 1. So, when applying the set test (Eq. (1)), N is the number of samples in the calibration set plus the number of selected samples. When the test is applied in this way, the plot of 1-α vs. the number of samples is shown in Fig. 3(b) for the Cab model. The observation of those plots shows that, for all the models, the variations in 1-α are not so large when the predictions of selected samples are added

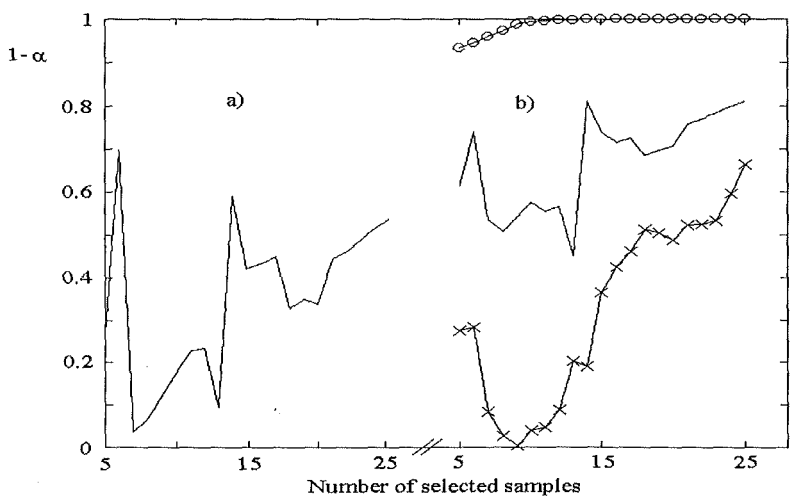


Fig. 3. Plot of the result of joint confidence test application vs. the number of selected samples for the Cab model. -: initial conditions, x: conditions 2, o: conditions 3. In (a) the test is carried out with all samples, and in b) with the selected samples added to all samples in initial condition.

to the initial predictions for all samples and in general, 1-α tends to increase as the number of samples selected increases. As a result, from the values of 1-α obtained using this procedure, we can make acceptable estimates of the behaviour of the model considering all the samples in the new conditions. Therefore, it can be considered that the selection algorithm of the subset of samples gives rise to N calibration samples which adequately represent the behaviour of the total set of samples as far as accuracy is concerned. The values of 1-α in Table (1b) were calculated in this way.

The plot of log(Det($\mathbf{M}_N$)) vs. N, for the Cab model, is shown in Fig. 4, with the maximum indicating the number of samples selected, since it gives the most information per sample. From the concentrations predicted by the model of these selected samples and the ones determined by chemical analysis, the prediction errors, parameters of the regression line and the results of the joint confidence test of slope and intercept are calculated. The number of samples selected and the results are shown in Table 1(b).

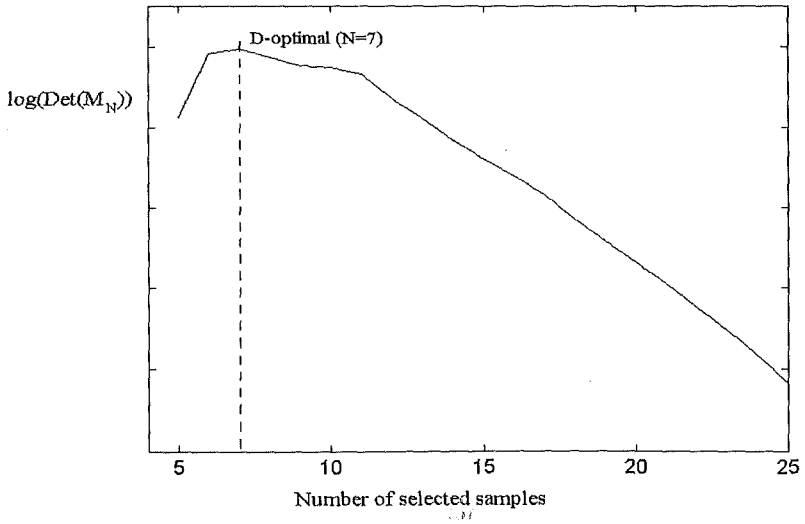


Fig. 4. Plot of log(Det($\mathbf{M}_N$)) versus N for the Cab model.

UNIVERSITAT ROVIRA I VIRGILI
EXPERIMENTAL DESIGN APPLIED TO THE SELECTION OF SAMPLES AND SENSORS IN MULTIVARIATE CALIBRATION
Joan Ferré Baldrich
ISBN:978-84-691-1875-7/DL: T-337-2008

4.3 *Application of the PCR models for all the samples measured in the new instrumental conditions*

To check whether a subset of N selected samples gives the same information about the validity of the model as the set of I calibration set samples, all the samples of this latter set must be tested in the new conditions. Of course, this is not the process which will be followed in the real sets in which only one subset of N samples of the calibration set is measured again in the new condition. So, the models created and validated for the set of I calibration set samples in conditions 1 are applied to predict the concentrations of the analytes determined in the new conditions for all the water samples. The results obtained are shown in Table 2(a).

As far as the trueness of the new results obtained is concerned, it can be seen that the joint confidence interval test ($1-\alpha=0.89$, 0.50 and 0.56) did not find differences for the slope and the intercept for only three of the new conditions (Cab(2), Caa(2) and sim(2), respectively), whereas for the others the test did find differences at a 0.05 level of significance. The precision estimated by the $MSEP_{bc}$ can be seen to generally increase a great deal for the models that show bias according to the joint confidence test, whereas for the models which give a good trueness in the new conditions, such as Caa(2) and sim(2), precision is maintained, except for Cab(2) which increases quite a lot.

4.4. *Application of the PCR model in new conditions for the selected samples*

The PCR models developed initially are applied to obtain the concentrations from the spectra obtained in the new conditions by the selected N samples. The precision for the sets of predicted N concentrations is calculated and the joint interval confidence test is applied to them to detect bias. The results obtained are shown in Table 2(b).

The F- test is used to evaluate the magnitude of the $MSEP_{bc}$ obtained in the new conditions for the reduced set of N samples. The results in Table 2(c) indicate that, in all cases, the uncertainty obtained for the new conditions taking into account all the I samples (Table 2(a)) or only the selected subset of N samples (Table 2(b)) are

Table 2. Performance characteristics of the models in the new condition

PCR model (conditions)	Cab (2)	Cab (3)	Mgb (2)	Mgb (3)	Caa (2)	Mga (2)	SO_4 (2)	sim (2)	sim (3)
(a) All samples									
Joint confidence interval. test (1–α)	0.89	1.00	0.99	1.00	0.50	1.00	0.99	0.56	1.00
Slope	1.18	0.59	0.95	0.25	0.97	0.88	0.92	1.00	1.00
Intercept	-0.43	-11	-2.5	12.8	2.85	-2.2	13.8	0.06	-0.88
$MSEP_{bc}$	36.1	64.1	20.5	82.9	25.8	49.9	610	0.06	0.06
$RRMSEP_{bc}$ (%)	44.4	59.2	40.3	81.1	8.7	31.6	9.4	2.4	2.5
(b) Selected samples									
Number of samples selected by D-optimal	7	7	7	7	7	7	8	7	7
Joint confidence interval. test (1–α)	0.08	0.96	0.85	0.93	0.40	0.96	0.80	0.24	0.80
Slope	1.00	0.93	0.94	0.81	0.96	0.84	0.96	1.00	1.00
Intercept	0.28	-3.5	-0.3	3.41	2.62	2.47	7.83	0.03	-.03
$MSEP_{bc}$	26.5	52.0	21.1	80.0	26.4	73.2	739	0.04	0.03
$RRMSEP_{bc}$ (%)	41.7	58.3	44.2	85.9	8.4	40.4	13.2	2.2	1.9
(c) *F*-test for $MSEP_{bc}$ (Eq.(9))									
1–α	0.64	0.59	0.56	0.48	0.56	0.77	0.68	0.62	0.83

statistically comparable at a 0.05 level of significance. So, we have the same information in relation to the precision using only the *N* samples selected.

As far as validating trueness, it can be seen that in all the cases in which statistical differences have not been detected between the calculated values of the slope and the intercept, and the theoretical values of unity slope and zero intercept using all the samples, neither of them are detected using the N samples selected at the same or lower levels of significance. However, in cases in which the joint confidence test did find differences in trueness using all the samples but not using the selected samples (Mgb(2), Mgb(3), sul(2) and sim(3)), this value of 1-α turns out to be much larger than in the initial conditions, and given that the trend is to increase with *N*, there is sufficient evidence to question the validity of the model.

The results are summarized in Table 3. In this table, the *F*-test value of 1-α to evaluate the precision corresponds to comparing the $MSEP_{bc}$ from all the samples in initial conditions with the $MSEP_{bc}$ from the samples selected in the new conditions. A

Table 3. Summary of the results obtained for all the models in all the conditions

PCR model (conditions)	Cab (2)	Cab (3)	Mgb (2)	Mgb (3)	Caa (2)	Mga (2)	SO_4 (2)	sim (2)	sim (3)
Evaluation of the accuracy and precision									
Joint conf. test (1–α), initial conditions, all samples	0.54	0.54	0.30	0.30	0.45	0.95	0..22	0.28	0.28
Joint conf. test (1–α), new conditions, selected samples	0.08	0.96	0.85	0.93	0.40	0.96	0.80	0.24	0.80
(1–α) for *F*-test	0.96	1.00	1.00	1.00	0.62	0.90	0.77	0.80	0.90
Validity of model	yes[a]	no	no	no	yes	no[b]	no	yes	no

[a] Bias is not detected, and the model is considered accurate if the decrease in the precision can be accepted.

[b] Bias is detected, and the model is considered inaccurate in spite of the precisions not being significantly different.

comparative analysis between the results in Table 2(b) and the ones in Table 1(a) indicates that conclusions can be drawn from the conjoint behaviour of precision and trueness:

1. Comparable $MSEP_{bc}$ while the value of 1-a decreases: this is the typical case we find if there have been no significant change in the conditions, and so the model constructed is still valid. It is obtained, for example, in the Caa models in condition 2 (1-α goes from 0.45 to 0.40) and sim in condition 2 (1-α goes from 0.28 to 0.24). Fig. 5(a) shows how the predictions in conditions. 2 by the selected samples of the Caa set are very similar to the predictions in initial conditions.

2. Loss of precision, to the point of not being comparable with the initial conditions, while the value of 1-α decreases: in this case the model gives us results that would be accurate but with a considerable decrease in precision which may be one of the reasons for not finding differences in accuracy. It must be decided for each model whether we can accept the increase in uncertainty. The Cab model in condition 2 is an example of this case (1-α goes from a value of 0.54 to 0.08), and we could continue using the model if we can accept that the $RRMSEP_{bc}$ increases from 24% (Table 1(a), initial conditions) to 42% (Table 2(b), new conditions). Fig. 5(b) shows how the predictions in conditions 2 by the samples selected with the Cab model

are distributed around the regression straight line in a similar way as in condition 1 but in general, the residuals are high which causes a decrease in precision.

3. Comparable precision, while the 1-α value of the trueness test increases considerably. This is the case for the Mga and SO_4 models in condition 2 and the sim model in condition 3. For the Mga model, bias is not detected according to the joint confidence test in the initial conditions (1-α=0.95) while it is detected in the new conditions, whereas there is a considerable increase in the value of 1-α, from 0.22 to 0.80 for the SO_4 model and from 0.28 to 0.80 for the sim model, and so it can be concluded that bias is detected for both models. Fig. 5(c) shows how the predictions in conditions 2 made by the samples selected with the Mga model are quite similar to the predictions in initial conditions, but as in the initial model bias was not detected according to the joint confidence test at a very low significance level, differences in trueness were detected in the new conditions.

4. Non comparable precision and 1-α in the trueness test increases considerably: in this case it should be taken into account that the constructed PCR model is not applicable to the new conditions since some instrumental changes have been detected which make it unusable for obtaining acceptable results. This is the case for the Cab model in conditions 3, and the Mgb model in conditions 2 and 3. Fig. 5(d) shows how the predictions in conditions 3 by the models selected with the Mgb model are much worse than the predictions in initial conditions and so the model cannot be used.

5. Conclusions

The methodology proposed, which includes sample selection and statistical tests on trueness and precision, enables the validity of a PCR model to be checked when it is suspected that a considerable change in the working conditions may jeopardize the results obtained with this model. This methodology is applicable as a check, assessing the trueness and the precision of the method even though there has been no changes in the experimental conditions. In many cases, the conclusions as to the validity of the model are quite clear, and when there is any doubt, the results give the

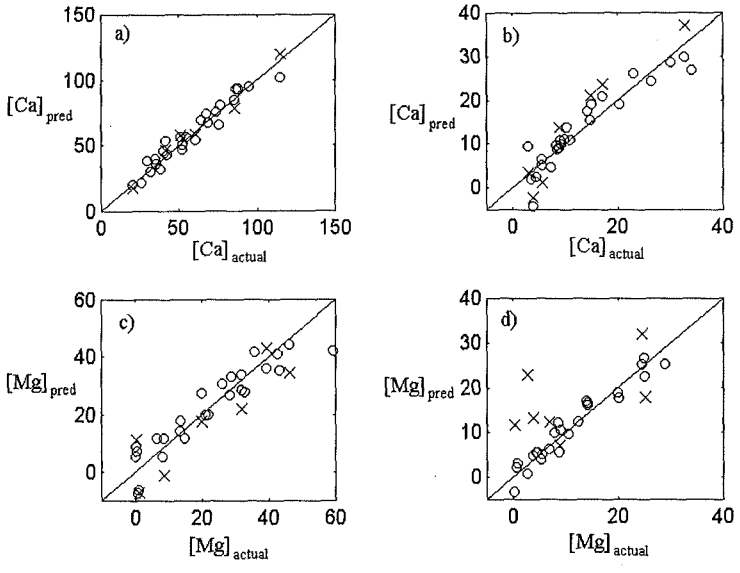


Fig. 5. Plot of the predicted concentrations vs. actual concentration for different models: (a) Caa in condition 2, (b) Cab in condition 2, (c) Mga in condition 2, and (d) Mgb in condition 3. o: predicted concentration for the initial conditions (cross-validation), x: predicted concentrations for the new conditions.

analyst sufficient information to decide whether to continue using the model, apply a standardization procedure or construct the multivariate regression model again.

This methodology can also be used for other related purposes such as validating the results obtained from the multivariate model after applying a standardization procedure. In this latter case, the test results would enable us to validate the standardization itself.

It should be pointed out that the statistical tests that allow the validity of the model to be decided on are applied to a reduced number of samples, a fundamental aspect from the point of view of cost because multivariate models are generally constructed from a great deal of samples. Selecting this selected sample set may be a good starting point for the process of standardizing the model, should this stage be necessary.

Acknowledgements

We would like to acknowledge the economic support from the Spanish Ministry of Education and Science (DGICyT project BP93-0366). J Ferré would like to thank the Comissionat per a Universitats i Recerca of the Generalitat de Catalunya for providing a doctoral fellowship (FI/94-7001).

References

1. F. McLennan and B. R. Kowalski, Process Analytical Chemistry, 1st edn., Chapman & Hall, London, New York, 1995.
2. Onno E. de Noord, Chemom. Intell. Lab. Syst., 25 (1994) 85-97.
3. Y. Wang, D. J. Veltkamp and B. R. Kowalski, Anal. Chem., 63 (1991) 2750-2756.
4. J. S. Shenk, M. O. Westerhaus and W. C. Templeton, Jr., Crop Science, 25 (1985) 159-161.
5. J. S. Shenk and M. O. Westerhaus, Crop Science, 31 (1991) 1694-1696.
6. Y. Wang and B. R. Kowalski, Appl. Spectrosc., 46 (1992) 764-771.
7. Y. Wang and B. R. Kowalski, Anal. Chem., 65 (1993) 1301-1303.
8. M. Forina, G. Drava, C. Armanino, R. Boggia, S. Lanteri, R. Leardi, P. Corti, P. Conti, R. Giangiacomo, C. Galliena, R. Bigoni, I. Quartari, C. Serra, D. Ferri, O. Leoni and L. Lazzeri, Chemom. Intelll. Lab. Syst., 27 (1995) 189-203.
9. E. Bouveresse, D. L. Massart and P. Dardenne, Anal. Chem., 67 (1995) 1381-1389.
10. Y. Wang, M. J. Lysaght and B. R. Kowalski, Anal. Chem., 64 (1992) 562-564.
11. Z. Wang, T. Dean and B. R. Kowalski, Anal. Chem., 67 (1995) 2379-2385.
12. Y. Wang and B. R. Kowalski, Anal. Chem., 65 (1993) 1174-1180.
13. E. Bouveresse and D. L. Massart, Chemom. Intell. Lab. Syst, 32 (1996) 201.
14. E. Bouveresse and D. L. Massart, Vibrational Spectrosc., 11 (1996) 3-15.
15. N. Draper and H. Smith, Applied Regression Analysis, 2nd edn., Wiley-Interscience, New York, 1981.
16. Mandel, J., Linnig, F. J., Anal. Chem., 29 (1957) 743-749.
17. Atkinson, A.C. ; Donev.A.N. Optimum Experimental Designs, Oxford Statistical Science Publications, Oxford, 1992
18. V.V. Fedorov, in W.J. Studden and E.M. Klimko (Eds.), Theory of Optimal Experiments, Academic Press, New York, 1972

19. D. Mathieu, Contribution de la Méthodologie de la Recherche Experimentale à l'étude des relations Structure-Activité, Thèse Sciences, Marseille, 1981
20. Ryuei Nishii, Discrete Mathematics, 116 (1993) 209-225.
21. D.V. Steinberg and W.G. Hunter, Technometrics, 26 (1984) 71-130.
22. A. Rius, M. P. Callao and F. X. Rius, Anal. Chim. Acta, 316 (1995) 27-37.
23. A. Rius, M. P. Callao and F. X. Rius, Non-published results.
24. MATLAB, The Mathworks, South Natick, MA, USA.
25. Statistique appliquée a l'exploitation des mesures, 2e édition, Cetama, Masson, 1986.

Chapter 4

Wavelength Selection in Multivariate Calibration Models

4.1 Introduction

4.1.1 Aim of the chapter

The aim of this chapter is the study of the wavelength selection criteria in CLS. First, the concepts of the experimental design are used for interpreting the effect of the criteria used in the literature on the confidence interval of the predicted concentrations. Guidelines are given to improve trueness and precision of the analytical results by selecting the calibration wavelengths used in the multivariate calibration model. A second aim is to propose a new methodology for detecting the presence of bias in unknown samples in CLS and select the best subset of wavelengths for quantification.

4.1.2 Structure of the chapter

The introduction (§4.1) contains the aim of the chapter (§4.1.1), its structure (§4.1.2)and a bibliographic revision (§4.1.3) of the different approaches and criteria for wavelength selection used mainly in CLS. The following sections, §4.2 to §4.7, contain the main contents dealing with wavelength selection in CLS structured in papers. The theoretical part of CLS was considered in the section §2.2.4.

§4.2 is the paper *A graphical criterion to examine the quality of multicomponent analysis. Implications for wavelength selection.* J. Ferré and F.X. Rius *Trends Anal. Chem.* 16 (1997) 155-162. Here, the ideas of the experimental design are used to interpret wavelength selection criteria in CLS on the basis of their effect on the volume, shape and orientation of the confidence region of the predicted concentrations. This confidence region is an (hyper)ellipsoid in the *K*-dimensional space of the concentrations and can be used as a graphical criterion. New guidelines for wavelength selection in multicomponent analysis are given. These ideas are applied to the wavelength selection in a mixture of chlorophenols.

§4.3 is the paper *Further considerations on the sensitivity and selectivity of multicomponent*

systems. J. Ferré and F.X. Rius . *In preparation*. This paper complements the section §4.2. The effect of sensitivity, selectivity, variance proportion decompositions and condition number on the precision of the estimated concentrations are interpreted from the confidence ellipsoid. The effect on these measures of adding a new sensor to the calibration matrix confidence is considered. Simulations confirm the discussed effects.

§4.4 is the paper *Equivalence between Selectivity and Variance Inflation factors in multicomponent analysis*. J.Ferré, F.X. Rius *Química Analítica* 15 (1996) 259-262, which demonstrates the mathematical equivalence between selectivity and variance inflation factors. Both can be used as measures of collinearity in CLS.

§4.5 is a tutorial where the definitions of accuracy, trueness and precision and their relationship with the wavelength selection criteria in CLS are revisited. This is motivated by the confusion in the literature about the effect of these criteria on the accuracy, trueness and precision of the results.

§4.6 is the paper *Figures of Merit in Multivariate Calibration. Determination of Four Pesticides in Water by FIA and Spectrophotometric Detection.* J. Ferré, R. Boqué, B. Fernández-Band, M.S. Larrechi and F.X. Rius. *Anal. Chim. Acta* 348 (1997) 167-175. The variance proportion decompositions and the effects of the selectivity and sensitivity in CLS are studied in the analysis of four pesticides with a FIA system. The part referred to the detection limits in this paper is not considered a part of this thesis, it has been included here since it is a part of the published paper.

§4.7 is the paper *Detection and correction of biased results of individual analytes in multicomponent spectroscopic analysis* J.Ferré, F.X. Rius. *Submitted for publication.* This work is motivated by the fact that large selectivity and sensitivity values of the analytes in the paper in §4.6 did not agree with the prediction errors observed. It was supposed that the large errors were not due to the instability of the system due to collinearity but to an erroneous preparation of the validation samples, with a deficient assigned value of the concentration. In the present paper, a tool for internal validation of the standards and of the validation samples is developed based on the net analyte signal. It enables the bias in CLS models to be detected taking advantage of the multivariate signal. A wavelength selection procedure and the error indicator presented to select the wavelength with less prediction error.

4.1.3 Bibliographic revision and comments

4.1.3.1 Using all the recorded spectrum in multivariate calibration

When setting up a multivariate calibration model, the simplest option is to use the entire spectra, easily collected by modern analytical instruments (such as photodiode arrays detectors), and calibration approaches that can deal with over-determined systems [1-7]. Full-spectrum methods such as PLS and PCR enable to build models with little or no knowledge of the spectra of the constituents of interest. These methods are able to discern, from the information contained in the spectra of the training set, the spectral regions that are most important for calibration. Moreover, it is generally agreed that using a large number of wavelengths has an error-averaging effect that is beneficial for the accuracy and precision and makes more robust calibrations [8-10]. Lorber and Kowalski[11] gave a mathematical proof that increasing the number of sensors improves the prediction performance of multivariate calibration models[7,12,13]. The proof was based on a expression where the prediction error is determined by the norm of the regression coefficients. This norm decreases when wavelengths are added, which leads to the conclusion that the addition of wavelengths always improves the quality of the prediction.

4.1.3.2 Reasons for wavelength selection in multivariate calibration

The range of wavelengths at which the absorbances are measured influence the trueness and precision of the spectrophotometrically determined concentrations. Wavelength selection is important even in methods capable of using a very large number of measurements. The performance of the calibration model can be improved by excluding the spectral regions that do not contain information correlated with the concentration of the constituents of interest (e.g. where either the detector, the spectrometer source or the optics are not effective, where none of the constituents absorb, or regions in the spectrum that deviate from Beer's Law or where impurities expected in unknown samples absorb appreciably). The inclusion

of these non-informative spectral measurements can seriously degrade performance[3].

Wavelength selection is also concerned in physical and economical constraints of the measurement apparatus where a reduced number of wavelengths for measurement must be used (e.g. compact on-line analyzers in an industrial process). In such cases, the accuracy of the quantitatively analyzed data largely depends on the selection of wavelengths. Sometimes the prediction ability of the models can be sacrificed at expense of a reduction of the number of sensors and thus of the cost of the instrumentation. In such cases, the sensors must be carefully selected so that the prediction ability of the model is degraded the minimum.

Wavelength selection is also of special concern to avoid collinearity in inverse least-squares modeling, where the samples are characterized by many variables (spectral absorbances).

The results presented by several authors[14,15,] evidence that calibration employing selected wavelength regions rather than the entire spectrum improves the predictive ability of CLS [16-23], weighted least squares (WLS)[24], ILS [22,25,26], PCR [13,22,27] and PLSR[12,22,25,28-31] models. Gemperline[32] reviewed some of the work on wavelength selection before 1989. In CLS, Frans and Harris[16] found that replicate measurements at a smallest number of selected wavelengths improves concentration precision with respect the acquisition of a complete spectra. Rossi and Pardue[18] used an empirical selection of wavelengths to reduce the spectral overlap and improve the accuracy of predictions. Liang *et al.*[23] showed that larger number of wavelengths do not always lead to higher precision of analytical results. In some cases, the use of a large number of wavelengths in PCR and PLS provides more noise than accuracy in the analytical results[22]. Using a reduced number of wavelengths that carry most information may give PLS models easier to interpret, with fewer factors and with better precision of the predictions, specially if very noisy wavelengths or with irrelevant information or non-linearity are avoided. Garrido Frenich *et al.* [29]found wavelength ranges that yield superior or equal prediction results to those obtained for the whole wavelength range in PLS. Jouan-*Rimbaud et al.*[33] presented results that feature selection with NIR data can improve the performance of ILS, PCR and PLS. Navarro et al.[22] made a comparative study to select calibration mixtures and wavelengths in CLS, ILS, PLS,PCR and Kalman filter. For ILS, PCR and PLS, the best

results in prediction were obtained with ranges of wavelengths selected with the condition number of the calibration matrix. Recently, Xu and Schechter[13] showed that the assumptions that lead to Lorber and Kowalski's conclusion[11] (see §4.1.3.1) are not always valid and demonstrated by both experimental and theoretical considerations that better results can be obtained by a proper selection of the spectral range used in simultaneous multicomponent analysis and that wavelength selection is essential when considerable spectral overlap exists.

Very recently, Faber and Kowalski[34] commented the wavelength and sample selection problem in ILS, PCR and PLS. They concluded, considering errors in the dependent and independent variables that, contrary to the stated by Lorber and Kowalski[11], a large size of the regression vector is unfavorable. Wavelength selection is potentially favorable, for example, when some wavelengths are noisier than others, or wavelengths that do not relate to the concentration in a linear way. They also proposed that wavelength selection must be guided by a prediction criterion.

4.1.3.3 The wavelength selection problem

An important problem in multivariate calibration is selecting at which wavelengths the sample should be measured to obtain an unbiased and precise prediction Selecting a representative wavelength set for calibration is not trivial. When the number of wavelengths in the analytical process is not predetermined, the analyst has to decide the optimal wavelengths to use. This is often an empirical and subjective choice[18,28,35] based on the chemical and spectroscopic knowledge of the samples, the visual inspection of available spectra and considering tabulated spectral bands of the constituents of interest. However the use of the spectroscopic knowledge may miss the optimal subset of wavelengths (OSW). Several methods have been proposed in the literature to aid the experimenter in selecting the best regions, mainly for multicomponent spectrophotometric determinations (CLS) [13,16,19-21,36-40] although also for weighted least squares (WLS)[24], ILS[25,41] or PCR[13] and PLSR[12,25,29,42].

The wavelength selection problem consists of selecting, according to some predefined criterion, the optimal subset of wavelengths among all the wavelengths of the spectrum. Requirements for the general strategy of selection are indicated in

§2.3.4. The optimization criterion to evaluate each candidate subset, the optimization procedure and the quantitative comparison of optimal subsets with a different number of wavelengths are considered below. Few methods in the bibliography use all these requirements. The most complete[12,19,38] use search-based strategies in which combinations of wavelengths in all the spectral range are evaluated with the selection criteria. In many occasions, however, some regions of the spectra are selected empirically and used to evaluate the correlation of a determined criterion with the prediction errors[17,28]. This latter are not optimal optimization procedures, since they do not guarantee that the optimal wavelength range is used.

4.1.3.3.1 Criteria for wavelength selection

Searching the OSW is a sequence of improving steps (the candidate subsets of wavelengths) with the aim of achieving the global optimal. A decision criterion (also called *objective function* or *optimality criterion*) evaluates the quality improvement of the candidate subsets through the optimization. Different criteria have been described in the literature and used either for searching the set that optimized the criterion or just for checking the relationship between the criteria and the actual prediction errors for already selected spectral ranges.

Optimality criteria in CLS. They are mainly related to properties of the calibration matrix **S** and include:

1. Minimizing the sum of the diagonal elements of the variance-covariance matrix (the variance factors) [16], which is $Tr(\mathbf{S}^T\mathbf{S})^{-1}$
2. Minimizing the mean square error between the true concentrations of the mixture components and their estimates [20,23,37]. This criterion is $Tr(\mathbf{S}^T\mathbf{S})^{-1}$ if noise obeys an uncorrelated process with zero mean and a constant variance.
3. Minimizing the condition number of the calibration matrix [17, 21-23, 35,43-47]
4. Maximizing $Det(\mathbf{S}^T\mathbf{S})$ [21,23,35,46]
5. Optimizing figures of merit derived by Lorber and coworkers [48,(49)] such as maximizing the selectivity [17,37,43,50], the norm of the net analyte signal[11,17,43] and accuracy (ACC)[17,19,37]or minimizing the limit of detection (LOD)[17]or the total error propagation (TEP) [17].

The most preferred criteria for prediction of the best wavelength combination are the condition number the determinant of the calibration matrix (sometimes considered as measures of matrix orthogonality [12]). Other optimality criteria or selection approaches not considered above are:

1. Maximizing the signal-to-noise (S/N) ratio[22] (ILS, PCR, PLS). Salamin *et al.*[24] selected wavelengths for which the ratio of the absorbance to the standard deviation of the error of measurement was above some optimal value. This optimal value is chosen such as to maximize the 'quality' of the estimate of concentration.
2. Minimizing the condition number in ILS [22] or PCR/PLS [22,28]
3. Selecting the wavelengths with high loadings on some selected PCs[33]
4. Maximizing the prediction error in ILS [25,41,53] or PCR/PLS [12,25] measured as predicted residual error sum of squares (PRESS).
5. Selection of wavelengths with the largest correlation coefficient with the concentration[33].
6. The covariance between each independent variable and the concentration[33].
7. Identification of the individual wavelengths based on the linear regression between the concentration and the absorbances at individual wavelengths[14].
8. Error indicator function[13] developed to predict the performance under given experimental conditions, using a certain spectral range. This function is applied for the location of the most informative spectral ranges to be utilized in multicomponent analysis.
9. A feature selection method based on the regression coefficients of the closed form of the PLS model[29].
10. Selection of wavelengths to be used in the monitoring of absorbance ratiograms during liquid chromatographic separations applying the key set factor analysis (KSFA) [51]. Warren *et al.*[40] compared four alternatives for the selection of representative wavelengths sets from a collection of 101 UV-vis spectra. The key set factor analysis (KSFA) technique provided the best overall performance.
11. Maximum differences between molar absorptivities[36].

Although the researchers agree in the mathematical expressions of the criteria, their considerations about the relationship of these criteria with the precision and accuracy of the analytical result is sometimes different and even difficult to understand since they do not indicate what they consider as *precision* and *accuracy*.

ISBN:978-84-690-8034-1

The modern concept of accuracy (trueness and precision) is rarely used. Moreover, different conclusions about the adequacy or not of the criteria to improve the prediction results can be found in the literature so it is really difficult to evaluate which criterion is the best.

4.1.3.3.2 Optimization procedures for wavelength selection

Due to its complexity, the problem of determining the OSW cannot be solved in an analytical way. The space of all wavelength subsets is discrete, so the optimization problem has a combinatorial nature. Search-based strategies based on optimization algorithms are required to avoid an exhaustive search of all possible combinations of wavelengths of a given number. The optimization methods most used in the literature include:

1. Simplex optimization[19].
2. Branch and bound combinatorial optimization techniques [20,23].
3. Stepwise selection [11,21]. Although the forward stepwise selection has given optimal wavelength sequences[23], this technique may be unsatisfactory[41,44] since wavelengths are selected one at a time and may miss the optimal wavelength set. So that the resulting models may have a poor predictive ability[41]. For multivariate calibration, approaches that select the whole subset at the same time might be preferable.
4. Genetic algorithms (GAs)[52]. GAs have been used in feature variable selection problems [37,25]. They are able to find acceptable solutions in a reasonable time either in CLS[37,44], ILS[25,41], PCR [27,53] or in PLS regression[12,25,31,42].
5. generalized simulated annealing (GSA)[19,38,50].

Lucasius *et al.*[37] studied comparatively the ability of GSA, GA and stepwise elimination to locate OSW for quantitative multicomponent analysis. They found that the GA generally performed the best and GSA performed worst. Later, Hörchner and Kalivas[38] showed that GSA is also able to find the optimal solutions optimizing the same criteria employed by Lucasius *et al.*[37] but a more adequate operating configuration of GSA for the wavelength selection optimization.

It is clear that the main ideas about wavelength selection criteria, precision and accuracy must be clarified. In addition, many of these criteria do not consider the possibility of bias introduced by an unknown sample. In this way, even the optimal subset of wavelengths can produce wrong predictions if the spectrum of the unknown sample does not follow the model postulated. This is dealt in the sections §4.2 to §4.7

4.1.4 References

1. Martens H.; Naes T. *Multivariate Calibration*, Wiley: New York, 1989.
2. Kowalski, BR.; Seasholtz M.B. *J. Chemom.* 5 (1991)129-145.
3. Thomas E.V. *Anal. Chem.* 66 (1994) 795A-804A.
4. Geladi P.,Kowalski B.R. *Anal Chim Acta* 185 (1986) 1-17.
5. Beebe, K.R.; Kowalski B.R. *Anal. Chem.* 59 (1987) 1007A-1017A.
6. Haaland D.M., Thomas E.V. *Anal. Chem.* 60 (1988) 1193-1202.
7. Booksh K.S. , Kowalski B.R *Anal. Chem.* 66 (1994) 782A-804A.
8. Kisner H.J., Brown C.W., Kavarnos G.J. *Anal. Chem.* 55 (1983) 1703-1707.
9. Otto M., Thomas J.D.R. *Anal. Chem.* 1985 (57) 2647-2651.
10. Zscheile Jr., F.P; Murray, H.C; Baker, G.A; Peddicord, R.G *Anal. Chem.* 34 (1962) 1776-1780.
11. Lorber A. , Kowalski B.R. *J. Chemom.* 2 (1988) 67-79.
12. Bangalore A.S., Shaffer R.E., Small G.W., Arnold M.A. *Anal. Chem.* 68 (1996) 4200-4212.
13. Liang Xu , Israel Schechter *Anal. Chem.* 68 (1996) 2392-2400.
14. Brown P.J. *J. Chemom.* 6 (1992) 151-161.
15. Baroni M., Clementi S., Cruciani G., Costantino G., Riganelli D. *J. Chemom.* 6 (1992) 347-356.
16. Frans S.D., Harris J.M., *Anal. Chem.* 57 (1985) 2680-2684.
17. Juhl L.L , Kalivas J.H. *Anal. Chim. Acta* 207 (1988) 125-135.
18. Rossi D.T., Pardue H.L. *Anal. Chim. Acta* 175 (1985) 153-161.
19. Kalivas J.H., Roberts N., Sutter J.M. *Anal. Chem.* 61 (1989) 2024-2030.
20. Sasaki K, Kawata S., Minami S. *Appl. Spectrosc.* 40 (1986) 185-190.

21. Smeyers-Verbeke J., Detaevernier M.R., Massat D.L. *Anal. Chim. Acta* 191 (1986) 181-192.
22. Navarro-Villoslada F.; Pérez-Arribas, L.V.; León-González M.E.; Polo-Díez L. M. *Anal. Chim. Acta* 313 (1995) 93-101.
23. Liang Y., Xie Y ., Yu R. *Anal. Chim. Acta* 222 (1989) 347-357.
24. Salamin P.A., Bartels H., Foster P. *Chem. Intell. Lab. Syst.* 11 (1991) 57-62.
25. Leardi R. *J. Chemom.* 8 (1994) 65-79.
26. Brown C.W., Lynch P.F., Obremski R.J , Lavery D.S. *Anal. Chem.* 54 (1982) 1472-1479.
27. Li T., Lucasius, C.B; Kateman, G. *Anal. Chim. Acta* 268 (1992) 123-134.
28. Otto M., George T. *Anal. Chim. Acta* 200 (1987) 379-385.
29. Garrido Frenich A., Jouan-Rimbaud D., Massart D.L., Kuttatharmmakul S., Martínez Galera M., Martínez Vidal J.L. *Analyst* 120 (1995) 2787-2792.
30. Lindgren F., Geladi P., Rännar S., Wold S. *J. Chemom.* 8 (1994) 349-363.
31. Shaffer R.E. Small, G.W. Arnold M.A. *Anal. Chem.* 68 (1996) 2663-2675.
32. Gemperline P.J. *J. Chemom.* 3 (1989) 549-568.
33. Jouan-Rimbaud D., Walczak B., Massart D.L., Last I.R., Prebble K.A. *Anal. Chim. Acta* 304 (1995) 285-295.
34. Faber K., Kowalski B.R. *J. Chemom.* 11 (1997) 181-238.
35. Jochum C., Jochum P., Kowalski B.R., *Anal. Chem.* 53 (1981) 85-92.
36. Costadinnova L., Nedeltcheva T. *Analyst* 120 (1995) 2217-2220.
37. Lucasius, C.B; Beckers M.L.M; Kateman, G. *Anal. Chim. Acta* 286 (1994) 135-153.
38. Hörchner U., Kalivas J.H. *Anal. Chim. Acta.* 311 (1995) 1-13.
39. Bergmann G., von Oepen B., Zinn P. *Anal. Chem.* 59 (1987) 2522-2526.
40. Warren F.V. Bidlingmeyer B.A., Delaney M.F. *Anal. Chem.* 59 (1987) 1890-1896.
41. Jouan-Rimbaud D., Massart D.L., Leardi R., De Noord O.E. *Anal. Chem.* 67 (1995) 4295-4301.
42. Arcos M.J., Ortiz M.C., Villahoz B., Sarabia L.A. *Anal. Chim. Acta* 339 (1997) 63-77.
43. Bauer, G.; Wegscheider, W.; Ortner, H.M. *Spectrochimica Acta 47B* (1992) 179-188.
44. Lucasius, C.B; Kateman, G. *Trends Anal. Chem.* 10 (1991) 254-260.
45. Otto M., Wegscheider W. *Anal. Chem.*57 (1985) 63-69.
46. Otto M., Wegscheider W. *Anal. Chim. Acta* 180 (1986) 445-456.
47. Pérez-Arribas, L.V.; Navarro-Villoslada F.; León-González M.E.; Polo-Díez L. M. *J. Chemom.* 7 (1993) 267-275.
48. Lorber A *Anal. Chem.* 58 (1986) 1167-1172.

49. Lorber A., Harel A., Goldbart Z., Brenner. B. *Anal. Chem.* 59 (1987) 1260-1266.
50. Hörchner, U.; Kalivas J.H. *J. Chemom.* 9 (1995) 283-308.
51. Warren F.V. Bidlingmeyer B.A., Delaney M.F. *Anal. Chem.* 59 (1987) 1897-1907.
52. Leardi R., Boggia R., Terrile M. *J. Chemom.* 6 (1992) 267-281.
53. Tong-Hua Li, Lucasius C.B., Kateman G. *Anal. Chim. Acta* 268 (1992) 123-134.

4.2 A Graphical criterion to examine the quality of multicomponent analysis. Implications for wavelength selection

Trends Anal. Chem. 16 (1997) 155-162

Joan Ferré*, F.Xavier Rius

Departament de Química. Universitat Rovira i Virgili.
Pl. Imperial Tarraco, 1, 43005-Tarragona. SPAIN

A graphical criterion approach has been developed to examine the quality of a set of sensors in multicomponent analysis. Criteria such as sensitivity and selectivity, used in wavelength selection problems can be explained in terms of the confidence interval of the estimated concentrations. These confidence intervals describe (hyper)ellipsoids whose volume, shape and orientation are related to the optimiztion criteria. The effect of sensor selection on these criteria is discussed and guidelines for wavelength selection are given. The usefulness of the graphical criterion is shown in the simultaneous determination of 2-chlorophenol and 2,4-dichlorophenol in water.

1.Introduction

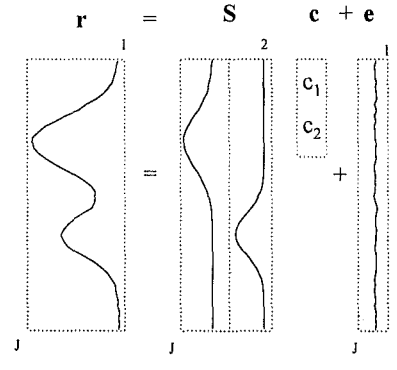
Spectroscopic multicomponent analysis consists of determining the concentrations of the K components in a mixture ($\mathbf{c}_{true}$) from the spectra of the mixture measured at J wavelengths ($\mathbf{r}$) and the calculated molar absorption coefficients of the pure components ($\mathbf{S}$). The linear additive model, based on Beer's law and described by Eq. (1), is often used:

$$\mathbf{r}_{J\times 1} = \mathbf{S}_{J\times K}\,\mathbf{c}_{true} + \mathbf{e}_{J\times 1} \qquad (1)$$

where the subscripts indicate the dimensions of the matrices and $\mathbf{e}_{J\times 1}$ is a vector of error terms. This model is represented in Scheme 1 for a K=2 component system.The least squares estimation of $\mathbf{c}_{true}$ can be obtained using Eq. (2):

$$\mathbf{c} = (\mathbf{S}^T\mathbf{S})^{-1}\mathbf{S}^T\,\mathbf{r} \qquad (2)$$

where the superscript ^{T} indicates a transposition. Eq. (1) can be evaluated rigorously only if all elements that contribute to the signals in the spectral regions investigated are known. This model is adequate to analyse samples prepared synthetically or artificially such as the quality control of pharmaceuticals.



Scheme 1. Simulated spectra for a two-component system and multicomponent analysis formulation.

Wavelength (also called 'sensor') selection has been reported to improve the prediction ability of the model in Eq. (1). Since the spectra in Eq. (1) are column vectors, selecting sensors corresponds to selecting rows of the $\mathbf{S}$ and $\mathbf{r}$ matrices. Sensors have been selected to optimize criteria such as selectivity [1], the condition number of $\mathbf{S}^T\mathbf{S}$ [2], the determinant of $\mathbf{S}^T\mathbf{S}$ [2] , or minimize the mean squared error (MSE) [2-4], among others [5-6]. The large number of studies available, which sometimes draw different conclusions about the performance of these criteria, makes the situation rather confusing for users who

want to select sensors by applying these definitions to experimental data. The actual merit of certain criteria might not be directly obvious, and very often the optimal number of sensors is not clearly indicated by the criterion used.

Here we provide a means to interpret, from the point of view of experimental design theory, the relationship between the most commonly used criteria by considering the confidence region of the estimated concentrations. This region defines, in the *K*-dimensional space of the concentrations, a (hyper)ellipsoid that can be used as a graphical criterion to better understand the relationship between these criteria and gives a better description of the guidelines for sensor selection.

2. Theoretical background

2.1 The confidence hyperellipsoid

The estimated analyte concentrations in the unknown sample from Eq. (2) are random quantities. The bounds of the 100(1-α)% confidence region where the true values of concentrations ($\mathbf{c}_{true}$) lie with α probability of committing a type I error are described by Eq. (3) [7]:

$$(\mathbf{c}_{true} - \mathbf{c})^T \mathbf{S}^T\mathbf{S} (\mathbf{c}_{true} - \mathbf{c}) = p^2 \qquad (3)$$

where $p^2 = Ks^2F_{K,J-K,1-\alpha}$, s^2 being is an estimate of the variance of instrumental response errors on *J-K* degrees of freedom and $F_{K,J-K,1-\alpha}$ is the α per cent point of the F distribution on *K* and *J-K* degrees of freedom. In the *K*-dimensional space of the concentrations, Eq. (3) defines a (hyper)ellipsoid centred on **c**. The length of the half-axes is equal to $p\sqrt{\lambda_j}$ where λ_j are the eigenvalues of the matrix $(\mathbf{S}^T\mathbf{S})^{-1}$. The extreme confidence intervals (the ends of the confidence ellipsoid) are given by Eq. (4) [8]:

$$c_k - p\sqrt{(UVIF_k)} \leq c_{true,k} \leq c_k + p\sqrt{(UVIF_k)} \qquad (4)$$

where $UVIF_k$ is the *k*th element in the diagonal of $(\mathbf{S}^T\mathbf{S})^{-1}$. An example of a confidence ellipse for a two-component system is given in Fig. 1 where the

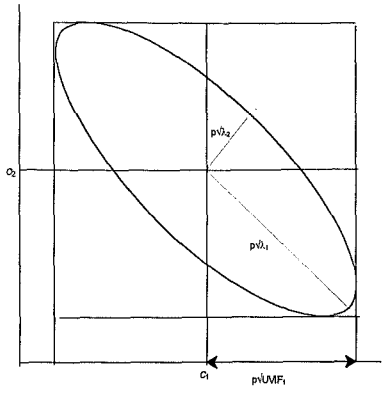


Fig. 1. Confidence ellipsoid for a two-component system. The individual confidence intervals are shown.

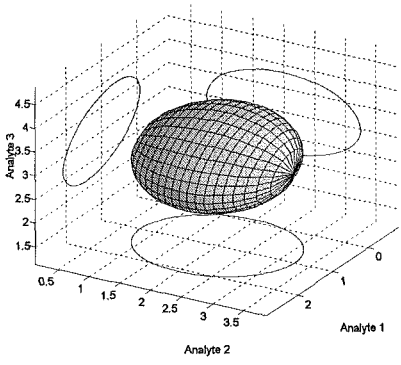


Fig. 2. Confidence ellipsoid for a three-component system.

geometrical characteristics are shown. Fig. 2 shows the ellipsoid of a three component determination. For easier interpretation, only two components are considered below and the corresponding ellipses are plotted by considering p=1 in Eq. (3).

2.2 *Volume, shape and orientation of the confidence ellipsoid*

The volume, shape and orientation of the ellipsoid are of interest if the analyte concentrations are to be correctly estimated.

2.2.1 *Volume*

For a given α, the smaller the volume of the ellipsoid, the more globally precise are the concentrations estimated. As the volume is proportional to the product of the ellipsoid axes (thus proportional to $[\mathrm{Det}(\mathbf{S}^T\mathbf{S})]^{-1/2}$), the highest global precision is obtained when the J sensors in $\mathbf{S}$ maximise $\mathrm{Det}(\mathbf{S}^T\mathbf{S})$. This is known as the D-optimality criterion in experimental design theory.

2.2.2 *Shape*

The shape is related to the length of the ellipsoid axes, and thus to $Tr(\mathbf{S}^T\mathbf{S})^{-1}$. For a given volume of the confidence region, the minimum (optimal) value of $Tr(\mathbf{S}^T\mathbf{S})^{-1}$ is attained when the length of the ellipsoid axes is the same, i.e. when a hypersphere is obtained. This is known as the A-optimality criterion. This criterion is equivalent to minimizing the maximum eigenvalue of $(\mathbf{S}^T\mathbf{S})^{-1}$ (E-optimality) which makes the eigenvalues as similar as possible [9].

2.2.3 *Orientation*

The orientation depends on the overlap of the pure component spectra. It can be described by multicollinearity measures for each analyte such as the variance inflation factors (VIF_k) [9] and Lorber's selectivity ($LSEL_k$) [10-11] or the correlation matrix of the variables. Lorber's selectivity for the kth analyte is the ratio between the Euclidean norm of the part of the analyte's spectrum that is actually used for quantification (the 'net analyte signal') and the Euclidean norm of the analyte's spectrum. When there is no spectral overlap, the spectra are said to be orthogonal and $VIF_k=LSEL_k=1$, and the axes of the ellipsoid are parallel to the axes defined by the kth component concentration and the concentrations are independently estimated. With increasing spectral overlap, $LSEL_k$ decreases, VIF_k becomes larger and the inclination of the ellipsoid approaches 45°, so increasing the correlation of the estimated concentrations. A maximum value of $VIF_k=10$ has been proposed by several authors [12] to prevent large errors from being obtained in the predicted concentrations. To obtain independent concentration estimates, the J selected sensors in $\mathbf{S}$ should be nearer to $LSEL_k=1$ than any other set of J sensors.

2.3 *Effect of wavelength selection criteria on the confidence region*

Of all the sets of J sensors, the optimal set for calibration is the one that simultaneously maximizes $Det(\mathbf{S}^T\mathbf{S})$, minimizes $Tr(\mathbf{S}^T\mathbf{S})^{-1}$ or the largest eigenvalue of $(\mathbf{S}^T\mathbf{S})^{-1}$ and maximizes $LSEL_k$. The optimisation of one criterion alone does not necessarily guarantee very precisely estimated concentrations. It has been found that the expected improvement in the prediction errors of the concentrations determined

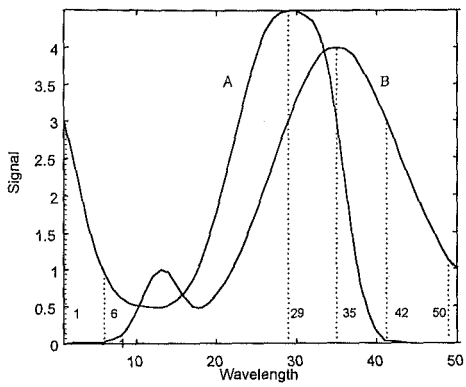


Fig. 3. Simulated absorbance spectra for a system with two components. The spectra incorporates sensors with both high and low selectivities and sensitivities. Matrices S_1 , S_2 , S_3 and S_4 are composed of the absorbance values measured at sensors 6 and 50, 1 and 42, 6 and 42 and 29 and 35, respectively.

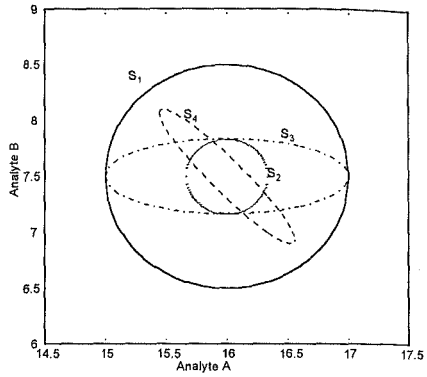


Fig. 4. Ellipses corresponding to four different pairs of wavelengths selected from Fig. 3 and listed in Table 1.

Table 1. Individual absorbance values for four different pairs of selected wavelengths for analytes A and B in Fig. 3

	S_1		S_2			S_3			S_4		
selectivity	high		high			high			low		
sensitivity	low		high			low/ high			high		
sensor	A	B	sensor	A	B	sensor	A	B	sensor	A	B
6	1	0	1	3	0	6	1	0	29	4.5	3.0
50	0	1	42	0	3	42	0	3	35	3.0	4.0

using a particular set of wavelengths is not only correlated to the improvement in $LSEL_k$ [13] or $Det(\mathbf{S}^T\mathbf{S})$ [14]. The performance of these selection criteria can be interpreted from the confidence ellipsoid.

Fig. 3 shows the simulated absorbance spectra for a system with two components. The spectra incorporate sensors with both high and low selectivities and sensitivities and Fig. 4 shows the ellipses of four different pairs of wavelengths selected from Fig. 3 and listed in Table 1. The ellipses are arbitrarily centred at $\mathbf{c}^T=(16,7.5)$.

Matrices S_1 - S_3 are made up of pairs of absorbance values which are measured at completely selective sensors with different sensitivity values (Table 1). The two sensors 1 and 42 (the absorbance values of which are the components of matrix S_2) would probably have been selected by an experienced analyst because of their high selectivity and sensitivity. S_4 corresponds to the sensors of maximal absorbance for the two analytes (maximum sensitivity). Table 2 lists, for each selected subset, the Det($\mathbf{S}^T\mathbf{S}$), Tr($\mathbf{S}^T\mathbf{S}$)$^{-1}$, $UVIF_k$, $LSEN_k$, $LSEL_k$ and VIF_k values for the two analytes (A and B). $LSEN_k$ is a global measure of the absorbance of kth component at the selected wavelengths and was calculated as the Euclidean norm of each column in **S**, and $LSEL_k$ and VIF_k according to the expressions given in Ref. [15].

Table 2. Values of several wavelength selection criteria for the analytes A and B in matrices from Table 1.

Selected sensor	Det	Tr	UVIF		LSEN		LSEL		VIF	
			A	B	A	B	A	B	A	B
S_1	1	2	1	1	1	1	1	1	1	1
S_2	81	0.22	0.11	0.11	3	3	1	1	1	1
S_3	9	1.11	1	0.11	1	3	1	1	1	1
S_4	81	0.67	0.31	0.36	5.41	5.00	0.33	0.33	9.03	9.03

2.3.1 *Volume*

The largest volume corresponds to set S_1, since it has the lowest determinant value. Sets S_2 and S_4 have the highest determinant and thus the confidence ellipses have the smallest volume. Both these sets have Det($\mathbf{S}^T\mathbf{S}$)=81 so other criteria other than D-optimality should be considered to decide between them. It can be observed in the ellipse that the length of the confidence intervals for S_4 is larger than for S_2. This is due to spectral overlap. The comparison of the ellipses from sets S_2, S_3 and S_4 with S_1, shows that the volume can be minimised in two ways: either by having a sphere of smaller radius (set S_2) or by increasing the eccentricity of the ellipse. If orthogonal sensors are selected, the 'flattening' is done parallel to the axes (S_3) and independent estimations are still obtained but with a different precision. If sensors with spectral overlap are selected (set S_4), the 'flattening' is done diagonally to the axes and the covariance of the estimated concentrations increases. Sets S_2 -S_4 are examples of sets

which have the same value of Det($\mathbf{S}^T\mathbf{S}$) but different shape, due to different sensitivity and selectivity.

2.3.2 *Shape*

According to the trace criterion, S_3 is preferable to S_1 since this has a smaller $Tr(\mathbf{S}^T\mathbf{S})^{-1}$ value. This can be graphically interpreted from the shorter length of the second ellipsoid axis. But S_4 still has a smaller $Tr(\mathbf{S}^T\mathbf{S})^{-1}$ value and according to this criterion should be preferred. Although this latter set provides a smaller confidence interval for analyte A, this is not true for analyte B, which is larger. Should analyte B be predicted with a small variance, S_3 should be preferred to S_4 despite having a larger $Tr(\mathbf{S}^T\mathbf{S})^{-1}$ value. This criterion can be used to select set S_2 instead of S_4, the two sets with the same Det($\mathbf{S}^T\mathbf{S}$) value. The large eccentricity of the ellipse in S_4 , due to multicollinearity effects, makes one eigenvalue much larger than the other and thus $Tr(\mathbf{S}^T\mathbf{S})^{-1}$ large.

2.3.3 *Orientation*

Sets S_1,S_2 and S_3 have orthogonal sensors (VIF_k=$LSEL_k$=1) so that $(\mathbf{S}^T\mathbf{S})^{-1}$ is diagonal and the axes of the ellipses are parallel to the co-ordinate axes. The smaller volume in S_2 is due to the added effect of high sensitivity. Note that S_3 estimates the precisions differently, despite having $LSEL_k$=1 as S_1 does. In such cases, the criterion $LSEL_k$ does not identify which set of sensors (or even which number of them) is preferable.

3. Determining 2-chlorophenol and 2,4-dichlorophenol in water

3.1 *Equipment*

A Hewlett-Packard 8452 diode array spectrophotometer equipped with a quartz cell with 1 cm path and interfaced to a Hewlett-Packard Vectra AT computer was used.

3.2 *Reagents*

2-Chlorophenol (Aldrich, 98% pure) and 2,4-dichlorophenol (Aldrich, 99% pure) were used to prepare 1000 ppm stock solutions in 0.05 *M* NaOH. Standard working solutions of the chlorophenols were prepared by diluting the stock solutions with deionised water so that the resulting solutions were 10^{-3} *M* NaOH. Each sample was prepared 3 times.The composition of the prepared solutions is shown in Table 3.

Table 3. Concentration values (in ppm) for the mixtures of two chlorophenols. Calibration samples are printed in bold type

Sample number	2CP	2,4-DCP
1	0	5
2	**0**	**10**
3	2.5	2.5
4	2.5	7.5
5	5	0
6	5	5
7	5	10
8	7.5	2.5
9	7.5	7.5
10	**10**	**0**
11	10	5
12	**10**	**10**

3.3 *Spectra and selection of analytical wavelengths*

The spectra of the twelve samples were recorded between 226 and 346 nm. The two sensors that optimized the following criteria were selected by checking all possible combinations of wavelengths: maximum Det($\mathbf{S}^T\mathbf{S}$), minimum Tr $(\mathbf{S}^T\mathbf{S})^{-1}$, maximum selectivity for analyte 1 ($LSEL_1$), maximum sensitivity for analyte 1 ($LSEN_1$) and the minimum prediction error for Test Set 1.

3.4 *Calibration and validation sets*

Sample numbers 2,10 and 12 were used as calibration mixtures to calculate the S matrix. The remaining samples were used for validation purposes grouped in three validation sets, each of which contained one set of replicates of each sample. For each optimally selected set of two sensors, the prediction error was evaluated for the validation samples by computing the total root mean squared error of prediction, RMSEPT (Eq. (5)). The mean value of RMSEPT for the three validation sets was computed.

$$\text{RMSEPT} = \sqrt{\frac{\sum_{i=1}^{I}\sum_{k=1}^{K}(c_i - \hat{c}_i)^2}{I \times K}} \quad (5)$$

where c_i and $\hat{c}_i$ are the real and predicted concentrations, I is the number of test samples and K the number of analytes.

3.5. Results and discussion

Fig. 5 shows the spectra of the pure components resulting from the **S** matrix along with the sensors that optimize different selection criteria. Fig. 6 depicts the confidence ellipsoids for two selected sensors using different criteria. Table 4 shows the values of the different criteria for the selected pairs of sensors. The D-optimal (sensors a and b) and A-optimal sensors (sensors c and d) correspond to near wavelengths in the spectra and their confidence regions are very similar, so similar prediction errors

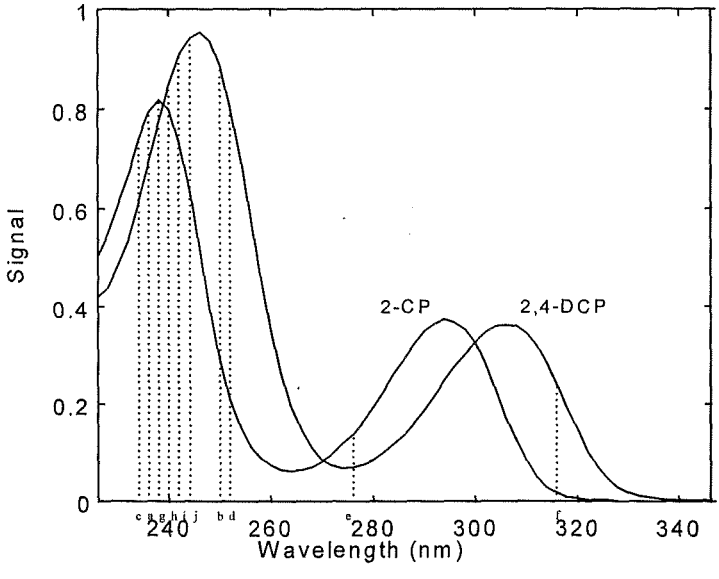


Fig. 5. Calculated absorbance spectra for 2-chlorophenol (12.85 ppm) and 2,4-dichlorophenol (16.3 ppm). Pairs of sensors that optimise different selection criteria are shown.

Table 4. Values of several criteria for the matrices with two selected sensors optimizing DET, TR, Selectivity for analyte A, sensitivity for analyte A, RMSEPT and all sensors. (analyteA: 2CP, analyteB: 2,4DCP)

	DET	TR	UVIF		LSEN		LSEL		VIF		sensors
Analyte			A	B	A	B	A	B	A	B	
DET	0.253	7.82	4.99	2.82	0.84	1.12	0.53	0.53	3.57	3.57	6 13
TR	0.219	7.37	4.66	2.71	0.77	1.01	0.60	0.60	2.77	2.77	5 14
$LSEL_1$	0.001	78.8	60.1	18.6	0.14	0.25	0.92	0.92	1.18	1.18	26 46
$LSEN_1$	0.005	483	242	240	1.14	1.15	0.06	0.06	316	316	7 8
RMSEPT	0.014	188	121	67.2	0.97	1.31	0.09	0.09	114	114	9 10
	18.91	0.93	0.57	0.36	2.60	3.28	0.51	0.51	3.86	3.86	ALL

should be expected for the models built with these sensors. Also note the small determinant values, i.e. large area ellipses, corresponding to sensors which optimize $LSEL_1$ (sensors e and f) and $LSEN_1$ (sensors g and h). Also notice the large VIF values for the set which optimizes $LSEN_1$, indicating that it may produce unreliable estimations of the concentrations. The large ellipse corresponding to the sensors that minimize the prediction error (sensors i and j) indicates that although these sensors have been selected by using test set 1, they may produce large errors for new samples, or even for repetitions of the same samples, since small measuring errors could considerably affect the predicted concentrations.

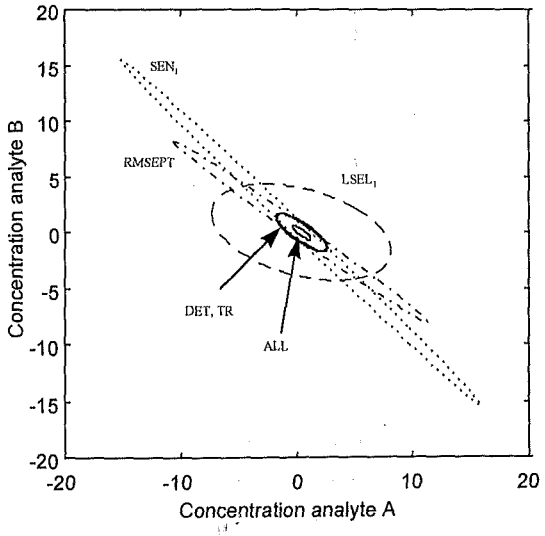


Fig. 6. Ellipses corresponding to the pairs of wavelengths selected from Fig. 5 and listed in Table 4.

UNIVERSITAT ROVIRA I VIRGILI
EXPERIMENTAL DESIGN APPLIED TO THE SELECTION OF SAMPLES AND SENSORS IN MULTIVARIATE CALIBRATION
Joan Ferré Baldrich
ISBN:978-84-691-1875-7/DL: T-337-2008

Table 5. RMSEPT (ppm) for the test samples and two selected sensors. Mean value of RMSEPT for the three validation sets

Selection criteria	RMSEPT
DET	0.100
TR	0.094
$LSEL_1$	0.177
$LSEN_1$	0.311
Prediction error	0.167
All sensors	0.089

Table 5 shows the prediction results using the mean value of RMSEPT corresponding to the three validation sets and using the sensors selected by the different selection criteria used. The prediction errors are in good agreement with what it is expected from the confidence ellipsoids. Sensors which optimize $Det(S^TS)$ and $Tr(S^TS)^{-1}$ provide the smallest prediction errors, which agrees with the small area of their confidence ellipse.

The prediction error increases in the sensors that optimise selectivity ($LSEL_1$), which is to be expected because of the larger volume of the corresponding ellipse. Observe, however, that this latter ellipse has the smallest inclination with respect to the co-ordinate axes, which agrees with the maximization of the selectivity, thus providing less correlated estimated concentrations.

The sensors which maximize sensitivity for analyte 1 have the highest error. These are the sensors with maximum absorbance for this analyte. Observe that not only is a large confidence interval for analyte 2 computed but also for analyte 1, which is the one to be determined.

4.Conclusions

The graphical representation of the confidence ellipsoid gives a better understanding of the effect of several wavelength selection criteria on the prediction ability of the model. The limitations of applying criteria such as sensitivity or selectivity are better understood from the confidence intervals derived. The need for a global optimization criterion is therefore apparent. This criterion should also take into account the response of the unknown sample. This global criterion would require optimization algorithms to find the optimal set of sensors, such as genetic algorithms

or generalized simulated annealing. Other approaches would include the optimum selection of wavelengths to predict specific analytes. Nevertheless, we must stress the importance of the intrinsic limitations of this classical regression approach: models usually show a lack of robustness due to common effects such as different errors in the instrumental responses (heteroscedasticity) or in concentrations, and interaction terms that are supposed to be absent.

Acknowledgements

J. Ferré thanks the Comissionat per a Universitats i Recerca of Generalitat de Catalunya, for providing a doctoral fellowship (FI/94-7001). Financial support from the Spanish Ministry of Education and Science (DGICyT project BP93-0366) is gratefully acknowledged.

References

[1] C.B. Lucasius, M.L.M. Beckers and G. Kateman, Anal. Chim. Acta 286 (1994) 135.
[2] Y. Liang, Y. Xie and R. Yu, Anal. Chim. Acta, 222 (1989) 347.
[3] K. Sasaki, S. Kawata and S. Minami, Appl. Spectrosc., 40 (1986) 185.
[4] S.D. Frans and J.M. Harris Anal. Chem. 57 (1985) 2680.
[5] U. Hörchner and J. H. Kalivas, J. Chemom. 9 (1995) 283.
[6] L. Costadinnova and T. Nedeltcheva, Analyst, 120 (1995) 2217.
[7] A.C. Atkinson and A.N. Donev, Optimum Experimental Designs. Oxford Statistical Science Publications: Oxford 1992
[8] M. Meloun, J. Militký and M. Forina, Chemometrics for analytical chemistry. Vol2. PC-aided Regression and Related Methods. Ellis Horwood: Great Britain 1994.
[9] J.H. Kalivas and P.M. Lang, Chem. Intell. Lab. Syst., 32 (1996) 135.
[10] A. Lorber, Anal. Chem., 58 (1986) 1167.
[11] A. Lorber and B.R. Kowalski. J. Chemom. 2 (1988) 67.
[12] Harrison M. Wadswoth, Handbook of Statistical Methods for Engineers and Scientists. McGraw Hill, USA, 1990.
[13] G. Bauer, W. Wegscheider and H.M. Ortner. Spectrochim. Acta 46B (1991) 1185.
[14] M. Otto and W. Wegscheider Anal. Chim. Acta, 180 (1986) 445.
[15] J. Ferré and F.X. Rius, Quim. Anal., in press.

UNIVERSITAT ROVIRA I VIRGILI
EXPERIMENTAL DESIGN APPLIED TO THE SELECTION OF SAMPLES AND SENSORS IN MULTIVARIATE CALIBRATION
Joan Ferré Baldrich
ISBN:978-84-691-1875-7/DL: T-337-2008

4.3 Further considerations on the sensitivity and selectivity of multicomponent systems

(in preparation)

Joan Ferré and F. Xavier Rius.

Departament de Química. Universitat Rovira i Virgili.
Pl. Imperial Tarraco, 1, 43005-Tarragona. SPAIN

Condition number, variance proportions decomposition, sensitivity and selectivity and the effect of adding a new sensor to the model are interpreted based on their effect on the ellipsoid defining the confidence interval for the concentration values of the different analytes to be determined.The understanding of the relationship between these criteria enables guidelines for variable selection criteria to be proposed.

1.Introduction

The confidence ellipsoid of the predicted concentrations in multicomponent analysis has been shown useful for interpreting the quality of the selected sensors in wavelength selection problems (see §4.2). Expressions related to the classical least-squares (CLS) models and not considered in §4.2 are shown here. The present work contains two parts: a theoretical part that introduces the mathematical expressions of the variance of the estimated concentrations in CLS and selectivity and sensitivity measures in CLS, not explained in §2.4.1 and a second part with the interpretation, using the ellipsoid, of these measures and the effect of adding a new sensor to the model. The understanding of these criteria enables guidelines for variable selection criteria to be proposed. The notation used here (§2.2) and the mathematical expressions of the CLS calibration (§2.4.1), the ellipsoid and the confidence intervals (§2.4.2) have been presented in the sections indicated.

2 Theoretical background

2.1 Mathematical model and singular-value decomposition of **S**

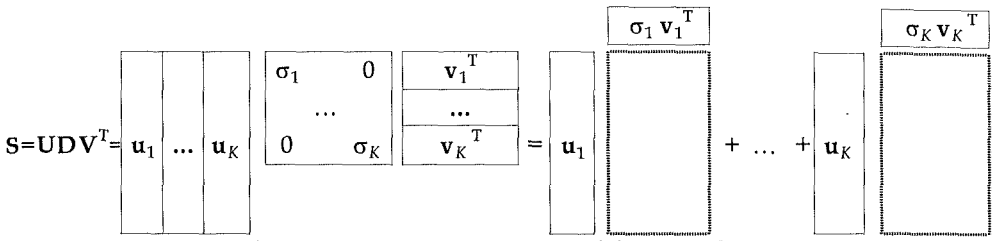
The model considered is:

$$\mathbf{r}_{un} = \mathbf{S}\,\mathbf{c}_{un} + \mathbf{e} \qquad (1)$$

where $\mathbf{r}_{un}$ is the spectra of the mixture measured at J wavelengths, $\mathbf{S}$ is the $J \times K$ calibration matrix, $\mathbf{c}_{un}$ is the vector of concentrations of the K components in the mixture and $\mathbf{e}$ is a vector of error terms. The *singular value decomposition* (SVD) of $\mathbf{S}$ can be written as (see Scheme 1):

$$\mathbf{S} = \mathbf{U}\mathbf{D}\mathbf{V}^T \qquad (2)$$

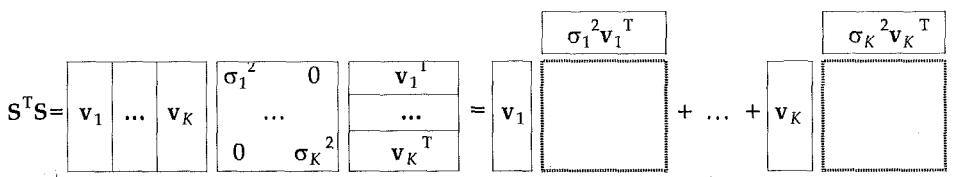
where $\mathbf{U}_{J\times K}=[\mathbf{u}_1, ..., \mathbf{u}_k, ..., \mathbf{u}_K]$ and $\mathbf{V}_{K\times K}=[\mathbf{v}_1, ..., \mathbf{v}_k, ..., \mathbf{v}_K]$ have orthogonal columns $\mathbf{u}_k$ and $\mathbf{v}_k$, and $\mathbf{D}_{K\times K}=\{\sigma_{ii}\}$ is the matrix whose entries are all zero except for σ_{ii}, i=1,2, ..., K. The values σ_{ii}, from now on denoted by σ_i, are the *singular values* of $\mathbf{S}$ and they satisfy $\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_K$. σ_i^2 are the eigenvalues of $\mathbf{S}^T\mathbf{S}$ and the inverse of the



Scheme 1. Matrix representation of the SVD of **S**.

eigenvalues λ_j of $(\mathbf{S}^T\mathbf{S})^{-1}$: $\lambda_j=1/\sigma_i^2$. Since the eigenvalues of $\mathbf{S}^T\mathbf{S}$ and $(\mathbf{S}^T\mathbf{S})^{-1}$ are decreasingly ordered, their indexes and those of the associated eigenvectors are interchanged (e.g. *if* $\lambda_1=10$ and $\lambda_2=1$ are the *two* eigenvalues of $(\mathbf{S}^T\mathbf{S})^{-1}$, the first eigenvalue of $\mathbf{S}^T\mathbf{S}$ *is* $\sigma_1^2=1/\lambda_2=1$ and the *second* is $\sigma_2^2=1/\lambda_1=0.1$). Below *i* is the subscript of the eigenvectors and eigenvalues of $\mathbf{S}^T\mathbf{S}$ and that of the eigenvectors and eigenvalues of $(\mathbf{S}^T\mathbf{S})^{-1}$ is *j* .From eq 2 (see also Scheme 2):

$$\mathbf{S}^T\mathbf{S} = \mathbf{V}\mathbf{D}^2\mathbf{V}^T = \sum_{i=1}^{K} \mathbf{v}_i\mathbf{v}_i^T\sigma_i^2 \qquad (3)$$



Scheme 2. Matrix representation of the SVD of $\mathbf{S}^T\mathbf{S}$.

The diagonal element that corresponds to the *k*th analyte is:

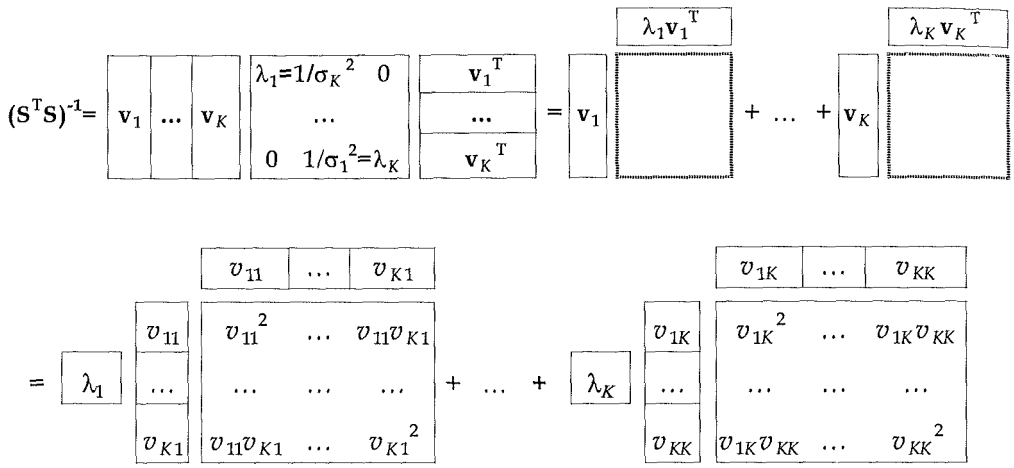
$$(\mathbf{S}^T\mathbf{S})_{kk} = \sum_{i=1}^{K} v_{ki}^2\sigma_i^2 \qquad (4)$$

where v_{ki} is the *k*th element of $\mathbf{v}_i$. The determinant $\mathbf{S}^T\mathbf{S}$ and $(\mathbf{S}^T\mathbf{S})^{-1}$ is the product of their eigenvalues $\mathrm{Det}(\mathbf{S}^T\mathbf{S})^{-1}=\prod_j \lambda_j$ and $\mathrm{Det}(\mathbf{S}^T\mathbf{S})=\prod_i \sigma_i^2$.

The variance-covariance matrix of the predicted concentrations is:

$$\mathrm{var}(\mathbf{c}_{un}) = \sigma^2(\mathbf{S}^T\mathbf{S})^{-1} = \sigma^2\mathbf{V}\mathbf{D}^{-2}\mathbf{V}^T = \sigma^2\sum_{i=1}^{K}\frac{\mathbf{v}_i\mathbf{v}_i^T}{\sigma_i^2} = \sigma^2\sum_{j=1}^{K}\mathbf{v}_j\mathbf{v}_j^T\lambda_j \qquad (5)$$

where σ^2 is the variance of the instrumental response (usually calculated as the standard deviation of the residuals between measured and fitted values of the absorbance vector[1]). Scheme 3 represents the SVD of $(\mathbf{S}^T\mathbf{S})^{-1}$.



Scheme 3. Matrix representation of the SVD of $(\mathbf{S}^T\mathbf{S})^{-1}$. In this scheme, $\mathbf{v}_1,\ldots\mathbf{v}_K$ of $(\mathbf{S}^T\mathbf{S})^{-1}$ correspond to $\mathbf{v}_K,\ldots\mathbf{v}_1$ of $(\mathbf{S}^T\mathbf{S})$ respectively in the Scheme 2. The elements of the vectors are also shown for an easier understanding of the eq 5 and eq 7.

The variance of the estimated concentration for the *k*th analyte is[1,2]:

$$\mathrm{var}(c_{un,k}) = \sigma^2(\mathbf{S}^T\mathbf{S})^{-1}_{kk} = \sigma^2\mathrm{UVIF}_k \qquad (6)$$

where the *k*th diagonal element of $(\mathbf{S}^T\mathbf{S})^{-1}$ is called *unscaled variance inflation factor*[2] for the *k*th analyte (UVIF_k) or just *variance factor* and can be written as (see below for the parameters not defined yet)[1-4]:

$$\mathrm{UVIF}_k = (\mathbf{S}^{\mathrm{T}}\mathbf{S})^{-1}_{kk} = \sum_{i=1}^{K} \frac{v_{ki}^2}{\sigma_i^2} = \sum_{j=1}^{K} v_{kj}^2 \lambda_j = \sum_{j=1}^{K} (v_{kj}\sqrt{\lambda_j})^2 \qquad (7a)$$

$$\mathrm{UVIF}_k = \left\|\mathbf{S}^{+}_{k\text{-row}}\right\|^2 = \frac{1}{\left\|\mathbf{s}_k^*\right\|^2} = \frac{1}{\mathrm{SEN}_k^2} = \frac{1}{\mathrm{LSEL}_k^2\,\mathrm{LSEN}_k^2} \le \frac{1}{\sigma_K^2} \qquad (7b)$$

Therefore, the variance of the estimated concentration follows:

$$\mathrm{var}(c_{\mathrm{un},k}) = \frac{\sigma^2}{\left\|\mathbf{s}_k^*\right\|^2} = \frac{\sigma^2}{\mathrm{SEN}_k^2} = \frac{\sigma^2}{\mathrm{LSEL}_k^2\,\mathrm{LSEN}_k^2} = \sigma^2 \sum_{i=1}^{K} \frac{v_{ki}^2}{\sigma_i^2} \le \frac{\sigma^2}{\sigma_K^2} \le \mathrm{MSE}(\mathbf{c}_{\mathrm{un}}) \quad (8)$$

and the MSE error (the difference between the predicted value and the expected value):

$$\mathrm{MSE}(\mathbf{c}_{\mathrm{un}}) = \sigma^2 \mathrm{Tr}\left[(\mathbf{S}^{\mathrm{T}}\mathbf{S})^{-1}\right] = \sigma^2 \sum_{j=1}^{K} \lambda_j = \sigma^2 \sum_{i=1}^{K} \frac{1}{\sigma_i^2} = \sigma^2 \sum_{k=1}^{K} \mathrm{UVIF}_k \le \frac{K\sigma^2}{\sigma_K^2} \qquad (9)$$

The confidence intervals of the predicted concentrations are represented in Figure 1[5,6]. The confidence interval for an analyte is inversely proportional to its net sensitivity.

2.2 Measures of sensitivity and selectivity

The influence of sensitivity and selectivity on the predictive ability of the CLS model has been considered in several works[2,3,7]. Local (for each analyte) and global (for the complete system) measures of sensitivity and selectivity have been proposed [2, 8-10]. They have in common of being determined by the calibration matrix only. In the following paragraphs these measures are reviewed and interpreted from the confidence ellipsoid.

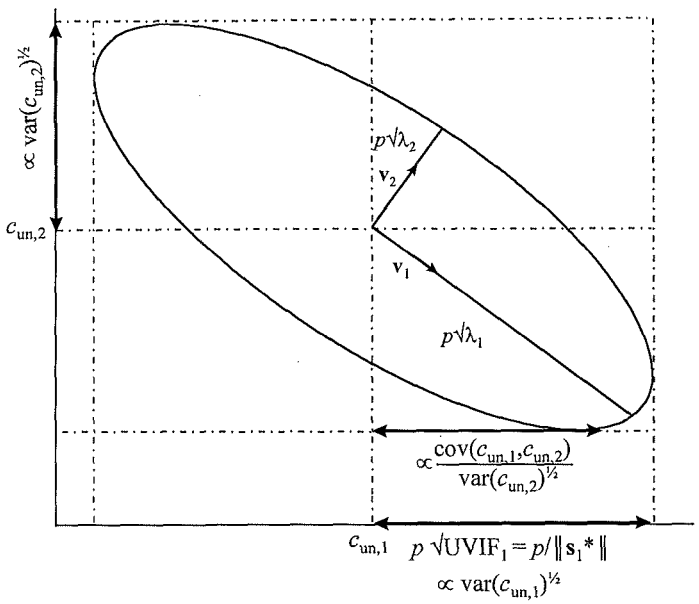


Figure 1. The confidence ellipse for a two component system. "∝" means *proportional.*

2.2.1 *Measures of sensitivity*

Sensitivity refers to the intensity of the instrumental response due to the components under consideration at the measured wavelengths. Different definitions of sensitivity are:

- *local sensitivity* of the *k*th component given by the Euclidean norm of the pure component spectra [2]:

$$LSEN_k = \|\mathbf{s}_k\| \tag{10}$$

- *local net sensitivity* of the *k*th component[2,3,11] (called *sensitivity* SEN_k by Lorber [8] and Kalivas and Lang [2]):

$$LSEN_k^* = \|\mathbf{s}_k^*\| = 1/\sqrt{UVIF_k} = 1/\ \|\mathbf{S}^+_{k\text{-row}}\| = 1/\|\mathbf{b}_{k,CLS}\| \tag{11}$$

Two recent papers have discussed which is the best the definition of sensitivity[12,13].

- *global sensitivity* of **S**[2]:

$$GSEN(\mathbf{S}) = \sigma_1 \qquad (12)$$

The larger GSEN, the better the global sensitivity. The problem of this measure is that spectral overlap makes σ_1 increase.

2.2.2 *Measures of selectivity*

Selectivity refers to the extent that one component responds at selected wavelengths compared to other components, i.e. the degree of spectral overlap of the pure component spectra. Mathematically it corresponds to the degree of orthogonality between the columns of **S**. The following are measures of selectivity:

- *local selectivity* of the *k*th component [1,2,8,16] (called *selectivity* for Lorber and Kowalski[17])

$$LSEL_k = \sin \alpha_k = \frac{\|\mathbf{s}_k^*\|}{\|\mathbf{s}_k\|} = \frac{\|\mathbf{a}_k^*\|}{\|\mathbf{a}_k\|} = \frac{1}{UVIF_k^{1/2}\, LSEN_k} = \frac{1}{\|\mathbf{b}_{k,CLS}\|\|\mathbf{s}_k\|} \qquad (13)$$

where $0 \leq LSEL_k \leq 1$. This is the most used measure of selectivity.

- *variance inflation factors* (VIF_k) of the *k*th component:

$$VIF_k = (\mathbf{S}^T\mathbf{S})_{kk}^{-1} = \sum_{i=1}^{K} \frac{v_{ki}^2}{\sigma_i^2} \qquad (14)$$

where the numbers making up **S** in eq 14 had been scaled so that $\mathbf{S}^T\mathbf{S}$ is a correlation matrix. This measure is equivalent to $LSEL_k$ (see §4.4):

$$VIF_k = 1/\ LSEL_k^2 \qquad (15)$$

- *global selectivity* of the *K*-component mixture, which measures the selectivity of the complete system. This criterion has been used for wavelength selection[14-16] in CLS. It is given by:

$$SEL = K / \sum_{k=1}^{K} LSEL_k^{-1} \qquad (16)$$

- *variance-decomposition proportions (VDP)*[18-20].This diagnostic tool determines the proportion of the variance of each coefficient in a model that is attributed to the linear dependencies in the calibration matrix. It evaluates, in CLS, the extent of collinearity (and therefore of selectivity) in the columns of **S** due to spectral overlap of the pure component spectra[21]. The components whose concentration estimates could be misinterpreted owing to spectral overlap and which will not be extensively affected can be determined. The variance coefficient of the *k*th analyte can be decomposed as a sum of squared terms (right-hand term in eq 7a):

$$UVIF_k = \sum_{j=1}^{K} (v_{kj}\sqrt{\lambda_j})^2 = (v_{k1}\sqrt{\lambda_1})^2 + (v_{k2}\sqrt{\lambda_2})^2 + \ldots + (v_{kK}\sqrt{\lambda_K})^2 \qquad (17)$$

the *j*th element of the sum divided by the total variance:

$$\pi_{jk} = \frac{(v_{kj}\sqrt{\lambda_j})^2}{UVIF_k} = \frac{\frac{v_{kj}^2}{\sigma_j^2}}{UVIF_k} \qquad (18)$$

is the proportion of variance of the concentration of the *k*th analyte attributed to the collinearity characterized by the *j*th eigenvalue. These elements, calculated for each analyte and each eigenvalue make up a $K \times K$ matrix Π of variance- decomposition proportions, shown in Table 1.

		Π		
Associated eigenvalue	Associated condition index	$var(c_{un,1})$	...	$var(c_{un,K})$
λ_1		π_{11}	...	π_{1K}
λ_2		π_{21}	...	π_{2K}
...		...	...	...
λ_K		π_{K1}	...	π_{KK}

Table 1. Matrix Π of the variance-decomposition proportions.

The entries in the *k*th column of Π are the terms of eq 18 for the *k*th analyte. They sum to unity since they are proportions of the total variance of $c_{un,k}$. Each row of Π is the contribution of the *j*th singular value to the variance of each concentration divided by the total variance. A row with two o more large variance proportions (π_{jk} >0.5) and a large condition index (defined as $\mu_j=\sigma_{max}/\sigma_j$) identify the particular terms that contribute to the collinearity and indicates spectral overlap between the spectra of the analytes involved and a motive to suspect of the concentration estimates of the respective components.

- *condition number*. The condition number is the ratio of the largest singular value of a matrix to the smallest one. Cond(S) and Cond(S^TS) have been widely used as collinearity measure in wavelength selection problems (see §4.1.3). It can be calculated in three equivalent ways:

$$\text{Cond}(S^TS) = \max(\sigma_i^2) / \min(\sigma_i^2) = \sigma_1^2/\sigma_K^2 \quad (19)$$

$$\text{Cond}(S) = (\text{Cond}(S^TS))^{1/2} = \sigma_1/\sigma_K \quad (20)$$

$$\text{Cond}[(S^TS)^{-1}] = \max(\lambda_j) / \min(\lambda_j) = \lambda_1/\lambda_K = (1/\sigma_K^2)/(1/\sigma_1^2) = \sigma_1^2/\sigma_K^2 \quad (21)$$

UNIVERSITAT ROVIRA I VIRGILI
EXPERIMENTAL DESIGN APPLIED TO THE SELECTION OF SAMPLES AND SENSORS IN MULTIVARIATE CALIBRATION
Joan Ferré Baldrich
ISBN:978-84-691-1875-7/DL: T-337-2008

2.3 *Interpretation considering the confidence ellipsoid*

- *interpretation of the variance coefficients* ($UVIF_k$). The length of the confidence interval of the predicted concentration of the analyte k is proportional to $UVIF_k$ (Figure 1). Although $UVIF_k$ is a function of both selectivity and sensitivity [2], it is only function of the norm of the net analyte signal of the analyte (eq 11).

- *interpretation of the net sensitivity* $LSEN_k^* = \|\mathbf{s}_k^*\|$. The length of the confidence interval is inversely proportional to the net sensitivity $\sqrt{UVIF_k} = 1/\|\mathbf{s}_k^*\|$ (from eq 11). Therefore, the more different the spectrum of the analyte k is from the spectra of the other analytes, the larger is norm of the net analyte signal and the shorter the length of the confidence interval for the analyte.

- *interpretation of the VDP.* The kth element of the eigenvector $\mathbf{v}_j$, v_{kj}, is the cosine of the angle between the jth ellipsoid axis and the kth concentration axes. Therefore, the term $v_{kj}\sqrt{\lambda_j}$ is the projection of the half-axis j onto the concentration axis of the analyte k divided by p, as shown in the Figure 2. The sum of the squared projections is proportional to the variance of $c_{un,k}$. The variance depends on the inclination of the ellipsoid axes (that gives the proportion (v_{kj}) that the eigenvalue contributes to the variance) and their length (that gives the extent that it contributes to the variance), both of them associated with collinearity and sensitivity. Therefore, a large eigenvalue of $(\mathbf{S}^T\mathbf{S})^{-1}$ does not necessarily involve a large variance for $c_{un,k}$; the eigenvalue must coincide with a large eigenvector (that is to say, a large projection). Figure 3 shows the VDP in a ellipsoid with a different inclination and in a sphere. In the sphere, the eigenvectors are in the direction of the concentration axes and the contribution to the total variance is only due to one eigenvalue.

- *interpretation of the condition number.* The condition number is the largest ellipsoid axis divided by the shortest one. The higher the spectral overlap, the more inclined the ellipsoid is, the larger the ratio between the largest and shortest ellipsoid axes and the concentration estimates numerically degrade [21].

UNIVERSITAT ROVIRA I VIRGILI
EXPERIMENTAL DESIGN APPLIED TO THE SELECTION OF SAMPLES AND SENSORS IN MULTIVARIATE CALIBRATION
Joan Ferré Baldrich
ISBN:978-84-691-1875-7/DL: T-337-2008

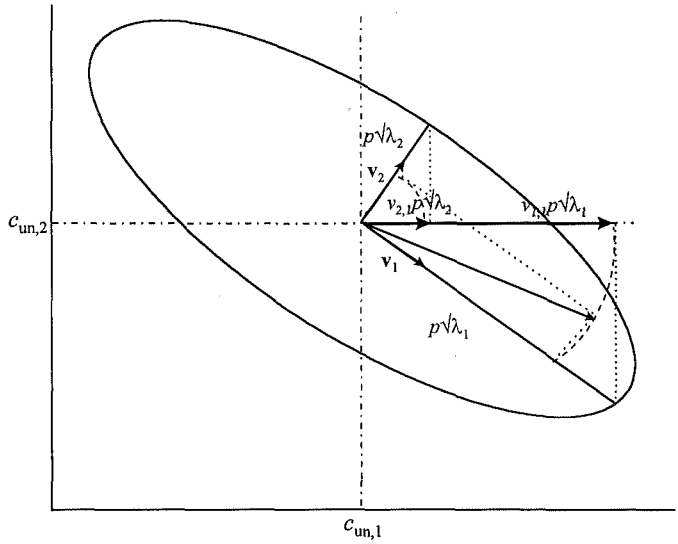


Figure 2. Geometrical representation of the variance-decomposition proportions for $c_{un,1}$ in a system of two components. The projections for $c_{un,2}$ are in the vertical axis.

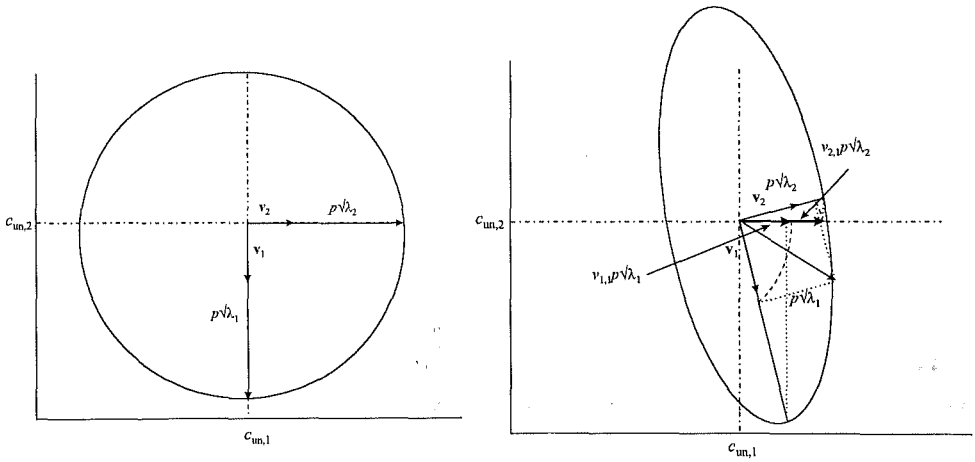


Figure 3. Geometrical representation of the variance-decomposition proportions in different confidence regions: a sphere and a ellipse.

2.4 Criteria for sensor selection

Numerous approaches have been reported in the recent years to improve the accuracy and precision of the predicted concentrations in spectrophotometric determinations using wavelength selection. The criteria used to evaluate the performance of a selected subset of sensors include $LSEL_k$ [1,3,7,14-16], variance coefficients[22], sensitivity[3,7,16], $Det(\mathbf{S}^T\mathbf{S})$[4,23,24], $Cond(\mathbf{S}^T\mathbf{S})$[3,4,7,23-28], Lorber's accuracy[16,29] and minimum squared error (MSE)[4,16,30] among others[14,23,31]. In addition, experimentation showed that the expected improvement of the prediction errors in the concentrations determined using a particular set of wavelengths did not correspond with the only improvement of $Det(\mathbf{S}^T\mathbf{S})$[23], $LSEL_k$[7,15] or $Cond(\mathbf{S}^T\mathbf{S})$ [7,28]. The large number of studies available makes the situation rather confusing when the experimenter has to decide the best criterion to use. Kalivas and Lang[2], showed the relationship among some of the criteria quoted above. Below, some of these criteria for sensor selection are considered and explained in terms of their effect on the confidence ellipsoid.

- *maximum local sensitivity* $LSEN_k=\|\mathbf{s}_k\|$. It corresponds to selecting the sensors with largest absorbance for the analyte of interest (e.g. the peak in the spectrum). Its effect on the confidence ellipsoid is not evident since this measure is not directly related to the parameters of the ellipsoid. $\|\mathbf{s}_k\|$ does not considers the spectral overlap with the spectra of the other analytes so that a high absorbance of the analyte is not enough to obtain precise predictions (an extreme case is the presence of another analyte with an spectra very similar to that of the *k*th analyte; the collinearity is large independently of the amount of absorbance of the analyte of interest). Thus, the absorbance must be high compared to the absorbance of the other analytes. This is considered in the next criterion.

- *maximum* $LSEN_k^*=\|\mathbf{s}_k^*\|$, *minimum* $UVIF_k$. Maximizing $LSEN_k^*$ is equivalent to minimizing $UVIF_k$ .The higher the norm of the net sensitivity of the method with respect to the components($\|\mathbf{s}_k^*\|$), the shorter the individual confidence interval and the more precise are the estimated concentrations. The sensors that maximize this criterion should lead to the smallest prediction errors. This was confirmed by Bauer *et al.*[3,7]. They found that $1/\|\mathbf{S}^+_{k\text{-row}}\|$ (which is the inverse of $UVIF_k$ since $\|\mathbf{S}^+_{k\text{-row}}\|=1/\|\mathbf{s}_k^*\|$) correlated quite well with the prediction errors and that it could be used to

predict errors in the concentration under the assumption of equal variances for all signals independent of their magnitude. However, they considered this as a measure of selectivity, when in reality is a measure of net sensitivity.

- $Tr(\mathbf{S}^T\mathbf{S})^{-1}$. This is the sum of the variance factors $UVIF_k$ (eq 9). Its improvement should lead to a reduction of the confidence interval since it reduces the length of the axes of the ellipsoid, and thus their projections on the coordinate axes. Since this is a global measure, the improvement for a particular analyte is not evident. Frans and Harris[22] and Liang *et al.*[4] used this criterion to select wavelengths that yield superior concentration precision. They also found that the selection of wavelengths affects the variance less as the number of measurements increases. This is a logical result of a selection process where the best sensors are selected at the beginning so that each new sensor considered has a smaller net sensitivity. This criterion has been considered in §4.2.

- *maximum* $Det(\mathbf{S}^T\mathbf{S})$. The sensors selected according to this criterion make the volume of the ellipsoid as small as possible. However, if the data is highly collinear, the D-optimal set is also collinear and gives a *thin* but inclined ellipsoid with correlated estimations of the concentrations and unfavorable prediction errors (despite being D-optimal!). This agrees with Frans and Harris[22], who indicated that $Det(\mathbf{S}^T\mathbf{S})$ provides no prediction as to the expected improvement in accuracy or precision of the concentrations determined using a particular set of wavelengths and that the wavelengths selected using the determinant minimized the concentration error. Otto and Wegscheider[23] concluded that the determinant gives a wrong picture of the selectivity. This agrees with Kalivas[21], who indicated that it is possible for a well-conditioned matrix to have a small determinant and, likewise, for an ill-conditioned matrix to have a large determinant. In addition, it is a global measure and it does not disclose which components are involved. This only indicates that the value of the determinant, by itself, is not indicative of the conditioning of the system. Kalivas[32]suggested that the circumstances that maximize the $Det(\mathbf{S}^T\mathbf{S})$ will also minimize the condition number. *Liang et al*[4] considered that MSE and $Det(\mathbf{S}^T\mathbf{S})$ were almost identical criteria for the use of experimental design and analytical wavelength selection although $Det(\mathbf{S}^T\mathbf{S})$ could be preferable to the MSE since it is easy to calculate.

- *the condition number of* $\mathbf{S}^T\mathbf{S}$ *or* $\mathbf{S}$. Cond(S) has been used in wavelength selection to characterize the global selectivity of the multicomponent system [4,7,23,26,28,33,35]. The condition number gives an upper limit of the relative error in solving matrix equations [8,24,36] and thus measures the error propagation to the estimated vector of concentrations in CLS due to the spectral overlap. Usually, the lower the selectivity, the higher is the condition number[28] and the larger the concentration errors can be. A Cond(S) equal to one has been said to represent complete orthogonality of the calibration spectra (fully selective system, where every component has a specific absorption band) while larger values indicate spectral overlap[21,24,26,28,34]. However this is only true if the matrix S has been scaled so that $\mathbf{S}^T\mathbf{S}$ is a correlation matrix or for fully selective measurements with equal sensitivity. Orthogonal spectra (the axes of the ellipse are parallel to the concentration axes) may have a condition number larger than 1 (one ellipse axis divided by the other) if each analyte has a different sensitivity (Figure 4). In addition, Cond(S) is not directly related with selectivity. For example, in the Figure 4 $\mathbf{s}_1$ is fully selective despite the condition number being large.

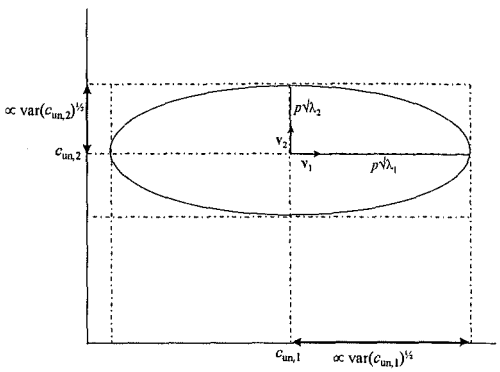


Figure 4. Confidence ellipsoid for an orthogonal system with a different net sensitivity.

Since Cond(S) is a global measure (all components are examined together with the same matrix [8,21,31,34]) it does not disclose which components may have degraded concentration estimates[8,19] . Its value can be influenced by analytes in the system different from that of interest. Figure 5 shows the ellipsoid for a three component mixture where the spectra of two of them are similar and that of the 3rd is orthogonal. The axes of the ellipses in the planes $c_{un,1}$-$c_{un,3}$ and $c_{un,1}$-$c_{un,2}$ are parallel to the concentration axes and indicates that $c_{un,1}$ is uncorrelated with $c_{un,2}$ and $c_{un,3}$. The ellipse in the plane $c_{un,2}$-$c_{un,3}$ is inclined and indicates correlation between $c_{un,2}$ and $c_{un,3}$. The third eigenvalue of the system (proportional to the shortest axis of the ellipsoid) is small because two spectra are similar and the condition number is large. Nevertheless, the component with the orthogonal spectrum may be well-

determined with a low uncertainty and the large confidence regions only affect two of the three analytes. Thus, Cond(**S**) may be not directly related to the analyte of interest. This could also explain why Bauer *et al.*[7] found a significant divergence between calculated errors in concentrations and the values of Cond(**S**) for different wavelength combinations. Other drawbacks of minimizing the Cond(**S**) is that it is more time-consuming than optimizing the determinant[24] in wavelength selection problems to optimize the precision. Moreover, Cond(**S**) does not asymptotically approach a limiting value when increasing the number of selected sensors[4].

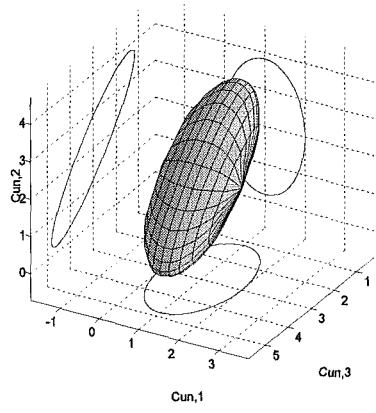


Figure 5. Confidence ellipsoid for a three component system with one completely selective spectrum ($\mathbf{s}_1$) and two not selective ($\mathbf{s}_2$ and $\mathbf{s}_3$). The projections of the ellipsoid on the planes are also shown.

- *selectivity.* Selectivity is related to the orientation of the ellipsoid axes with respect to the concentration axes. Bauer *et al.*[3,7] found a poor coincidence between concentration errors and selectivity evaluated as $1/(\|\mathbf{S}^+_{k\text{-row}}\|\,\|\mathbf{s}_k\|)$ (which is the same as $LSEL_k$, in eq 16 since $\|\mathbf{S}^+_{k\text{-row}}\|=1/\|\mathbf{s}_k^*\|$). They also indicated the difficulty in obtaining adequate definitions of selectivity that are useful for the direct prediction of errors in the concentration. Hörchner and Kalivas[15] indicated that wavelength selection based on high selectivity may not necessarily improve accuracy and precision. The confidence ellipsoid shows why a direct relationship between selectivity and error in the concentrations cannot be expected. The sensors that optimize the selectivity have an ellipsoid whose axes are as parallel as possible to the concentration axes and the concentrations are the maximum possible uncorrelated. However, this does not considers the length of the confidence interval (or the volume of the ellipsoid). Then, the $LSEL_k$ optimal sensors may have either a low or a high net sensitivity. In the first case, but the confidence interval is large (large variance, and possibly a large ellipsoid volume) and the prediction error may not improve. In the second case, it is small and the prediction ability of the model is good.

2.5 The ellipsoid in a completely selective system

In a completely selective system for the kth analyte, the spectrum of the analyte is orthogonal to the spectra of the other components and $\mathbf{s}_k^* = \mathbf{s}_k$. Hence $VIF_k = LSEL_k = 1$ and the confidence interval of the kth analyte has the shortest length $\sqrt{UVIF_k} = 1/\|\mathbf{s}_k\| = 1/LSEN_k$ since always $\|\mathbf{s}_k\| \geq \|\mathbf{s}_k^*\|$. In addition, if the system is made up of orthogonal spectra, $(\mathbf{S}^T\mathbf{S})^{-1}$ is a diagonal matrix, the ellipsoid axes are parallel to the axes of the component concentration and the concentrations are independently estimated. The matrix Π is the identity matrix. Complete selectivity does not imply the confidence region be a sphere. Orthogonal spectra with a different net sensitivity give an ellipse (Figure 5) and the concentration are estimated with a different precision. A sphere is the particular case where the concentrations are estimated with the same precision. In this case, $VIF_k = LSEL_k = 1$ despite the ellipsoid having a different shape.

In the non-orthogonal (not completely selective) case, the increase in the length of the confidence interval due to collinearity among the columns of $\mathbf{S}$ with respect to the length of the interval for orthogonal spectra with the same number of sensors is $(1/\|\mathbf{s}_k^*\|)/(1/\|\mathbf{s}_k\|) = 1/LSEL_k$. As the spectral overlap becomes higher, the values of the off-diagonal elements of $(\mathbf{S}^T\mathbf{S})^{-1}$ become higher.

2.6 Addition of a new sensor to the matrix S

The addition to the matrix $\mathbf{S}$ of a new sensor with absorbance different from zero makes the diagonal elements of $(\mathbf{S}^T\mathbf{S})^{-1}$ decrease (the mathematical proof is given by the Sherman-Morrison-Woodbury theorem; see e.g. Meyers[18] page 459 or Weisberg[37] page 293) and $\det(\mathbf{S}^T\mathbf{S})$ increase* . The orthogonality among the columns of $\mathbf{S}$ can either increase or decrease. All this has the following effects in the confidence ellipsoid:

* Proof: if the sensor with absorbances $\mathbf{s}_i$ ($\mathbf{s}_i$ is here a row vector) is added to $\mathbf{S}_J$ so that $\mathbf{S}_{J+1} = [\mathbf{S}_J ; \mathbf{s}_i]$, then $\mathbf{S}_{J+1}^T\mathbf{S}_{J+1} = \mathbf{S}_J^T\mathbf{S}_J + \mathbf{s}_i^T\mathbf{s}_i$ and $\det(\mathbf{S}_{J+1}^T\mathbf{S}_{J+1}) = \det(\mathbf{S}_J^T\mathbf{S}_J)\cdot(1+d_{s0})$ where $d_{si} = \mathbf{s}_i^T(\mathbf{S}_J^T\mathbf{S}_J)^{-1}\mathbf{s}_i$. Since always $d_{si} \geq 0$, $\det(\mathbf{S}^T\mathbf{S})$ always increases when a sensor (with not all the absorbances equal to zero) is added to the $\mathbf{S}$.

- $UVIF_k$ and thus the variance of the concentration of the *k*th component and the length of the individual confidence interval decrease, and $\|\mathbf{s}_k^*\|=LSEN_k^*=1/\sqrt{UVIF_k}$ increases. The magnitude of the change depends on the amount of net analyte signal that the new sensor has. Hence in CLS the precision of the procedure [25,38] (concentrations) increases with an increasing number of measurements [4,16,25] even if the added sensor only has noise. The best precision is attained when all wavelengths in the spectrum are selected [16,25]. This fact is the result of the approximations in the CLS of considering the **S** matrix as determined without error, homocedasticity for all the sensors and that the assumed model is true. Lorber and Kowalski[LOR1] considered different error sources and also concluded that the variance of the predicted concentration decreased with an increasing number of sensors. However, in reality the improvement strongly depends on the balance between the information content and the noise contributions of the additional wavelengths[3]. As a result, the analytical error of each component depends on the partial sensitivity, selectivity and noise of measured signal[1]. This has been recently shown by Xu and Schechter[31].

- $Tr(\mathbf{S}^T\mathbf{S})^{-1}=\sum_j \lambda_j$ decreases and thus, the variance of the estimations of the coefficients decreases depending on the degree of orthogonality of the new sensor with respect to the already existing sensors. The more orthogonal, the larger the decrease. This measure is related to the sum of the length of the ellipsoid axes, therefore, the axis of the ellipsoid decrease.

- $LSEN_k=\|\mathbf{s}_k\|$ increases, since the vector has one more term.

- the volume of the confidence region (inversely proportional to $Det(\mathbf{S}^T\mathbf{S})$) decreases and hence, for a given α, the concentrations are globally more precise with an increasing number of measurements[25]. The volume can be reduced by obtaining a more 'flat' ellipsoid (with axes either more parallel or more diagonal to the concentration axes) or an ellipsoid more similar to a sphere of smaller radio. If the inclination of the ellipsoid increases, the covariance of the concentrations estimated increase.

- the effect in the orientation corresponds to a change of $LSEL_k$. If the added sensor increases orthogonality, the axes of the ellipsoid tend to be more parallel to the concentration axes. On the contrary the ellipsoid is more inclined.

ISBN:978-84-691-1875-7/DL: T-337-2008

2.7 Considerations for wavelength selection

Since the precision of the estimated concentrations depends on the volume, shape and orientation of the ellipsoid, the selection of sensor based on the optimization of one criterion alone does not guarantees a high precision.

$Det(\mathbf{S}^T\mathbf{S})$ and $Tr(\mathbf{S}^T\mathbf{S})^{-1}$ always improve as more sensors are considered[4] so that they cannot be used as a criterion to decide on the sufficient number of sensors to use. The Det(M), where $\mathbf{M}=\mathbf{S}^T\mathbf{S}/J$ could be used, since it measure the information content per sensor, and the variance inflation factors (VIF) could be used to indicate if the selected sensors have the necessary information to estimate the concentration. A VIF_k value larger than 10 has been proposed [39,40] to indicate when multicollinearity can be a serious problem for the reliable estimation of concentrations. This can be applied here to the analyte concentrations, which are the coefficients of the model given in eq 1.However, if the experimenter is not concerned about using the minimum number of sensors necessary but the number that produces the best predictions, this criteria cannot be used. These criteria state that even the sensors that only have noise contribute to improve the confidence interval of the predicted concentration.

To obtain the best precision and accuracy for concentration predictions by means of wavelength selection, the optimal subset of J sensors should simultaneously maximize $Det(\mathbf{S}^T\mathbf{S})$, minimize $Tr(\mathbf{S}^T\mathbf{S})^{-1}$ or the largest eigenvalue of $(\mathbf{S}^T\mathbf{S})^{-1}$ and maximize $LSEL_k$ among all possible subsets of J sensors. It has also been suggested that the wavelength selection criterion should also consider spectral and concentration noise information, correlate with actual prediction errors and include information on prediction samples[14,15]. A hypersphere of small radio is the optimal confidence region that simultaneously optimizes the three parameters and gives independent concentration estimates with a high global precision and with similar variances for each analyte. Finding A-optimal, E-optimal, SEL-optimal sensors or combinations of them need optimization algorithms such as generalized simulated annealing (GSA) or genetic algorithms (GA) to be used, to avoid checking all possible combinations of J wavelengths, and multicriteria functions to simultaneously optimize several responses at the same time. Since a subset of sensors that accomplishes all these conditions may not exist, subsets which represent a compromise between them should be used.

ISBN:978-84-691-1815-7/DL:T-337-2008

A new selected sensor should improve the mentioned properties as much as possible. The magnitude of the change depends on whether the added sensor increases or decreases the orthogonality among the columns of **S** and on the instrumental response values in the added sensor. If the added sensor increases orthogonality, the ellipsoid tends to a small radio sphere. On the contrary, if it decreases orthogonality, the ellipsoid is more inclined, although the predictions are globally more precise. In most cases, the optimal conditions for the determination of each component is different and should be optimized separately.

4. Conclusions

The effect of several measures on the confidence ellipsoid of the estimated concentrations has been shown. Representing graphically the confidence ellipsoid helps to understand the effect of the wavelength selection criteria such as sensitivity and selectivity on the prediction ability of the model. The optimization of one criterion alone does not necessarily guarantees low prediction errors. This paper helps to understand the already used criteria and to develop a more general criterion for wavelength selection based on selectivity, sensitivity, noise and information on prediction samples. This shows that a global optimization criterion is needed. This criterion should take into account the uncertainty in the spectra. Other approaches include the optimization to best predict a certain analyte. Probably, the cited global criterion would need optimization algorithms to find the best set of sensors, such as GA or GSA.

The used criteria only affect to the precision of the estimated concentrations. Thus, the technique proposed is limited by the assumptions of multicomponent analysis: the additivity of the responses according to the Beer's Law, that **S** can be determined errorless, that the model is correct and that the r variables have experimental error that is distributed following a normal law $N(0,\sigma^2)$. Although in reality, the **S** variables do have error, their uncertainty can be reduced by averaging repetitions of the spectra of the pure components until the error can be considered be smaller than the error in the spectrum of the unknown sample. These assumptions enable the least-squares expression to be used to estimate the concentrations in the

ISBN:978-84-691-1875-7/DL: T-337-2008

unknown sample. Under these hypothesis, the variance of the coefficients always decreases when a new sensor is added.

Another problem of wavelength selection is the lack of measures that truly indicate the ability of the selected sensors to provide unbiased and precise predictions. The algorithms select the best wavelengths among the set of wavelengths given. If none of the wavelengths given is good enough (e.g. when all the wavelengths are highly correlated), the solutions of the algorithm would have little sense.

5. References

1. Bergmann G., von Oepen B., Zinn P. *Anal. Chem.* 59 (1987) 2522-2526.
2. Kalivas J.H., Lang P. M. *Chem. Intell. Lab. Syst.* 32 (1996) 135-149.
3. Bauer, G.; Wegscheider, W.; Ortner, H.M. *Spectrochimica Acta 46B* (1991) 1185-1196.
4. Liang Y., Xie Y ., Yu R. *Anal. Chim. Acta* 222 (1989) 347-357.
5. Fedorov, V.V. *Theory of Optimal Experiments.*(translated and edited by W.J. Studden and E.M. Klimko) Academic Press: New York, 1972. FED2.
6. Papakyriazis P.A. *J. Econometrics* 7 (1978) 351-372.
7. Bauer, G.; Wegscheider, W.; Ortner, H.M. *Spectrochimica Acta 47B* (1992) 179-188.
8. Lorber A *Anal. Chem.* 58 (1986) 1167-1172.
9. Xie Y., Kalivas J.H. *Anal. Chim. Acta* 348 (1997) 19-27.
10. Xie Y., Kalivas J.H. *Anal. Chim. Acta* 348 (1997) 29-38.
11. Booksh K.S. , Kowalski B.R *Anal. Chem.* 66 (1994) 782A-804A.
12. Faber K., Lorber A., Kowalski B.R. *Chem. Intell. Lab. Syst.*, 38 (1997) 89-93.
13. Kalivas J.H., Lang P. M. *Chem. Intell. Lab. Syst.* 38 (1997) 95-100.
14. Hörchner, U.; Kalivas J.H. *J. Chemom.* 9 (1995) 283-308.
15. Hörchner U., Kalivas J.H. *Anal. Chim. Acta.* 311 (1995) 1-13.
16. Lucasius, C.B; Beckers M.L.M; Kateman, G. *Anal. Chim. Acta* 286 (1994) 135-153.
17. Lorber A. , Kowalski B.R. *J. Chemom.* 2 (1988) 67-79.
18. Myers R.H *Classical and modern regression with applications 2nd edition.* Duxbury

ISBN:978-84-693-1337-2/DL:T-337-2008

Press: Belmont 1990.

19. Belsley, D.A.; Kuh E.; Welsch R.E.. *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity.* Wiley: New York 1980 Chapter 3.
20. Harrison M. Wadswoth. *Handbook of Statistical Methods for Engineers and Scientist.* McGraw Hill: USA 1990.
21. Kalivas, J.H. *J. Chemom.* 3 (1989) 409-418.
22. Frans S.D., Harris J.M., *Anal. Chem.* 57 (1985) 2680-2684.
23. Otto M., Wegscheider W. *Anal. Chim. Acta* 180 (1986) 445-456.
24. Smeyers-Verbeke J., Detaevernier M.R., Massat D.L. *Anal. Chim. Acta* 191 (1986) 181-192.
25. Massart D.L., Vandeginste B.G.M., Deming S.N., Michotte Y., Kaufman L. *Chemometrics: a textbook.* Elsevier. Amsterdam. 1988.
26. Otto M., Wegscheider W. *Anal. Chem.*57 (1985) 63-69.
27. Rossi D.T., Pardue H.L. *Anal. Chim. Acta* 175 (1985) 153-161.
28. Otto M., George T. *Anal. Chim. Acta* 200 (1987) 379-385.
29. Kalivas J.H., Roberts N., Sutter J.M. *Anal. Chem.* 61 (1989) 2024-2030.
30. Sasaki K, Kawata S., Minami S. *Appl. Spectrosc.* 40 (1986) 185-190.
31. Liang Xu , Israel Schechter *Anal. Chem.* 68 (1996) 2392-2400.
32. Kalivas J.H. *Anal. Chem.* 55 (1983) 565-567.
33. Kalivas J.H. *Anal. Chem.* 58 (1986) 989-992.
34. Juhl L.L , Kalivas J.H. *Anal. Chim. Acta* 207 (1988) 125-135.
35. Jochum C., Jochum P., Kowalski B.R., *Anal. Chem.* 53 (1981) 85-92.
36. Woordford C. *Solving Linear and non-linear Equations* , Ellis Horwood 1992, England.
37. Weisberg S. *Applied Linear Regression* 2nd Edition, Wiley, New York 1985.
38. Salamin P.A., Bartels H., Foster P. *Chem. Intell. Lab. Syst.* 11 (1991) 57-62.
39. Snee R.D. *Technometrics* 19 (1977) 415-428.
40. Marquardt D.W. *Technometrics* 12 (1970) 591-612.

ISBN:978-84-691-1875-7/DL: T-337-2008

4.4 Equivalence between Selectivity and Variance Inflation Factors in Multicomponent Analysis

Química Analítica 15 (1996) 259-262

J. Ferré, F.X. Rius

Departament de Química. Universitat Rovira i Virgili.
Pl. Imperial Tàrraco, 1, 43005-Tarragona. SPAIN

Diagnostic parameters for detecting collinearity in multicomponent analysis have been known for a long time. Variance inflation factors (VIF's) for least-squares estimates, derived from multivariate statistical modelling, and selectivity are both used as guidelines for deciding when multicollinearity is such that the spectral chemical analysis results should be questioned. Their equivalence is shown in the present article.

Received December 11th 195/ Accepted April 26th 1996

ISBN:978-84-691-1875-7/DL: T-337-2008

Introduction

The term multicomponent analysis (MCA) is used for techniques in which several components in a sample are determined simultaneously. In such cases the linear additive model is frequently used, e.g. models based on the Lambert-Beer law. For a mixture of K components, the response measured at each sensor (wavelength) can be described with Equation (1):

$$r_j = \sum_{k=1}^{K} s_{jk}\ c_k + e_j \qquad (1)$$

where r_j is the measured absorbance of the K-component system at jth wavelength, s_{jk} is the molar absorptivity of the kth component at jth wavelength, c_k denotes the concentration of kth component in the mixture and e_j represents the noise or error in measuring r_j. When J responses are measured, the linear model is described by Equation (2):

$$\mathbf{r} = \mathbf{S}\,\mathbf{c} + \mathbf{e}_r \qquad (2)$$

where $\mathbf{r}$ is the $J\times1$ vector of responses, $\mathbf{S} = [\mathbf{s}_1, \mathbf{s}_2, \ldots \mathbf{s}_k \ldots \mathbf{s}_K]$ is the $J\times K$ matrix of sensitivities, whose columns ($\mathbf{s}_k$) are the spectra of the pure components present in the system when their concentration is equal to 1, $\mathbf{c}$ is the vector of component concentrations and $\mathbf{e}_r$ is the vector of error terms. After determining $\mathbf{S}$ either from standard solutions of individual components or their mixtures, the vector of unknown concentrations of the various analytes contained in an unknown sample whose spectrum is $\mathbf{r}_{un}$ can be obtained with least squares by using Equation (3):

$$\mathbf{c}_{un} = (\mathbf{S}^T\mathbf{S})^{-1}\mathbf{S}^T\,\mathbf{r}_{un} \qquad (3)$$

and the uncertainty in the predicted concentrations is given by Equation (4):

$$\mathrm{var}(\mathbf{c}_{un}) = \sigma^2(\mathbf{S}^T\mathbf{S})^{-1} \qquad (4)$$

where σ^2 is the variance in the measured responses of the unknown sample, which is usually determined from the residuals calculated using the calibration sample set. The kth diagonal element of $\sigma^2(\mathbf{S}^T\mathbf{S})^{-1}$ contains the concentration variance for the kth component.

4.4 *Química Analítica* 15 (1996) 259-262

ISBN:978-84-691-1875-7/DL:T-337-2008

The prediction error expressed in Equation (4) is largely affected by the collinearity among the columns of **S**. Collinearity (also called multicollinearity or ill-conditioning) exists when the columns of **S** are approximately linearly dependent due to overlapped calibration spectra. Collinearity causes numerical instability when solving Equation (3); small relative changes in $\mathbf{r}_{un}$ and **S** can cause large relative changes in $\mathbf{c}_{un}$. This makes the variances and covariances for concentration estimates large. Thus, a high degree of spectral overlap can severely affect the sample concentration estimates. Disappointing results have been obtained (e.g. negative values for components known to be present [1]) because of linearly related absorption curves leading to an extremely unstable system of equations.

Before predicting with the regression model, Equation (3), diagnostic tools can be used to evaluate the extent to which the concentration estimates can be degraded by the collinearity. Some collinearity diagnostics commonly used in regression analysis have been reported [2-4]: the correlation matrix of the regressor variables, the condition number of **S**, variance-decomposition proportions and variance inflation factors (VIF) of the regressors, etc... The condition number of the calibration matrix **S** has been used to select the most suitable wavelength range in Kalman multivariate calibration and classical least squares (CLS) calibration in order to optimize accuracy and precision [5-6]. This measure, however, has been criticized due to its statistical properties[7]. Kalivas [8] used the variance-decomposition proportions to test for the presence of spectral overlap among the pure-component spectra in K-matrix calibration. VIF values have been recommended by Harrison *et al.* [4] as a general measure of collinearity in regression. VIFs can be calculated for each estimated value in a regression equation (here the estimated concentrations) and give information about the error propagation for the respective component. High VIF values point to the presence of high collinearity. Also, VIF is the low limit of the condition number [9].

On the other hand, in spectroscopic determinations, Lorber [10,11] defined selectivity for the *k*th component (SEL_k) as a measure of the degree of non-overlap between the spectrum of the *k*th component and the spectra of the other components. Some wavelength selection procedures [12,13] used the sum of selectivities for all components as the criterion for optimization.

ISBN:978-84-691-1875-7/DL: T-337-2008

This paper shows that the VIF value and selectivity for the *k*th component are two equivalent measures of collinearity. So, optimizing selectivity for a multicomponent system by wavelength selection is nothing more than ensuring that the collinearity among the calibration spectra will not degrade the concentration estimates.

Theory

The VIF for the estimated *kth* component concentration [2,4] is given by Equation (5):

$$VIF_k = (1 - R_k^2)^{-1} \quad (5)$$

where R_k is the multiple correlation coefficient obtained from the regression of the *kth* component spectrum $\mathbf{s}_k$ on the spectra of the rest of the components. By defining the **S** matrix without the *kth* column $\mathbf{s}_k$ as $\mathbf{S}_k = [\mathbf{s}_1, \ldots \mathbf{s}_{k-1}, \mathbf{s}_{k+1}, \ldots \mathbf{s}_K]$, the regression model can be described by Equation (6)

$$\mathbf{s}_k = \mathbf{S}_k\beta_k + \mathbf{e}_k \quad (6)$$

The $\mathbf{S}_k$ columns geometrically define a hyperplane, L, in the *J*-dimensional Euclidean space E*J*. If $\mathbf{s}_k$ is not a linear combination of the $\mathbf{S}_k$ columns it does not lie on this hyperplane, which is represented for two components in Figure 1. The least-squares estimation of β_k is given by Equation (7):

$$\mathbf{b}_k = \mathbf{S}_k^+\mathbf{s}_k \quad (7)$$

where $^+$ indicates pseudoinverse. Vector $\mathbf{b}_k$ is chosen so that the length of the residual vector $\mathbf{e}_k = \mathbf{s}_k - \hat{\mathbf{s}}_k$ is minimal, where $\hat{\mathbf{s}}_k = \mathbf{S}_k\mathbf{b}_k$ is the predicted vector. $\hat{\mathbf{s}}_k$ lies in the L hyperplane since it is a linear combination of the $\mathbf{S}_k$ columns and corresponds to the orthogonal projection of vector $\mathbf{s}_k$ onto hyperplane L. On the other hand, the residual vector $\mathbf{e}_k$ is orthogonal to the hyperplane L, which is precisely the definition of net analyte signal, $\mathbf{s}_k^*$ given by Lorber.

Lorber defined the net analyte signal $\mathbf{s}_k^*$ as the part of the spectrum of the *kth* component that is orthogonal to the spectra of the other components in the mixture

4.4 *Química Analítica* 15 (1996) 259-262

ISBN:978-84-691-1875-7/DL: T-337-2008

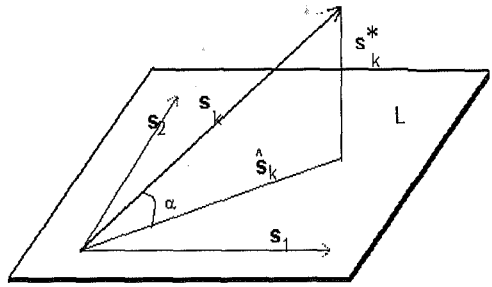


Figure 1. Geometrical representation of the regression of s_k on the spectra of two components defined by Eq (6). The L plane is generated by the S_k columns. s_k* is the residual vector of the regression model.

and is calculated using Equation (8):

$$s_k^* = (I - S_k S_k^+) s_k \tag{8}$$

It is clear that:

$$s_k^* = (I - S_k S_k^+) s_k = s_k - S_k S_k^+ s_k = s_k - S_k b_k = s_k - \hat{s}_k \tag{9}$$

as can be seen in Figure. 1. The coefficient of determination R_k^2 (the squared correlation coefficient) for this model is given by Eq (10) (see e.g. Belsley *et al.* [2]):

$$R_k^2 = 1 - \frac{e^T e}{y^T y} = 1 - \frac{s_k^{*T} s_k^*}{s_k^T s_k} = 1 - \sin^2\alpha \tag{10}$$

where ^{T} means transposed, **e** is the residual vector and **y** is the dependent variable vector. These vectors correspond to s_k and $\hat{s}_k$ respectively. α is the angle between s_k and $\hat{s}_k$ (see Figure.1). So the VIF value for the *k*th component is:

$$VIF_k = \frac{1}{1 - R_k^2} = \frac{1}{\sin^2\alpha} \tag{11}$$

ISBN:978-84-691-1875-7/DL: T-337-2008

The selectivity for the *k*th component is defined according to Equation (12):

$$SEL_k = \|s_k^*\| / \|s_k\| \tag{12}$$

where $\|\cdot\|$ designates the Euclidean norm. Following the geometrical interpretation, it is apparent that:

$$SEL_k = \frac{\|s_k^*\|}{\|s_k\|} = \sin\alpha \tag{13}$$

therefore

$$VIF_k = \frac{1}{SEL_k^2} \tag{14}$$

which shows the equivalence between the diagnostic parameters. When a spectrum corresponding to a calibration sample is orthogonal to the others, VIF=1 and selectivity=1, which is the optimal situation. As the spectral overlap gets larger, VIF increases and selectivity decreases. Depending on the authors, it is assumed [4] that VIF values higher than 7 to 10 indicate that the corresponding least-squares estimates may be so poorly estimated that one should attempt to fit a different multivariate model to the experimental calibration data (e.g. by introducing quadratic terms), select the wavelengths used in the analysis or use an alternative estimation technique (e.g. ridge regression). The equivalent selectivity values range between 0.38 and 0.32 as shown in Table 1. Therefore, selectivity values for a component which are lower than 0.32 indicate that the concentration might be erroneously estimated.

Table 1. Numerical relationship between selectivity and VIF for a component in a mixture.

VIF	selectivity
1	1
4	0.5
7	0.38
10	0.32

ISBN:978-84-691-1875-7/DL: T-337-2008

Acknowledgements

J. Ferré thanks the Comissionat per a Universitats i Recerca of Generalitat de Catalunya, for providing a doctoral fellowship (FI/94-7001). Financial support from the Spanish Ministry of Education and Science (DGICyT project BP93-0366) is gratefully acknowledged.

References

1. Sterans E I, **1953**. Eng. Chem. Anal. Ed 25: 1004
2. Belsley D A, Kuh E, Welsch R E, **1980** "Regression Diagnostics: Identifying Influential Data and Sources of Collinearity". Wiley: New York
3. Myers R H, **1990** "Classical and modern regression with applications" 2nd edition. Duxbury Press: Belmont
4. Harrison M. Wadswoth. **1990** "Handbook of Statistical Methods for Engineers and Scientist" McGraw Hill: USA
5. Pérez-Arribas LV, Navarro-Villoslada F, León-González M E, Polo-Díez L M, **1993**. J. Chemom. 7: 267-275.
6. Navarro-Villoslada F, Pérez-Arribas L V, León-González M E, Polo-Díez L M,**1995**. Anal. Chim. Acta 313: 93-101.
7. Liang Y, Xie Y, Yu R,. **1989**, Anal. Chim. Acta 222:347-357.
8. Kalivas J, **1989**. J. Chemom. 3: 409-418.
9. Berk K N, **1970**. J. Amer. Statist. Assoc. 12: 863-866
10. Lorber A, **1986**. Anal. Chem. 58: 1167-1172.
11. Lorber A, Kowalski B R, **1988**. J. Chemom. 2: 67-79.
12. Lucasius C B, Beckers M L M, Kateman G, **1994**. Anal. Chim. Acta 286: 135-153.
13. Hörchner U, Kalivas J H **1995**. J. Chemom. 9: 283-308.

ISBN:978-84-691-1875-7/DL: T-337-2008

4.5 The effect of wavelength selection in the trueness and precision of analytical results. A tutorial

(in preparation)

Joan Ferré and F. Xavier Rius.

Departament de Química. Universitat Rovira i Virgili.
Pl. Imperial Tarraco, 1, 43005-Tarragona. SPAIN

The ISO definitions of trueness, precision and accuracy are reviewed in general terms and related to the effect on the predicted concentration of different wavelength selection criteria in CLS used in the literature.

ISBN:978-84-691-1875-7/DL: T-337-2008

1. Introduction

One aim of analytical determinations is to give unbiased and precise estimations of the analyte concentrations in future unknown samples. Since this is influenced by each step of the analytical procedure, improving the ability of the calibration model (if used) to give unbiased estimations and reasonable confidence intervals will improve the quality of the analytical result. The performance of the model depends, among others, on the quality of the data used for calibration (e.g. the trueness and precision of the concentrations and instrumental measurements in the calibration samples), on which sensor (or sensors) are used and on the adequacy of the mathematical expression to the measurements of the unknown sample. Since multivariate calibration models based on spectroscopic data are increasingly used, criteria for wavelength selection are constantly studied in the literature. However, the information about the performance of these criteria to improve the precision and trueness of the result is disperse in a considerable number of papers and may be contradictory. For example, wavelength selection based on the condition number of the calibration matrix in CLS has been said to improve either the precision[1], the precision and the accuracy[2], the prediction ability[3], or be uncorrelated with the prediction errors[4] or with the accuracy and the precision[5]. Moreover the authors rarely specify what they consider *accuracy* and *precision* so it is difficult to compare their results. The modern concept of accuracy (trueness and precision) is seldom used and some authors use the term *accuracy* when they are probably referring to *trueness*. This causes confusion to the experimenter who wants to use the most appropriate criterion in his/her wavelength selection problem.

The purpose of this paper is to review the ISO definitions of trueness, accuracy and precision [6-7] and to relate them with the wavelength selection criteria in CLS used in the literature. The capacity of these criteria to improve the prediction results is indicated. Guidelines to achieve optimal trueness and precision of the model building step through wavelength selection in CLS models are given. Although this paper is focused on spectroscopy as a quantitation method and the instrumental responses are spectra, most conclusions can be extrapolated to many analytical methods.

ISBN:978-84-691-1375-1/DL:T-1337-2008

2. Definition of trueness, precision and accuracy

2.1 Trueness

"Trueness refers to the closeness of agreement between the arithmetic mean of a large number of test results and the true or accepted reference value" [6].

2.1.1 How is trueness measured?

The trueness of the measurement method can be investigated by comparing the accepted reference value with the measured or estimated value given by the measurement method[6]. The trueness is normally measured in terms of bias. Bias is the difference between the expectation of the test results and an accepted reference value:

$$\text{bias} = E(c_{un}) - c_{true} \qquad (1)$$

Bias is the total systematic error[6]. Unless the value accepted as *true* is known the trueness cannot be estimated; this means that certified reference materials (CRM), reference methods or interlaboratory collaborative studies among other references are necessary to assess the trueness of the result.

2.1.2 Which factors affect trueness?

Trueness is affected by systematic error components[6]. Some possible sources of bias in chemical analysis are incorrectly calibrated laboratory material, looses or increments of the measurand, the presence of one element that interferes with the determination of another or inadequacy of the calibration model to predict future unknown samples.

2.2. Precision

"Precision refers to the closeness of agreement between independent test results

ISBN:978-84-691-1875-7/DL: T-337-2008

obtained under stipulated conditions" [6]. Precision depends only on the distribution of random errors and does not relate to the true value or the specified value [6]. Precision is the general term for variability between repeated measurements. Reference values are not needed to estimate the precision [8]. Repeatability and reproducibility are the two extremes of precision, the first describing the minimum and the second describing the maximum variability in results.

2.2.1 *How is precision measured?*

The precision with which a given concentration of the component k, c_k, can be obtained using a given analytical method is normally expressed in terms of standard deviation $s(c_k)$, variance $s^2(c_k)$ or relative standard deviation s_k/c_k. with respect to the concentration c_k of component k [8,9].

2.2.2 *Which factors affect precision?*

Many factors may contribute to the variability of results from a measurement method , including[7,10]: the operator, the equipment used, the calibration of the equipment, the environment (temperature, humidity, air pollution, etc.), the batch of a reagent or the time elapsed between measurements. Within the multivariate calibration methodologies the standard deviation $s(c_k)$ is expected to decrease with a growing sensitivity of the method with respect to the components and to increase with the noise of the analytical signal [9]. The standard deviation due to the error propagation depends on the choice of the sensors[11] (wavelengths for the case of spectral data) and is affected by the selectivity of one component with regard to the others[9].

2.3. *Accuracy*

"Accuracy is the closeness of agreement between a test result and the accepted reference value". The term accuracy, when applied to a set of test results, involves a combination of random components and a common systematic error or bias component.

ISBN:978-84-691-1875-7/DL: T-337-2008

2.3.1 How is accuracy measured?

Accuracy is measured according to several statistics. One of the most common is the Mean Squared Error (MSE) between the true concentrations and their estimates for different test samples, defined as $E[(c_{true} - c_{un})^2]$ where $E[\cdot]$ is the expectation operator. This corresponds to the expected squared Euclidean distance between the estimated concentration c_{un} and the true concentration vector c_{true}. The smaller the distance, the closer is c_{un} to c_{true}. This measure can be decomposed into[12]:

$$MSE=E(c_{true}-c_{un})^2 = E[(c_{true}-c_{un})^T(c_{true}-c_{un})] = E[c_{un} - E(c_{un})]^2 + [E(c_{un})-c_{true}]^2 \quad (2)$$

The first term is the variance of c_{un} and the second term is called the bias squared. MSE takes into account systematic deviations (bias) and variance of the prediction errors[13]:

$$MSE= \text{variance} + \text{bias}^2 \quad (3)$$

In practice, MSE is computed as the average squared difference between actual and predicted concentration values for a validation set of I samples:

$$MSE = \sum_{i=1}^{I} (c_{true} - c_{un})^2 / I \quad (4)$$

To obtain a good estimate of the average prediction ability, the set of measurands on which it is based must be representative for the whole population of future unknown measured quantities in question. As c_{un} is the result of the complete analytical procedure, MSE takes into account systematic deviations (i.e. bias) and variance of the prediction errors[13] due to the complete analytical procedure. If the method is unbiased , MSE evaluates the precision.

2.3.2 Which factors affect accuracy?

Since the term accuracy refers to both trueness and precision[6] , accuracy is affected by all the factors that influence trueness and precision.

ISBN:978-84-691-1875-7/DL: T-337-2008

3. The influence of the wavelength selection criteria on the trueness, precision and accuracy of the calibration model.

The objective of many wavelength selection studies is to improve the precision of the model and to assure that no bias is introduced in the result due to a model incorrectly specified for the unknown samples. Using the considerations given above, the next section discusses how different wavelength selection criteria can influence the trueness and precision of the predicted result using a CLS model and thus, of the result of the analysis. The criteria considered are selectivity (SEL) and sensitivity (SEN), accuracy (ACC) and minimum squared error of Sasaki[15] , Det($\mathbf{S}^T\mathbf{S}$) and Tr($\mathbf{S}^T\mathbf{S}$)$^{-1}$ and condition number Cond($\mathbf{S}$). If the word 'accuracy' or 'precision' is followed by the reference to an author, it means that the author used that word to refer to the performance of the criterion. It is possible that the meaning that the author attributed to that word does not correspond to the ISO concepts of accuracy and precision considered above.

3.1 Trueness

Bias in the final result can be introduced in any step of the analytical procedure. In the step which involves predicting with a calibration model, the trueness of the result can be prejudiced by measured data that not accomplish the assumptions of the model. Possible sources of bias are blank problems, baseline shift, spurious peaks, interaction between components (non-additivity of the pure analyte responses), non-linearities (deviation from Beer's law at high absorbance values) and spectral overlap.

3.1.1 Spectral overlap and trueness

Spectral overlap has been considered a source of bias. Wavelength selection has tried to reduce the overlap by optimizing some measure of selectivity. Although overlap does introduce bias in univariate calibration (which requires specific measurements and for this reason the selectivity is important since it is the main reason for using multivariate instead of univariate calibration), this is not necessarily the case in CLS and needs further considerations.

ISBN:978-84-691-1875-7/DL:T-337-2008

Two situations may arise: (a) when the overlap is due to the spectra of known analytes that are included in the model and (b) when the spectra of analytes in the unknown sample not considered in the model overlap the spectra of the analytes of interest. In the first case spectral overlap does not introduce bias since the model is able to discern the signal from the different analytes. The use of not selective sensors does not affect the trueness as long as the analytes that give rise to a signal are considered in the model. However, the larger the spectral overlap and the similarity between the spectra, the smaller is the net signal of the analyte of interest and thus the precision. Frans and Harris[14] considered the CLS model to be accurate if the component spectra in **S** are those actually present in the mixture spectrum. This really corresponds to the definition of trueness. In the case (b) overlap is a source of bias. The trueness of the result might be kept through wavelength selection based on chemical knowledge about the sensors where the possible interferents absorb and not including that spectral range in the model. If the interferent absorbs in all the spectral range, bias can be avoided by including the spectrum of the interferent in the model or using separation techniques to eliminate the interferent.

According to the indicated above, the wavelength selection criteria based only on the calibration matrix do not measure nor necessarily improve the trueness of the result. This is because they only consider the analytes included in the model and cannot guarantee that the unknown samples will be free of any unexpected interferents that absorb in the selected spectral region. In addition, the degree of bias depends on the degree of spectral overlap of the interfering species and usually influences the analytes in a different degree.

Usually, the prediction errors with respect to some reference values are used to prove the efficiency of a selection procedure based on the calibration matrix. In reality, this prediction error evaluates the complete analytical procedure not only the modeling step since the predicted value is the result of all the steps of the analysis. Then, in the wavelength selection in CLS, it must be assumed that that all the steps in the chemical procedure are free from bias and that the model describes correctly the system under study, i.e. there are no uncalibrated constituents that can introduce bias in the final result.

In the bibliography, selectivity and spectral overlap have usually been associated to *accuracy*[16]. This is based on the idea that the propagation of the errors into the

ISBN:978-84-691-1875-7/DL: T-337-2008

analytical result is much more marked for non-selective than for selective procedures, and thus, the wavelengths with the lowest overall overlap (highest selectivity) should yield the most accurate component concentration estimation[17] (minimum errors) in the determined concentrations[18]. Consequently, wavelength selection based on criteria related to the selectivity have been proposed as a strategy to improve the accuracy. Several measures of selectivity have been reported in the literature. The three measures considered below only consider the analytes in the calibration matrix **S** and cannot assure that analytes not considered in **S** will not be present in the unknown sample. Thus one must consider that all the analytes present in future samples are known, and that the postulated model is correct:

- *selectivity after Lorber* ($LSEL_k$)[19]. It measures the degree of non-overlap between the spectrum of the *k*th analyte and the spectra of the other pure components in the system. It has been said to improve the accuracy[17].

- *condition number.* Different results have been presented for the condition number. The set of wavelengths that produces the minimum Cond(**S**) has been said to optimize the accuracy [20,21] or the accuracy and precision of the multicomponent determinations [2,5,22,23]. Moreover, the wavelengths selected according to the condition number has been reported to give good results[3] while other investigations indicate that they do not necessarily result in low prediction errors [1,4,5,24,25] and that the condition number must be regarded as a qualitative tool for error estimation[19].

- Det ($\mathbf{S}^T\mathbf{S}$). A large value of Det($\mathbf{S}^T\mathbf{S}$) has been considered a global measure of accuracy and precision[5] , of selectivity between the vectors in the calibration matrix [5,16,26] and of sensitivity[1].

$LSEL_k$, Det($\mathbf{S}^T\mathbf{S}$) and Cond(**S**) are measures related to the precision of the estimated concentrations in CLS, not to the trueness.

3.2 *Precision*

The precision of the estimated concentration is affected by three sources of error: the random error in the spectra of the unknown sample and on the precision of the

ISBN:978-84-691-1875-7/DL: T-337-2008

concentration and the measurements in the wavelengths used to build the model. The influence of these errors in the precision depends on their propagation through the calibration model, which in turn, depends on the sensors used (the number and which ones), on the mathematical expression of the model and is worsened by the collinearity problem due to spectral overlap. The objective of the wavelength selection is to improve the precision of the estimated concentration. Many wavelength selection criteria try to improve the accuracy based on an improvement in the precision since the trueness of the result is assumed. The variance of the least-squares predicted concentrations in CLS is:

$$\mathrm{var}(c_{un}) = (\mathbf{S}^T\mathbf{S})^{-1}\sigma^2$$

This expression only considers the propagation through the model of the error in the measured spectrum of the problem sample. The criteria for wavelength selection are based on this expression to improve the precision. Some examples are:

- *Variance coefficients.* They are the diagonal elements of $(\mathbf{S}^T\mathbf{S})^{-1}$. They have been said to identify analytical wavelengths which yield superior concentration precision, especially in the case of severe spectral overlap[14].

- *MSE.* The minimum mean square error of Sasaki *et al.*[15], defined as:

$$\mathrm{MSE}= E(c-c_{pred})^T(c-c_{pred})=\sigma^2\,\mathrm{Tr}[(\mathbf{S}^T\mathbf{S})^{-1}]$$

where Tr denotes trace has been used as a criterion for wavelength selection [15,17]. It calculates the difference between the theoretical component concentrations and their estimates, hence the variance in the concentration vector, that originates from the expectation value of the noise. This has been considered a measure of accuracy[15] or accuracy and precision of the multicomponent determinations[5]. The assumption that no bias is present reduces the expression to $\sigma^2\mathrm{Tr}[(\mathbf{S}^T\mathbf{S})^{-1}]$ (the sum of the variances of each estimated concentration) which is a measure of precision. This criterion is the same as the sum of the variance factors, used by [14].Larger subsets of wavelengths will always yield a smaller MSE, i.e. better precision in the concentration estimates. For Liang *et al.* [5] this measure is related to the collinearity of the calibration matrix. The trueness is not involved in this criterion if it is evaluated as $\sigma^2\mathrm{Tr}[(\mathbf{S}^T\mathbf{S})^{-1}]$.

ISBN:978-84-691-1875-7/DL: T-337-2008

Table 1 summarizes the different criteria for wavelength selection used in the literature the comments of the different authors about the capacity of each criterion to improve the accuracy, the precision, the selectivity or the sensitivity. The last column gives the performance of each criterion according to the definitions of trueness, precision and accuracy given in the section 1.

Table 1. The different criteria used in the literature and the measure that has been associated to the criteria. The authors that found that the criteria were not related with the associated measure are indicated by the asterisk(*) (ACC:accuracy, PRE:precision, SEL:selectivity, SEN: sensitivity)

	As a measure of:					
Criterion	ACC	PRE	ACC and PRE	SEL	SEN	New consideration
selectivity			16			PRECISION
sensitivity s_k*				18,4		PRECISION
Lorber's selectivity	17			18*,4*		ERROR PROPAGATION
Cond(**S**) or Cond($\mathbf{S}^T\mathbf{S}$)		1	5*2	4*,22 16,14		PRECISION
Det($\mathbf{S}^T\mathbf{S}$)		1	14*,5	16*,5,26	9	PRECISION
variance factors		14				PRECISION
MMSE	15	17	5			PRECISION
more spectral measurements	5					ACCURACY

The last column indicates our interpretation of the effect of the criteria according to the actual knowledge.

ISBN:978-84-691-1875-7/DL: T-337-2008

3. Classification of the criteria for wavelength selection

The large number of criteria used for wavelength selection has lead us to classify them and to clarify their paper in the wavelength selection problem.

a) *Depending on the number of analytes to predict.*

- *global criterion.* The property (e.g. prediction ability) of the sensors that optimize this criterion is a compromise for all the analytes in the sample (e.g. $Det(\mathbf{S}^T\mathbf{S})$, $Tr(\mathbf{S}^T\mathbf{S})^{-1}$, $Cond(\mathbf{S}^T\mathbf{S})$, RMSEPT).
- *local criterion.* The property (e.g. prediction ability) of the sensors that optimize this criterion are optimal for a specific analyte although may be not for the other analytes (RMSEP for one analyte, $LSEL_k$, VIF_k, variance coefficients of a particular analyte).

b) Depending of the data used.

- *criteria based on the calibration matrix:* $Det(\mathbf{S}^T\mathbf{S})$, $Tr(\mathbf{S}^T\mathbf{S})^{-1}$, $Cond(\mathbf{S}^T\mathbf{S})$, maximize the minimum eigenvalue?, VIF_k
- *criteria that use a validation set*: RMSEP.

At the same time, these criteria can be either global (RMSEPT) or local ($LSEL_k$, $RMSEP_k$).

c) Depending on the number of criteria used

- *individual criterion*: only one criterion is optimized. The sensors that optimize one individual criterion may be different from the ones that optimized another criterion (e.g. $Det(\mathbf{S}^T\mathbf{S})$ and $Tr(\mathbf{S}^T\mathbf{S})^{-1}$
- *composite criterion (multicriteria)* : the selected sensors are an acceptable compromise for several criteria at the same time. These sensors may be not optimal for a given criterion, but sub-optimal. Composite criteria have not been found in the literature of wavelength selection yet.

ISBN:978-84-691-1875-7/DL: T-337-2008

4. Conclusions

The effect of different wavelength selection criteria in CLS on the precision, trueness and accuracy of the predicted concentration have been considered. These criteria are an important improvement with respect to using a experimenter's subjective choice of wavelengths. However, the criteria based only on the calibration matrix **S** cannot guarantee the trueness of the final result since they cannot prevent the absorbances in selected sensors from being free of systematic errors in future unknown samples. Therefore, the efficiency of these criteria is limited to the cases when all the steps in the analytical procedure in addition to the modeling step give unbiased results. Previously to using criteria such as the $\mathrm{Det}(\mathbf{S}^T\mathbf{S})$ or the $\mathrm{Tr}(\mathbf{S}^T\mathbf{S})^{-1}$ the sensors that do not follow the model due to an interferent absorbs in future unknown samples must be eliminated. Otherwise the criteria could select these sensors. Diagnostics are necessary to detect if the sample to be predicted follows the model in the selected sensors or otherwise may contain a systematic error. A criterion for wavelength selection not considered here is the prediction error (e.g. RMSEP). Although this has the advantage that may discard sensors that contain more error, it is highly dependent on the validation samples used.

References

1. Smeyers-Verbeke J., Detaevernier M.R., Massat D.L. *Anal. Chim. Acta* 191 (1986) 181-192.
2. Kalivas J.H. *Anal. Chem.* 55 (1983) 565-567.
3. Navarro-Villoslada F., Pérez-Arribas, L.V., León-González M.E., Polo-Díez L. M. *Anal. Chim. Acta* 313 (1995) 93-101.
4. Bauer G., Wegscheider W., Ortner H.M. *Spectrochimica Acta 47B* (1992) 179-188.
5. Liang Y., Xie Y ., Yu R. *Anal. Chim. Acta* 222 (1989) 347-357.
6. ISO 5725-1 (1994) *Accuracy (trueness and precision) of measurement methods and result. Part 1. General principles and definitions.* International Organization for Standardization, Geneva, Switzerland.

ISBN:978-84-691-1875-7/DL:T-237-2008

7. Eurachem 1994. *Quantifying Uncertainty in Analytical Measurement.* Available from the EURACHEM Secretariat, P.O. BOX 46, Teddington, Middlesex, TW11 0NH, UK.
8. Zscheile Jr. F.P., Murray H.C., Baker G.A., Peddicord R.G *Anal. Chem.* 34 (1962) 1776-1780.
9. Bergmann G., von Oepen B., Zinn P. *Anal. Chem.* 59 (1987) 2522-2526.
10. Green J.M. *Anal. Chem.* ?? (1996) 305A-309A.
11. Massart D.L., Vandeginste B.G.M., Deming S.N., Michotte Y., Kaufman L. *Chemometrics: a textbook.* Elsevier. Amsterdam. 1988.
12. Frank I.E. *Trends Anal. Chem.* 6 (1987) 271-275.
13. Marbach R., Heise H.M. *Chem. Intell. Lab. Syst.* 9 (1990) 45-63.
14. Frans S.D., Harris J.M., *Anal. Chem.* 57 (1985) 2680-2684.
15. Sasaki K., Kawata S., Minami S. *Appl. Spectrosc.* 40 (1986) 185-190.
16. Otto M., Wegscheider W. *Anal. Chim. Acta* 180 (1986) 445-456.
17. Lucasius, C.B., Beckers M.L.M., Kateman G. *Anal. Chim. Acta* 286 (1994) 135-153.
18. Bauer G., Wegscheider W., Ortner H.M. *Spectrochimica Acta 46B* (1991) 1185-1196.
19. Lorber A. *Anal. Chem.* 58 (1986) 1167-1172.
20. Kalivas J.H., Lang P. *J. Chemom.* 3 (1989) 443-449.
21. Juhl L.L., Kalivas J.H. *Anal. Chim. Acta* 207 (1988) 125-135.
22. Otto M., Wegscheider W. *Anal. Chem.*57 (1985) 63-69.
23. Kalivas J.H. *Anal. Chem.* 58 (1986) 989-992.
24. Otto M., George T. *Anal. Chim. Acta* 200 (1987) 379-385.
25. Kalivas J.H., Lang P. M. *Mathematical Analysis of Spectral Orthogonality,* Marcel Dekker, New York (1994).
26. Warren F.V., Bidlingmeyer B.A., Delaney M.F. *Anal. Chem.* 59 (1987) 1890-1896.

ISBN:978-84-691-1875-7/DL: T-337-2008

4.6 Figures of merit in multivariate calibration. Determination of four pesticides in water by FIA and spectrophotometric detection

Anal. Chim. Acta 348 (1997) 167-175

J. Ferré, R. Boqué, B.Fernández-Band[1], M.S. Larrechi and F.X. Rius*

Departament de Química. Universitat Rovira i Virgili de Tarragona
Pça Imperial Tàrraco, 1, 43005 Tarragona, Spain

[1] On leave from Universidad Nacional del Sur. Bahía Blanca. Argentina*A. Rius*, M.P.*

The accuracy, trueness and determination limit of a FIA method are evaluated in the simultaneous determination of the pesticides Carbaryl (RYL), Carbofurane (CBF), Propoxur (PPX) and Isoprocarb (IPC) in water by multicomponent analysis. Calibration is based both on the spectra of artificially made samples according to the experimental design theory and the spectra of pure pesticides. Prediction errors in the range 0.1-1.4 evaluated as RMSEP are obtained. The absence of bias is evaluated from the joint confidence interval test for the regression line obtained from measured and predicted concentrations taking into account errors in both axes. Multivariate determination limits were found to be between 0.03 and 1.0 ppm.

Received in revised form 3 February 1997; accepted 7 February 1997

ISBN:978-84-691-1875-7/DL: T-337-2008

1. Introduction

Pesticides of the carbamate family are widely used in agriculture because of their powerful biological activity [1]. Since they are also serious environmental pollutants, a considerable number of analytical procedures have been proposed to determine and control their presence in surface waters [2-12]. Of these procedures, the spectrophotometric methods use the reaction between the pesticide, previously hydrolysed to its naphtol, and different reagents to produce strongly coloured species. These methods have also been used coupled with a FIA system, due to the simple instrumentation and high analysis speed. Khalaf et al. developed a flow system with initial liquid-liquid extraction to spectrophotometrically determine Carbaryl [13] and Propoxur [14] with *p*-aminophenol in natural waters. Fernández-Band et al. [15] simultaneously determined three pesticides using a FIA system with in-situ pre-concentration in a C18 stationary phase of the complexes formed. Espinosamansilla et al. [16-17] also describe a stopped-flow detection system for the determination of Carbaryl, which is based on the degradation speed of the pesticide in an alkaline medium. García et al [18] describe the simultaneous determination of Propoxur, Carbaryl, Ethiofencarb and Formetanate by using *p*-aminophenol and partial least-squares regression.

In the above mentioned papers, the figures of merit determined are not associated to the multivariate nature of the analysis but to the response of the analytes measured on a single channel. So, the limits of detection are either associated to the lower limit of the linearity obtained for each pesticide at a single wavelength or calculated as the concentration derived from a response equivalent to three times the standard deviation of the method. Also, precision is evaluated in terms of standard deviation calculated in conditions of repeatability and associated to replicated analyses of samples with the same concentration of analytes. In addition, the absence of systematic errors is usually detected by measuring the recovery percentage in samples which have been spiked with the analyte under study.

In this paper the accuracy, trueness and determination limit of a FIA method are evaluated when it is used to determine four pesticides (Carbaryl, Carbofurane, Propoxur and Isoprocarb) in water with the classical least-squares regression method (also called the K-matrix approach). Calibration is based on the spectra of artificially

ISBN:978-84-691-1875-7/DL: T-337-2008

made samples, either the pure pesticides or mixtures prepared according to the experimental design theory. The methodology has been validated with 9 artificially prepared drinking waters and 6 samples of ground and river waters spiked with pesticides.

2. Theoretical background

2.1 Calibration model

The classical least-squares model in spectrophotometric analysis of several components is based on measurements of absorbances at selected wavelengths according to Beer's law (eq 1)

$$\mathbf{r} = \mathbf{S}\,\mathbf{c} + \mathbf{e} \qquad (1)$$

where the vector $\mathbf{r}_{J\times 1}$ represents the absorbances measured at J wavelengths, $\mathbf{S}_{J\times K} = [\mathbf{s}_1, \mathbf{s}_2, \ldots \mathbf{s}_K]$ is the matrix of molar absorptivities for J wavelengths and K components, $\mathbf{c}_{K\times 1}$ is the concentration vector for K components and $\mathbf{e}_{J\times 1}$ is the vector of error terms. The unknown concentrations of the various analytes contained in a sample whose spectrum is $\mathbf{r}_{un}$ can be obtained by using eq 2.

$$\mathbf{c}_{un} = (\mathbf{S}^T\mathbf{S})^{-1}\mathbf{S}^T\,\mathbf{r}_{un} \qquad (2)$$

where T means transposed. The variance associated to the estimated concentrations is given by:

$$\mathrm{var}(\mathbf{c}_{un}) = s^2\,(\mathbf{S}^T\mathbf{S})^{-1} \qquad (3)$$

where s^2 is the variance of the spectral measurements , evaluated as:

$$\sigma^2 = \mathbf{r}^T_{un}(\mathbf{I} - \mathbf{S}\mathbf{S}^+)\mathbf{r}_{un}/(J - K) \qquad (4)$$

where + means pseudoinverse. The matrix S is determined by recording the spectra of standard solutions of individual components or mixtures:

ISBN:978-84-691-1875-7/DL: T-337-2008

$$\mathbf{S}^T = (\mathbf{C}^T\mathbf{C})^{-1}\mathbf{C}^T\mathbf{R} \quad (5)$$

where $\mathbf{C}_{I\times K}$ is the concentration matrix of the K components in I standard solutions of individual components or mixtures and $\mathbf{R}_{I\times J}$ is the absorbance matrix of the standard solutions with J wavelengths (columns) and I rows. If pure component samples are used, $(\mathbf{C}^T\mathbf{C})^{-1}$ is a diagonal matrix.

Application of Eq. 2 requires matrix **S** to be error-free. This is not so because the matrix is estimated from Eq. (5); however, the error can be decreased by selecting the mixtures according to a designed plan. A Hadamard matrix can be used since it provides minimum variance estimators for a given number of samples. Its performance can later be compared with the one evaluated from the pure component spectra.

2.2 *Selectivity and sensitivity measures*

The spectral selectivity and sensitivity in the **S** matrix are known to influence the prediction ability of the model. The following measures were considered here:

Selectivity for the kth component [19] is evaluated from Eq. (6):

$$SEL_k = \|\mathbf{s}_k^*\| / \|\mathbf{s}_k\| \quad (6)$$

where $\mathbf{s}_k^* = (\mathbf{I} - \mathbf{S}_k\mathbf{S}_k^+)\mathbf{s}_k$ is the net analyte signal and $\|\cdot\|$ denotes the Euclidean norm. $\mathbf{S}_k$ is the matrix of the spectra of the pure constituents except the k analyte, $\mathbf{S}_k\mathbf{S}_k^+\mathbf{s}_k$ is the null component of $\mathbf{s}_k$, i.e. the part of spectrum $\mathbf{s}_k$ that is contained in the spectra of the other constituents present in the sample. This measure has been shown to be equivalent to the variance inflation factors (VIFs) [20] often used as multicollinearity performance characteristics. Low selectivity values are associated to unstable estimations of the concentrations and large prediction errors for new samples.

Sensitivity for the kth component [19] refers to how large the analyte responses are at each sensor in terms of net analytical signal. It is defined as:

ISBN:978-84-691-1875-7/DL: T-337-2008

$$SEN_k = \|s_k^*\| \quad (7)$$

This measure is directly related to the confidence interval for the estimated concentrations in CLS [22].

2.3 Accuracy

In order to take into account all the causes of variability of the method, accuracy in the concentrations was assessed by root-mean-squared error of prediction (RMSEP) of a set of I validation samples not included in the calibration set, which were analysed in reproducible conditions:

$$RMSEP = \sqrt{\frac{\sum_{i=1}^{I} (c_{i,\text{known}} - \hat{c}_{i,\text{un}})^2}{I}} \quad (8)$$

2.4 Trueness

The absence of bias in the model [23] was assessed by performing a joint statistical test for the slope and the intercept in the regression of c_{un} versus c_{known} for the test samples taking into account the errors in both axes [24]. This test needs to know the variance of the true concentrations, which was considered to be constant for all samples and was evaluated by error propagation in the preparation of the standard solutions, and the variance of the predicted concentrations, evaluated from eq 3.

2.5 Limits of determination

The limits of determination were calculated with the expression derived by Boqué and Rius [25], which is based on the theory of the hypothesis tests applied to the concentration domain. The null hypothesis, H_0: $c = 0$ (analyte not present in the sample) is tested against the alternative hypothesis, H_1: $c > 0$ (analyte present in the

sample), where c is the true but unknown concentration of the analyte in the sample.

The procedure is to reject H_0 when the statistic $t = (\hat{c} - c_0) / \hat{\text{var}}(c_0)^{1/2}$ is higher than t_a, the a-percentage point of the Student's t-distribution with v degrees of freedom. $\hat{c}$ is the estimated value of c and $\hat{\text{var}}(c_0)$ is the estimated variance at 'zero concentration level'.

The power of the above Student's t-test is given by $1 - b = \text{pr}\{ t(\Delta) > t_a \}$, where b is the probability of committing a false negative (i.e. erroneously accepting H_0) and 1Δ is the noncentrality parameter of the noncentral t-distribution, $t(\Delta)$, with v degrees of freedom.

If the probabilities a and b are fixed then the noncentrality parameter can be computed and a multivariate determination limit, MDL, for the kth analyte can be calculated:

$$(\text{MDL})_k = \Delta(\alpha,\beta)\ \text{var}(c_{0,k})^{1/2} \qquad (9)$$

The noncentrality parameter, $\Delta(\alpha,\beta)$, can either be computed or obtained from the tables [26]. The variance at 'zero concentration level', $\text{var}(c_{0,k})$, can be estimated from any expression of the variance of the predicted concentration (eq 3), but only under the null hypothesis, i.e. the kth analyte is not present in the sample. If this equation is used to estimate $\text{var}(c_{0,k})$, the degrees of freedom in the calculation of t_a and $\Delta(\alpha,\beta)$ are $v = J - K$, where J is the number of wavelengths and K is the number of analytes. σ^2 is in this case the variance of the spectral measurements at zero concentration level. However, if the measurement errors are assumed to be homoscedastic, σ^2 can be estimated from Eq (4) or calculated from replicates on different samples.

3. Experimental

3.1 Instrumentation

The following equipment was used to build the FIA system: a Hewlett-Packard 8452A diode array spectrophotometer controlled by an HP Vectra 386s/20 computer equipped with an HP-IB IEEE 488 interface for communications, two Gilson

4.6 *Anal. Chim. Acta* 348 (1997) 167-175

ISBN:978-84-691-1875-7/DL: T-337-2008

Minipuls-3 peristaltic pumps, a Rheodyne 5041 injection valve, a Hellma 178.713 QS (10 mm optical path) flow cell and OMNIFIT tubing.

3.2 Reagents

Standard solutions of Carbofurane, Propoxur, Carbaryl and Isoprocarb (from Riedel-deHaën) were prepared by dilution of stock solutions with Millipore water. Aqueous solutions of 0.2% $NaNO_2$, 2M NaOH and 0.2% sulfanilic acid in 30% of acetic acid were used. Millipore water was used as carrier.

3.3 Manifold

The FIA system is shown in Fig. 1, according to the optimal parameters described in reference 15. The time for measuring the spectrum at the FIA peak maximum was found to be 85 seconds. The spectrum of a blank sample was recorded and substracted from each sample spectrum.

3.4 Procedure for determining the pesticides

The sample was inserted into the carrier merging with a basic stream to favor hydrolysis of the analytes along reactor R1. Solutions of $NaNO_2$ and sulfanilic acid

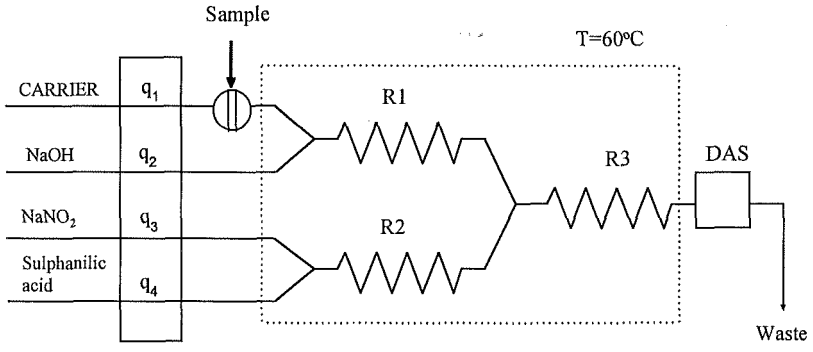


Fig. 1. FIA manifold used for on-line determination of carbamate compounds based on hydrolysis of the analytes and dye formation. (q = peristaltic pump, R = reactor, DAS = diode array spectrophotometer).

ISBN:978-84-691-1875-7/DL: T-337-2008

were also merged to facilitate formation of diazotized sulfanilic acid along reactor R2. The subsequent confluence of R1 and R2 resulted in the formation of the corresponding dyes along R3. Hydrolysis and derivative reactions were boasted by immersing the reactors in a thermostated bath at 60°C. The colour appears instantaneously and, in the flow conditions used [15], the reaction is developed enough for the signal recorded in the detection cell to be perfectly quantifiable.

3.5 Software

All calculations were made using Matlab home-made subroutines.

3.6 Samples

Data were collected to enable the model built from the pure component spectra and the spectra of the mixtures to be compared. Spectra were recorded between 340-650 nm every 2 nm and grouped into the following calibration and validation sets:

- *Calibration set 1* consisted of 5 replicates of the pure pesticide spectra at a concentration of 10 ppm of each pesticide. The value of 10 ppm was chosen in order to obtain high sensitivity.
- *Calibration set 2* consisted of 4 replicates of mixtures designed according to a Hadamard matrix of 4 variables and 8 samples between 2 and 8 ppm
- *Validation set 1* consisted of 3 replicates of 9 mixtures of the spiked four pesticides in tap water between 2 and 8 ppm. The mixtures correspond to a 3 level Hoke design D1 plus a point in the centre. This design was selected because it covered the experimental domain with 3 levels of concentrations and required a small number of samples. The mixtures that had already used in the Hadamard matrix were not considered here.
- *Validation set 2* consisted of the spectrum of 6 real samples from ground and river waters spiked with 5 ppm of each pesticide.

The designed mixtures (ppm of each pesticide) are shown in Table 1.

ISBN:978-84-691-18[illegible]

Table 1. Experimental designs. A) Calibration set 1 Each sample was prepared 5 times. B) Calibration set 2. Hadamard design. Each sample was prepared 4 times. C) Validation set according to the Hoke design D1 plus a point in the centre. Each sample was prepared 3 times. Concentrations of each pesticide in ppm. (RYL = Carbaryl, CBF = Carbofurane, PPX = Propoxur, IPC = Isoprocarb).

Sample	(A) Calibration set 1				(B) Calibration set 2				(C) Validation set 1			
Number	RYL	CBF	PPX	IPC	RYL	CBF	PPX	IPC	RYL	CBF	PPX	IPC
1	10	0	0	0	8	8	8	2	8	5	5	5
2	0	10	0	0	8	8	2	8	5	8	5	5
3	0	0	10	0	8	2	8	2	5	5	8	5
4	0	0	0	10	2	8	2	2	5	5	5	8
5					8	2	2	8	8	2	8	8
6					2	2	8	8	8	8	2	2
7					2	8	8	8	2	8	8	2
8					2	2	2	2	2	8	2	8
9									5	5	5	5

4. Results and discussion

Two models were built with calibration set 1 (pure component samples) and calibration set 2 (mixture samples) according to eq 5. Fig. 2 shows the pure component spectra for each pesticide at 10 ppm evaluated from the two calibration sets. The considerable difference for the spectra of carbaryl (RYL) at low wavelengths estimated from both calibration sets may be because of interactions between the analytes, which cannot be detected from the pure analyte samples, and kinetical effects. The same can be said of propoxur (PPX). Because several analytes are present in real samples, calibration from the samples of mixtures using the Hadamard matrix is expected to give estimates of **S** with less uncertainty than the ones estimated from the pure component samples.

Table 2 shows the selectivity and sensitivity values associated to the matrix **S** (calculated from eq 5) for the calibration sets. As it is expected from the Fig. 2, Carbaryl is the component with highest selectivity and sensitivity while Carbofurane, Propoxur and Isoprocarb have lower values due to their similar spectra (collinearity).

ISBN:978-84-691-1875-7/DL: T-337-2008

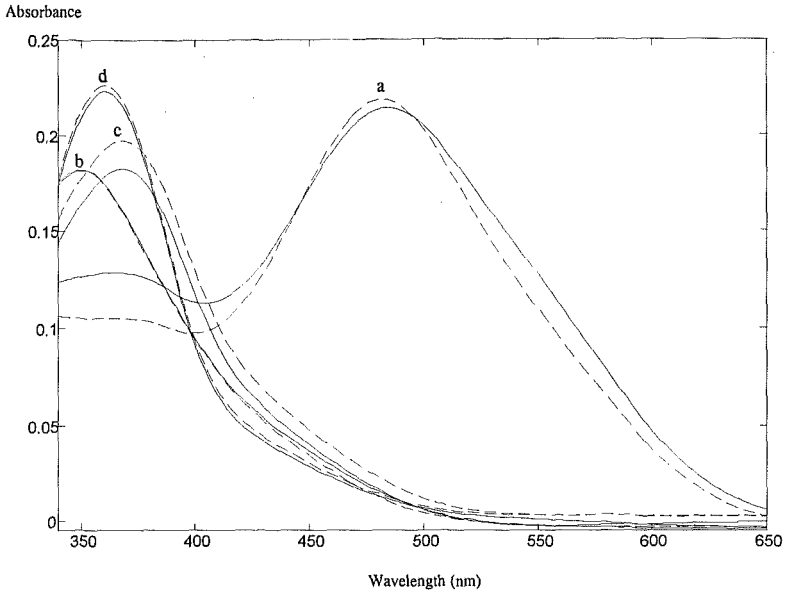


Fig. 2. UV absorption spectra of the four pesticides: (a) Carbaryl, (b) Carbofurane, (c) Propoxur and (d) Isoprocarb, measured from solutions of 10 ppm of the pure pesticides (solid line) and calculated from the calibration mixtures (dashed line).

Table 2. Individual values of sensitivity , SEN_k , and selectivity , SEL_k , for the pesticides in calibration sets 1 and 2 and accuracy values (RMSEP) for validation set 1. (RYL = Carbaryl, CBF = Carbofurane, PPX = Propoxur, IPC = Isoprocarb).

	Calibration set 1		*Validation Set 1*	*Calibration set 2*		*Validation Set 1*
	SEN_k	SEL_k	RMSEP	SEN_k	SEL_k	RMSEP
RYL	0.122	0.764	0.38	0.135	0.797	0.41
CBF	0.011	0.117	2.29	0.010	0.111	0.46
PPX	0.015	0.144	0.74	0.014	0.142	0.32
IPC	0.014	0.137	2.08	0.015	0.138	0.82

ISBN:978-84-691-1875-7/DL: T-337-2008

Collinearity can also be assessed from the variance-decomposition proportions [27] given in Table 3, where each column decomposes the variance of each pesticide as a function of the eigenvalues of the S^TS matrix. 90.4% (0.202 + 0.702) of the variance of RYL (first and second rows of Table 3) is associated to the first and second largest eigenvalues and has no important collinearity with the other pesticides since in these rows the contributions from the other pesticides are practically zero. This agrees with the selectivity observed in the RYL spectra with respect to the other pesticides. The fourth row in the table shows that the smallest eigenvalue considerably contributes to the variance of CBF (99.8%), PPX (38.1%) and IPC (40.1%). These large variance proportions associated to this small eigenvalue are due to the collinearity between the three pesticides. Similarly, the third row indicates that a considerable part of the PPX and IPC variance (61.6% and 59.4% respectively) is only due to collinearity between them.

According to the sensitivity and selectivity values as well as the variance-decomposition proportions, the estimated concentrations for RYL should be more accurate than the estimations for CBF, PPX and IPC.

Table 2 shows the RMSEP values for validation set 1. RMSEP values are considerably smaller for the models based on mixtures than for the models evaluated from pure components (except for a slight increase in the prediction error of RYL that cannot be considered significant). This can be explained by the effect of the Hadamard matrix on the calibration step and the effect of interactions that cannot be taken into account when calibrating with pure pesticide samples. The prediction results from calibration set 1 are in agreement with the pesticides' sensitivity and selectivity values, since the lowest prediction error corresponds to RYL, which has the

Table 3. Variance-decomposition proportions evaluated from the calibration matrix (RYL = Carbaryl, CBF = Carbofurane, PPX = Propoxur, IPC = Isoprocarb)

eigenvalue	RYL	CBF	PPX	IPC
0.0446	0.2020	0.0003	0.0007	0.0009
0.0131	0.7024	0.0010	0.0023	0.0035
0.0002	0.0659	0.0006	0.6163	0.5941
0.0001	0.0297	0.9981	0.3807	0.4014

ISBN:978-84-691-1875-7/DL: T-337-2008

largest SEN and SEL values. The large prediction errors of CBF and IPC may therefore be associated to spectral collinearity. Although this is true for calibration set 1, the results from calibration set 2 cannot be explained so straightforwardly from their sensitivity and selectivity values, suggesting that these measures alone do not explain the final prediction error. This has already been noticed in previous works [28]. Nevertheless, due to its better prediction ability, the calibration model made with the mixtures was used to calculate the figures of merit in the following sections.

Table 4 shows the values of the different parameters of the straight lines obtained by regressing the actual concentrations on the predicted ones for each of the pesticides taking into account uncertainties in both axes. The individual variance values calculated for the true concentrations are considered constant at 10^{-3} ppm while the mean values of the variances for the predicted concentrations, calculated from the multivariate calibration model using eq. (3), were 10^{-3} (RYL), 2×10^{-1} (CBF), 10^{-1} (PPX) and 10^{-1} (IPC) respectively.

The joint confidence interval test of the slope and the intercept, taking into account errors in both axes, proved that the methodology is free from bias for the four pesticides, at a 90%, 97.5%, 90% and 99.99% level of significance for RYL, CBF, PPX and IPC respectively. Fig. 3 shows the confidence ellipse for the pesticide Carbaryl, where the center corresponds to the coordinates intercept = 0.24 and slope = 0.95 (see Table 4). It can be seen that the theoretical point (0,1) is within the joint confidence interval of the ellipse for $\alpha = 0.10$

Table 4. Values of intercepts, slopes, corresponding standard deviations and correlation coefficients for the regression lines of true concentrations versus predicted concentrations for each pesticide. (RYL = Carbaryl, CBF = Carbofurane, PPX = Propoxur, IPC = Isoprocarb

	Slope	s.d. slope	Intercept	s.d. Intercept	r
RYL	0.95	0.03	0.24	0.18	0.9847
CBF	1.02	0.04	-0.31	0.23	0.9849
PPX	1.00	0.02	-0.05	0.12	0.9921
IPC	0.95	0.05	0.59	0.26	0.8600

ISBN:978-84-691-1875-7/DL: T-337-2008

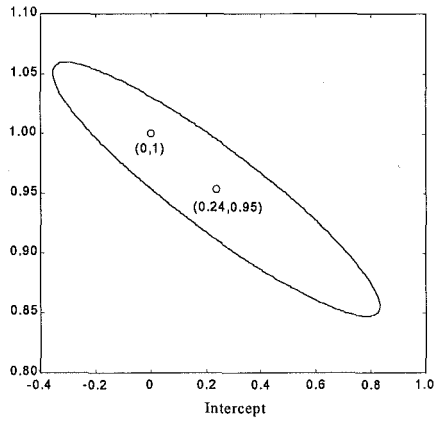


Fig. 3. Confidence ellipse for the slope and the intercept of the straight line obtained regressing c_{un} on c_{known} taking into account the uncertainties in both axes for the pesticide Carbaryl.

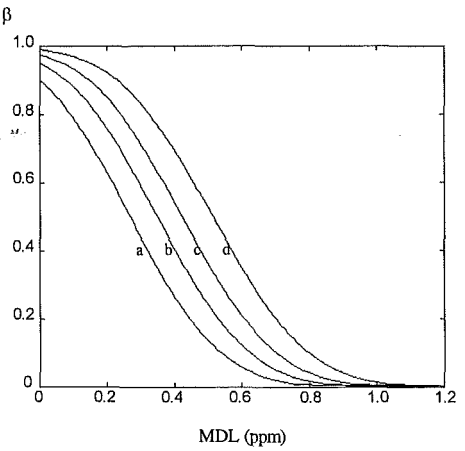


Fig. 4. Characteristic curve of determination of the pesticide isoprocarb (IPC) at different probabilities of type I error, a: (a) 0.10, (b) 0.05, (c) 0.025 and (d) 0.01.

The limits of determination, MDL, for each pesticide were computed according to Eq (9). The term σ^2 was obtained experimentally using validation set 2 by calculating the variance of the spectra of the 10 replicates and selecting the maximum in order to estimate the error in the response vector. Table 5 shows the MDL's calculated with different α and β probabilities of error. It can be observed that the MDL values are highly correlated with the selectivity values so that for a given value of α and β, Carbaryl has the lowest determination limits and Carbofurane has the highest values.

Table 5. Limits of determination (ppm) of the 4 pesticides computed according to Eq (8) for different a and b probabilities of error. (RYL = Carbaryl, CBF = Carbofurane, PPX = Propoxur, IPC = Isoprocarb).

	$\alpha = 0.05$				$\alpha = 0.10$			
	$\beta = 0.05$	$\beta = 0.10$	$\beta = 0.20$	$\beta = 0.50$	$\beta = 0.05$	$\beta = 0.10$	$\beta = 0.20$	$\beta = 0.50$
RYL	0.08	0.07	0.06	0.04	0.07	0.06	0.05	0.03
CBF	1.04	0.92	0.78	0.52	0.91	0.80	0.66	0.40
PPX	0.75	0.67	0.56	0.37	0.66	0.58	0.47	0.29
IPC	0.71	0.63	0.53	0.35	0.62	0.54	0.45	0.27

ISBN:978-84-691-1875-7/DL: T-337-2008

Fig. 4 shows, as an example, the characteristic curve of determination for Isoprocarb, i.e. the representation of MDL as a function of the β probability of error at different probabilities of a type I error. It can be observed that for a given probability, the MDL increases when a probabilities decrease. On the other hand, given a fixed probability of type I error, α, low MDL can only be obtained by increasing the β probabilities of error.

4.1 Validation with real data

The method has also been validated using real river and ground water samples spiked with 5 ppm of each pesticide. The mean concentrations found for three replicated measurements per sample and the percentages of resulting recoveries are shown in Table 6. It can be observed that the RMSEP values calculated according to eq. (8) are correlated with the values obtained from validation set 1 (Table 2). Propoxur gives rise to the most accurate results while Isoprocarb can be determined with a certain degree of inaccuracy.

Table 6. Mean value of the concentrations found from 3 replicates and percentage of recovery (% R) for the six river samples and ground water spiked with 5 ppm of each pesticide (validation set 2). (RYL = Carbaryl, CBF = Carbofurane, PPX = Propoxur, IPC = Isoprocarb)

	[RYL]	%R	[CBF]	%R	[PPX]	%R	[IPC]	%R
F1 - river	5.19	103.8	4.65	92.9	4.95	99.0	6.05	120.9
F2 - river	5.12	102.3	4.70	93.8	5.04	100.8	6.31	126.1
VE - river	5.46	109.1	4.47	89.4	5.16	103.3	6.47	129.3
J1 - ground	5.70	114.0	3.93	78.6	5.04	100.8	6.81	136.2
J2 - ground	5.44	108.9	4.45	89.0	4.97	99.5	6.14	122.7
J3 - ground	5.57	111.4	4.52	90.4	5.09	101.9	6.59	131.8
RMSEP (global)	0.49		0.61		0.13		1.43	

ISBN:978-84-691-0021-2/DL:T-337-2008

5. Conclusions

Multivariate analysis is gaining importance nowadays and methods using this technique should be validated accordingly. In the present paper, important performance characteristics such as accuracy, trueness and determination limits are reported for a FIA method using classical least squares regression. The calculation and interaction of these figures of merit with other performance measurements of the calibration model such as selectivity, sensitivity and the variance-decomposition proportions are shown. The main problem found was collinearity. The specific methodology proposed to determine pesticides in water requires a preconcentration step to reach the concentration level necessary for spectrophotometric determination of real samples within the legal limits. By accepting a global error of approximately 10%, and considering that 60% of the error corresponds to the previous steps of the analysis and 40% to the instrumental determination itself, the statistical tests performed indicated that only 3 of the four pesticides could be determined with acceptable accuracy.

Acknowledgements

J. Ferré thanks the Comissionat per a Universitats i Recerca of Generalitat de Catalunya, for providing a doctoral fellowship (FI/94-7001). B. Fernández-Band thanks the Spanish Ministry of Education and Science for the Intercampus fellowship received.

References

1. J.H. Ruzicka, Proc. Soc. Anal. Chem., 10 (1973) 32.
2. Ballesteros, M. Gallego and M. Valcárcel, J. Chromatography, 633 (1993) 169.
3. Daghbouche, S. Garrigues and M. de la Guardia, Anal. Chim. Acta, 314 (1995) 203.
4. D. Barceló, S. Chiron, S. Lacorte, E. Martínez, J.S. Salau and M.C. Hennion, Trends Anal. Chem., 13 (1994) 352.
5. M.T. Tena, M.D. Luque de Castro and M. Valcárcel, J. Chromatographic Science,

ISBN:978-84-691-1875-7/DL: T-337-2008

30 (1992) 276.

6. M.R. Driss, M.C. Hennion and M.L. Bouguerra, J. Chromatography, 639 (1993) 352.
7. J.A. Perez Lopez, A. Zapardiel, E. Bermejo, E. Arauzo and L. Hernandez, Fresenius J. Anal. Chem., 350 (1994) 620.
8. A. Guiberteau, T.G. Diaz, F. Salinas and J.M. Ortiz, Anal. Chim. Acta, 305 (1995) 219.
9. M.C. Quintero, M. Silva and D. Pérez-Bendito, Talanta, 36 (1989) 717.
10. K. M. Appaiah, R. Ramakrishna, R.R. Sabbarao and O. Kapur. J. Assoc. off Anal. Chem. 65 (1982) 32.
11. C.S.P. Sastry and D. Vijaya, Talanta, 34 (1989) 372.
12. C.S.P. Sastry, D. Vijaya and D.S. Mangala, Analyst, 112 (1987) 75.
13. K.D. Khalaf, J. Sancenón and M. de la Guardia, Anal. Chim. Acta, 266 (1992) 119.
14. K.D. Khalaf, A. Morales Rubio and M. de la Guardia, Anal. Chim Acta, 280 (1993) 231.
15. B. Fernández Band, P. Linares, M.D. Luque de Castro and M. Valcárcel, Anal. Chem. 63 (1991) 1672.
16. A. Espinosamansilla, F. Salinas and A. Zamoro, Mikrochim. Acta, 113 (1994) 9.
17. A. Espinosamansilla, F. Salinas and A. Zamoro, Analyst, 119 (1994) 1183.
18. J.M. García, A.I. Jimenez, J.J. Arias, K.D. Khalaf, A. Morales Rubio and M. de la Guardia, Analyst, 120 (1995) 313.
19. A. Lorber and B.R. Kowalski, J. Chemom., 2 (1988) 67.
20. J. Ferré and F.X. Rius, Química Analítica, accepted for publication.
21. J.H. Kalivas and P.M. Lang, Chemom. Intell. Lab. Syst., 32 (1996) 135.
22. J. Ferré and F.X. Rius, in preparation.
23. J. Fleming, B. Neidhart, H. Albius and W. Wegscheider, Accreditation and Quality Assurance, 3 (1996) 135.
24. J. Riu and F.X. Rius, Anal. Chem., 68 (1996) 1851.
25. R. Boqué and F.X. Rius, submitted for publication.
26. C.A. Clayton, J.W. Hines and P.D. Elkins, Anal. Chem., 59 (1987) 2506.
27. J.H. Kalivas. J. Chemom, 3 (1989) 409.
28. G. Bauer, W. Wegscheider and H.M. Ortner, Spectrochim. Acta, 46B (1991) 1185.

ISBN:978-84-691-1875-7/DL: T-337-2008

4.7 Detection and correction of biased results of individual analytes in multicomponent spectroscopic analysis

(submitted)

Joan Ferré, F.Xavier Rius

Departament de Química Analítica. Universitat Rovira i Virgili.
Pl. Imperial Tarraco, 1, 43005-Tarragona. SPAIN

Simultaneous spectroscopic multicomponent analysis based on Beer's law requires the test sample to follow the hard model, independently of whether this model is built with the full spectrum or only with a few sensors selected according to an optimality criterion. We have developed a graphical method based on the net analytical signal concept to detect bias in the predicted results of individual analytes in test samples. When an interferent, or other causes which produce bias, are detected, a moving window approach is used to select the subset of sensors that minimize the bias of the predicted results. The method has been validated with UV-Vis spectra of binary clorophenol mixtures and of mixtures of four pesticides in water analysed with a FIA system with diode array detection.

ISBN:978-84-691-1875-7/DL: T-337-2008

1. Introduction

Multicomponent analysis (MCA) applied to spectroscopic data is a multivariate calibration method based on Beer's law that enables all the components in mixtures of a well defined qualitative composition to be determined. Most statistical properties of this technique are well known although very recent research has provided some figures of merit such as the detection limit[1]. The selection of the best sensors for quantification has been the subject of much recent interest. Criteria based on the condition number[2] of $\mathbf{S}^T\mathbf{S}$, (^{T} means 'transposed' and $\mathbf{S}$ is the calibration matrix), the determinant[2] of $\mathbf{S}^T\mathbf{S}$, the trace[2-4] of $(\mathbf{S}^T\mathbf{S})^{-1}$, the accuracy[5] and selectivity[5] among others [6-7] have been optimised with different methods that include genetic algorithms[5] and simulated annealing[8]. Recently, Xu et al.[9] showed that the sensors that give more noise than signal may spoil the prediction results and they developed a new criterion for wavelength selection.

Several limitations make MCA less popular than the factor-based calibration methods such as principal component regression (PCR) or partial least squares (PLS). MCA is generally more susceptible to noise, baseline effects and spurious peaks. Moreover, all the analytes in the test sample that absorb in the spectral region of interest must be included in the model to prevent biased predictions. This combination of constraints means that MCA is not suitable for analysing natural samples, but it can be used with synthetic samples (e.g. pharmaceutical samples) and in some process monitoring where the analytes that make up the samples are known (e.g. gas phase spectroscopy).

The presence of non-modelled chemical components in a test sample is a major problem in these models since they may lead to biased predictions of the concentrations. This may happen even when the model is built, not with the full spectrum, but with only a few sensors which have been selected as optimal by the above mentioned criteria. These criteria have been shown to improve the precision and accuracy of the multicomponent analysis models. Their main drawback is that most of them are based on the calibration matrix, and they do not guarantee that a new test sample will be free of any unexpected interferent or baseline shift. The prediction error, often used in the calibration and/or validation steps to check the adequacy of selected sensors, cannot be used to check the presence of interferents in

ISBN: 978-84-691-1875-7/DL: T-1237-2010

the test sample since the true value of the concentration is unknown. Moreover, the comparison of the measured spectrum and the one reproduced with the predicted concentrations and the calibration matrix does not indicate the degree to which the concentration of a particular analyte is affected since the predicted concentration of all the analytes is used to reconstruct the spectrum. An interferent may affect the predicted concentration of each analyte differently depending on the degree of spectral overlap.

As the presence of interferents must be checked before the prediction is made, the idea here is to take advantage of the multivariate signal to detect how appropriate the test sample is for the model. The purpose of this paper is twofold. First, it presents a graphical criterion to detect bias in MCA for any individual analyte in test samples. The criterion is based on the plot, for the analyte of interest, of the net analyte signal in the test sample versus the net analyte signal of the pure analyte spectrum. This plot shows if the concentration of the analyte in the test sample can be predicted with any guarantee or whether interferents are present. The main importance of this graph is that it represents the multivariate signal in a two-dimensional plot, thus making it easier to interpret the calibration and prediction process. In the second place, a wavelength selection procedure is devised to eliminate the effect of the interfering species when they do not absorb in the full recorded spectrum. If the interferents absorb in the full spectrum, this selection procedure can minimise the bias in the predicted concentration.

2. Theory

2.1 Theoretical Background

Conventional notation has been used: matrices are represented by bold capital letters, column vectors by bold lower-case letters and scalars by cursive characters. The superscripts T and + indicate transposition and the pseudoinverse respectively. $\|\cdot\|$ stands for the Euclidean norm. A 'hat' ('^'), that should be present in the 'calculated' matrices, has been dropped from the symbols to simplify the notation; if the magnitude is measured or calculated can be deduced from the context.

ISBN:978-84-691-1875-7/DL: T-337-2008

Spectroscopic multicomponent analysis based on the linear additive model from Beer's law is given in eq 1:

$$\mathbf{r} = \mathbf{S}\,\mathbf{c}_{true} + \mathbf{e} \quad (1)$$

where $\mathbf{r}$ is the Jx1 vector of measured responses at J wavelengths for the test sample, $\mathbf{S}=[\mathbf{s}_1,\ldots\ \mathbf{s}_k\ \ldots\ \mathbf{s}_K]$ is a JxK matrix whose columns $\mathbf{s}_k$ are the Jx1 vectors of each analyte's sensitivities (responses divided by the concentration of the analyte in a pure sample), $\mathbf{c}_{true}$ are the concentrations of the K analyte in the test sample and the vector $\mathbf{e}$ is the contribution to the measured response not modelled in $\mathbf{S}$ such as measurement error, the absorbance of interfering constituents or baseline shift. The true (but unknown) response of the test sample is $\mathbf{r}_{true}= \mathbf{S}\,\mathbf{c}_{true}$ and spans the column space of $\mathbf{S}$ since it is a linear combination of the columns of $\mathbf{S}$. This is no longer true for $\mathbf{r}$ because $\mathbf{e}$ can be any vector in the J-dimensional space of the sensors. The least-squares estimation of $\mathbf{c}_{true}$ is given by eq 2:

$$\mathbf{c} = \mathbf{S}^{+}\,\mathbf{r} \quad (2)$$

where $\mathbf{S}^{+}$ is often calculated as $(\mathbf{S}^{T}\mathbf{S})^{-1}\mathbf{S}^{T}$. To predict the concentration of the kth analyte only the kth row of $\mathbf{S}^{+}$ is needed and is given by:

$$c_k = \mathbf{S}^{+}_{k\text{th-row}}\,\mathbf{r} \quad (3)$$

2.2. Prediction using the net analyte signal

The net analyte signal[10-11] of the kth analyte is the part of the signal that is orthogonal to the subspace spanned by the spectra of all the analytes except the kth. The net analyte signal of the spectrum of the pure kth analyte, of the kth analyte in the test sample and of the vector of errors can be calculated with eq 4 to 6 respectively:

$$\mathbf{s}_k^{*} = (\mathbf{I}-\mathbf{S}_k\mathbf{S}_k^{+})\mathbf{s}_k \quad (4)$$

$$\mathbf{r}_k^{*} = (\mathbf{I}-\mathbf{S}_k\mathbf{S}_k^{+})\mathbf{r} \quad (5)$$

$$\mathbf{e}_k^{*} = (\mathbf{I}-\mathbf{S}_k\mathbf{S}_k^{+})\mathbf{e} \quad (6)$$

ISBN:978-84-691-1875-7/DL:T-1230-2008

where $\mathbf{S}_k$ is the matrix $\mathbf{S}$ without the *k*th column. Since $\mathbf{s}_k^*$ (or $\mathbf{r}_k^*$) is a vector of residuals of the regression of $\mathbf{S}_k$ on $\mathbf{s}_k$ (or $\mathbf{r}$), the value of each element of $\mathbf{s}_k^*$ ($s_{k,j}^*$)(or $r_{k,j}^*$) associated to the *j*th sensor depends on the number of sensors considered in the model and their absorbance. The vector $\mathbf{e}_k^*$ cannot be calculated because $\mathbf{e}$ is unknown, but it is important below for understanding how the error is propagated to the predicted concentration.

In the prediction step, $\mathbf{s}_k^*$ is different for each analyte but the same for all the test samples and $\mathbf{r}_k^*$ is different for each analyte and test sample. The relationship between the concentration of the *k*th analyte in the test sample and the calculated $\mathbf{r}_k^*$ and $\mathbf{s}_k^*$ can be derived by inserting eq 1 into eq 5:

$$\mathbf{r}_k^* = \mathbf{r}_{k,\text{true}}^* + \mathbf{e}_k^* \qquad (7)$$

where $\mathbf{r}_{k,\text{true}}^*$, the net analyte signal of the test sample without errors, is proportional to the vector $\mathbf{s}_k^*$, as shown in eq 8:

$$\mathbf{r}_{k,\text{true}}^* = (\mathbf{I}-\mathbf{S}_k\mathbf{S}_k^+)\mathbf{S}\mathbf{c}_{\text{true}} = (\mathbf{I}-\mathbf{S}_k\mathbf{S}_k^+)[\mathbf{s}_1,\ldots\mathbf{s}_k\ldots\mathbf{s}_K]\mathbf{c}_{\text{true}} = [\mathbf{0},\mathbf{0},\ldots,\mathbf{s}_k^*\ldots,\mathbf{0}]\mathbf{c}_{\text{true}} = \mathbf{s}_k^* c_{k,\text{true}} \qquad (8)$$

where $\mathbf{0}$ is a vector of zeros corresponding to the net analyte signal of the spectrum of each analyte except the *k*th in the subspace spanned by the spectra of these analytes. When errors are not present, $\mathbf{e}_k^* = 0$ and eq 8 enables the concentration of the *k*th analyte to be calculated either from the net analyte signal of any sensor *j* or from the norm of the net analyte signals:

$$c_{k,\text{true}} = r_{k,\text{true},j}^* / s_{k,j}^* = \|\mathbf{r}_{k,\text{true}}^*\| / \|\mathbf{s}_k^*\| \qquad (9)$$

Thus, the expression $c_k = \|\mathbf{r}_k^*\| / \|\mathbf{s}_k^*\|$, as used by Xu *et al.*[9], is only true for error-free samples (with $\mathbf{e}_k^*=0$). When spectral error is present, the net signal associated to the *j*th sensor is:

$$r_{k,j}^* = c_{k,\text{true}}\, s_{k,j}^* + e_{k,j}^* \qquad (10)$$

Due to $e_{k,j}^*$, each sensor would predict a different concentration value:

$$c_{k,j} = r_{k,j}^* / s_{k,j}^* = (r_{k,\text{true},j}^* + e_{k,j}^*) / s_{k,j}^* = r_{k,\text{true},j}^* / s_{k,j}^* + e_{k,j}^* / s_{k,j}^* = c_{k,\text{true},j} + c_{k,\text{false},j} \qquad (11)$$

ISBN:978-84-691-1875-7/DL: T-337-2008

The sensors with small $s_{k,j}^*$ (e.g $s_{k,j}^* \approx 0$ in zones where the analyte does not absorb or the net analyte signal changes from positive to negative values) may have large prediction errors if $e_{k,j}^*$ is large since $r_{k,j,\text{true}}^*$ is also small. In the multivariate model, the concentration is calculated, not from only one sensor, but from all the sensors selected. Since $\mathbf{r}_k^*$ no longer lies on the subspace spanned by $\mathbf{s}_k^*$, the least-squares estimation of $c_{k,\text{true}}$ using all the sensors can be used:

$$c_k = \mathbf{r}_k^{*T}\mathbf{s}_k^* / \|\mathbf{s}_k^*\|^2 = \cos\alpha \|\mathbf{r}_k^*\| / \|\mathbf{s}_k^*\| \quad (12)$$

where the cosine of the angle between two vectors $\cos\alpha = \mathbf{r}_k^{*T}\mathbf{s}_k^* / \|\mathbf{s}_k^*\| \|\mathbf{r}_k^*\|$ has been used. Thus, the predicted concentration corresponds to how many times the norm of $\mathbf{s}_k^*$ is in the norm of the projection of $\mathbf{r}_k^*$ along the $\mathbf{s}_k^*$ direction. c_k can be divided into two parts: one for the true signal and the other for the error (see eq 13):

$$c_k = (\mathbf{r}_{k,\text{true}}^* + \mathbf{e}_k^*)^T \mathbf{s}_k^* / \|\mathbf{s}_k^*\|^2 = \|\mathbf{r}_{k,\text{true}}^*\| / \|\mathbf{s}_k^*\| \pm \|\mathbf{e}_{k,\text{proj}}^*\| / \|\mathbf{s}_k^*\| = c_{k,\text{true}} \pm c_{k,\text{false}} \quad (13)$$

where $\mathbf{e}_{k,\text{proj}}^*$ is the projection of $\mathbf{e}_k^*$ along the direction of $\mathbf{s}_k^*$. Eq 13 shows that the larger $\|\mathbf{s}_k^*\|$ is, the smaller the prediction error for a given concentration. Moreover, a large $\|\mathbf{e}_k^*\|$ does not necessarily imply large prediction error because, in the least-squares formulation, only the norm of its projection is used in prediction. This is different from the concentration evaluated using the net analyte signal of only one sensor (given in eq (11)), where the error in the concentration depends directly on the net analyte signal of the error. In the concentration evaluated with the least-squares method, only the projection of the net analyte signal of the error influences the prediction. The residuals of the regression of $\mathbf{r}_k^*$ vs. $\mathbf{s}_k^*$ are given by:

$$\mathbf{e}_{k,\text{res}} = \mathbf{r}_k^* - \mathbf{s}_k^* c_k \quad (14)$$

and can be used as a measure to test whether the test sample is appropriate for the model. The relationship between $\mathbf{s}_k$, $\mathbf{r}$, $\mathbf{s}_k^*$,$\mathbf{r}_k^*$, $\mathbf{r}_{k,\text{true}}^*$, $\mathbf{e}$, $\mathbf{e}_k^*$, $\mathbf{e}_{k,\text{proj}}^*$ and $\mathbf{e}_{k,\text{res}}$ is shown in Figure 1 for a simplified system of three sensors and two analytes with $\mathbf{s}_A = [0.5\ 1\ 0]^T$, $\mathbf{s}_B = [1\ 0\ 0]^T$, $\mathbf{e} = [0.4\ 0.5\ 0.3]^T$, $\mathbf{c}_{\text{true}} = [1.5\ 2.5]$. Notice the importance of $\|\mathbf{e}_{k,\text{proj}}^*\|$ relative to $\|\mathbf{s}_k^*\|$ in the prediction error, since the predicted concentration is given by $(\|\mathbf{r}_{k,\text{true}}^*\| \pm \|\mathbf{e}_{k,\text{proj}}^*\|) / \|\mathbf{s}_k^*\|$.

ISBN:978-84-691-1875-7/DL: T-337-2008

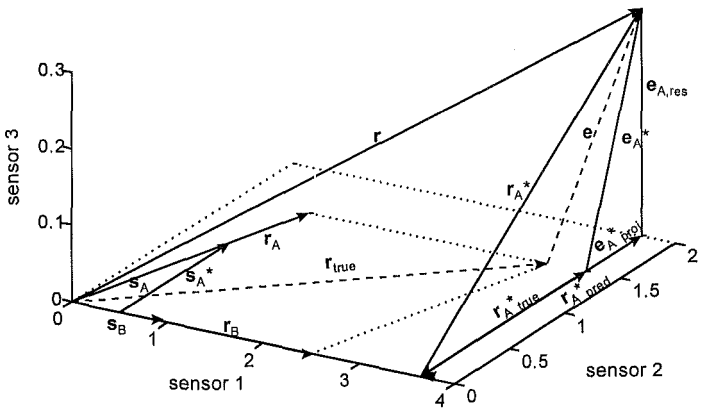


Figure 1. Geometrical representation of $\mathbf{s}_k$, $\mathbf{r}$, $\mathbf{s}_k^*$,$\mathbf{r}_k^*$, $\mathbf{r}_{k,\text{true}}^*$, $\mathbf{e}$, $\mathbf{e}_k^*$, $\mathbf{e}_{k,\text{proj}}^*$ and $\mathbf{e}_{k,\text{res}}$ for three sensors and two analytes A and B with $\mathbf{s}_A$=[0.5 1 0]^T, $\mathbf{s}_B$=[1 0 0]^T, $\mathbf{e}$=[0.4 0.5 0.3]^T, $\mathbf{c}_{\text{true}}$=[1.5 2.5] (see text for details).

According to eq 13, the quality of the predictions improves when sensors with a small signal but with large error (thus large $e_{k,j}^*/s_{k,j}$) are not used. Moreover, using one sensor instead of another in a set of selected sensors (as done in the wavelength selection procedures) changes the direction of $\mathbf{s}_k^*$ (e.g. from lying on the plane [1 0 0] to [0 1 0] in the case above). If, with the new sensor, $\|\mathbf{s}_k^*\|$ increases and the associated error decreases, the prediction error may decrease. Thus, wavelength selection can improve the prediction ability of the multicomponent analysis model.

2.3. The Net Analyte Signal Regression Plot (NASRP) for bias detection in test samples

A plot is presented to assess the degree to which the test sample follows the postulated model and if only random error is present. The plot of $r_{k,j}^*$ vs. $s_{k,j}^*$ (from now on called 'Net Analyte Signal Regression Plot', NASRP) should fit a straight line

ISBN:978-84-691-1875-7/DL: T-337-2008

through the origin with random residuals and slope c_k given by eq 12. Large and correlated residuals in this plot reveal discrepancies between the measured spectrum (and thus in r_k^*) and the model and, possibly, bias in the estimated concentration. The discrepancies may be due to non-modelled effects such as interferents or baseline shift. Only an interferent with the same spectrum as the analyte considered would not be detected (but in such cases, neither continuum regression nor particular cases such as PCR or PLS would produce correct estimations). The plot is shown in the experimental section.

2.4. *Methodology for wavelength selection*

When an interferent is detected using the NASRP, the prediction ability of the model can be improved if the sensors with the largest error are not used, according to eq 13. Although the sensors that best fit the straight line represented by eq 12 could be selected intuitively and those sensors where the interferent absorbs could be left-out, selecting sensors using the plot indicated above is misleading. The reason is that the sensors with the largest residuals do not necessarily correspond to those with a systematic error. 'Appropriate' sensors can have large residuals if the sensors where the interferent absorbs have a great effect on the calculation of r_k *. For this reason, the usual tests for outlier detection[12-13] in straight line models based on the size of the residuals are not suitable here. In addition, most of these tests consider the model built with and without the point studied but maintaining the original values of the regressor variables of the other points; this is not appropriate here because $r_{k,j}$* and $s_{k,j}$* may change when a sensor is deleted so would they always have to be calculated again. In addition, most tests are suitable for detecting only one outlier. Here multiple outliers may be present and if they are to be detected several models must be built, each of which omits a different number of possible outliers. All these difficulties may be overcome with a moving-window strategy with different widths and positions of the starting sensor of the window. The representation of the criterion described below versus the window width and the first selected sensor will give rise to a surface where the best sensors will be located at the minima. If the interferent only absorbs in one region of the spectrum of the analyte of interest, the bias can be completely eliminated. On the other hand, if the interferent absorbs in the whole spectrum, the bias in the predicted concentration can only be partially reduced.

ISBN:978-84-691-1875-7/DL: T-337-2008

2.4.1 *Criterion for wavelength selection*

For each window, the criterion for wavelength selection is evaluated. This criteria should not only measure the quality of fit of the data in the NASRP (or simply the correlation) but also take into account the norm of the net analyte signal $\|\mathbf{s}_k^*\|$ and the error. The reason is that the relative prediction error, derived from eq 13:

$$\%\text{error}=c_{k,\text{false}}/c_{k,\text{true}}=\|\mathbf{e}_{k,\text{proj}}^*\|/\|\mathbf{r}_{k,\text{true}}^*\|=\|\mathbf{e}_{k,\text{proj}}^*\|/c_{k,\text{true}}\|\mathbf{s}_k^*\| \qquad (15)$$

shows that the error in $\mathbf{r}$ propagates to the concentration depending on the norm of the net analyte signal $\|\mathbf{s}_k^*\|$ and the analyte concentration. Steep or shallow slopes will have different effect since the error is relative to $\|\mathbf{r}_{k,\text{true}}^*\|$. The error is not necessarily minimised when the sensors with maximum $\|\mathbf{s}_k^*\|$ or minimum $\|\mathbf{e}_{k,\text{proj}}^*\|$ are selected. A set of sensors with small $\|\mathbf{s}_k^*\|$ can predict correctly if $\|\mathbf{e}_{k,\text{proj}}^*\|$ is small enough. The sensors with minimum $\|\mathbf{e}_{k,\text{proj}}^*\|/\|\mathbf{s}_k^*\|$ should be used to reduce the prediction error. Since $\|\mathbf{e}_{k,\text{proj}}^*\|$ cannot be calculated, some sort of estimate of noise/signal must be minimised. The criterion for wavelength selection is deduced here from the Error Indicator (EI) used by Xu *et al.*[9]:

$$\text{EI}=\text{var}(\|\mathbf{r}_k^*\|-\|\mathbf{r}_{k,\text{true}}^*\|)^{1/2}/\|\mathbf{r}_{k,\text{true}}^*\| \qquad (16)$$

where the variance of the error in the norm of the net analyte signal is

$$\text{var}(\|\mathbf{r}_k^*\|-\|\mathbf{r}_{k,\text{true}}^*\|)=[(2\|\mathbf{r}_{k,\text{true}}^*\|\,s)^2+(Js^2)^2]/(\|\mathbf{r}_k^*\|+\|\mathbf{r}_{k,\text{true}}^*\|)^2 \qquad (17)$$

with the standard deviation calculated as $s=(\mathbf{r}^T(\mathbf{I}-\mathbf{S}\mathbf{S}^+)\mathbf{r}/J-K)^{1/2}$. Since $\mathbf{r}_{k,\text{true}}$ * cannot be known, Xu et al. proposed to replace it with $\mathbf{r}_k^*$, which leads to the eq 18:

$$\text{EI}=[\,s^2(1+J^2s^2/4\|\mathbf{r}_k^*\|^2)\,]^{1/2}/\|\mathbf{r}_k^*\| \qquad (18)$$

Among other things, Xu et al.[9] assumed that the error in the net analyte signal can be approximated by the errors in $\mathbf{r}$ and that the errors in all the sensors are normally distributed with the same standard deviation. This standard deviation takes into account all the analytes because $\mathbf{r}$ is used. In our methodology, which only considers one analyte, we suggest evaluating s as the standard deviation of the straight line in the NASRP, as given in eq 19, which measures the quality of fit. This is particular for

ISBN:978-84-691-1875-7/DL: T-337-2008

the analyte of interest and it is related to the uncertainty in the slope (and thus in the predicted concentration):

$$s_k = (\mathbf{e}_{k,res}{}^T \mathbf{e}_{k,res} / J\text{-}1)^{1/2} \quad (19)$$

The fact that statistic s_k, given in eq 19, is from a sample that contains an interferen may violate the assumptions of 'same standard deviation' for all the sensors However, it has proved to be useful because it amplifies the spectral error.

2.4.2 *Procedure for bias correction*

Thus, the methodology proposed for correcting bias by selecting wavelengths firs considers the NASRP of the *k*th analyte in the test sample. If the points for the rang of wavelengths considered to measure **r** do not approximately fit a straight line wit random residuals, the moving-window strategy is used to select the subset of sensor that optimise the selection criterion. Finally, the NASRP and occasionally the plot o the residuals vs. sensor number derived from the NASRP for the selected sensors ar considered again to decide if the residuals are acceptable and if the prediction can b made.

3. Experimental section

Measured and simulated data were used to assess the validity of the propose approach. The first data set consists of UV-visible spectra of mixtures of 2 chlorophenol (2-CP) and 2,4-dichlorophenol (2,4-DCP) in water recorded every 2 nm between 226 and 346 nm (61 sensors) in the conditions described in Ferré and Rius[15] Three samples were used for calibration and nine for validation. Here, as an example only three of the validation mixtures were used. Two simulated spectra o interferents were added to the spectrum of one of the validation mixtures. The tw interferents were simulated as Gaussian peaks with a standard deviation of 10, an maximums at sensor numbers 10 and 50, respectively. In addition, each interferen

ISBN:978-84-691-1875-7/DL: T-337-2008

peak was multiplied by 0.1 and 0.9 respectively to simulate a small absorbance for the first interferent and a large absorbance for the second one. The second data set consists of UV-visible spectra of mixtures of the pesticides Carbaryl (RYL), Carbofuran (CBF), Propoxur (PPX) and Isoprocarb (IPC) in water analysed with a FIA system and recorded between 340 and 650 nm every 2 nm (156 sensors) as described in Ferré *et al.*[16].

3.1 Computer programs

All computations were performed with home-made Matlab[14] subroutines.

4. Results and discussion

4.1.The chlorophenol system

Figure 2a shows the spectra of the two pure chlorophenols at concentration 10^{-4} M and the net analyte signal of 2-CP ($s_{2\text{-CP}}^*$) calculated using all the sensors, sensors 1 to 12 and sensors 1 to 9, respectively. It can be seen how the value of the net analyte signal in each sensor ($s_{2\text{-CP},j}^*$) depends on the number of sensors used to build the model, as has been mentioned in the theoretical section. Figure 2b shows three mixtures of the two chlorophenols with concentrations 0 *M* (2-CP) and $3.1 \cdot 10^{-5}$*M* (2,4-DCP), $3.9 \cdot 10^{-5}$*M* (2-CP) and $3.1 \cdot 10^{-5}$*M* (2,4-DCP) and $5.8 \cdot 10^{-5}$*M* (2-CP) and $4.6 \cdot 10^{-5}$*M* (2,4-DCP) respectively. Figure 2c shows the $r_{2\text{-CP}}$ * versus wavelength for both the 2-CP in these mixtures, and the pure analyte (dashed line). In each sensor, $r_{2\text{-CP}}$ * is a multiple (the concentration) of $s_{2\text{-CP}}^*$. This can be readily seen in the NASRP for 2-CP (Figure 2d). The slope of each regression line that fits the points is the concentration predicted by the model. Since the least-squares method is used, the spectral error contained in each sensor is averaged in the final calculation. In this case, it also seems that all the sensors are not needed to estimate the line, particularly the ones near $s_k^* \cong 0$. However, the larger the number of sensors with $s_k^* \neq 0$ and low spectral error, the higher the precision of the estimated slope c_k, of the straight line.

ISBN:978-84-691-1875-7/DL: T-337-2008

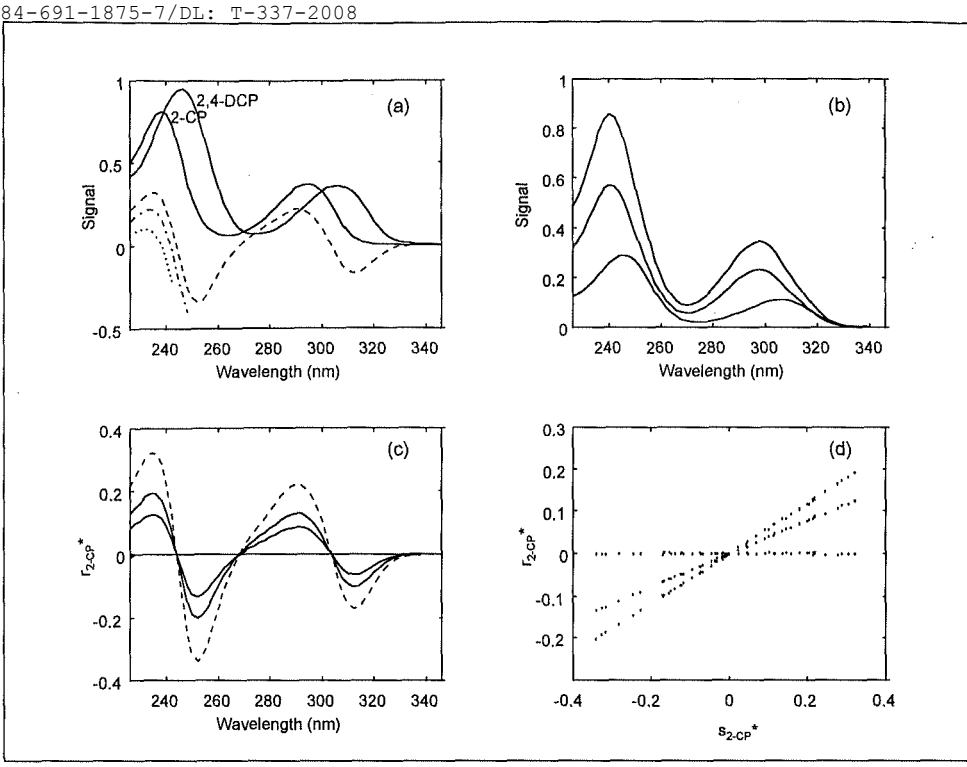


Figure 2. (a) Spectra of two chlorophenols and the net analyte signal for 2-CP ($s_{2\text{-CP}}^*$) obtaine for all the sensors and for two different subsets of sensors (——), 12 sensors (-.-) and 9 senso (···). (b) Three mixtures with a different concentration of chlorophenols at 0 *M* (2-CP) ar $3.1 \cdot 10^{-5}M$ (2,4-DCP) (bottom), $3.9 \cdot 10^{-5}M$ (2-CP) and $3.1 \cdot 10^{-5}M$ (2,4-DCP) (middle) and $5.8 \cdot 1$ ^{-5}M (2-CP) and $4.6 \cdot 10^{-5}M$ (2,4-DCP) (top) (c) $r_{2\text{-CP}}^*$ for the 2-CP and $s_{2\text{-CP}}^*$ (dashed line) for th mixtures in Figure 2b. (d) NASRP for 2-CP of the three mixtures.

In order to evaluate the usefulness of the proposed methodology and th wavelength selection procedure, simulated spectra of interferents were added to th spectra of the two-component system of 2-CP and 2,4-DCP. Figure 3a shows th spectrum of the mixture $3.9 \cdot 10^{-5}M$ (2-CP) and $3.1 \cdot 10^{-5}M$ (2,4-DCP) used in Figure 2 Fig 3b and Fig 3c show the spectra of this sample plus a simulated interferent. Fo comparison, the original spectra is also shown (dashed line). Figures 3d to 3f sho the NASRPs for 2-CP which correspond to these spectra. The non-random larg residuals in Figures 3e and 3f denote the presence of the spectral overlappir interferent. In order to find which sensors provide the lowest analytical errors, th moving-window strategy was used. For each window studied, the Error Indicat was plotted as the third dimension on a grid as a function of the width and th location of the first sensor of the spectral window. Figure 4a to 4c show the valu

ISBN:978-84-691-1875-7/DL: T-337-2008

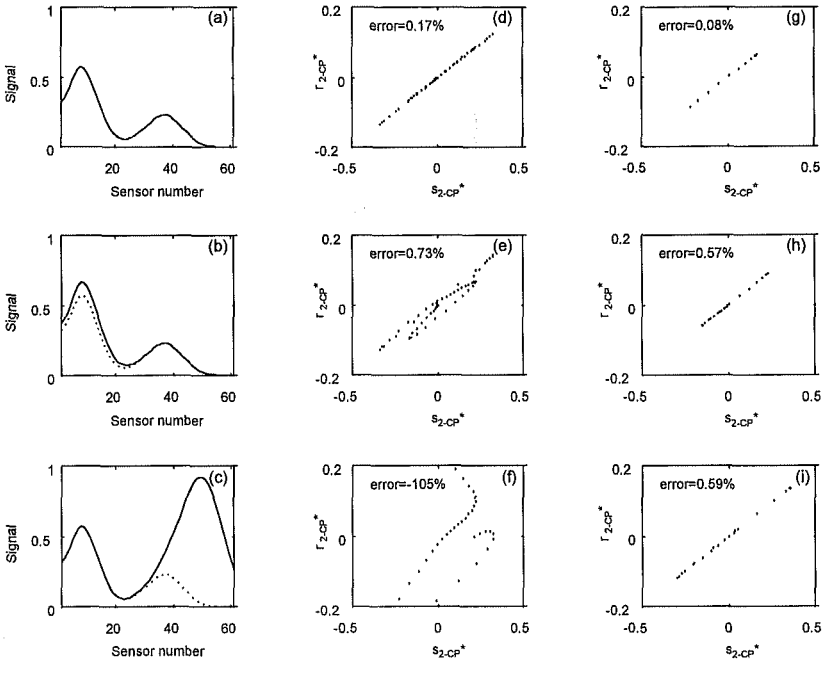


Figure 3a-3i. Two different simulated interferents added to a mixture of concentration $3.9 \cdot 10^{-5}$ *M* (2-CP) and $3.1 \cdot 10^{-5}$*M* (2,4-DCP) . (a) no interferent, (b) interferent centred on sensor number 10, (c) interferent centred on sensor number 50. Figures 3d-3f, NASRP of the mixtures with simulated interferents corresponding to Figures 3a-3c respectively. Figures 3g-3i, NASRP with the sensors selected according to the wavelength selection criteria corresponding to the spectra in Figures 3a-3c.

produced by this selection method for the samples in Figures 3a to 3c. As this mixture was prepared to validate the methodology, the true concentration was known and the prediction error of the 2-CP was evaluated in each case as the absolute value of $c_{2\text{-}CP} - c_{2\text{-}CP,true}$. Figures 5a to 5c show the prediction error surface on the same axes as Figures 4a to 4c. In order to better appreciate the lowest values in the surfaces, they have been truncated at Error Indicator and prediction error values higher than 0.005. Although there are some differences in the fine structures, the close agreement between the general features of the two surfaces shows that the proposed methodology can actually correct bias. Figures 4a to 4c enable the optimal window to be selected. The lowest values of the Error Indicator (the 'less dark' windows) clearly identify the windows with the smallest systematic errors.

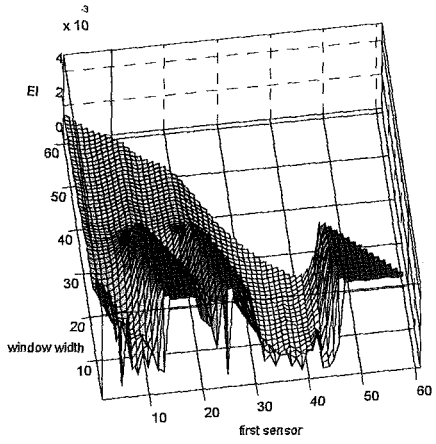
ISBN:978-84-691-1875-7/DL: T-337-2008

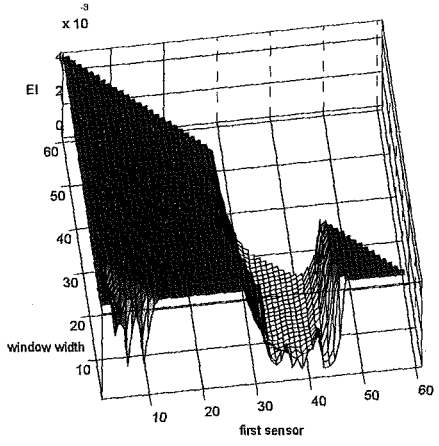
Fig 4a shows that a zone of minima of EI can be found for windows which start between sensor number 31 and 40 and with a width of 3 to 17 sensors. Although the optimal window should be the one with the lowest Error Indicator value, our experience has shown that the minimum may sometimes contain too few sensors, while some windows, with only a slightly larger Error Indicator value, can have more sensors. As this affects the norm of the net analyte signal, which is related to the uncertainty of the predicted concentration, in such cases the window with the absolute minimum of the criteria is not recommended; it is preferable to use a larger number of sensors. Fig 3g shows the NASRP obtained with sensors 31 to 42. In this case, the relative prediction error calculated as $100*((c_{2\text{-CP}} - c_{2\text{-CP,true}})/c_{2\text{-CP,true}})$ obtains a value as low as 0.08% compared to 0.17% for a pure mixture sample considering all the sensors (Fig 3d). Notice that in this case, despite the fact that the sample is considered to be free of systematic error, the criterion indicated windows where the prediction error was lower than when all the sensors are used

According to this criterion, plotted in figures 4a and 4b, the windows that only contain the highest wavelengths (on the right-hand of the spectra) should not be used either, even in the interferent-free mixture. In these wavelengths, the absorbance due to the analyte and the norm of the net analyte signal $\|s_k^*\|$ are very small. This degrades the analytical results since the uncertainty of the net analyte signal and the predicted concentration is affected by this norm, and the smaller the norm, the larger the error in the predicted concentration. In the spectrum of the interferent-free mixture (Figure 4a), the large values of the criteria in the narrow windows near the sensors 20 and 30 can be accounted for in a similar way. In these zones, the similarity between the spectra of the two chlorophenols results in a large collinearity and small norm of the net analyte signal. Accordingly, these windows have large Error Indicator values and should not be used since this would result in higher prediction errors. Observe the large agreement between the Error Indicator and the actual calculated error in these zones.

Figure 4b shows that the shape of this surface is the same as the shape in Figure 4a for those spectral windows whose first sensor is larger than 31. For windows with sensors between 1 and 30 the interferent absorbs in the way shown in Figure 3b and the Error Indicator has a large value which we truncated. We have shown this truncated zone in black. The windows whose first sensor is between 30 and 45 and

ISBN:978-84-691-1875-7/DL: T-337-2008





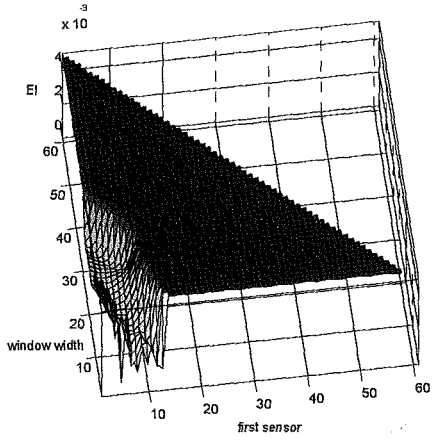
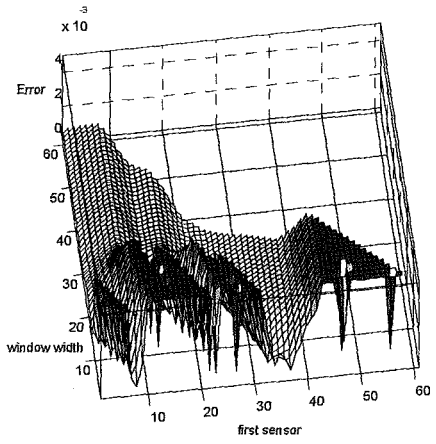
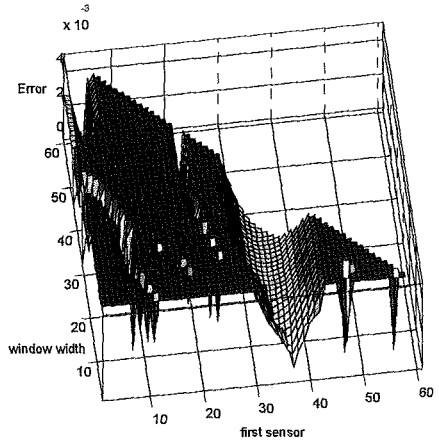


Figure 4a-4c.





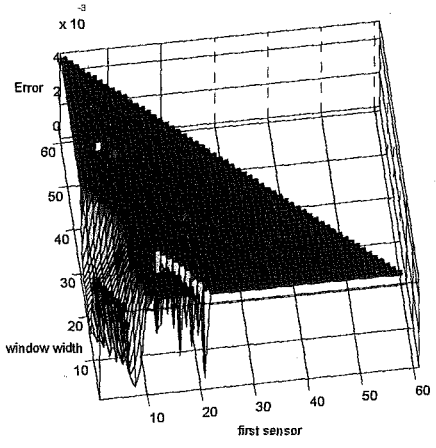


Figure 5a-5c.

ISBN:978-84-691-1875-7/DL: T-337-2008

Figure 4a-4c (previous page, from top to bottom). The selection criteria as a function of the window width and the origin for the spectra in the Figures 3a-3c.

Figure 5a-5c (previous page, from top to bottom). The prediction error for all the windows corresponding to the spectra in the Figures 3a-3c.

with a width between 5 and 30 correspond to the zone with the lowest values of the Error Indicator. These are the windows in which the interferent does not absorb and they can all be used for prediction. As in Figure 4a, the windows at the highest wavelengths are not selected due to the small norm of the net analyte signal. Fig 3h shows the NASRP obtained with sensors 34 to 54. The relative prediction error was lower, 0.57%, compared to the 0.73% obtained for a mixture sample when all the sensors were considered (Fig 3e).

Figure 4c corresponds to the presence of an interferent with a large absorbance in sensors 25 to 61. The criterion indicates that the windows which are most useful for regression are the ones between sensors 2 and 10 with a window width of 3 to 35, in which the interferent does not absorb. Fig 3i shows the NASRP obtained with sensors 4 to 24. Here, the relative prediction error was lower, 0.59% in contrast to the -105% obtained for a mixture sample when all the sensors were considered (Fig 3f). Clearly, the presence of an interferent with a large absorption (fig 3c) does not prevent a prediction error from being obtained that is similar to the one in figure 3h.

4.2. Application to a multicomponent system of pesticides

The methodology developed was applied to the four-component system of pesticides determined with a FIA system as described in the experimental section. Due to the similar spectra of three of the pesticides (Figure 6) this can be considered as an example of ill-conditioning. The columns of **S** are highly collinear, which produces small values of $s_{k,j}^*$ for the three least selective pesticides as can be seen in the plot of their net analyte signal in Figure 7. On the other hand Carbaryl has a high NAS value due to its well resolved peak and should not be difficult to

ISBN:978-84-691-1875-7/DL: T-337-2008

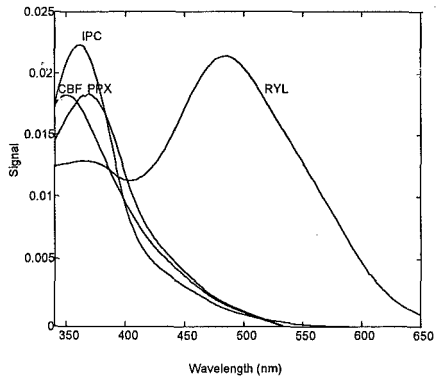


Figure 6. The spectra of four pesticides at unit concentration.

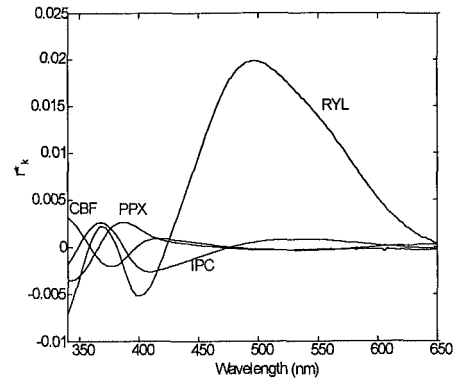


Figure 7. Net analyte signal of the four pesticides RYL, PPX, CBF and IPC.

quantify. To show the capacity of the proposed methodology to detect biased results in the validation step, a validation sample was deliberately prepared with a Carbaryl concentration which was lower than the nominal one. Only 7ppm was added instead of the 8ppm that should be in the validation sample according to the validation set designed in Ref. 16. The predicted concentration in this sample was 7.15 ppm, which was a prediction error of about 10% considering 8ppm as the reference value for this sample. Had this sample been included in a test set to validate the model it would have been difficult to explain the large prediction error for RYL since, because of its high selectivity, there should be no quantification problems. One possible explanation could be that this prediction error is due to the presence of an interferent. Figure 8 shows the NASRP of Carbaryl for this test mixture. The fact that the NASRP is quite a straight line suggests that no interferents are present and that the sample could have been wrongly prepared, with a true concentration that was smaller than the one assigned.

Finally, Figure 9a shows the NASRP of PPX in a test mixture containing 2ppm RYL, 8ppm CBF, 2ppm PPX and 8ppm IPC. Notice the large number of sensors with $s_{k,j}^*$ values near zero that correspond approximately to the zone between 420 nm and 650 nm where the net analyte signal is near zero. The sensors with the largest $s_{k,j}^*$ are the most influential in the slope (i.e. the predicted concentration) while the ones with $s_{k,j}^*$ values near zero do not influence the slope to a great extent,

ISBN:978-84-691-1875-7/DL: T-337-2008

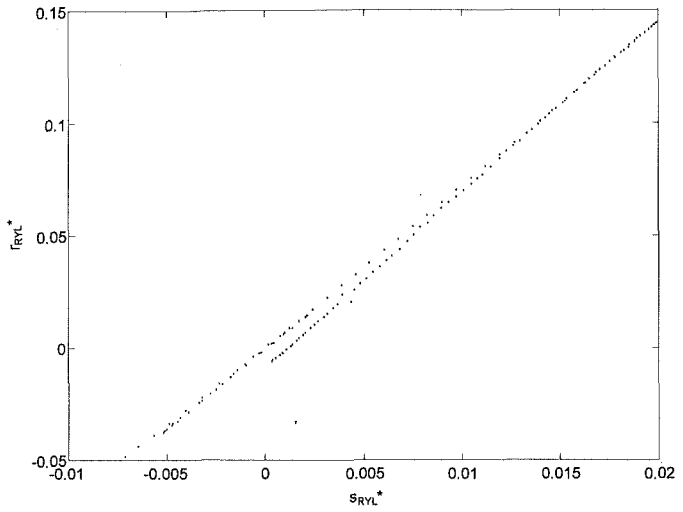


Figure 8. NASRP for Carbaryl in one of the test mixtures of pesticides.

even if they have large residuals. Thus, an interferent that absorbs in zones with relatively small $s_{k,j}^*$ does not significantly influence the predictions. Although they are not important for prediction, they can be important to evaluate the NAS when the model is built.

The trend of the points, a long way from the linearity, shows that there is some sort of problem in quantifying this sample. In fact, the relative prediction error for PPX in this sample was 52% of 2 ppm. The window-moving method was used to attempt to improve the prediction error for the mixture. All the windows between sensors 5 and 156 were considered, starting at all the possible points. Figure 9b shows the plot of the Error Indicator. The surface has been truncated for values of the Error Indicator larger than 0.05. In this way, many of the windows in the plot had a constant value and there was only one minimum. Only the subset of the plot with the minimum is shown to better appreciate the lowest values in the figure. In this case, the minimum (the 'white' coloured part of the surface) corresponds to the window whose first sensor is number 3 and whose width is 70, so the optimal window used to analyse PPX is the one that contains sensors 3 to 72. The final NASRP is shown in Figure 9c, where a series of well correlated points with a more

ISBN:978-84-691-1875-7/DL: T-337-2008

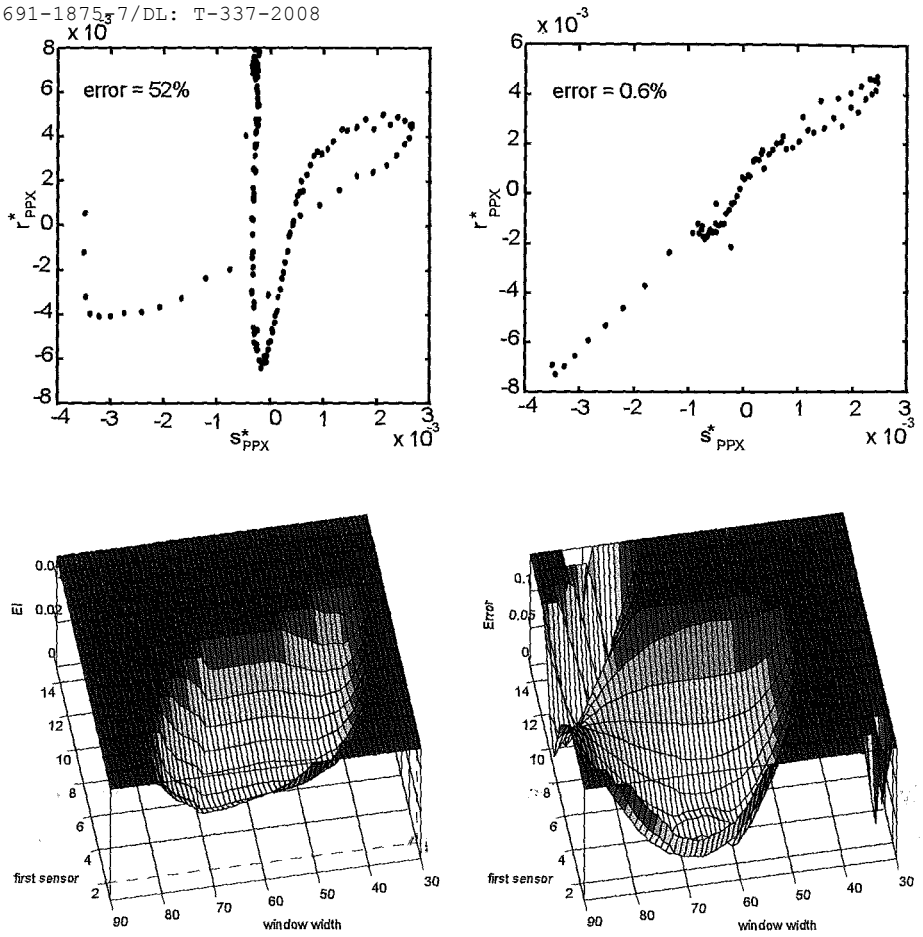


Figure 9. (a, top left) NASRP of PPX for a test mixture that had a large prediction error. (b, bottom left) Error Indicator for PPX, (c, top right) NASRP for PPX with the selected sensors 3 to 73. (d, bottom right) Error surface for PPX.

random distribution of the residuals can be observed. The regression using these sensors reduced the prediction error to 0.6 %. As the true concentration of this sample was available, the error surface was also plotted as a function of the window width and starting sensor. The agreement between this surface and the one predicted by the Error Indicator can be observed.

ISBN:978-84-691-1875-7/DL: T-337-2008

Figures 8 and 9a show that the range in the abscissa axis spanned by s_{PPX}^* is smaller than the range for s_{RYL}^*. This is due to the high selectivity of the Carbaryl and the low selectivity of Propoxur. Although they have a similar absorbance at unit concentration (Figure 6) the low selectivity of Propoxur gives small values of the net analyte signal. This could produce a larger uncertainty in the slope (the predicted concentrations) for Propoxur than for Carbaryl.

5. Conclusions

In this paper a methodology for detecting bias in multicomponent analysis has been proposed based on the net analyte signal concept and a graphical criterion. This approach leads to a better understanding of the multivariate calibration. Applied to the calibration, validation and prediction steps of the multivariate model, the approach can be used to check the internal coherence of calibration and test samples (e.g. to check whether the prediction error of a validation sample is due to an interferent or to the sample being prepared wrongly). It also allows the sensors where interferents absorb to be detected. To minimise bias, a wavelength selection procedure has been used based on the deletion of these latter sensors. This solves one of the main limitations of multicomponent analysis which is the presence of non-modelled interferents in the test samples. One limitation of this approach is when the interferent absorbs throughout the spectrum. Should an interferent be present, a new model can be calculated which includes this new analyte with the help of spectral libraries of the residual spectra. Moreover, the net analyte signal regression plot makes it possible to use robust regression methods or non-linear relationships to estimate the concentration. Elements such as spurious peaks, etc. can also be clearly seen in the NASRP.

The approach offers the additional advantage of selecting the most suitable sensors to minimize the error for each analyte in every test sample. Unlike other wavelength selection techniques this is not a process of calibration and validation for all the analytes together. The computational time could be a drawback for specific purposes such as control analysis since a large number of calculations are necessary to predict a test sample.

ISBN:978-84-691-0[illegible]

Nowadays, the same approach is being extended to ILS calibration based on the calculation of the NAS proposed by Lorber *et al.*[17]

Acknowledgement

J. Ferré thanks the Comissionat per a Universitats i Recerca of the Generalitat de Catalunya for providing a doctoral fellowship (FI/94-7001). Financial support from the Spanish Ministry of Education and Science (DGICyT project BP93-0366) is gratefully.

References

[1] R. Boqué, F.X. Rius *J. Chemom.* In press.
[2] Y. Liang, Y. Xie, R. Yu, *Anal. Chim. Acta* 222 (1989) 347.
[3] K. Sasaki, S. Kawata, S. Minami, *Applied Spectroscopy* 40 (1986) 185.
[4] S.D. Frans, J.M. Harris *Anal. Chem.* 57 (1985) 2680.
[5] C.B. Lucasius, M.L.M. Beckers, G. Kateman, *Anal. Chim. Acta* 286 (1994) 135.
[6] P.A. Salamin, H. Bartels, P. Foster *Chem. Intell. Lab. Syst.* 11,1991, 57.
[7] L. Costadinnova, T. Nedeltcheva, *Analyst* 120 (1995) 2217.
[8] U. Hörchner, J. H. Kalivas, *J. Chemom.* 9 (1995) 283.
[9] Liang Xu , Israel Schechter. *Anal. Chem.* 68 (1996) 2392.
[10] A. Lorber, *Anal. Chem.* 58 (1986) 1167.
[11] A. Lorber , B.R. Kowalski. *J. Chemom.* 2 (1988) 67.
[12] Meloun, M.; Militký, J.; Forina M.; *Chemometrics for analytical chemistry. Vol 2. PC-aided Regression and Related Methods.* Ellis Horwood: Great Britain 1994
[13] Myers R.H *Classical and modern regression with applications 2nd edition.* Duxbury Press: Belmont 1990
[14] Matlab .The Mathworks, South Natick, MA, USA.
[15] J. Ferré, F.X. Rius. *Trends Anal. Chem.* 16 (1997) 155.
[16] J. Ferré, R. Boqué, B.Fernández-Band , M.S. Larrechi, F.X. Rius. *Anal. Chim. Acta. In press.*
[17] A. Lorber, K. Faber , B.R. Kowalski *Anal. Chem.* 69 (1997) 1620.

UNIVERSITAT ROVIRA I VIRGILI
EXPERIMENTAL DESIGN APPLIED TO THE SELECTION OF SAMPLES AND SENSORS IN MULTIVARIATE CALIBRATION
Joan Ferré Baldrich

ISBN:978-84-691-1875-7/DL: T-337-2008

Chapter 5

Conclusions

ISBN:978-84-691-1875-7/DL: T-337-2008

5.1 Introduction

The aim of this chapter is to present the conclusions of the work reported in this thesis and to give some considerations for future work with respect to sample and sensor selection in multivariate calibration.

The conclusions of the thesis are in §5.2. The general conclusions about the algorithms and criteria used in the chapters 3 and 4 for selecting calibration samples and sensors are in §5.2.1. Conclusions concerning specific points of the chapters 3 and 4 are in §5.2.2 and §5.2.3 respectively. More explicit conclusions can be found at the end of each section in the chapters 3 and 4. The considerations for future work are in §5.3.

5.2 Conclusions

5.2.1 General conclusions

Multivariate calibration models, specially the ones based on spectroscopic data, are increasingly used in chemical analyses. The ability of these models to give precise and unbiased predictions influence decisively the quality of the analytical result. The calibration samples and sensors must be carefully selected so that the models can represent properly the phenomenon under study and assure the quality of the predictions. These are some general conclusions of the work reported in the chapters 3 and 4 (below, the term *points* may refer either to calibration samples in ILS, PCR and MLR or to wavelengths in CLS):

1. *Optimality criteria.* Optimality criteria derived from the experimental design in MLR (§2.3.4.2) have been applied for selecting calibration wavelengths in CLS (§4.2 and 4.3) and the minimum number of calibration samples in MLR and PCR from the instrumental responses (§3.4) or principal component scores (§3.2 and

ISBN:978-84-691-1875-7/DL: T-337-2008

§3.5) of a list of candidates. These criteria are an alternative (and/or a complement) to the experimenter's subjective criterion. The major assumption is that the error in the regressor variables is negligible in front of the error of the dependent variable. In such a way, the precision of the model coefficients depends only on the information in the independent but not on the values of the dependent variable. In this way the best points to construct a model can be determined before measuring the dependent variable. Although this may not be strictly accomplished (e.g. when the independent variables are absorbances in ILS or scores in PCR and the dependent variable is concentration determined with a reference method), the models built with the points selected with the proposed criteria had smaller variance of the estimated coefficients or concentrations and better predictive ability than the models built with the samples selected randomly.

2. *The D- and M- criteria.* The D-criterion has been successfully used for selecting calibration samples in PCR (§3.2, §3.3 and §3.5) and MLR (§3.4), for selecting a reduced set of samples to assess the validity of PCR models before standardization (§3.6) and in the selection of wavelengths in CLS from the matrix of sensitivities. The optimal number of calibration points is indicated by the M-criterion which selects, of the D-optimal sets with a different number of points, the set with the largest information content per point. To our knowledge, these criteria were used in PCR with spectroscopic data for first time.

3. *Other criteria.* In addition to the D- and the M- criteria, other criteria also characterize the performance of a design. The trace of the dispersion matrix, the maximal variance function, G-efficiency, the variance coefficients (UVIF) and the variance inflation factors (VIF) were considered for selecting a calibration set which is not M-optimal (§3.5). The trace of the dispersion matrix, the selectivity, the condition number, sensitivity and the variance-decomposition proportions can be interpreted considering their effect on the confidence ellipsoid of the estimated concentrations in CLS (§4.2 and §4.3).

4. *Optimization algorithms.* Selecting an optimal subset of I points from a list of N candidates by checking all the possible subsets may require a prohibitive time for the present personal computers if N and I are large. Optimization algorithms are needed to find the optimal subsets. One of these is the Fedorov's algorithm

ISBN:978-84-691-1875-7/DL: T-337-2008

(explained in §3.2), for the search of D-optimal sets. This algorithm written in Matlab language and running under Windows 95 on a Hewlett-Packard personal computer series Vectra with a processor Pentium 120 MHz and 16 Megabytes RAM found a D-optimal set in less than 5 seconds for any of the data sets considered in this thesis. However, this time depends on the number of candidates, the number of points to be selected and, mainly, on the computation power. Faster personal computers in the next years will enable more complex problems to be studied in a reasonable time. The Fedorov's algorithm favors the use of the D-criterion since it is easy to use and program. To our knowledge specific algorithms do not exist for A- and E-optimality so that these criteria are not so easily applicable and require general optimization algorithms. The genetic algorithms (GAs) can be used for any optimality criterion and are able to generate a list of sub-optimal solutions of a given number of points (§3.5). In this way, other solutions than the optimal can be selected according to the user's needs. A drawback of GAs is that the search may continue even when the optimum has been found. Therefore, Fedorov's algorithm can be faster when the number of candidates is small. Possibly GA would be more advisable for selecting many points from a large list of candidates but this was not proved here.

5.2.2 Conclusions of the chapter 3

1. *Methodology.* A methodology for selecting calibration samples for PCR using only the instrumental responses of the candidate samples when they are difficult to prepare according to an established design has been developed. Only the selected samples are analyzed with the reference or well-established method. This methodology overcomes the limitations of the other procedures commented in §3.1.3.

2. *Performance.* The PCR (§3.2) and MLR (§3.4) models whose calibration samples were D-optimal had generally a better predictive ability than those whose calibration samples were randomly selected or using the Kennard-Stone's algorithm (§3.4).

ISBN:978-84-691-1875-7/DL: T-337-2008

3. *Objections to D-optimal designs.* An objection usually made to the D-optimal designs is that the selected samples are optimum for a given model decided beforehand and, therefore, they can be very unsuitable for other models. In §3.2, the only assumption made apart from selecting the PCR technique is the number of initially important factors included in the screening model. The final regression is evaluated from experimental results.

4. *Regression domain.* In the experimental designs the limits of the experimental domain are defined by the maximum and minimum values of the independent variables. Similarly, these limits were defined in PCR by scaling the scores with the range-midrange transformation (§3.2). Since the D-criterion selects the most external samples, the range of variation of the scores is fully covered and ensures that unknown samples are mainly predicted by interpolation. This is an advantage over the randomly selected calibration samples that may not span adequately the experimental domain and future unknown samples may fall outside the calibration range.

5. *Predictive ability of PCR.* PCR models with the factors ordered according to their modeling ability perform better than the usual PCR with the top-down selection (§3.2 and §3.3). The factors which are not related to the analyte of interest can increase the prediction error of the PCR model despite being associated with a large percentage of the variance of the response data matrix.

6. *Selection of factors in PCR.* Since a CR model with the optimal number of factors must be postulated before selecting the samples requires, an initial selection of factors is needed using the minimum number of samples possible. Regressing the scaled scores against their analyte concentrations enable the factors to be selected rapidly in order of their modeling ability (§3.2, §3.3). These factors can also be selected using not all but a reduced number of samples selected with the M-criterion, which gives the best precision of the coefficients with a minimum number of samples.

7. *The D-criterion in QSAR studies.* D-optimal calibration subsets with a different number of points have been found for PCR in QSAR studies (§3.5). This variety of solutions can be better adapted to the experimenter's needs than selecting the points by their similarity to a classical experimental design (e.g. a fractional

ISBN:978-84-691-1875-7/DL: T-337-2008

factorial design). Unfortunately the expected superior predictive ability of the models built with the D-optimal subsets over the ones selected according to a classical design could not be shown since the values of the dependent variables were not available.

8. *Experimental results.* The discussed examples used spectroscopic data. However, the procedure could be applied to most methods of analysis that use multivariate detection. The examples confirmed that a calibration derived from a subset is capable of predicting the content of analytes in all samples of the initial population as long as the subset in the factor space spans correctly the experimental domain. The methodology substantially reduces the effort for reference analyses in spectroscopic calibration. While the cost of the model increases with the number of calibration samples, the prediction ability of the model does not increase proportionally above a certain number of samples. Therefore the use of a large number of calibration samples can be questioned if the relationship quality-price of the model is of interest.

5.2.3 Conclusions of the chapter 4

1. *Summary.* This chapter has focused on optimal wavelength selection in CLS. The different conclusions in the bibliography about the relationship of the criteria for sensor selection with trueness, precision and accuracy motivated a more profound study. The confidence ellipsoid of the estimated concentrations and the experimental design theory offers the framework for interpreting the effect of these criteria in the prediction results of the CLS model built with the selected sensors (sections §4.2, §4.3, §4.4) and for new guidelines for wavelength selection.

2. *The best criterion for wavelength selection.* Although advances were made in clarifying the performance of the different criteria, it is still difficult to decide on the best criterion for wavelength selection (classified in §4.5). Global criteria (such as $Det(\mathbf{S}^T\mathbf{S})$, $Tr(\mathbf{S}^T\mathbf{S})^{-1}$, $Cond(\mathbf{S}^T\mathbf{S})$) may be not correlated with the prediction errors of a particular analyte. Moreover, the optimization of only one criterion is not enough to ensure good quality predictions.

ISBN:978-84-691-1875-7/DL: T-337-2008

3. *Selection of the optimal number of sensors.* Many procedures for optimal wavelength selection described in the literature do not consider the selection of the best number of sensors, which is decided a priori. This is not a guarantee that the number is adequate for obtaining good predictions.

4. *Trueness, precision and accuracy.* The concepts of accuracy and precision used in many papers must be reconsidered to check their agreement with the ISO definitions (§4.5). The effect on the trueness was considered in the wavelength selection criteria in §4.5. The criteria in CLS that are based on the calibration matrix are related to the precision of the estimated concentrations, not to the trueness.

5. *Influence of bias in the selection criteria in CLS.* The efficacy of the selection criteria in CLS based on the calibration matrix is strongly dependent on the absence of bias in the response at the selected sensors. A validation set is necessary to check this assumption. Otherwise, different conclusions can be obtained about the performance of the wavelength selection criteria. For example, a model built with D-optimal sensors will predict incorrectly an unknown sample if an interferent absorbs in these sensors. A model with the sensors selected randomly that, by chance, are free of the interference would have a smaller prediction error. The conclusion would be that the selection based on the D-criterion does not offers any improvement versus the random selection. Therefore, checking the quality of the data is of primary concern before a wavelength selection method is used.

6. *The A-criterion.* In CLS, the A-criterion gave best prediction errors than the D-criterion for the example considered (§4.2). Its drawback is that no specialized algorithms are known for optimizing the A-criterion and general optimization algorithms must be used.

7. *Collinearity.* Collinearity is a common problem in spectroscopic data and should be considered in the wavelength selection process. The effects of collinearity have been indicated in §2.6. The most studied criteria are collinearity diagnostics such as variance inflation factors (VIF), the condition number, and the selectivity after Lorber. The criteria based on reducing the collinearity should improve the prediction ability of the model although this is not general (§4.5).

ISBN:978-84-691-1875-7/DL: T-337-2008

8. *VIF's.* The variance inflation factors (VIF), a measure of collinearity used in MLR, were shown to be equivalent to the selectivity after Lorber (§4.6), which measures spectral overlap and has been used as a criterion to be optimized in wavelength selection problems.

9. *CLS versus ILS, PCR and PLS.* CLS requires knowledge of all the analytes that give response and this limits its application. Although PLS and PCR are more used at the moment, CLS has the advantage of being an easy and well understood and it is recommendable in the cases where the products present are well known. On the other hand, ILS almost necessarily involves using optimization algorithms to reduce the number of wavelengths due to the collinearity problem and for limitations on the number of the calibration samples. Hence, this is difficult to use before the optimization algorithms have been developed. On the other hand, factor-based methods are mathematically more complex than CLS.

10. *The net analyte signal in CLS.* The net analyte signal (NAS) is important to understand the quantification process in CLS and the error propagation to the predicted concentrations. Diagnostics based on the NAS for each particular analyte (instead of global measures) such as sensitivity, selectivity (§4.3) and net analyte signal regression plots (NASRP) have been used. The NASRP (§4.5) is a tool for detecting graphically if the measured response of the unknown sample follows the calculated model. The points of this plot fits a straight line whose slope is the estimated concentration. The residuals of the points should be randomly distributed on both sides of the line. The sensors which have bias can be easily detected and the sensors that best follow the model can be selected using the error indicator function and a moving window method. This criterion takes into account the noise of the data and the sensors with the smaller net analyte signal-to-noise ratio are not selected. A cut-off value for this criterion could not be deduced; the sensors with a small signal-to-noise ratio are recognized by comparison with the rest of the sensors. This criterion, different from others that are based on the calibration matrix, takes into account the sample to be predicted.

ISBN:978-84-691-1874-7/DL:T-337-2008

5.3 Considerations for future research

This thesis has focused on wavelength selection in CLS and calibration sample selection in PCR. Future work should extend the experimental design concepts to select the best subset of samples and sensors to other regression methods such as ILS, PLS and continuum regression. In any case, the position of the samples and sensors in the calibration space should be considered before selecting them for calibration. The following are some ideas for future work:

5.3.1 General considerations

1. *The confidence ellipsoid.* The confidence ellipsoid should be integrated in a calibration software as graphical diagnostic tool for examining the quality of the selected points the same as other diagnostics are.

2. *The selection of subsets that are a compromise for different criteria.* The points that optimize one criterion may not be optimal for the other criteria and the optimization of one criterion may not be sufficient to ensure a good predictive ability. The experimenter has to decide which criterion to optimize .Subsets that are not optimal for a particular criterion but a compromise with sufficient 'good' values for different criteria would be desirable. The procedure , called *multicriteria optimization,* is similar to that of sample or sensor selection but the criterion to optimize is the product of *desirability values* (Deming S.N. *J. Chromatography* 550 (1991) 15-25 / Hendriks M. M.W.B., de Boer J.H., Smilde A.K., Doornbos D.A. *Chem. Intell. Lab. Syst.* 16 (1992) 175-191). The criteria of interest are calculated for each candidate subset (e.g. the trace of the dispersion matrix, the selectivity for one of the analytes and the prediction errors). The value of each criterion is transformed into a desirability value between 0 and 1 using the desirability functions defined by the experimenter. The product of these values is the value to be optimized. The search of the solution requires optimization algorithms, such as GAs.

ISBN:978-84-691-1875-7/DL: T-337-2008

3. *Optimization algorithms.* Many sample and sensor selection problems are of combinatorial nature and need optimization algorithms. The Fedorov's algorithm is recommended for its speed in the search of D-optimal designs. Other criteria such as A- or E-optimal designs for sample or sensor selection requires optimization algorithms such as GA (Lucasius, C.B; Kateman, G. *Chem. Intell. Lab. Syst.* 1993, 19, 1-33 / Lucasius, C.B; Kateman, G. *Chem. Intell. Lab. Syst.* 1994, 25, 99-145 / Hibbert D.B. *Chem. Intell. Lab. Syst.*, 19 (1993) 277-293 / Shaffer R.E. Small, G.W., *Anal. Chem.* 1997 236A-242A) or generalized simulated annealing (GSA) (Kalivas J.H. *Chem. Intell. Lab. Syst.* 15 (1992) 1-12). To our experience, GAs are more easily applicable and versatile that GSA. These algorithms are used increasingly in the last years (the most recent publications seem to deal with GA for wavelength selection in PLS). However, the speed of convergence to the optimum of GAs need be optimized as well as their ability to stop the search when the optimum has been found. The fact the spectroscopic data are correlated can be used to implement new operators in the GA that produce a guided search by, once an acceptable solution has been found, searching systematically the wavelength near the optimum.

4. *Quality of the selected subsets.* Models using the same number of samples can give different results depending on the division of the overall data set into calibration and evaluation set (§3.2). In ILS and CLS the collinearity of the selected sensors must be checked. VIFs values can be used in wavelength selection in CLS.VIF values indicate whether the information contained in the calibration set is uncorrelated enough to find adequate estimates of the coefficients of the model. This has also a sense in PCR since the matrix of scores loses its orthogonality after selecting a few samples. VIF values can be calculated from the instrumental responses for a given model before measuring the analyte concentrations.

5.3.2 Considerations from the chapter 3

1. *More examples are needed.* Although the proposed methodology for sample and factor selection worked well on the real data sets considered (§3.2, §3.3 and §3.5), more data sets of different types of data should be used to fully validate and improve the selection methodology. Further studies on how to determine the

ISBN:978-84-691-1875-7/DL: T-337-2008

PCs that possibly have good predictive ability in the initial step of the methodology in §3.2 and the optimal number of PCs for PCR from the screening model could be made.

2. *Outliers in the calibration data.* Outliers are of special concern when a reduced number of calibration samples is used. Outliers with extreme values in the independent variables that could be selected by the Fedorov's algorithm, and incorrect determinations of the dependent variable in the few selected samples can lead to incorrect models. Therefore the data must be studied before it is submitted to the algorithm, for example, by plotting the points on the variable or factor space.

3. *The methodology applied to ILS and PLS.* Sample selection methodologies should also be considered in ILS, PLS and continuum regression as alternative to the random selection of calibration samples. However, the applicability of the methodology for ILS is limited by the wavelength selection step to reduce the number of coefficients to a number that enables the least-squares fitting. This requires optimization algorithms and a criterion which cannot be based on the prediction ability of the model, since the concentration is not known at the initial stage of the methodology. PLS and continuum regression use the analyte concentration of the calibration samples to calculate the factors. The methodology should solve this problem to identify the best samples without the knowledge of the analyte concentration.

4. *Kennard-Stone's algorithm.* The ability of the Kennard-Stone algorithm to select samples spread all over the experimental domain could be considered in further studies in combination with the Fedorov's algorithm. One can select the most influent samples to build the model and afterwards the Kennard-Stone algorithm can add additional samples to cover the experimental domain. The criterion to decide how many samples should be selected by each algorithm needs further studies.

ISBN:978-84-691-1875-7/DL: T-337-2008

5.3.3 Considerations from the chapter 4

1. *Sensitivity and selectivity in CLS.* Further studies about sensitivity in multivariate calibration and its utility to relate to prediction errors are necessary. A large sensitivity is not directly related to a good prediction. The net sensitivity is related to the length of the confidence interval and is used in the prediction process. Relating local and global measures of selectivity is difficult and must also be considered. For example, although the determinant has been said to measure orthogonality, the D-optimal sensors do not correspond to those that are $LSEL_k$ optimal nor the sum of selectivities (other combination were not checked).

2. *Expression for the variance of the predicted concentrations.* The selection of the wavelengths with a good predictive ability must be further studied. Criteria are needed to determine the optimal number of wavelengths. It the $Det(\mathbf{S}^T\mathbf{S})$ and $Tr(\mathbf{S}^T\mathbf{S})^{-1}$ are considered, the optimal number of sensors is the largest possible since they always improve when a sensor is added. The apparent contradiction that the precision in CLS improves when a new sensor is considered in the model even when the measured values in the sensor are noise indicates that more advanced expressions for the prediction error are needed. The optimal criterion should include information about the noise. Each analyte may have an optimal set of sensors, which was seen from the net analyte signal. The criteria for wavelength selection should consider the three sources of error. Now, new expressions for error propagation are available and they could be considered for optimal sample and sensor selection (Faber K., Kowalski B.R. *J. Chemom.* 11, (1997) 181-238.).The criterion should take into account the errors in the dependent and independent variables (which is not considered by the experimental design). The variance inflation factors (VIF) could be used to indicate if the sensors have the sufficient information for an adequate estimation of the concentrations in CLS. The utility of the VIFs in this field has not been demonstrated yet. Another possibility is the use of the net analyte signal regression plots and criteria of fit in this plot.

ISBN:978-84-691-1875-7/DL: T-337-2008

3. *The net analyte signal in inverse models.* Advances can be made in the NASRP by using robust regression methods such to relate the NAS of the calibration samples and the NAS of the unknown samples. The NASRP could also be used with the NAS calculated for the inverse models (Lorber A., Faber K. and Kowalski B.R. *Anal. Chem.* 69 (1997) 1620-1626) to detect bias in unknown samples.

4. *Software.* All the algorithms and quality criteria should be integrated in programs for its practical use in a laboratory. Different regression methods (MLR, PCR and PLS), selection algorithms (e.g. Kennard-Stone, the Fedorov or GA) and criteria optimization for sample and wavelength selection, in addition to collinearity diagnostics such as the VIFs, the variance-decomposition proportions, the NASRP, etc. should be included.