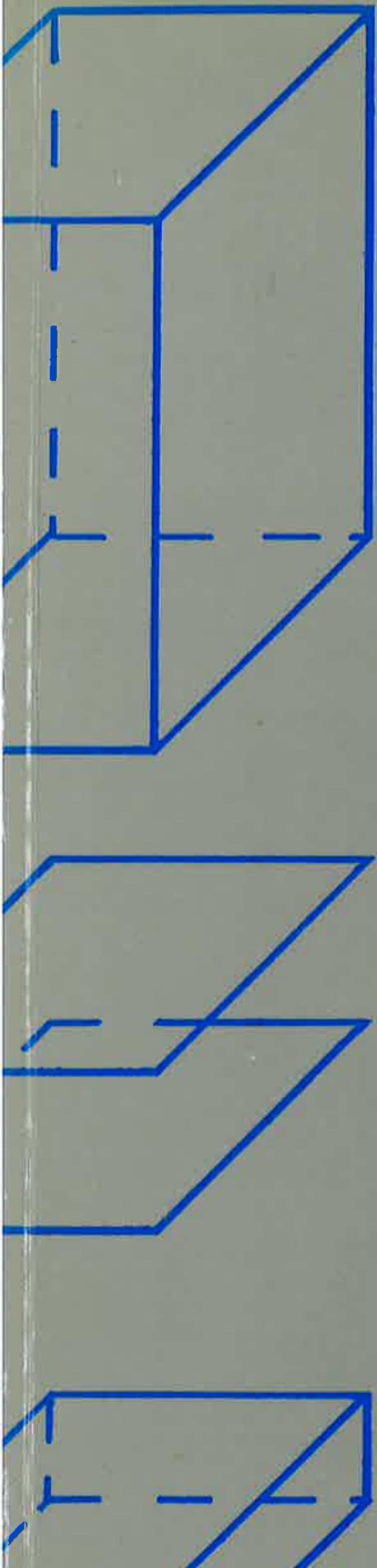


Multivariate Calibration of Reversed-Phase Chromatographic Systems

Age Smilde

PART I STATISTICAL INTRODUCTION

Chapter 1 The choice of the markers

Chapter 2 The linear model

Chapter 3 Biased estimation

Chapter 4 Experimental design considerations

PART II REVERSED-PHASE CHROMATOGRAPHIC INTRODUCTION

Chapter 5 The mobile phase

Chapter 6 The stationary phase

12. Het unieke van de televieserie "Medisch-Centrum West" is dat de hoeveelheid leed per vierkante meter is gemaximaliseerd.

13. Het dameshockey is nog niet erg geëmancipeerd: er moet nog steeds een "mannetje" worden gedekt.

14. Ook een proefschrift zónder chromatogram kan over chromatografie gaan.

Groningen, 16 maart 1990 Age Smilde

Stellingen

1. De R^2 waarde, zoals gebruikt door Héberger, is niet correct voor het door hem gebruikte exponentiële model. Dit verklaart waarom Héberger R^2 waarden groter dan 1 als uitkomsten verkrijgt.

   (K. Héberger; Emperical correlations between gas-chromatographic retention data and physical or topological properties of solute molecules, Anal.Chim.Acta, 223 (1989) 161-174)

2. De rechtvaardiging van het gebruik van PLS in response surface methodologie, zoals in de CARSO procedure, is zeer mager. Het beroep dat door de auteurs wordt gedaan op een beter voorspellend vermogen van PLS t.o.v. de gebruikelijke kleinste kwadraten methode, is onjuist.

   (S. Clementi, G. Cruciani, G. Curti and B. Skagerberg; PLS response surface optimization: the CARSO procedure, J. of Chemometrics, 3 (1989) 499-509)

3. De "intermediate least squares" regressie methode staat op gespannen voet met het "consistency at large" idee van H.Wold.

   (I.E. Frank; Intermediate Least Squares Regression Method, Chemom. and Intelligent Laboratory Systems, 1 (1989) 233-242.
   H. Wold; Soft Modeling: The Basic Design and Some Extensions, in: Systems under Indirect Observation, K. Jöreskog and H. Wold (eds), North-Holland Publ.Co., 1982)

4. Met het gebruik van partial least squares en principale componenten regressie wordt het probleem van variabele selectie niet vermeden.

5. De chromatografische begrippen "solvent strength" en "solvent selectivity" zijn niet exact gedefinieerd.

6. Een voorgeschreven vorm heeft in de muziek het creatieve proces nooit gedoofd, integendeel. Aio's en oio's komen dan ook niet per definitie tot hun recht door het geven van onbeperkte onderzoeksvrijheid.

7. De weerstand tegen onzuivere schattingsmethoden is te vergelijken met de weerstand tegen atonale muziek.

8. Strengere sancties op ontoelaatbaar spelgedrag van de voetballer sorteren een positief effect. Dit geldt niet voor strengere sancties op crimineel gedrag.

9. De snelheid waarmee tweede geldstroom projecten door de betreffende instanties wordt beoordeeld, dwingt de universitaire indieners van dergelijke projecten tot het verleggen van hun planningshorizon.

10. Ongerichte ambitie werkt destructief.

11. Het karakter van een citation index is meer incestueus dan serieus.

Promotor : Prof. Dr. D.A. Doornbos
Referent : Dr. P.M.J. Coenegracht

Promotiecommissie : Prof. Dr. P. Geladi, Universiteit van Umeå
Prof. Dr. D.L. Massart, Vrije Universiteit Brussel
Prof. Dr. T.J. Wansbeek, Rijksuniversiteit Groningen

Colofon :

Omslag-ontwerp: Wouter Daane
Druk: Febodruk, Enschede

RIJKSUNIVERSITEIT GRONINGEN

MULTIVARIATE CALIBRATION OF REVERSED-PHASE CHROMATOGRAPHIC SYSTEMS

Proefschrift

ter verkrijging van het doctoraat in de
Wiskunde en Natuurwetenschappen
aan de Rijksuniversiteit Groningen
op gezag van de
Rector Magnificus Dr. L.J. Engels
in het openbaar te verdedigen op
vrijdag 16 maart 1990
des namiddags te 14.45 uur precies

door

Age Klaas Smilde

geboren op 14 februari 1957

te

Leeuwarden

MULTIVARIATE CALIBRATION OF REVERSED-PHASE CHROMATOGRAPHIC SYSTEMS

Chapter 11 Two-way approach: the determinant criterion

Chapter 12 Two-way approach: other marker choices

Chapter 13 Three-way approaches

PART IV CALIBRATION OF OCTADECYL STATIONARY PHASES OF DIFFERENT BATCHES

Chapter 14 Experimental design

List of Symbols

$\mathrm{x}, \mathrm{y}, \mathrm{z}$	data vectors (first order arrays); sometimes a variable, depending on the context
x, y, z	multivariate variables
$\mathrm{X}, \mathrm{Y}, \mathrm{Z}$	data matrices (second order arrays)
$\mathbf{X}, \mathbf{Y}, \mathbf{Z}$	data cubes (third order arrays)
n	sample size
m, r	number of variables
i, j, k, s	running indices
Σ	population covariance matrix
R	sample correlation matrix
S	sample covariance matrix
P	matrix containing the right eigenvectors of a matrix (the loadings of the variables on the principal components)
D	diagonal matrix
T	matrix containing the left eigenvectors of a matrix (the scores of the objects on the principal components)
t_i	*i*th left eigenvector of a matrix (the scores of the objects on the *i*th principal component)
$\lambda_i^{\frac{1}{2}}$	*i*th singular value
λ_i	*i*th eigenvalue
p_i	*i*th right eigenvector of a matrix (the loadings of the variables on the *i*th principal component)
Q, W	special loading matrices used in PLS
$\mathrm{diag}(z_1, .., z_m)$	diagonal matrix with diagonal elements z_1 to z_m
I_k	Identity matrix of order k
$\det(X)$	determinant of matrix X
$\mathrm{tr}(X)$	trace of matrix X

ϵ	random error
β	population model coefficients
b_{ols}	an estimate of β with the ordinary least squares method
$V(b_{ols})$	variance-covariance matrix of the estimator b_{ols}
a_i , b_i , c_i	vectors containing loadings (scores) of the variables (objects) on the *i*th PARAFAC component
A, B, C	matrices containing loadings and scores of the variables and objects on the PARAFAC components
e, f	vectors (or scalars) of residuals
E, F	matrices of residuals
$\mathbf{E}$, $\mathbf{F}$	cubes of residuals
g	number of components in a PLS, PCA or PARAFAC model
δ_{ij}	Kronecker symbol: $\delta_{ij} = 1$ if i=j, $\delta_{ij} = 0$ otherwise
ι	the vector (1,...,1)'

Preface

One of the most widespread techniques in analytical chemistry is high-performance liquid chromatography, especially the reversed-phase mode (RP-HPLC). In the stationary phases of RP-HPLC a ligand is covalently bonded to the silanol groups at the silica surface. Different types of ligands give different selectivities towards the group of solutes that is to be separated. In this way optimization of a separation can be performed, besides mobile phase manipulation, by changing the type of stationary phase.

A problem that arises when using RP-HPLC is the occurence of between-batch variation. When the mobile phase composition of a separation is optimized on, e.g., an octadecyl stationary phase, the optimal mobile phase composition will no longer be optimal when a fresh column is used. This change is due to small irreproducible differences between the old and the fresh octadecyl stationary phase (assuming that they are from different batches). Similarly, during an operation, the stationary phase deteriorates and the optimal mobile phase composition has to be updated.

A first step in the optimization of a separation by varying the type of stationary phase, is the transfer of retention values of a set of solutes on a stationary phase to a stationary phase of a different type. This problem is called: "Calibration of various types of stationary phases" and is dealt with in Part III.

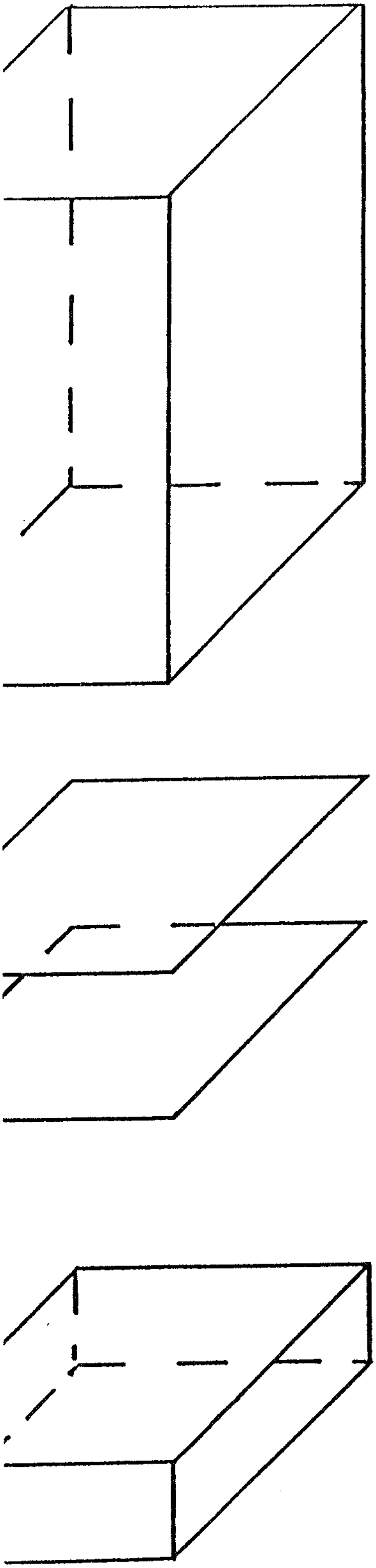
A first step in correcting the optimal mobile phase composition when batch differences arise, consists of transferring the retention values of a set of solutes on a stationary phase of one batch to a stationary phase of another batch. This topic is dealt with in Part IV and is called: "Calibration of octadecyl stationary phases of different batches".

The problem of the deterioration of a stationary phase is not dealt with, but some general ideas are given in Chapter 8.

The above mentioned calibration strategies demand their own statistics, in particular multivariate statistics. A review of the (multivariate) statistical techniques used in this thesis are given in Part I.

Part II deals with some recent relevant developments in RP-HPLC. Of particular interest is Chapter 8: the core of this thesis.

In order to avoid extensive phrases, some abbreviations are used instead. The test solutes are abbreviated with capitals. Thus TOL stands for the test solute toluene. The **solute** TOL refers to the solute toluene as such, whereas the **variable** TOL refers to the retention values (ln k values) of the solute toluene. TOL, without the addition of the noun "solute" or "variable" can have both meanings, but the exact meaning of the abbreviation will be clear from the context.



PART I
Statistical Introduction

Chapter 1 The choice of the markers

1.1 Description of the problem

Let x be an m-variate variable with mean zero and covariance matrix Σ of order $m \times m$. A number of n measurements $x_1, \ldots, x_n$ is made on this variable. These measurements are gathered in $X=(x_1, \ldots, x_n)'$ of order $n \times m$. The matrix X may contain retention measurements of a number of solutes (the variables) at different mobile phase compositions (the objects) on different stationary phases (clustering of the objects). The matrix Σ can be in correlation form, this depends on the kind of scaling of x.

If the data matrix X is available, the question arises whether it is possible to discard a number of variables (columns) of X without loosing too much information. Especially in cases where X is highly structured this is possible indeed. The notion "loss of information" should be defined sharply and depends on the problem definition for which X is used. A possible measure of information is variation and accordingly some variables in X are retained which explain a large part of the variation. Another possibility is to retain variables that can predict the discarded variables as well as possible. A third measure of information arises from the observation that X can be understood as the notation of n objects (data points, rows) in the m-dimensional Euclidian space. Discarding variables can then be seen as a projection of the objects to a space of lower dimension, spanned by the retained variables. In that case those variables should be retained which preserve the "distance structure" within the original m-dimensional space in a satisfactory way (the concept "distance structure" has to be defined exactly).

Obviously, the maximum of information is retained if all variables are used, but using all variables has two disadvantages: a statistical one and a practical one. The selected variables are used as predictors in models. For a given sample size there is an optimal number of predictors for a given model, optimal in the sense of prediction performance of the model. Increasing the number of predictors (variables) beyond this point only worsens the predictions. This notion is called the peak-phenomenon and is treated more extensively in Chapter 2. The intuitive reason is clear: additional variables give additional information and additional uncertainty. Beyond a certain point the additional uncertainty overwhelms the additional information. From a practical point of view a low number of variables is profitable. The variables represent solutes of which retention measurements have to be made in order to calibrate a new stationary phase. Minimizing the number of experiments is therefore equivalent to a low number of selected variables.

Throughout this thesis the retained variables are called markers. The meaning of this name will become clear in the following.

1.2 Principal component analysis (PCA)

It is assumed that a data matrix X is available which consists of n measurements (objects) on an m-variate variable x and X is column-mean centered (see Section 1.7). An unbiased estimate of Σ is then:

$$S = \frac{1}{n-1} X'X \qquad (I.1)$$

If the columns of X are scaled to variance one, this S is in correlation form. In the following analysis it is not necessary to make a distinction between an S matrix in correlation or in covariance form. It must be stressed, however, that there is no straightforward relationship between the principal components (PC's) obtained from a correlation matrix and those based on the corresponding covariance matrix.

A thorough description of PCA is given by Anderson[1], Mardia *et al.*[2], and Jolliffe[3]. Some results are stated. The matrix S is symmetric and positive semidefinite, therefore S has eigenvalues $\lambda_1,\ldots,\lambda_m \geq 0$ and eigenvectors $p_1,\ldots,p_m$ with $p_i'p_j=0$, $\|p_i\|=(p_i'p_i)^{-\frac{1}{2}}=1$. The matrix $P=(p_1,\ldots,p_m)$ is orthogonal ($P'P=PP'=I$), so from SP=PD follows

$$S = PDP' \qquad (I.2)$$

where $D=\text{diag}(\lambda_1,\ldots,\lambda_m)$ and $\lambda_i>0$ (each i) if rank(S)=m. The *i*th principal component (PC) is given by $t_i=Xp_i$ with p_i the *i*th column of P. Equation (I.2) can also be written in the spectral decomposition form:

$$S = \lambda_1 p_1 p_1' + \lambda_2 p_2 p_2' + \ldots + \lambda_m p_m p_m' \qquad (I.3)$$

This spectral decomposition is a successive rank one approximation of S because $\text{rank}(\lambda_i p_i p_i')=1$ for every i unless $\lambda_i=0$. If an eigenvalue is zero, this means that rank(S)<m, so that there is a linear dependence in the columns of S.

Some properties of PCA are readily shown:

i) $t_i't_j = p_i'X'Xp_j = (n-1)p_i'Sp_j = (n-1)p_i'PDP'p_j = (n-1)\delta_{ij}\lambda_i$, where $\delta_{ij}=1$ when i=j and 0 otherwise. This can be shown by making use of P'P=I.

ii) $(1/n)\iota' t_i = (1/n)\iota' Xp_i = (1/n)\iota'(x_1,..,x_m)p_i = (1/n)(\iota' x_1,..,\iota' x_m)p_i = (1/n)(0,..,0)p_i = 0$, where $\iota=(1,..1)'$ and $X=(x_1,..,x_m)$. Use the fact that X is column-mean centered.

Property *i)* means that two PC's are orthogonal to each other. Each PC has mean zero, when the original variables are centered, which is shown in property *ii)*. The variance of the *i*th PC is seen to be λ_i, by using property *i)* and *ii)*. It is customary to arrange the eigenvalues in D in non-increasing order, so $\lambda_1 \geq ... \geq \lambda_m \geq 0$. The PC's are arranged in the same order. The first PC is then the linear function $t_1 = p_1' x$ of *x* which has maximum variance λ_1. The second PC is a linear function $t_2 = p_2' x$ of *x* which has maximum variance under the condition that $t_1' t_2 = 0$, this variance is λ_2 and so on. The vectors p_i are called the loadings of the m variables on the *i*th PC and the vectors t_i are called the scores of the objects on the *i*th PC. The PC's are strictly defined as the scores, but in practice the term "PC" is used for either the scores or the loadings vector, usually this is clear within the context.

The total variance in X is trace(S), which equals the sum of variances of the m variables in *x*. It can be shown[4] that $tr(S) = tr(PDP') = tr(DP'P) = tr(D) = \Sigma_i \lambda_i$. The sum of variances in X is decomposed in parts related to the PC's. The percentage of variation accounted for by the first k PC's can be expressed as:

$$100 \sum_{1}^{k} \lambda_i \ / \sum_{1}^{m} \lambda_i \qquad (I.4)$$

The usefulness of PCA as a dimension reducing technique is seen with the use of formula (I.4). If a lot of variation in X can be accounted for by a few PC's, the transformation of the original x-variables to PC's is worth it to be considered. While the primary aim of PCA is the search for linear combinations of the original variables in such a way that the sum of variances of those original variables (tr(S)) is satisfactorily reallocated, formula (I.3) shows that the PC's explain successively also the off-diagonal elements of S. This can be seen by realizing that the eigenvalues are ordered in a non-increasing way and therefore the elements in the successive parts of the spectral decomposition tend to become smaller[3]. This distinguishes PCA from factor analysis (FA), because FA focusses on the off-diagonal elements of S.

The geometrical interpretation of PCA is shown in Figure I.1. From data matrix X, representing measurements made on two variables for eight objects, the first PC is visualized by a straight line through the origin (assuming that X is column-mean centered) of the data cloud and the vector t_1 contains the projections of the original data points on this line. This straight line is selected in such a way that the projections on this line have a maximum spread. The second PC is orthogonal to the first. There are only two PC's because there are not more than two variables. The PC's transform the data with an orthogonal transformation of the original variable axis. Detailed explanations of geometric interpretations have been given, e.g. by Wold[6] and Davis[7]. It should be noted that the first PC gives a good representation of the data. The data can be approximated or explained rather good by the first PC. If PCA is used to reduce an m-dimensional data structure to three dimensions, 3-D plots of the three first PC's can be seen as a view in that m-dimensional space.

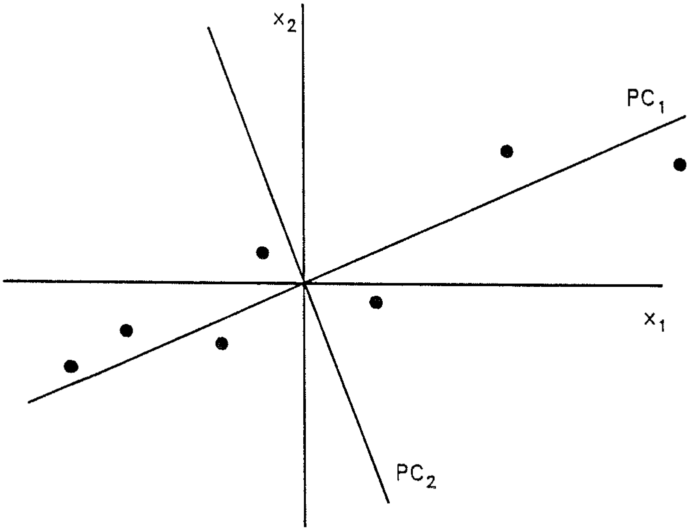


Figure I.1. Geometrical interpretation of PCA.

The loading vectors p_i and the eigenvalues λ_i are important in the diagnosis of the influence of the x variables. Near-zero eigenvalues give valuable information regarding linear dependence between the variables. To show this, suppose $\lambda_m \approx 0$. Then $Xp_m \approx 0$ because $t_m = Xp_m \Rightarrow t_m' t_m = (Xp_m)'(Xp_m) \approx 0$, with the use of property *i*). This indicates a near_linear dependence between the X-columns. Suppose $x=(x_1, \ldots, x_m)$ and the loading of x_1 on the final PC, the first value in p_m, is high, say 0.90, and the loading of x_2 on this PC is 0.44 (the loadings of the other variables are 0 because 0.90x0.90+0.44x0.44=1 as should be since P'P=I). One of the two variables is redundant, since the variation of one of them can be explained almost totally by the variation in the other because a near-linear dependence exists. Which one should be thrown away? For the sake of simplicity assume that S is in correlation form. The fraction of variation in x_i "explained" by t_j is $p_{ij}^2 \lambda_j$. This value can also be understood as the contribution of x_i to t_j. Likewise, the contribution of x_i to the first m-1 PC's is $\Sigma_{j=1}^{m-1}\ p_{ij}^2\lambda_j=1-p_{1m}^2\lambda_m$. Then $1-p_{2m}^2\lambda_m$ is the contribution of x_2 to the first m-1 PC's. Because p_{2m}^2 is smaller than p_{1m}^2, the contribution of x_2 to the first m-1 PC's is higher than the contribution of x_1. It seems, therefore, reasonable to discard variable x_1, the variable most associated with the redundant PC. Analogous reasoning can be used for all those PC's with near-zero or small eigenvalues. Note that in case of exact linear dependence, $\lambda_m=0$, it does not matter which variable with a non-zero loading on that PC is discarded.

Another way to look at the eigenvalues and loadings is by investigating the PC's with high eigenvalues and select the variables which are highly associated with these PC's. This can be done by investigating the loadings of the variables on the PC's with high eigenvalues and choosing those variables which span the PC-loading space at best.

Both lines of reasoning were followed by Jolliffe[7,8]. He concludes that both strategies give reasonable results in practice. A similar approach is adopted by S.Wold *et al.*[9] in the context of "Multivariate Design". S.Wold *et al.* use the scores of the objects on the first PC's to choose the objects which span the PC-score space as good as possible. Analogously, the loading plots can be examined and the variables be chosen which span the loading space.

The spectral decomposition (see formulas (I.2) and (I.3)) can be generalized for arbitrary nxm matrices, not necessarily square. This is called the singular value decomposition (SVD) of matrix X (see Jolliffe[3], Graybill[4], Rao and Mitra[10]). This results in:

$$X = \tilde{T}\tilde{D}^{\frac{1}{2}}P' \qquad (I.5)$$

where X is an (nxm) data matrix. $\tilde{T}$, P are (nxr), (mxr) matrices, respectively, in such a way that $\tilde{T}'\tilde{T}=I_r$, $P'P=I_r$. The matrix $\tilde{D}^{\frac{1}{2}}$ is a (rxr) diagonal matrix with diagonal elements $\tilde{\lambda}_1^{\frac{1}{2}}$ to $\tilde{\lambda}_r^{\frac{1}{2}}$ and r is rank(X). If X has full rank m then r=m. The $\tilde{D}^{\frac{1}{2}}$ matrix is arranged so that $\tilde{\lambda}_1^{\frac{1}{2}} \geq\ldots\geq\tilde{\lambda}_m^{\frac{1}{2}} \geq 0$. These values are called singular values. The SVD of X written in a slightly different form gives:

$$X = \tilde{\lambda}_1^{\frac{1}{2}}\tilde{t}_1p_1' + \ldots + \tilde{\lambda}_m^{\frac{1}{2}}\tilde{t}_mp_m' \qquad (I.6)$$

with $\tilde{T}=(\tilde{t}_1,\ldots,\tilde{t}_m)$, $P=(p_1,\ldots p_m)$, and rank(X)=m. The symbol "˜" is used to make a distinction between two closely related formulas, viz. (I.6) and (I.7) (see the following). Formula (I.6) is called the singular value decomposition of X. This singular value decomposition is a successively rank one approximation of X because each term $\tilde{\lambda}_i^{\frac{1}{2}}\tilde{t}_ip_i'$ has rank one. If rank(X)=r<m then the final m-r singular values are zero. In practice this is never the case, because X

contains measurement errors, so an exact linear dependence between the columns of X is not likely to occur. In cases where some of the singular values are nearly equal to zero, the effective rank of X becomes less than m and X can be approximated by, say, the first $r<m$ terms in the SVD.

The notion of approximation can be sharpened by using the concept of the Frobenius or Euclidian norm of a matrix. The Frobenius norm of X ($\|X\|$) is given by $\Sigma_i \Sigma_j x_{ij}^2 = tr(X'X)$. Straightforward calculations show that $\|x\| = \Sigma\tilde{\lambda}_i$ and $\tilde{\lambda}_i = \|\tilde{\lambda}_i^{\frac{1}{2}} \tilde{t}_i p_i'\|$, where $\tilde{\lambda}_i$ is the square of $\tilde{\lambda}_i^{\frac{1}{2}}$. This means that the matrix X can be approximated successively by matrices of non-increasing norms, the first matrix $\tilde{\lambda}_1^{\frac{1}{2}} \tilde{t}_1 p_1'$ accounts for the largest variation in X within the class of rank one matrices (Rao[10], Graybill[4]). The second matrix $\tilde{\lambda}_2^{\frac{1}{2}} \tilde{t}_2 p_2'$ accounts for the largest variation in $X - \tilde{\lambda}_1^{\frac{1}{2}} \tilde{t}_1 p_1'$ within the class of rank one matrices under the constraint that $\tilde{t}_1' \tilde{t}_2 = p_1' p_2 = 0$ and so on.

The similarity between the SVD and PCA becomes clear by realizing that $X'X = (\tilde{T}\tilde{D}^{\frac{1}{2}}P')'(\tilde{T}\tilde{D}^{\frac{1}{2}}P') = P\tilde{D}^{\frac{1}{2}}\tilde{T}'\tilde{T}\tilde{D}^{\frac{1}{2}}P' = P\tilde{D}P'$, because of $\tilde{T}'\tilde{T} = I$ and $\tilde{D}^{\frac{1}{2}}\tilde{D}^{\frac{1}{2}} = \tilde{D}$. This is the spectral decomposition of X'X. The eigenvalues of X'X are $\tilde{\lambda}_1, \ldots, \tilde{\lambda}_m$ and it is clear that the squares of the singular values are, besides a constant (n-1), equal to the eigenvalues of S ($\tilde{D} = (1/(n-1))D$). If the columns of X are scaled-to-length-one, the squares of the singular values equal exactly the eigenvalues of S. The P matrix as obtained in the SVD of X equals the matrix of loadings of the PCA on the x variables. The scores of the objects on the *i*th PC were $t_i = Xp_i$, see Section 1.2. With the use of (I.5) it follows that $t_i = Xp_i = \tilde{T}\tilde{D}^{\frac{1}{2}}P'p_i = \tilde{T}\tilde{D}^{\frac{1}{2}}\gamma_i = \tilde{\lambda}_i^{\frac{1}{2}}\tilde{t}_i$ where $P'P = I$ and γ_i is the *i*th column of the identity matrix. When the matrix $\tilde{D}^{\frac{1}{2}}$ is absorbed in the matrix $\tilde{T}$, the matrix T is the result, so $T = \tilde{T}\tilde{D}^{\frac{1}{2}}$. Thus the $\tilde{t}_i$th column vector contains the scaled-to-length-one scores on the *i*th PC. Efficient algorithms are available to calculate the SVD of a matrix[11]. The loadings and scores on the PC's are then readily available.

If the effective rank of X is $g<m$, then (I.6) can be written as

$$X = t_1 p_1' + \ldots + t_g p_g' + E = T_g P_g' + E \qquad (I.7)$$

where E is the (nxm) matrix of residuals and $t_i p_i'$ the outer products associated with each dimension, $T_g = (t_1, \ldots, t_g)$ and $P_g = (p_1, \ldots, p_g)$. Formula (I.7) shows that X is approximated by $T_g P_g'$. The quality of this approximation, in terms of $\|E\|$, does not alter if both T_g and P_g' are multiplied by an orthogonal matrix Q, because $X = T_g P_g' + E = T_g QQ'P_g' + E = (T_g Q)(P_g Q)' + E$. This invariance under orthogonal transformations has its consequences for the variable selection procedure mentioned above. Variables were selected on the basis of P_g, but it is also justified to select variables on the basis of P_gQ, where Q is an arbitrary orthogonal matrix. This complicates the variable selection procedure on the basis of loadings on PC's.

In order to show the bilinear character of the approximation of X, formula (I.7) is written in a slightly different form:

$$x_{ij} = \sum_{k=1}^{g} t_{ik}p_{kj} + e_{ij} \quad i=1,\dots,n \; ; \; j=1,\dots,m \qquad \text{(I.8)}$$

where t_{ik} and p_{kj} are the typical elements of T_g and P_g, respectively. The word "bilinear" means that, with p_{kj} fixed, (I.8) is linear in t and vice versa. Formula (I.8) may be understood as a model for x_{ij}. The number of components, g, used to approximate X can be established by cross-validation[12-14], by Horn's test[15], by procedures proposed by Malinowski[16,17] or by comparing the standard deviation of the residuals in (I.8) with the measurement error.

A justification of the use of the SVD (and also PCA) in the analysis of data tables comes from the idea of latent structure. If an underlying process, unobservable as such, becomes manifest in the measured variables which are summarized in the data table X, the scores on the PC's are estimates of so-called latent variables. The ideas concerning latent variables and latent variable modelling have received much attention recently in the statistical literature[18,19]. A particular justification of SVD and PCA for the analysis of multivariate chemical data is given by S.Wold[20,21] and finds it roots in the observation that extra thermodynamic relationships (linear free energy relationships) are special cases of latent variable modelling. The use of SVD (or PCA) to model the data matrix X can be viewed as a local approximation: a generalization of polynomial Taylor-approximation. This observation puts a restraint on the use of PCA and SVD in the analysis of multivariate data: a latent structure is presumed.

A convenient way to picture the SVD and PCA approach is given by S.Wold *et al.*[5] In Figure I.2, a data table X is shown which is approximated by rank-one matrices written as outer products: $t_i p_i'$. Geometrically, the SVD can be visualized along the same lines as PCA. The scores t_i can be regarded as orthogonal projections of the data on the linear combination of the x variables whose weights are given by p_i. Stated otherwise: a new basis is defined for describing the data which consists of an orthogonal transformation of the original axes, see Figure I.1.

Some new developments in PCA research comprise the stability of PCA[22], warnings against PCA[23,24], leverage and influence measures for PCA[25], and the effect of sample design on PCA[26].

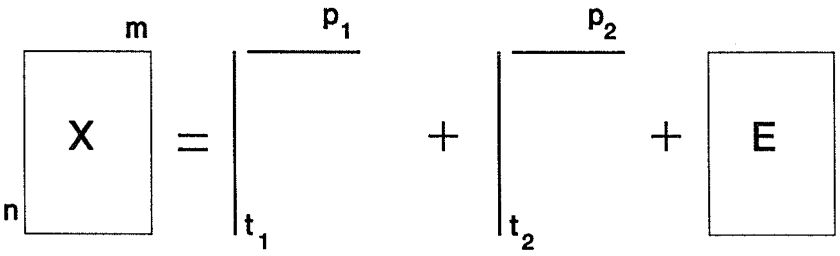


Figure I.2. Singular value decomposition of X.

1.3 Principal variables

Some of the multivariate techniques shown in this Section rely on the work of McCabe[27]. Let X be an (nxm) data matrix with zero column-means and let x_j be the *j*th column of X. Again no distinction is made between a correlation or covariance form of S=(1/(n-1))X'X. If r variables are selected, the X matrix can be partitioned as:

$$X = (X_1 , X_2) \qquad (I.9)$$

where X_1 is (nxr), X_2 is (nx(m-r)) and X_1 consists of the columns of X which contain the scores on the retained variables. The corresponding partition of S is:

$$S = \left(\frac{S_{11} \mid S_{12}}{S_{21} \mid S_{22}} \right) \qquad (I.10)$$

where S_{11} is (rxr) and S_{22} is (m-r)x(m-r). The generalized variance of the selected variables is given by det(S_{11}). The idea is to select the variables (for a fixed r) which maximize this generalized variance. All combinations of r variables are investigated and the combination with the highest det(S_{11}) is chosen. The determinant criterion for the selection of markers is suitable for situations where retention of variation is important. It should be stressed that the determinant criterion assumes a fixed r because determinants of matrices of different sizes cannot be compared in a simple way.

Other criteria for the selection of variables can be used. Let

$S_{22.1}$ be the conditional covariance of X_2 given X_1 or, in other terms, the covariance of X_2 corrected for the influence of X_1. Then:

$$S_{22.1} = S_{22} - S_{21}S_{11}^{-1}S_{12} \qquad (I.11)$$

For a proof of this formula, see Anderson[1]. Formula (I.11) can be explained by applying the following reasoning. Suppose the objective is to predict X_2 by X_1. The least squares predictor of X_2 is X_1B with $B=(X_1'X_1)^{-1}X_1'X_2$ (see Draper and Smith[28], Mardia *et al.*[2]). Now $S_{11}=(n-1)^{-1}X_1'X_1$ and $S_{12}=(n-1)^{-1}X_1'X_2$, so $B=S_{11}^{-1}S_{12}$ and $\hat{X}_2=X_1S_{11}^{-1}S_{12}$. The explained variation in X_2 is $(n-1)^{-1}\hat{X}_2'\hat{X}_2=(X_1S_{11}^{-1}S_{12})'(X_1S_{11}^{-1}S_{12})(n-1)^{-1}$ $= S_{12}'S_{11}^{-1}X_1'X_1S_{11}^{-1}S_{12}(n-1)^{-1} = S_{12}'S_{11}^{-1}S_{11}S_{11}^{-1}S_{12} = S_{21}S_{11}^{-1}S_{12}$. The variation to be explained in X_2 is $(n-1)^{-1}X_2'X_2=S_{22}$. Therefore, (I.11) represents the unexplained variation. It can be recognized as the covariance matrix of the (unexplained) residuals $E=X_2-\hat{X}_2$. A logical step seems to be the choice of the variables which minimize this covariance matrix in some respect. The minimization of $\det(S_{22.1})$ is the same as the maximization of $\det(S_{11})$, because $\det(S) = \det(S_{11})\times\det(S_{22.1})$ (Graybill[4]) and $\det(S)$ is fixed. Therefore, maximizing the retained variation, represented by $\det(S_{11})$ is the same as minimizing the lost variation, represented by $\det(S_{22.1})$.

Another way to minimize $S_{22.1}$ is the minimization of $\mathrm{tr}(S_{22.1})$, this is not the complement of the maximization of $\mathrm{tr}(S_{11})$. This $\mathrm{tr}(S_{22.1})$ is the sum of unexplained variances of the discarded variables after regression of these variables on the retained ones. The minimization of the trace aims at an optimal prediction of the discarded variables. It can be shown[27] that:

$$\sum_{j=1}^{m} s_{jj}R^2(x_j,X_1) = \mathrm{tr}(S) - \mathrm{tr}(S_{22.1}) \qquad (I.12)$$

where s_{jj} is the *j*th main-diagonal element of S, the variance of the *j*th variable, and R^2 is the squared multiple correlation coefficient if x_j is regressed on X_1. The term $s_{jj}R^2(x_j,X_1)$ is called the induced variance of X_1 on x_j. Therefore, by minimizing $\mathrm{tr}(S_{22.1})$, the sum of the induced variances is maximized because $\mathrm{tr}(S)$ is fixed for a given data matrix X. This marker choice criterion will be called the induced-variance criterion. Note that for a column x_j of X which is retained, the induced variance is exactly s_{jj}.

In PCA, results are often described by a percentage of explained variation. The induced-variance criterion leads to the definition of percent explained variation:

$$P = \frac{\sum_{j=1}^{r} s_{jj} + \sum_{j=r+1}^{m} s_{jj} R^2(x_j, X_1)}{\sum_{j=1}^{m} s_{jj}} \times 100\% \qquad (I.13)$$

This percentage can be recognized as the percentage of the total variation which is induced (it should be remembered that the first r variables are retained and their induced variances are the s_{jj}'s).

Another interpretation of P can be shown by making use of the Frobenius (or Euclidian) norm of a matrix X, which was defined earlier. From this definition it follows that $\|X\|^2 = tr(X'X)$. Thus $\|X\|^2 = tr((n-1)S) = (n-1)tr(S)$. The matrix $E = X_2 - \hat{X}_2$ has norm $\|E\|^2 = tr(E'E) = tr((n-1)S_{22.1}) = (n-1)tr(S_{22.1})$. The denominator of P is equivalent to (I.12) and becomes $(n-1)^{-1}\|X\|^2 - (n-1)^{-1}\|E\|^2$. The nominator of P is $(n-1)^{-1}\|X\|^2$. Therefore, P can also be written as:

$$P = \frac{\|X\|^2 - \|E\|^2}{\|X\|^2} \times 100\% \qquad (I.14)$$

Realizing that the norm of a matrix is a measure of its magnitude, the term $\|E\|$ in (I.14) can be interpreted as the magnitude of the amount of information which is thrown away. The interpretation of P becomes then: the relative magnitude of the amount of information which is retained.

Both criteria, the determinant- and the induced-variance criterion, can be used to select variables. Generally, the result of the selection will depend on the criterion which is used. The problem of the fixed r value in case of working with the determinant criterion can be solved by first using (I.13) for establishing the r value and for this choice of r the determinant criterion can be used. A measure of the percentage explained variation which can be attained maximally, using an r-dimensional subspace of the original m dimensions, is given by the percentage of explained variation by the first r principal components (note that by choosing r variables, a subspace of the original m-dimensional space is chosen). This is because of optimum properties of PCA[29]. Comparing the percentage explained variation by the first r PC's with the percentage explained variation of r specific variables, gives an idea of how efficient these variables are with regard to the dimension-reduction.

The complete procedure can be summarized as follows

1. Make a choice with regard to scaling the columns of X, or stated otherwise, use a correlation or a covariance approach (see Section 1.7).

2. Let r be 1,2,3,...
3. Choose a criterion (determinant or induced-variance) and select the optimal variables.
4. Calculate the explained variation of those variables and compare this with the explained variation of the first r PC's.
5. If the percentage explained variation is not high enough in comparison with the PCA result, then make r one unit higher and return to 3.

To illustrate a weakness of the induced-variance- and determinant criterion suppose the following. From a set of four variables two should be selected (assuming the variables scaled in such a way that S is in correlation form). The first three variables have reasonable high two and two correlation coefficients (say 0.9) and are therefore exchangeable. The final variable has no high correlation with any of the first three ones (say maximally 0.40), so this final variable can be regarded as "outlying". If two of the variables are chosen according to the induced-variance criterion, the final variable will be part of the optimal subset, because exclusion of that variable will give rise to a low P (see I.13) since its induced variance is low. The final variable is not chosen on the basis of its predictive power, but on the basis of its bad predictability. Tests on "outlying" variables are therefore needed prior to using the induced-variance criterion, because this criterion is not robust against deviating, or outlying, variables. The same holds for the determinant criterion.

An approach related to the use of the induced-variance criterion is used by Steigstra[30]. There are two differences. Steigstra selects variables successively, but if a variable is in the selected set, it remains in that set when the number of variables in the retained set is raised by one. This does not guarantee that, for a given size of the set of selected variables, the optimal ones of Steigstra's approach, equal the best ones of the induced-variance approach. A second difference is that Steigstra performs a Gram-Schmidt orthogonalisation procedure on the set of selected variables in order to obtain an orthogonal basis for the space spanned by the selected variables. This difference is not essential, but is related to the purpose of the variable selection.

A very nasty property of the above sketched procedures for choosing markers is the uncontrollable influence of chance. If four variables are chosen out of twenty, a number of 4845 possible combinations of four variables are evaluated. The outcome of the above sketched procedure is an optimal subset of four variables, but it may be purely chance that this particular subset shows up as the optimal one. This problem is thoroughly described for the cases of variable selection in multiple regression and discriminant analysis by Steerneman[31,32], Schaafsma and Steerneman[33] and is called data-mining or chance correlation (see also Section 2.4). Some examples of this phenomenon are also reported by Topliss[34,35] in the area of quantitative structure activity relationships.

In order to avoid the problem of data-mining a possible solution may be to use a jack-knife procedure. The idea is to leave one observation out of the training set and calculate the markers. This is done n times, where each observation is omitted once. A number of n subsets of markers is the result. By examining these subsets it is possible to get an idea of how reliable the selection procedure is. If the calculations are too excessive, this procedure can be adjusted (see Sections 10.1 and 11.1). This jack-knife procedure may be a solution of the data-mining problem because it is not very likely to obtain n times the same set of markers purely by chance. This is more likely when only one calculation is performed on the whole training set.

Another way to avoid the problem of data-mining is the reduction of the number of subsets of variables at the beginning of the procedure on the basis of chromatographic arguments. A limited number of variables (solutes) are then tested on their mark-power and there will be less chance correlation. Quantitative measures of the degree of chance correlations in this procedure are hard to derive (Steerneman[32]). The build-in restriction used by Steigstra[30] decreases the number of possible subsets and might have a favourable effect on decreasing chance correlations.

1.4 Procrustes analysis and related ideas

A method of selection of variables is outlined by Krzanowski[36,37]. Starting from the matrix X (n×m), which is assumed to be column-mean-centered, a principal component analysis is performed on this matrix. Let the scores on the first r principal components be gathered in Y (n×r). The multivariate structure of the set of points (in R^m) associated with X are supposed to be revealed by the PCA of X (in R^r). Suppose g of the original m variables are selected ($g \geq r$) and the scores on these g variables are gathered in X_{sel} (n×g). The multivariate structure, present in X_{sel}, is revealed by a PCA on X_{sel}, where again r components are supposed to be sufficient to describe this structure. The scores of this latter PCA are gathered in Z (n×r). The discrepancy between the original configuration (enclosed in Y) and the configuration after the selection (enclosed in Z) can be calculated as the sum of squared differences between corresponding points of the two configurations after they have been matched as well as possible under translation, rotation, and reflection. Matching under translation is already performed, because both X and X_{sel} are column-mean-centered, and therefore Y and Z are also mean-centered. This matching operation is called Procrustes analysis and the sum of squared differences (Gower[38]) after matching under rotation and reflection is

$$SSD = tr\ [YY'+ZZ'-2ZQ'Y'] \qquad (I.15)$$

where Q=TP' from Z'Y=TDP', the singular value decomposition of Z'Y, and SSD is the abbreviation of sum of squared differences.

The procedure to choose g markers is now easy. All possible subsets are subjected to the above sketched analysis. The subset with the lowest SSD value is the best one. Two remarks are appropriate. First, the danger of chance results is present in this procedure, so leave-one-out techniques may give insight in the stability of the solution. Second, Krzanowski[36] gives a backward elimination procedure to select the best subset, which partly reduces the problem of chance results, because less alternatives are tested.

A modification of the procedure described above is the removal of the necessity to calculate two principal component analyses. Suppose X (nxm) and X_{sel} (nxg) are as before. Only a PCA on X is performed extracting g components. The resulting scores are gathered in T (nxg). With Procrustes analysis the matrices X_{sel} and T, and the associated sets of points in R^g, are matched under rotation and reflection. The necessity to establish r is obsolete.

A related procedure (called DISNORM, <u>dis</u>tance <u>norm</u>) to Procrustes analysis is the following. The first step is the same as above. The matrix X is subjected to a PCA and the first r components are supposed to give a sufficient describtion of the multivariate structure in X. The scores on these r components are gathered in Y (nxr) and can be represented as n points in R^r. A quantitative measure of the configuration of these n points is the distance matrix DS (nxn) with typical element ds_{ij}, the Euclidian distance between points i and j, defined as

$$ds_{ij} = [\ \sum_{k=1}^{r}(y_{ik}-y_{jk})^2\ ]^{\frac{1}{2}} \qquad (I.16)$$

When the matrix X_{sel} (nxg) is subjected to a PCA, the resulting score matrix of the first r components is Z (nxr). From this matrix a distance matrix can be calculated also and will be called DS^*. Note that both DS and DS^* are rotation and reflection independent. A measure of the discrepancy of the configurations represented by Y and Z is $\|DS-DS^*\|$, where $\|.\|$ is a norm of a matrix, e.g. the Frobenius norm. The subset of variables is chosen that gives rise to the lowest $\|DS-DS^*\|$ value.

Two other alternatives are reported. Both redundancy analysis[39] and Escoufier's RV-coefficient[40] seem promising.

1.5 Variable selection in supervised pattern recognition (SPR)

If the objects in the data matrix X can be ordered in predefined groups and interest focusses on the description of differences between these groups, the way is open to use variable selection techniques for supervised pattern recognition (SPR). The variables that discriminate most between the groups should be chosen. Again the danger of data-mining shows up and care must be taken in the variable selection procedure[32,33]. In the case at hand, a natural group structure might be given by the stationary phases. Measurements performed on one specific stationary phase at varying mobile phase compositions of a set of solutes are then regarded as one group. A problem of this approach might be the strong overlap between groups. This can occur because within group variation is caused by changing the mobile phase composition, which is known to have a larger effect on retention than changing the stationary phase which accounts for the between group variation.

The SPR method SIMCA, developed by S.Wold[41], models each group with a few PC's. The discriminating power of a variable is a measure of the involvement of that variable in the PC class models. This discriminating power can be used to explore influential variables[41,42].

The SPR method linear discriminant analysis (LDA) is a technique developed by Fischer[43]. Its use is widely spread and there are numerous ways to select variables[44,32]. The SPR methods CLASSY and ALLOC can also be used (van der Voet and Hemel[45], Coomans and Broeckaert[46]). The SOLOMON method uses a variable selection method that is already described (Steigstra[29,47]).

1.6 The simultaneous choice of markers and mobile phase compositions: three-way methods

Extensions of PCA and factor analysis (FA) are described which pertain to three-way data tables (Carroll[48], Kroonenberg[49], Harshman[50], Tucker[51]). A survey of three- and multi-way data analyses is given by Law *et al.*[52]. Applications in chemistry are given by S.Wold[53] and de Ligny *et al.*[54]. The idea of tensorial calibration[55,56] is closely related[57] to the PARAFAC model which will be discussed later on.

A natural distinction in the data can be made in three directions. The first direction corresponds to the stationary phases, the second direction to the mobile phase composition variables, the fractions of MeOH, ACN, THF etc. The third direction comprises the solutes. This situation is depicted in Figure I.3. Each number in the three-way data table is a retention measurement of a solute at a specific mobile phase composition on a specific stationary phase. The stationary phases can be regarded as cases (objects), the solutes and mobile phase compositions as two categories of variables. Two approaches to generalize PCA and FA to three-mode data tables will be outlined.

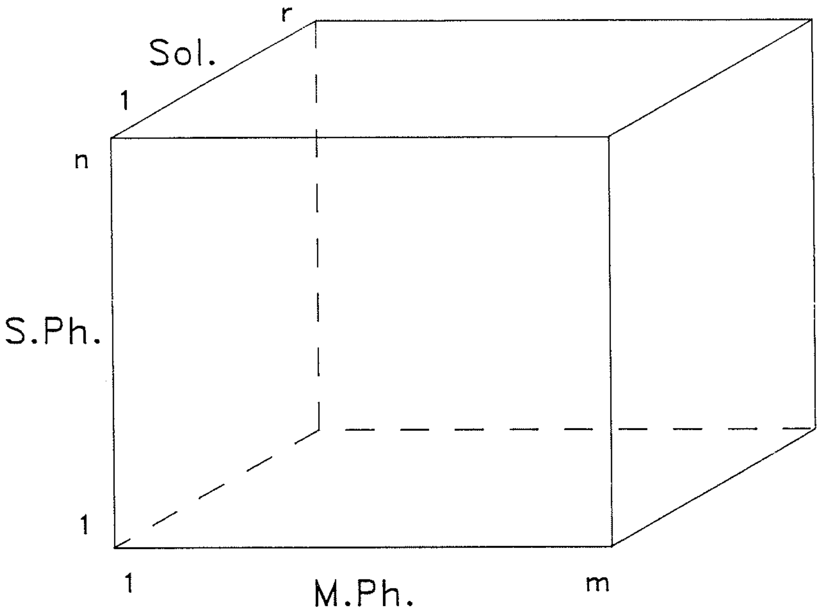


Figure I.3. Data cube of retention values; n is the number of stationary phases (S.Ph.), r the number of solutes (SOL.) and m the number of mobile phase compositions (M.Ph.).

The first generalization of model (I.8) is given by:

$$x_{ijk} = \sum_{s=1}^{g} t_{is}p_{sjk} + e_{ijk} \qquad \begin{array}{l} i=1,\ldots,n \\ j=1,\ldots,m \\ k=1,\ldots,r \end{array} \qquad (I.17)$$

or, alternatively,

$$\mathbf{X} = t_1 \otimes P_1 + \ldots + t_g \otimes P_g + \mathbf{E} = T. \otimes \mathbf{P} + \mathbf{E} \qquad (I.18)$$

where $t_s=(t_{1s},\ldots,t_{ns})'$, $T=(t_1,\ldots,t_g)$, **X** has typical element x_{ijk}, **E** has typical element e_{ijk}, P_s has typical element p_{sjk} and **P** is a three-way array with dimensions (gxmxr), with P_1 the uppermost horizontal slice in **P** and P_g the lowermost one. For notational details we refer to S.Wold *et al.*[53]. The typical element of $t_s \otimes P_s$ is $t_{is}p_{sjk}$. Note that the symbol "⊗" does not denote the usual Kronecker product. Since $t_s \otimes P_s$ is a three-way array, model (I.18) is again a successive approximation of **X**. The number of factors, g, used to decompose **X**, must be established. The same procedures as mentioned earlier (Section 1.2) can be used. In Figure I.4, model (I.18) is illustrated for the two-factor case.

In order to make the generalization consequent, t_s and P_s are calculated in such a way that T'T is diagonal, $P_s'P_s = I$ for every s=1,..,g and the Frobenius norm of **E** is minimized. With this generalization the idea of bilinearity is not extended: model (I.18) is not a trilinear model[52].

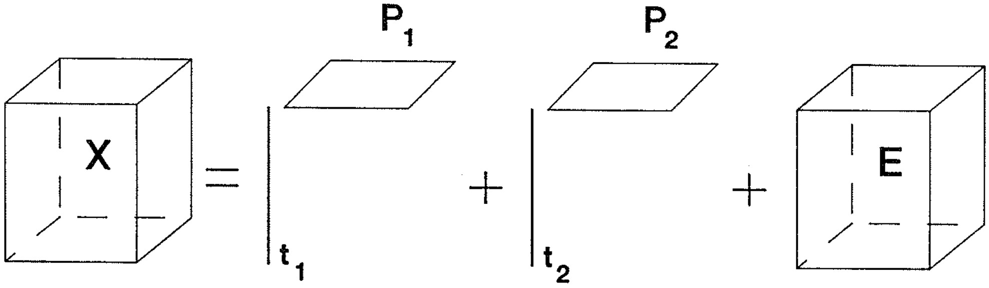


*Figure I.4. Two-factor decomposition of **X** given by model (I.18).*

S.Wold *et al.*[53] show that the estimates t_s and P_s can be obtained by unfolding the data cube **X**. The idea of unfolding is visualized in Figure I.5. The unfolding can, of course, be done in different directions. That direction which is related to the objects should remain intact and projected onto the t_s vectors. The unfolded data matrix X can be subjected to a singular value decomposition yielding the wanted estimates. For obvious reasons, this generalization (model (I.17)) of model (I.8) will be called unfold-PCA.

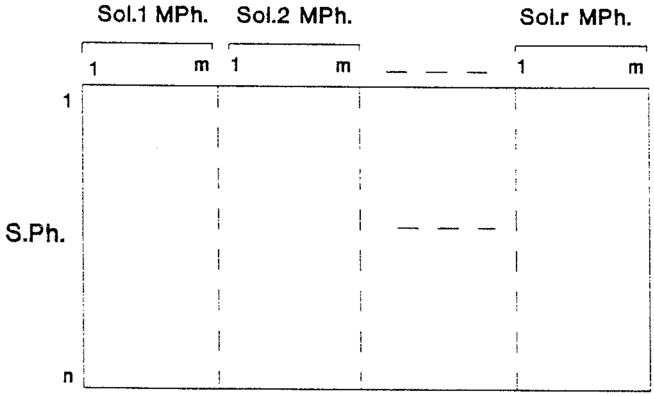


Figure I.5. Unfolded data cube of Figure I.3, for the abbreviations see legend Figure I.3.

The second generalization of (I.8) is given by

$$x_{ijk} = \sum_{s=1}^{g} a_{is}b_{js}c_{ks} + e_{ijk} \qquad \begin{array}{l} i=1,\ldots,n \\ j=1,\ldots,m \\ k=1,\ldots,r \end{array} \qquad (I.19)$$

where, again, g is the number of factors used, not necessarily the same as in (I.17). If X_i represents the m×r matrix, which is the *i*th slice of **X**, then model (I.19) can be written as

$$X_i = BA_iC' + E_i \qquad i=1,\ldots,n \qquad (I.20)$$

where B is a m×g matrix with typical element b_{js}, C is a r×g matrix with typical element c_{ks}, E_i is a m×r matrix with typical element e_{ijk}. The g×g matrix A_i is diagonal, with diagonal elements taken from the *i*th row of A, the n×g score matrix of the first mode with typical element a_{is}. The g diagonal elements thus represent the effect of the changes in the relative importance of the g factors on influencing retention on stationary phase i (when the stationary phases are treated as objects). For a given number of components in model (I.19) the coefficients a_{is}, b_{js}, and c_{ks} are estimated such that $\Sigma\Sigma\Sigma e^2_{ijk}$ is minimum, where the summation runs over i, j, and k. Each slice X_i is factor analysed, and this is performed in a parallel fashion. Model (I.19) is, therefore, referred to as the PARAFAC decomposition[50] (Parallel Factor Analysis).

The PARAFAC decomposition is shown in Figure I.6 for the two-factor case. Establishing the number of factors in model (I.19) can, e.g., be done by comparing the standard deviation of the residuals with the size of the measurement error. The vectors $a_s=(a_{1s},\ldots,a_{ns})'$ can be chosen orthogonal to each other or not. The same can be done for the b_s and c_s vectors. Note that the PARAFAC model is a trilinear model; the idea of bilinearity is extended in model (I.19), contrary to model (I.17).

For both the models (I.17) and (I.19) holds that scaling and centering is of particular importance, because the results of the decompositions depend on these aspects.

There is a conceptual difference between models (I.17) and (I.19). Model (I.19) is an unconstrained model: the values p_{sjk} are the factor loadings of the *s*th (component) across modes B and C of the data and no constraint is placed on these values. Such a constraint is present in model (I.19): $p_{sjk}=b_{js}.c_{ks}$. The meaning of this constraint can be explained as follows. Comparing $b_{1s}.c_{1s}$ and $b_{1s}.c_{2s}$ with $b_{2s}.c_{1s}$ and $b_{2s}.c_{2s}$, it is clear that the expression of the influence of factor s across mode C (as measured by c_{1s} and c_{2s}) does not depend on the level of mode B. The reverse is also true: the expression of the influence of factor s across mode B does not depend on the level of mode C. In order to illustrate the consequences of

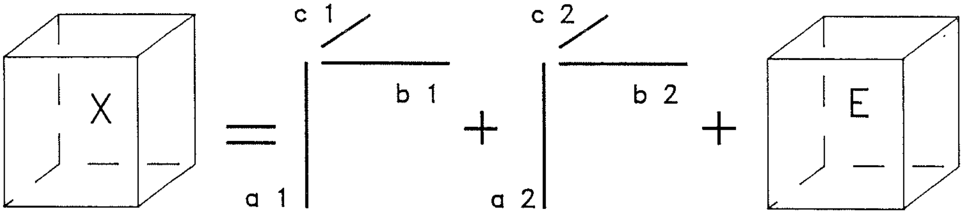


*Figure I.6. Two-factor PARAFAC decomposition of **X**.*

this difference between models (I.17) and (I.19), suppose that the stationary phases comprise the first mode (the objects), the solutes the second mode and the mobile phase compositions the third mode. PCA studies[58,59] on reversed-phase chromatographic data indicate that the loadings of the solutes on the first PC are related to the hydrophobic character of the solutes. Similarly, the loadings of the mobile phase compositions may be related to the polarity of the eluent. With respect to the first factor in the decomposition of X, the above mentioned constraint means that the way in which hydrophobic differences between solutes affect the differences in retention values of those solutes, does not depend on the mobile phase composition.

Whether to use unfold-PCA or PARAFAC is a tedious question, analogous to the question which model to use in linear regression. S.Wold *et al.*[53] claim that the PARAFAC decomposition is too restricted and propagate the unfolding strategy. Law *et al.*[52], however, show that the presupposed latent structure in X has its consequence for the kind of model to be used. An example of such a consideration was already given above. A careful examination of the kind of data in X is suggested, especially its background, and the decision which model to use should not only be made on the basis of the size of the residuals e_{ijk}. Assuming the same number of factors in the alternative models (I.17) and (I.19), it should be stressed that the number of parameters to be estimated is considerably larger for the unfold-PCA model than for the PARAFAC model. This is the pay-off for not imposing the constraint, as mentioned above, on the model. The result is a larger number of degrees of freedom for the PARAFAC model and therefore perhaps lower variances of the estimated loadings and scores of the PARAFAC model compared to the unfold-PCA ones.

The treatment of the different modes or directions in the three-way table, or data cube, is not symmetrical in the unfold-PCA model (note that the unfolding can be done on three different ways). A choice has to be made, therefore, when model (I.17) is used, which direction constitutes the one associated with the t vectors: the scores.

The variable selection problem as described in Section 1.1 can be generalized to the three-way case. The selection should encompass the solute/mobile phase combinations that will constitute the training set. Those solute/mobile phase combinations should be selected which "explain much of the variation" in the data cube **X**. A measure of variation of Z is $\|Z\|$, the Frobenius norm: the sum of squared elements of Z. This measure can be used to define the notion of explained variation. This selection problem is visualized in Figure I.7.

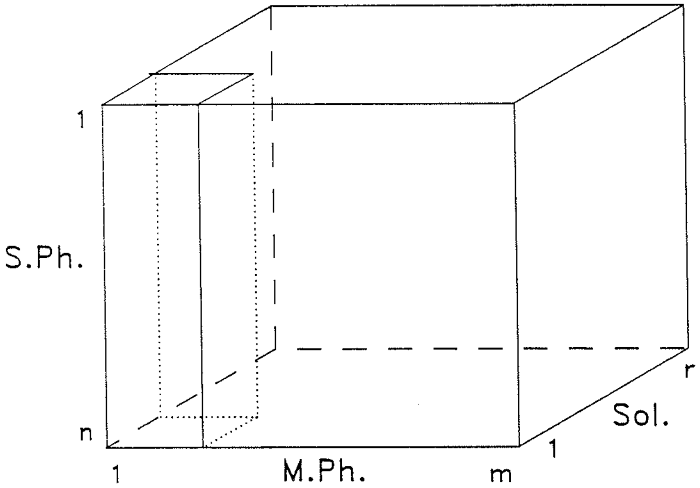


Figure I.7. Choice of the markers/mobile phase combinations which describe the systematic variation in X.

A generalization of the Jolliffe-approach (see Section 1.2) is readily available. Suppose **X** is decomposed by unfolding this data cube (in the direction so that the stationary phase mode is left intact) and two factors are obtained which explain enough variation in **X** to hold as a good description of **X**. The loading plot of these two factors of the unfolded data cube can be used to chose the solute/mobile phase combinations that load high on these two factors, or stated otherwise, which are mostly associated with the two factors. Examples which will illustrate that such a choice is not easy, will be given in the following chapters. When a decomposition of **X** with the use of (I.19) is available (suppose again that two factors are appropriate), loading vectors b_1, b_2, c_1, and c_2 are calculated which describe the contribution of the variables in the second and third mode, respectively, to the two factors (note that the first mode contains the stationary phases). Variables from the second mode (say, the mobile phases) can be chosen that mostly associate with the loadings b_1 and b_2. Analogous for the solutes, the third mode, variables can be chosen that mostly associate with c_1 and c_2. Examples will be given in Chapters 13 and 16.

Extensions of the principal variables approach with regard to the problem of selection in two directions, the simultaneous choice of variables related to two categories, are not described in the literature. A "quick and dirty" method is the unfolding of **X**, followed by the principal variables approach to the unfolded matrix X.

1.7 The effect of scaling

Different kinds of scaling can be applied to the data matrix $X=(x_1,\ldots,x_m)$. The different scaling features are illustrated by applying these to the scores on the first variable $x_1=(x_{11},\ldots,x_{n1})'$. Scaling of x_1 is an operation cx_1, where c is an arbitrary real number. Usually scaling is done to give $z_1=cx_1$ a predefined length d, hence c is chosen in such a way that $z_1'z_1=d$. Scaling-to-length-one is a special case of scaling so that d=1. This can be accomplished by choosing c as $1/\|x_1\|$, where $\|x_1\|=(x_1'x_1)^{\frac{1}{2}}$. Scaling influences the variance of x_1 but not the mean.

Centering is the operation on x_1 so that the resulting variable z_1 has mean zero. Vinod and Ullah[60] show that the centering operation on x_1 can be performed by using the matrix $[I-\iota\iota'/n]$, hence $z_1=[I-\iota\iota'/n]x_1$, with $\iota=(1,..,1)'$. Centering influences the mean of x_1 but not the variance.

The autoscaling operation involves two aspects, first the variable x_1 is centered and then scaled to obtain variance one. The variable $z_1=(1/s_1)[I-\iota\iota'/n]x_1$, where s_1 is the standard deviation of the scores on x_1 with divisor n-1, has mean zero and variance one.

If all variables in X are column-autoscaled $S=(n-1)^{-1}X'X$ gives the dispersion of X in correlation form. If the columns in X are centered and scaled-to-length-one, $S=X'X$ produces the correlation form dispersion of X. The last type of scaling is common in ridge- and Stein regression. If X is only column-centered, $S=(1/(n-1))X'X$ gives the dispersion of X in covariance form. Note that centering followed by scaling-to-length-one on the one hand and autoscaling on the other hand differ only the multiplicative constant $\surd(n-1)$.

It is already mentioned that PCA, Unfold PCA, and PARAFAC depend on scaling. Jolliffe[3] states that the modelling behaviour of PCA towards the off-diagonal elements of S depends crucially on the kind of scaling. The approaches sketched in Sections 1.3, 1.4, and 1.5 do also depend on scaling. Whether to use scaling, is a choice which has to be made and is part of the model structure which is postulated. The effects of using or ignoring scaling must be fully realized. The subject of scaling is given attention in the literature (Kvalheim[61], Brown[62]). One rule of thumb may be to use autoscaling or scaling-to-length-one when variables are involved measured on different measurement scales. The combination in one model of fractions of organic modifiers and solute capacity factors would, e.g., justify scaling.

The influence of scaling on results in ridge- Stein- and partial least squares regression will be discussed in the chapters where these methods are described.

Chapter 2 The linear model

2.1 Some basic issues

The issues stated here are readily available from text books concerning regression analysis (Vinod[60], Johnston[63], Judge *et al.*[64], Draper and Smith[28]). Consider the linear regression model:

$$\tilde{y} = \tilde{X}\beta + \tilde{\epsilon} \qquad \text{(I.21a)}$$

where $\tilde{\beta}'=(\tilde{\beta}_0,...,\tilde{\beta}_m)'$ is a vector of unknown regression coefficients; $\tilde{y}$ is an nx1 vector of an observable random (dependent) variable; $\tilde{X}$ is an nx(m+1) matrix of observable independent or predictor variables with the first column consisting of ones to account for the constant in the regression equation; $\tilde{\epsilon}$ is an nx1 vector of disturbances. It is assumed that $\tilde{\epsilon}$ has mean 0 and variance-covariance matrix $E(\tilde{\epsilon}\tilde{\epsilon}')=\sigma^2 I$. This assumption means, among other things, that the variance of the disturbances does not depend on the magnitude of X (homoscedasticity). The symbol "~" is used to make a distinction between the model with uncentered data, (I.21a), and the model with centered data, which is (I.21b). Measuring $\tilde{y}$ and $\tilde{X}$ about their column means, gives y of order nx1 and X of order nxm, respectively, where the column of zero's obtained after the column-mean centering of $\tilde{X}$ is omitted to obtain X. The matrix X has, therefore, full rank. To be more specific, $y=(I-\iota\iota'/n)\tilde{y}$ and $X=(I-\iota\iota'/n)\tilde{X}_1$ with $\tilde{X}=(\iota \ , \ \tilde{X}_1)$. In the following it is assumed that y and X have this form, unless stated otherwise. The linear regression model in terms of deviations from the column means is

$$y = X\beta + \epsilon \qquad \text{(I.21b)}$$

where $\beta'=(\tilde{\beta}_1,...,\tilde{\beta}_m)'=(\beta_1,...,\beta_m)'$ and $\epsilon=(I-(1/n)\iota\iota')\tilde{\epsilon}$, with ι as before. Scaling of y and X has no effect on the results of the least squares calculations[60,63]. The ordinary least squares (OLS) estimator of β is

$$b_{ols} = (X'X)^{-1}X'y \qquad \text{(I.22)}$$

This estimator minimizes SSE=e'e=(y-Xb)'(y-Xb), where b is an arbitrary estimator of β and SSE is the sum of squared residuals. Some familiar properties of b_{ols} are

$$E(b_{ols}) = \beta \qquad \text{(I.23)}$$

$$V(b_{ols}) = \sigma^2(X'X)^{-1} \qquad \text{(I.24)}$$

These formulas show that the b_{ols} estimator is unbiased and has variance-covariance matrix $V(b_{ols})$. Because of $X'e=X'(y-Xb_{ols})$

$=X'(y-X(X'X)^{-1}X'y)=X'y-X'X(X'X)^{-1}X'y=X'y-X'y=0$, it follows that

$$y'y = b'_{ols}X'Xb_{ols} + e'e \tag{I.25}$$

Formula (I.25) can be written as SST=SSR+SSE, where SST is the total sum of squares (which has to be explained), SSR is the sum of squares due to regression (explained sum of squares) and SSE is the sum of squared errors (unexplained sum of squares).

A well-known criterion for the performance of the model is the squared multiple correlation coefficient

$$R^2 = \frac{SSR}{SST} = 1- \frac{e'e}{y'y} \tag{I.26}$$

This is interpreted as the proportion of variation in y explained by the regression model. This performance criterion has some disadvantages as will be shown in Section 2.3.

The estimation of β in the linear model can also be obtained in the case that X consists of n realizations of an m-variate random variable (see Judge *et al.*[58]). For the moment no difference is made between the random and non-random case.

If the fractions of the mobile phase constituents are incorporated in a linear model, which is familiar in response surface modeLling with mixture variables[65], centering of the data is not appropriate. To see this, consider the model $\tilde{y}=\tilde{X}\tilde{\beta}+\tilde{\epsilon}$, in which $\tilde{y}$ and $\tilde{X}$ $(n\times(m+1))$ are not centered, $\tilde{X}=(\iota,\tilde{X}_1)=(\iota,\tilde{x}_1,\ldots,\tilde{x}_m)$ with $\tilde{x}_1+\tilde{x}_2+\tilde{x}_3=1$ or $\tilde{X}_1c=\iota$, $c=(1,1,1,0,..0)'$ and $\iota=(1,..,1)'$ as usual. Vinod[60] shows that column centering a matrix $\tilde{X}_1$ gives the matrix X, which can be calculated as $X=(I-\iota\iota'/n)\tilde{X}_1$. Now $Xc= (I-\iota\iota'/n)\tilde{X}_1c =(I-\iota\iota'/n)\iota=(I\iota-\iota\iota'\iota/n)=(\iota-\iota)=0$. The column-centering operation introduces a linear combination in the centered design matrix X, diminishing rank(X) by one. The calculation of the inverse of X'X, necessary for b_{ols}, is therefore not possible. One solution is to delete one of the variables $\tilde{x}_1$, $\tilde{x}_2$ or $\tilde{x}_3$. Another solution is to keep the variables $\tilde{x}_1$ to $\tilde{x}_3$ in the equation, but to leave the constant out.

2.2 Diagnostics for the linear model

If a linear model like $y=X\beta+\epsilon$ is used, it is wise to use some diagnostic tools in order to make the right interpretation of the results. Some results from Belsley *et al.*[66] are summarized. For the sake of convenience, it is assumed that y $(n\times1)$ and X $(n\times m)$, of full rank, are mean-centered and the columns of X are scaled-to-length-one. Special attention is paid to the nature of X because (near) linear dependence between the x-variables give rise to problems in the linear model. This phenomenon, which is called multicollinearity, is discussed in detail in Section 2.4.

A first step in measuring dependence between columns of X is the examination of the correlation matrix R=X'X. Now r_{ij} (the *ij*th element of R) is the correlation coefficient between x_i and x_j. High correlations indicate potential problems. This R, however, only reveals dependence between two variables and not between a set (three or more) of x-variables. The diagnostic value is, therefore, limited.

Because of (I.24), the diagonal elements of R^{-1}, where $R^{-1}=(X'X)^{-1}$, are often called the variance inflation factors (VIF's). Their diagnostic value rises from the observation that

$$VIF_i = \frac{1}{1 - R_i^2} \qquad (I.27)$$

where R_i^2 is the multiple correlation coefficient of x_i regressed on the remaining predictor variables. If R_i^2 is high, large variances of the b_{ols} can occur, see (I.24). In case of near linear dependence between columns in X the inversion of X'X becomes very unstable. The VIF's are, therefore, not very reliable in case of serious multicollinearity.

With the use of the SVD of $X = TD^{\frac{1}{2}}P'$ (omitting the "$\tilde{}$" notation as used in Section 1.2 for convenience), the variance-covariance matrix of b_{ols} can be written as:

$$V(b_{ols}) = \sigma^2 (X'X)^{-1} = \sigma^2 PD^{-1}P' \qquad (I.28)$$

The variance of the *k*th component of b_{ols} can then be written as

$$\text{var}(b_{ols,k}) = \sigma^2 \sum_j \frac{p_{kj}^2}{\lambda_j} \qquad (I.29)$$

where p_{kj} is the *kj*th element of P and $D^{\frac{1}{2}}=\text{diag}(\lambda_1^{\frac{1}{2}},\ldots,\lambda_m^{\frac{1}{2}})$. The variance of the *k*th component of b_{ols} is decomposed in parts associated with one and only one of the m singular values $\lambda_j^{\frac{1}{2}}$. A singular value near zero points to a near linear dependence between the columns of X and let the associated terms in (I.29) blow up. The columns of X which are not part of this dependence have loadings zero on the associated PC, so that for these variables the p_j values are zero. Consequently the near-zero value of the particular singular value does not blow up the variance of the estimated coefficients of these variables. An unusually high proportion of the variance of two or more coefficients concentrated in the same small singular value suggests that the corresponding near dependence may give rise to problems. If the *kj*th proportion of the variance decomposition is defined as the proportion of the variance of the *k*th regression coefficient associated with the *j*th component of its decomposition in (I.29), these proportions can be readily calculated. Let

$$\theta_{kj} = \frac{p_{kj}^2}{\lambda_j} \quad \text{and} \quad \theta_k = \sum_{j=1}^{m} \theta_{kj} \quad k=1,\ldots,m \tag{I.30}$$

then the variance decomposition proportions are

$$\pi_{jk} = \frac{\theta_{kj}}{\theta_k} \qquad k,j=1,\ldots,m \tag{I.31}$$

The π-values can be arranged conveniently in a Π-matrix providing a quick view on probable near dependence (Figure I.8). Examples of the use of these diagnostic tools will be given in Chapters 10 to 12.

$$\begin{array}{c|ccc|} & b_1 & \cdots & b_m \\ \hline & \pi_{11} & \cdots & \pi_{1m} \\ \Pi = & \vdots & & \vdots \\ & \pi_{m1} & \cdots & \pi_{mm} \\ \hline \end{array}$$

Figure I.8. Matrix of variance-decomposition proportions.

Another diagnostic tool which proves to be useful is latent-root singular value decomposition (LR-SVD) (Jolliffe[3], Mason[67], Vinod[60]). The LR-SVD is a singular value decomposition performed on the X matrix augmented with the y-column and indicates which variables in X are important for the prediction of y. Of special interest are the loading vectors associated with the low singular values because, as argued before, here the linear combinations show up. When y loads on such a singular vector and also one or more x variables then these x variables can be regarded as important in predicting y. In Section 2.4 it is shown that this LR-SVD can be used to investigate multicollinearity. Latent root regression is an estimation method based on this principle (Hawkins[68], Webster *et al.*[69]).

2.3 Validation of linear models

One of the most important issues in (linear) model building is the assessment of the quality of the model. The model may fail to describe reality or to predict future observations because the important variables are omitted, or because unimportant variables are retained, or because there is no linear relationship at all. Familiar criteria for the validation of linear models are t- and F-statistics, these are described in most textbooks for regression analysis. It should be kept in mind that these criteria are designed to test specific hypotheses, which may not be of interest for the user. Another familiar criterion is R^2 (see Section 2.1). Because R^2 is a non-decreasing function of the number of variables (Judge *et al.*[64], Hocking[70]), it will always advise the use of all variables. This is not always the optimal choice when good prediction and estimation is the aim (Allen[71], Breiman and Freedman[72], Steerneman[32]). This will be illustrated later. One way to cope with this problem is the use of

$$R^2_{adj} = 1-(n/(n-m-1))(1-R^2) \qquad (I.32)$$

the adjusted R^2, where m+1 is the number of variables in the non-centered regression equation (I.21a). A penalty is introduced for the use of too much variables. Both R^2 and R^2_{adj} do not consider the particular purpose for which the model will be used: for a description of the data or for the prediction of future observations (our purpose).

A different class of validation criteria, designed to meet the specific demands of the user with regard to the purpose of the model, follows from the concept of mean square error (MSE) as a part of statistical decision theory (see e.g. Amemiya[73], Breiman and Freedman[72], Vinod[60]). Let b be an estimator of β in the usual linear model $y=X\beta+\epsilon$. Then

$$\begin{aligned} MSE(b) &= E(b-\beta)'(b-\beta) = E(b-Eb+Eb-\beta)'(b-Eb+Eb-\beta) \qquad (I.33)\\ &= E(b-Eb)'(b-Eb) + (Eb-\beta)'(Eb-\beta)\\ &= tr[V(b)] + [bias(b)]'[bias(b)] \end{aligned}$$

where tr represents the trace of a matrix and E represents the expectation over b. The Euclidian length of a vector v is $(v'v)^{\frac{1}{2}}$ and MSE(b) measures, therefore, the average of the squared Euclidian distance between b and β. It was shown earlier that $E(b_{ols})=\beta$, consequently the bias of the OLS estimator is zero, so $MSE(b_{ols})= tr[V(b_{ols})]$.

Considering the prediction of mean values of y at the sample points ($\mu_y=X\beta$), the mean squared error of prediction (MSEP) is defined by:

$$MSEP(b,X) = E(\mu_y-Xb)'(\mu_y-Xb) = E(b-\beta)'X'X(b-\beta) \qquad (I.34)$$

and measures the average squared Euclidian distance between predicted and mean values of y, the expectation is again over b. This MSEP(b,X) can also be conceived as MSE($\hat{y}$), where $\hat{y}$=Xb.

If the prediction of the mean value of y at a point x_0 is considered ($\mu_{y0}=x_0'\beta$), the mean squared error of prediction is:

$$\begin{aligned} \mathrm{MSEP}(b,x_0) &= E(\mu_{y0}-x_0'b)'(\mu_{y0}-x_0'b) \\ &= E(b-\beta)'x_0x_0'(b-\beta) \end{aligned} \qquad (I.35)$$

which is a measure of the prediction error in an arbitrary point in the design space, again the expectation is over b. Stated otherwise, MSEP(b,x_0) is the conditional mean square error of prediction, conditional on a value of x_0. If the expectation is taken over x_0 the result is the unconditional mean squared error of prediction (UMSEP):

$$\mathrm{UMSEP}(b) = E(\mathrm{MSEP}(b,x_0)=E[E(b-\beta)'x_0x_0'(b-\beta)] \qquad (I.36)$$

where the expectations are taken over b and x_0. It can be observed that the MSE criteria are all expected values of quadratic functions of b-β. They all measure, therefore, the distance between b and β according to a metric corresponding with the positive semidefinite matrix involved in the quadratic term.

As a warning against using too much variables in the model Hocking[70] shows the following. Suppose $y=X_1\beta_1+X_2\beta_2+\epsilon$ is the true model, where $(X_1,X_2)=X$ and $\beta'=(\beta_1',\beta_2')$. Let $\hat{y}=x'b$ and $\hat{y}_1=x_1'b_1$ be the predictions of y at a point x using the full model or the truncated model, respectively, where $x=(x_1 , x_2)$ and b, b_1 are OLS estimates of β, β_1. Then var($\hat{y}$)≥var($\hat{y}_1$). That is, even if β_2 is not equal to zero the future response can be predicted with less variability if the truncated model is used. The penalty is in the bias, because $\hat{y}_1$ is not predicted with the right model. It can be shown that, if β_2 is small enough, then var($\hat{y}$)=MSE($\hat{y}$)≥MSE($\hat{y}_1$).

If the validation of a linear model, from the perspective of these MSE criteria, is wanted, it is necessary to have good estimators of these criteria. These estimators have attained much attention in the statistical literature (Hocking[70], Breiman and Freedman[72], Bunke and Droge[74,75], Allen[76], Picard[77], Berk[78], Mallows[79], and Amemiya[73]). The statistical properties of these estimators are sometimes hard (or not!) to obtain (Steerneman[32]).

Three estimators of the MSE criteria will be outlined: Allen's prediction sum of squares (PRESS), Amemiya's prediction criterion (PRC) and Mallows' C_q. The PRC of Amemiya[73] (see also Judge *et al.*[64]) is derived under the assumption that x_0 is regarded as a random vector that satisfies the condition $E(x_0x_0')=(1/n)X'X$. Under this assumption (I.36) becomes

$$\mathrm{PRC} =s_q^2(1 + q/n) \qquad (I.37)$$

where s_q^2 is the usual estimate of σ^2 in the restricted model and q is the number of predictor variables in the model (including the constant). This criterion is suited for the comparison of the predictive performance of linear models where a decision has to be made which variables to retain in the model. The model that generates the lowest PRC is desired. Often a peak-phenomenon occurs which is visualized in Figure I.9. The PRC can only be used to compare models where the parameters are estimated using the ordinary least squares (OLS) method. Generalizations to PLS, ridge and Stein esimation techniques (see Sections 3.2, 3.3 and 3.4) are not readily available.

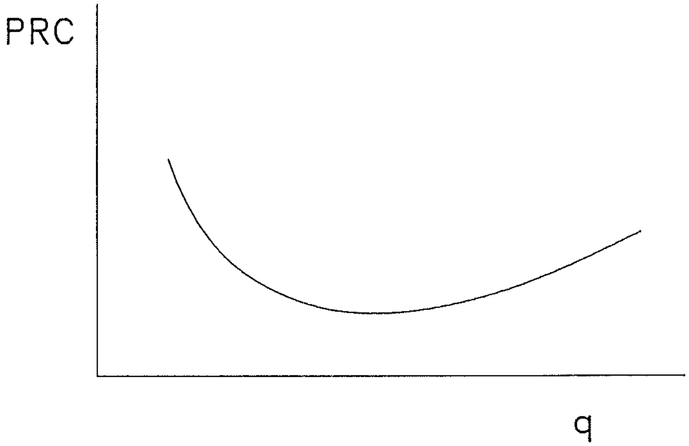


Figure I.9. Peak-phenomenon, q is the number of variables in the model (model complexity), for PRC see text.

The C_q criterion of Mallows is defined as

$$C_q = \frac{RSS_q}{s^2} + 2q - n \qquad (I.38)$$

where s^2 is the residual mean square for the full model, if all variables are used (the constant is understood as a variable), q is the number of variables (including the constant) used in the model whose performance is assessed and RSS_q is the residual sum of squares of this q-term model. A disadvantage of C_q appears immediately: the full model must be known. The justification[70] of C_q is that it is an estimate of the standardized total mean squared error of estimation for the current data $\Gamma_q=(1/\sigma^2)\Sigma\ MSE(\hat{y}_i)$ where the summation index i runs from 1 to n; $\hat{y}_i$ is the prediction of y in data point i with the q term model; σ^2 is the population variance of ϵ. Note that $\Gamma_q=(1/\sigma^2)MSEP(b_q,X)$, where the subscript q refers to the q-term model. Extensions of the C_q criterion for use in ridge regression are reported by Mallows[79].

The prediction error sum of squares (PRESS) is defined as

$$\text{PRESS} = \sum_{i} (y_i - \hat{y}_{(i)})^2 \qquad (\text{I.39})$$

where $\hat{y}_{(i)}$ represents the predicted value of the *i*th observation with the use of a model where the *i*th observation on (y,x) is omitted. This is done n times (for each object). The PRESS value measures the predictive performance of this model, it can be conceived as a sample estimate of MSEP(b,X), formula (I.34) (Golub[80]). The model which produces the lowest PRESS is preferred. Usually the peak-phenomenon, already discussed when introducing the PRC, is observed when plotting PRESS against the complexity of the model, ie. the number of terms in the model. The advantage of the PRESS criterion is in its flexibility to validate a range of models, whose parameters are estimated with a range of estimation methods. OLS estimates can thus be compared with RR, JSR and PLS estimates. Examples will be given in Chapters 10 and 11.

The idea behind PRESS is called cross-validation and stems from a class of related validation methods, which are based on a special use of the sample. These methods are the bootstrap, the jack-knife, and cross-validation. An overview is given by Efron and Gong[81]. The idea of the bootstrap method is to understand the sample (size n) as a population (size n). From this population random samples, say 1000, of size N are drawn and the parameters of interest are calculated. The moments of the parameters can be calculated from these repetitions and are then considered as bootstrap-estimates of the real moments. The jack-knife is closely related to the bootstrap and consists of leaving successively out one observation and calculating the parameters of interest of the model on the basis of the (n-1) remaining observations. This is repeated n times and in this way jack-knife estimates of the moments of the parameters are obtained.

Cross-validation (CV) is a very old method and is aimed at validating prediction rules (or predictive models in a more restricted sense). Reference should be made to Stone[82]. The idea of cross-validation is to leave out one observations and to use the predictive model to predict this observation. Similarly, this can be done for more than one observation. When each observation is omitted once, a value like PRESS can be used to assess the predictive performance of the model. Cross-validation can be used in the context of regression models[74,77] or PLS modeling[14,42]. Applications in the field of PCA are also reported[12,13]. One of the choices to be made in cross-validation is the number of observations to be omitted simultaneously. The validation results can differ depending on this choice (Osten[14]). When one observation is omitted, as advised by Stone[82], the decrease of the number of degrees of freedom in estimation of the model parameters, is minimal. This corresponds to an honest assessment of the predictive performance. A disadvantage of this leave-one-

-out (LOO) method is the heavy computational burden. A more fundamental criticism is stated by S.Wold *et al.*[83] by arguing that the LOO method opens the way for overfitting, which is avoided by leaving more observations out simultaneously.

Stone[82] advocates the use of a double-cross-validation, which consists of two imbedded LOO-methods: one for the choice of the predictor and one for the assessment of that choice. An extension of cross-validation is described by Golub *et al.*[80] and is called generalized cross-validation (GCV). This method is proposed for the choice of the k parameter in ridge regression (see Section 3.2), but can be used in a much broader sense. Its justification stems from the fact that in the extreme case when all entries of X are 0 except for the diagonal elements, the PRESS values as a function of the k parameter do not have a unique minimizer. The GCV solves this problem and can be viewed as a form of ordinary cross-validation invariant under rotations of the measurement coordinate system.

A different way to validate the predictive performance of models is to hold out a part of the data set (test set) and use predictions made for the test set observations for assessment. Note the resemblance of this idea and cross-validation, in fact CV can be viewed as using the training set also as a test set.

A very convenient and intuitively appealing advantage of cross-validation is the possibility it offers to validate a complex prediction rule. In our case the prediction rule consist of a few steps: the choice of the markers, the choice of a prediction method, the choice of k and c parameters for ridge and Stein regression (see Sections 3.2 and 3.3), and is complex indeed. Cross-validation makes it possible not only to validate the whole procedure, but also its parts. If the cross-validatory results are compared with the results of the "real" predictions in the test set, the behaviour of cross-validation in this particular prediction situation can be established. This is one of the goals of this thesis.

2.4 Multicollinearity

The term multicollinearity has been used already in the preceding sections. Multicollinearity is defined as the existence of (near) linear dependence between two or more columns of X. Most textbooks in the field of econometrics treat this subject, e.g. Judge *et al.*[64], Johnston[63]. A detailed treatment is given in Belsley *et al.*[66] and Vinod[60]. The effects of multicollinearity can be very serious. This can be illustrated by noting that

$$\mathrm{MSE}(b_{ols}) = \mathrm{tr}[V(b_{ols})] = \Sigma\ \mathrm{var}(b_{ols,i}) = \sigma^2\Sigma\lambda_i^{-1} \qquad (\mathrm{I}.40)$$

where λ_i is the *i*th eigenvalue of X'X (which is the correlation matrix of *x*, if scaling to column-length-one is adopted). In terms of eigenvalues, multicollinearity means that there exists at least one eigenvalue with a near zero value. Consequently, one or more of the variances of the regression coefficients may become very high. This can even result in regression coefficients with the wrong sign. The influence of multicollinearity on the variances of the regression coefficients can be understood by applying the VIF's (formula I.27). If one column in X can be explained by one or more of the others, then the R^2 becomes 1 and the VIF of the estimated coefficient of that specific x variable becomes very high. Marquardt[84] suggests as a rule of thumb that serious multicollinearity exists if the VIF of a coefficient is greater than five (the phrase "variance inflation factor" is now understood as the variance of a coefficient inflated by multicollinearity).

Furthermore, it can be shown that the regression coefficients become unstable towards slight distortions of the data, which is not desirable[66]. The same goes for t-values, commonly used to judge the significance of a regression coefficient[60].

The influence of multicollinearity on the predictive power of the model is not easy to evaluate. The following argument is valid, however (Judge *et al.*[64]). The distribution of b_{ols} is given by

$$b_{ols} \sim N_m\,(\beta,\ \sigma^2(X'X)^{-1}) \qquad (I.41)$$

under the assumption that ϵ has a normal distribution. The prediction of a new value y is done with the predictor $x_t'b_{ols}=\hat{y}$, where the x_t vector consists of the realization of the m variate x variable which is used for the prediction. The variance of this predictor is given by:

$$\mathrm{var}(\hat{y}) = \mathrm{var}(x_t'b_{ols}) = x_t'(\sigma^2(X'X)^{-1})x_t \qquad (I.42)$$

Making use of the eigenvalue decomposition of S=X'X (assuming that X is scaled such that S has the correlation form), which can be written as S=PDP' (see Section 1.2), formula (I.42) can be rewritten as $\sigma^2x_t'(PD^{-1}P')x_t = \sigma^2x_t'(\Sigma\lambda_i^{-1}p_ip_i')x_t$ because $(X'X)^{-1}=(PDP')^{-1}=$ $(P')^{-1}D^{-1}P^{-1} = PD^{-1}P'$ using $P'=P^{-1}=I$. Assume that multicollinearity exists and $\lambda_m\approx0$. Then $Xp_m\approx0$ (see Section 1.1).

What is the effect on the variance of $\hat{y}$? The problem arises in the last term of the summation in the rewritten form of (I.42) as shown in (I.43),

$$\mathrm{var}(x_t'b_{ols})=\sigma^2(\ldots\ldots.+(1/\lambda_m)(x_t'p_m)(p_m'x_t)) \qquad (I.43)$$

The λ_m value is (near) zero and blows up the variance of $\hat{y}$. Each row x_j in X satisfies $x_j'p_m\approx0$ because $Xp_m\approx0$. When x_t' has a similar structure as x_j' then $x_t'p_m\approx0$ and the effect of a near zero value of λ_m in the last term of the summation in I.43 is cancelled out. Hence the variance of $\hat{y}$ is not necessarily blown up by multicollinearity as long as the values of the explanatory variables for which predictions are desired satisfy the same near-exact linear dependence as the original design matrix X. Note that when σ^2 is small than the effect of multicollinearity is also diminished. A high signal-to-noise ratio, defined as $E(y)/\sigma^2$ where the expectation is over x and y, decreases the effect multicollinearity has on the prediction of y. A remark made by Hoerl *et al.*[85] is relevant in this context. They mention that $\text{MSEP}(b_{ols},X)$, see Section 2.3 formula (I.34), has a $\chi_m{}^2$ distribution regardless the level of multicollinearity. OLS, therefore, can be used at high levels of multicollinearity if prediction at the original design points is the primary concern.

This is only a part of the problem. A difficulty that obscures the reasoning above is the problem of misspecification. A model $y=X\beta+\epsilon$ is assumed but this might not be the true model, it may e.g. contain x-variables which are (in terms of the population) not influential. Misspecification is another problem encountered in regression analysis and has attained much attention[86]. One of the solutions of this problem is performing variable selection on the basis of the sample data, which is difficult in case of multicollinearity[31,66,70]. If the true model $y=X\beta+\epsilon$ is not known and a model $y=X\delta+\epsilon$ is postulated, then d_{ols} is the least squares estimator of δ and $x_0'd_{ols}$ the linear predictor of y_0. But then

$$\begin{aligned}\text{MSEP}(d_{ols},x_0) &= E(\mu_{y0} - x_0'd_{ols})^2 = E(x_0'\beta - x_0'd_{ols})^2 \\ &\text{var}(x_0'd_{ols}) + E(x_0'\beta - x_0'\delta_{ols})'(x_0'\beta - x_0'\delta_{ols}) \qquad (\text{I}.44)\end{aligned}$$

The term $E(x_0'\beta - x_0'\delta_{ols})'(x_0'\beta - x_0'\delta_{ols})$ can no longer be neglected, contrary to the case of predicting y_0 with $x_0'b_{ols}$, where b_{ols} is the OLS estimator of β, because then $E(x_0'\beta_{ols}-\mu_{y0})=0$. The influence of multicollinearity and misspecification on predictive performance is difficult to assess.

The detection of multicollinearity is already mentioned briefly in Section 2.2. VIF's can be used, but have the disadvantage of being the result of unstable calculations when multicollinearity is severe. An overall measure of the degree of multicollinearity is given by the condition index K of X which is defined as[66]

$$K = \frac{\lambda_{max}^{\frac{1}{2}}}{\lambda_{min}^{\frac{1}{2}}} \geq 1 \qquad (\text{I}.45)$$

where $\lambda_{max}^{\frac{1}{2}}$ and $\lambda_{min}^{\frac{1}{2}}$ are the highest and lowest singular values of X, respectively. The condition number of any matrix A with orthonormal columns (A'A = I) is unity. In case of multicollinearity the $\lambda_{min}^{\frac{1}{2}}$ becomes near zero and so K will increase. In Section 2.2 the variance-decomposition proportions were discussed and the phrase "small singular values" was used there. Now a yardstick is obtained against which smallness can be measured. Define the *k*th condition index as

$$\mu_k = \frac{\lambda^{\frac{1}{2}}_{max}}{\lambda^{\frac{1}{2}}_k} \qquad k = 1,\ldots,m \qquad (I.46)$$

with $\lambda^{\frac{1}{2}}_k$ the *k*th singular value of X. A singular value that is small relative to $\lambda^{\frac{1}{2}}_{max}$ has a high condition index. The variance-decomposition table (Π-matrix) can be augmented by a column which gives the condition index associated with each row. Relationships between variables, indicated by their variance-decomposition proportions, are more severe when associated with a high condition index. Unfortunately the condition number and the condition indices depend on the kind of scaling. To make comparisons between condition indices meaningful, it is wise to scale each column to length one. Whether a condition index is large is a matter of empirical determination. Belsley *et al.*[66] argue that condition indices around 5 or 10 indicate weak dependence whereas strong dependence is indicated by values of 30 to 100.

A distinction can be made between predictive and non-predictive multicollinearity (Vinod[60], Jolliffe[3], Mason[67], Mager[87]). If the latent root SVD of (X,y) is performed, the dimensions associated with low singular values are of special interest because these low values indicate linear dependence. If the loading of y on a singular vector associated with a low singular value is small, this singular vector reveals a "non-predictive multicollinearity". The removal of an x variable strongly associated with this singular vector might be appropriate, following the same lines of reasoning as in Section 1.2. In latent root regression this singular vector is removed from the model (Mason[67]). If, on the contrary, the loading of y is high on a singular vector associated with a near-zero singular value and two or more x variables load also on this singular value then there exists a predictive multicollinearity. Those variables that load on this singular vector are able to predict y and cannot be deleted without consequences for the prediction. Some critical remarks are appropriate. The "multicollinear-structure" (non-predictive and predictive multicollinearity) as evident in the training set does not necessarily resemble the relationship between the predictor variables and y for future observations. When the population is well defined and the training set is a representative sample then the above mentioned criticism is not completely valid. Another warning against the use of latent root SVD must be made. If an x variable, seemingly unrelated to the other x variables, is deleted and a latent root SVD of the new augmented design matrix is performed, new "multicollinearity patterns" emerge. These patterns are totally different from the patterns observed in the original design matrix. Examples will be given in Chapter 10.

Remedies against multicollinearity can be divided in three parts. The first kind of remedy is the inclusion of new observations in the training set. Silvey[88] and Judge *et al.*[64] show how (if this is possible) the values of the predictor variables should be chosen to obtain one optimal new observation. If it is not possible to control the values of the predictor variables, random observations should be taken. The second remedy is already discussed partly: the exclusion of "troublesome" variables. Of special interest are the variables which are involved in non-predictive multicollinearity. This should be done carefully, however, keeping in mind the remarks made on the representativeness of the training set. The remarks in Sections 1.3 and 1.5 regarding data-mining are valid here. A closely related problem to data-mining is overfitting. If variables are included in the regression model which are not relevant in the prediction of y (the true model is generally unknown) they add only noise and should be removed. But how can one be sure which variables are not relevant? A dilemma shows up because retaining all variables gives an overfit, which can mask the influence of really important variables, and on the other hand the discarding of variables makes the problem of data-mining manifest. So variable selection should be done with extreme care and with the use of all possible prior information. Cross-validation may help to validate the step of excluding variables. The third remedy against multicollinearity is the use of estimation techniques that differ from ordinary least squares and are designed to handle multicollinearity, e.g. ridge regression and partial least squares. These will be discussed separately in Chapter 3.

Chapter 3 Biased estimation

3.1 Background

If the MSE criteria are used to choose estimation methods, a much broader scala of methods becomes attractive. Throughout Chapter 3 the linear model $y=X\beta+\epsilon$ is assumed, with y centered and X (n×m) column-centered and of full rank in such a way that X'X=R, the correlation matrix of *x*. The OLS estimate of β, which is $b_{ols}=(X'X)^{-1}X'y$, is an unbiased estimate of β. This means that $E(b_{ols})=\beta$. It can be shown[89] that b_{ols} minimizes MSE(b) within the class of unbiased estimates b of β. If the restriction of unbiasedness is omitted, estimators can be found with a lower MSE than the MSE of the OLS estimate but with a bias. Estimators which do not share the property of unbiasedness are called biased estimators.

Generally speaking, the MSE(b)=tr[V(b)]+[bias(b)]'[bias(b)]. If b is the OLS estimator bias(b)=0, so that $MSE(b_{ols})=tr[V(b_{ols})]$. The advantage of the biased estimators is, therefore, in the reduction of the variance of the estimator. If the bias becomes too large, the biased estimator is not advantageous any more.

Note the correspondence of tr[V(b)] and bias(b) with the ideas of "precision" and "accuracy" in analytical chemistry, respectively. In Figure I.10 the idea of biased estimation is shown: an accurate but imprecise estimator (Fig. I.10a) is exchanged for a precise but inaccurate estimator (Fig. I.10b), resulting in an estimator with a lower MSE. Several estimation methods (and corresponding estimators) have been developed, three of these are discussed explicitly below. Reference should be made to Vinod[60], Mayer and Willke[90], Hocking[70], Judge and Bock[89], and Marquardt[84].

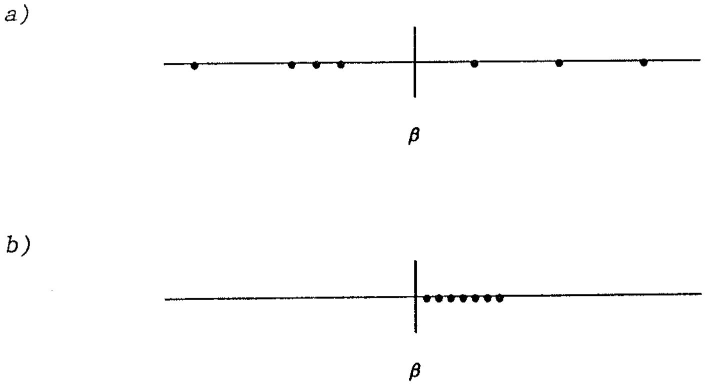


Figure I.10. Illustration of biased estimation; a single β is estimated accurate but imprecise (a) or precise but inaccurate (b).

In case of multicollinearity, biased estimation may present a special advantage. To see this note that $MSE(b_{ols})=\sigma^2\Sigma\lambda_i^{-1}$ (see (I.40)) and in case of multicollinearity one (or several) of the λ_i's become near zero. The MSE becomes very large and it may be very profitable to use an estimator b with lower variance than b_{ols} and with an inevitably larger bias.

3.2 Ridge regression

The idea of ridge regression is due to Hoerl and Kennard[91,92]. The ridge estimator is defined as

$$b_{rr}=(X'X + kI)^{-1}X'y \tag{I.47}$$

where $k\geq0$ (if k=0 then $b_{rr}=b_{ols}$). Originally[93], Riley[94] proposed a method for solving b in Ab=d, where A is a square ill-conditioned matrix and d is a vector. A is expanded in terms of F, where F=(A+kI) for some k. Then $b=A^{-1}d=F^{-1}d+kF^{-2}d+\ldots$ Note the resemblance between this idea and problem at hand: solving b in (X'X)b=X'y, where X'X is ill-conditioned caused by multicollinearity. The ridge solution is the first term in the series expansion and is specifically designed to handle the problem of multicollinearity.

The justification of the ridge estimator can be shown by the following [60,91,92]. Let b be an arbitrary estimator of β in the linear model $y=X\beta+\epsilon$. Then

$$\begin{aligned}SSE(b) = (y-Xb)'(y-Xb) &= (y-Xb_{ols})'(y-Xb_{ols}) + \\ &\quad (b-b_{ols})'X'X(b-b_{ols}) \\ &= SSE(b_{ols}) + \Phi(b)\end{aligned} \tag{I.48}$$

where $\Phi(b)=(b-b_{ols})'X'X(b-b_{ols})$ and $\Phi(b)$ can be assumed a function of b because b_{ols} is fixed (given X and y). Now X'X is a positive definite matrix (assuming X has full rank) and therefore $\Phi(b)\geq0$, with $\Phi(b)=0$ if and only if $b=b_{ols}$, then SSE(b) attains its minimum value. From (I.33) follows

$$\begin{aligned}MSE(b_{ols}) = E(b_{ols}-\beta)'(b_{ols}-\beta) &= E(b_{ols}'b_{ols}-\beta'b_{ols}+ \\ &+ \beta'\beta-b_{ols}'b) = -\beta'\beta + E(b_{ols}'b_{ols})\end{aligned}$$

but

$$MSE(b_{ols}) = \sigma^2\Sigma\lambda_i^{-1}$$

so

$$E(b_{ols}'b_{ols}) = \beta'\beta + \sigma^2\Sigma\lambda_i^{-1} \tag{I.49}$$

In words (I.49) means that, especially in case of multicollinearity, the length of b_{ols} overestimates the true length $(\beta'\beta)^{\frac{1}{2}}$ of β. It can be shown[91] that b_{rr} is the solution of the following problem:

$$\text{minimize } b'b \text{ with respect to } b \qquad \text{(I.50)}$$

$$\text{subject to } \Phi(b)=C$$

where C is a positive constant. The idea is to allow a small increase C on SSE and choose that vector b which is the shortest under that condition. This seems reasonable when observing (I.49). The MSE of b_{rr} is[60]

$$\begin{aligned} MSE(b_{rr}) &= \sigma^2\Sigma\delta_i^2\lambda_i^{-1} + \Sigma(\delta_i-1)^2\gamma_i^2 = \\ &= tr[V(b_{rr})] + [bias(b_{rr})]'[bias(b_{rr})] \end{aligned} \qquad \text{(I.51)}$$

with $\delta_i=\lambda_i/(\lambda_i+k)$ and $\gamma_i=(P'b)_i$, the *i*th element of the vector P'b. The matrix P and the numbers λ_i stem from the SVD of X: $X=TD^{\frac{1}{2}}P$, the λ_i's are the squares of the singular values. By comparing $MSE(b_{rr})$ with $MSE(b_{ols})=\sigma^2\Sigma\lambda_i^{-1}$, it is clear that large values of $1/\lambda_i$ are muted (because $\delta_i\leq1$) at the cost of bias and $tr(V(b_{ols}))\geq tr(V(b_{rr}))$.

Using the SVD of X, the linear model becomes

$$y = Xb+e = TD^{\frac{1}{2}}P'b + e = TD^{\frac{1}{2}}c + e \qquad \text{(I.52)}$$

with P'b=c. This reparametrization transforms the linear model in another linear model with uncorrelated components of b, because $(TD^{\frac{1}{2}})'(TD^{\frac{1}{2}}) = D^{\frac{1}{2}}T'TD^{\frac{1}{2}} = D$. The least squares estimate $c_{ols}=D^{-1}D^{\frac{1}{2}}T'y =D^{-\frac{1}{2}}T'y$. Obviously, $b_{ols}=Pc_{ols}$. The variance-covariance matrix of c_{ols} is $V(c_{ols})=\sigma^2D^{-1}$. The ridge estimator is $c_{rr}=(D+kI)^{-1}D^{\frac{1}{2}}T'y$. Again simple calculations show that $b_{rr}=Pc_{rr}$. Define Q as $(D+kI)^{-1}D = diag(\delta_1,\ldots,\delta_m)$ with δ_i as before. Then $c_{rr}=(D+kI)^{-1}DD^{-\frac{1}{2}}T'y=Qc_{ols}$. Written out fully this becomes

$$c_{rr} = (\delta_1c_1,\ldots,\delta_mc_m)' \qquad \text{(I.53)}$$

with $\delta_i=\lambda_i/(\lambda_i+k)$ as before and $c_{ols}=(c_1,\ldots,c_m)'$. If k>0, the c_{rr} is an estimator that shrinks c_{ols}, the estimator of the uncorrelated components of b, towards zero (δ_i<1). The components of c_{ols} which carry the most variance are the ones with low λ_i values, see $V(c_{ols})$. These components are shrunken mostly, because the associated δ_i values are the lowest. The length of the c_{rr} estimator, the square root of $\Sigma(\delta_ic_i)^2$, is of course smaller than the length of c_{ols} (the square root of Σc_i^2). The length of the b_{rr} vector is smaller than the length of b_{ols} because $b_{rr}'b_{rr}=c_{rr}'P'Pc_{rr}=c_{rr}'c_{rr}$. Thus an estimator is obtained with smaller length and smaller variance.

A generalization of the ridge estimator is given by

$$c_{rrg} = (D+K)^{-1}D^{\frac{1}{2}}T'y \tag{I.54}$$

where K is diag$(k_1,..,k_m)$ with $k_j \geq 0$ and distinct.

An important issue when using ridge regression (RR) is the choice of k. Hoerl and Kennard[91] proved that for every (β,σ^2) there exists a $k>0$ so that $\mathrm{MSE}(b_{rr})<\mathrm{MSE}(b_{ols})$, where b_{rr} is the ridge estimator with that k value. But $\mathrm{MSE}(b_{rr})$ depends on β and σ^2 which are unknown and have to be estimated, so that k cannot be chosen to minimize $\mathrm{MSE}(b_{rr})$. There are several possibilities to chose k[70,84,91,92]. Hoerl and Kennard[91] recommend the choice of $k=ms^2/b'_{ols}b_{ols}$ with $s^2=\mathrm{SSE}/(n-m)$ because $k=m\sigma^2/\beta'\beta$ minimizes $\mathrm{MSE}(b_{rr})$ if $X'X=I$. A very valuable feature of ridge regression is the ridge trace. This trace is a plot of the b_{rr}-estimates against the ridge parameter k. Large decrements in the absolute values of particular estimated coefficients indicate unstable estimation of these coefficients[70,91,92]. This trace plot allows one to chose a k-value that "just stabilizes all estimates". Another criterion to choose the ridge parameter k comes from Marquardt and Snee[95]. The VIF's, already discussed in Section 2.2, can be defined for the ridge estimator too. The variance inflation factors of b_{rr} are the diagonal elements of[70]

$$V(b_{rr})/\sigma^2 = (X'X + kI)^{-1}X'X(X'X + kI)^{-1} \tag{I.55}$$

The recommendation is to chose k in such a way that the VIF's are between one and ten and closer to one. Note that, assuming k fixed, $s\sqrt{\mathrm{VIF}_i}$ gives an estimate of the standard deviation of the ridge estimation of the *i*th coefficient (s is the square root of the estimated error variance). These standard deviations can be used in a modified t-test on the significance of the associated coefficient[85]. The fourth way to chose k is described by S.Wold *et al.*[83] and Golub *et al.*[80]. Cross-validation is used to take that k-value which has lowest $\mathrm{PRESS}(k)=\Sigma(y_i-\hat{y}_{(i)})^2$, where $\hat{y}_{(i)}$ is the ridge (with ridge parameter k) prediction of y_i without the use of y_i in calculating the model. S.Wold *et al.*[83] use a "leave-more-out" procedure and Golub *et al.*[80] use Generalized CV (see Section 2.3). The cross-validation with a leave-one-out evaluation instead of leave-more-out is adopted, because the predictive performance of the procedure is assessed at best when the maximum number of degrees of freedom is attained[82] (see Section 2.3). Cross-validation (CV) instead of GCV is adopted, because of the ability of CV to validate the whole predictive procedure (see Section 2.4). Moreover, the drawback of calculating a PRESS value with CV (see Section 2.3) does not exist in this situation because of multicollinearity.

A fifth method of choosing k is given by McDonald and Galarneau[96]. They choose k in such a way that $b'_{rr}b_{rr}=b'_{ols}b_{ols}-s^2\Sigma\lambda_i^{-1}$, where b_{rr} is the ridge estimator and $\lambda_1\geq...\geq\lambda_m$ are the eigenvalues of X'X. This procedure tries to "give b_{rr} the right length".

There have been performed a number of simulation studies regarding ridge regression (Hoerl *et al.*[85], McDonald and Galarneau[96], Dempster *et al.*[97]) Hoerl *et al.*[85] conclude that subset selection is not the best way to combat multicollinearity if prediction is the purpose and if most variables are related to some degree with the response. Better predictions are obtained if ridge regression is used. They also conclude that choosing $k=ms^2/b'_{ols}b_{ols}$ is reasonable and suggest that the choice of k according to McDonald and Galarneau[96] is particularly good with high signal-to-noise ratios. McDonald and Galarneau[96] state their results in terms of the MSE of the estimated coefficients, see Section 2.3 formula (I.33). They conclude that the performance of the evaluated ridge type estimators depends on the variance of the random error, the correlations among the explanatory variables and the unknown coefficient vector. To be more specific, they show that ridge estimation becomes less profitable with smaller error variances and lower correlations between the x variables. They state that there is no rule for choosing k which assures that the corresponding ridge estimator is better than least squares.

Berk[78] has validated some biased estimators and subset selection techniques with the use of real data sets. He splits the data sets in two parts: the training set and the validation set. As a yardstick of performance he uses the MSE of prediction in the validation set. If $k=ms^2/b'_{ols}b_{ols}$ is used, he concludes that ridge estimation performs well if the correlations between dependent and independent variables are all positive. Otherwise subset selection might be the best choice, especially if one or a few dominant predictors can be identified.

A nice view on the geometry of ridge regression is given by Swindel[98].

3.3 James-Stein Regression (JS)

Consider again the linear model $y=X\beta+\epsilon$, with X of full rank and in such a way that X'X is in correlation form. James and Stein[99] have shown that the estimator

$$b_{js} = (1- (\frac{c_1}{n-m})(\frac{e'e}{b'_{ols}X'Xb_{ols}}))b_{ols} = c_2 b_{ols} \qquad (I.56)$$

where c_1 is a positive constant ($0 \leq c_2 \leq 1$) and e is the estimated residual vector from OLS, has a better performance than b_{ols} in terms of $MSEP(b,X)/\sigma^2$, where $var(\epsilon)=\sigma^2$, for an appropriate choice c_1. It may be profitable to use the James-Stein estimator in case where prediction is the purpose. Reference should be made to Judge and Bock[89], Hocking[70], Draper and Van Nostrand[93], Jennrich and Oman[100]. A detailed overview is given by Vinod[60].

The James-Stein estimator shrinks the OLS estimator towards zero (note the resemblance with the ridge estimator). The James-Stein estimator is the solution of the following minimization problem:

$$\text{minimize } (b-b_{ols})'(b-b_{ols}) \text{ with respect to } b \qquad (I.57)$$

$$\text{subject to } b'b \le \delta^2$$

In words, the minimization problem comes down to the choice of a shorter b-vector while staying close to the OLS estimator (note again the resemblance with the ridge solution, particularly (I.50)). Stein[70] suggests to use c_2 as

$$c_2 = \max\ [(1-((m-2)/(n-m+2))(1-R^2)/R^2),0] \qquad (I.58)$$

where R^2 is the multiple correlation coefficient obtained from the OLS regression of y on X. Necessary and sufficient conditions[60,89] for the improvement of the James-Stein estimator over OLS, for all β, are $m \ge 3$; $n \ge m+2$ and $d = \lambda_m \Sigma \lambda_i^{-1} > 2$, where the summation runs from 1 to m. High multicollinearity can destroy the improvement of the JS estimator over the OLS estimator because then $d<2$. This does not imply, however, that there is no β for which improvement is possible. The derivation of the MSE of the JS estimator is tedious[60,90] to perform.

In the study of Berk[78], James-Stein estimation is shown to yield better predictions than OLS, except in one case were the validation set does not resemble the original data. Jennrich and Oman[100] give some hints about situations were James-Stein regression is appropriate.

In this thesis the value of c_2 is established by cross-validation in the same way as for the ridge parameter k (see Section 3.2).

3.4 Partial least squares (PLS)

Originally, the idea of partial least squares (PLS) stems from H.Wold. Descriptions of the original PLS modelling strategies are given by Jöreskog and H.Wold[19]. Extensive details on properties of the PLS estimators are given by Dijkstra[101,102]. The PLS estimation method is introduced by means of the algorithms which are used. These algorithms are described by Geladi and Kowalski[103,104], S.Wold *et al.*[42,83], Höskuldsson[105], and Manne[106]. A slightly different algorithm is used by Martens[107], but numerically the same results are obtained (in terms of predictions).

In order to get a good understanding of PLS the NIPALS algorithm is discussed firstly. This algorithm has been developed by H.Wold, and is used to calculate the principal components of X.

NIPALS:

```
1. initial guess of $t$, e.g. the column in $X$ with the largest variance
2. $p' = t'X/t't$
3. normalize $p$ to $\|p\| = 1$
4. $t = Xp$
5. check convergence: if $d = \|t_{new}-t_{old}\|$ is small then 6, else 2
6. residuals: $E = X-tp'$; use $E$ as $X$ in the next dimension
```

Step 2 as well as step 4 in the NIPALS algorithm can be recognized as least squares steps. The first t and p vectors in NIPALS are calculated in such a way that $X-tp'$ is minimized in least squares sense. The minimization of $X-tp'$ and normalization of p (step 3) provides that t is the first principal component of $X'X$, with associated loadings p (Gabriel and Zamir[108]). In the next dimension t_2 is the first principal component of $E'E$ and, therefore, the second principal component of $X'X$. If g dimensions are calculated successively, the matrices $T=(t_1,..,t_g)$ and $P=(p_1,..,p_g)$ are obtained and $X=TP'+E$. The matrix TP' has rank g and is therefore a lower rank approximation of X, TP' is the spectral decomposition of X truncated to g terms (see Section 1.2). NIPALS is recognized as a member of a family of algorithms which yield lower rank approximations of matrices by least squares methods[11,108].

This algorithm has the advantage that not the whole spectral decomposition of X has to be performed if only the first few principal components are of interest. It can be proven[105] that $t_i't_j=0$ and $p_i'p_j=0$, for i not equal j as should be.

The PLS algorithms are closely related to the NIPALS algorithm. Two versions of PLS are discussed. The PLS1 algorithm comprises one y variable and PLS2 comprises more y variables simultaneously.

PLS1:

```
1. $w' = y'X/y'y$
2. normalize $w$ to $\|w\| = 1$
3. $t = Xw$
4. $q = t'y/t't$
5. $p' = t'X/t't$
6. residuals: $E = X-tp'$; $f = y-tq$.
7. use $E$ and $f$ as $X$ and $y$ in next dimension
```

Step 1 in the PLS1 algorithm is recognized as the result of regressing X on y. To see this consider the linear equations $x_s=y\beta_s+\epsilon$, with $s=1,...,m$. Least squares theory gives estimates $b_s= (y'y)^{-1}y'x_s$. Then $w'=(b_1,...,b_m)'$. If y and X are column-autoscaled, w' is simply the vector of correlation coefficients between y and each x variable. The linear combination of the columns of X with weight factors w' produces t: the estimated latent variable of the X block (step 3).

The dependent variable y is then regressed, in step 4, on the estimate of the latent variable of the X block, t, which results in q. Step 5 is performed to make t'E=t'(X-tp')=0. When the matrix E and the vector f are used in the next cycle the result is t_2, w_2, q_2 and p_2. Because t_2 is a linear combination of the columns of E, the condition $t_1'E=0$ results in $t_1't_2=0$, where t_1 is the estimated latent vector of the first cycle. So the estimated latent vectors of the different cycles (dimensions) are orthogonal[105,106]. The number of dimensions (components) can be established by cross-validation, this will be described in more detail in Subsection 10.3.4. Note that the vectors p obtained in the different cycles are not orthogonal, in contrast with the NIPALS algorithm. The vectors w are mutually orthogonal, however[105].

Although not treated explicitly here, coefficients for b in the linear model $y=X\beta+\epsilon$ can be obtained if the dimensionality of the PLS model is established[106,107]. These b values are biased estimates of β whose properties are hard to derive[102]. Predictions for a new object can be obtained as follows:

```
1. y=0 : initialize predictions
2. for k = 1 to g
   2.1 $t_k = x'w_k/w_k'w_k$
   2.2 $y = y + t_k q_k$
   2.3 $e' = x' - t_k p_k'$
   2.4 x = e
   2.5 go to 2
```

In step 2.1, the score of x on the latent vector is calculated with the use of the model parameter w_k. In step 2.2 the contribution of the latent score t_k to y is calculated. The contribution of x to t_k is subtracted from x in step 2.3 and the whole procedure starts again with the e obtained in step 2.3. Note that the successive q values are estimated independent from each other in the PLS1 algorithm, which is possible because the t vectors are mutually orthogonal. This property of the estimated q values is used in the prediction procedure. If a constant is assumed in the linear model $y=X\beta+\epsilon$, y and X are column-mean-centered prior to the PLS calculation (see Section 2.1) and the predictions are initialized at the mean value of y.

The second PLS algorithm will be labeled PLS2 and is designed to handle more than one y variable simultaneously. Let all y variables be gathered in Y (n×r).

PLS2:

```
1. initial guess of u; a column in Y
2. w' = u'X/u'u
3. normalize w to $\|w\| = 1$
4. t = Xw
```

5. $q' = t'Y/t't$
6. $u = Yq/q'q$
7. check convergence on u (see NIPALS). If no convergence step 2
8. $p' = t'X/t't$
9. residuals: $E = X-tp'$; $F = Y-tq'$; use E, F as X, Y in the next dimension

Again the number of dimensions can be established with cross-validation. The t vectors are mutually orthogonal, so are the w vectors. This is not the case for the p vectors. The u vectors are estimates of the latent variables in the Y block. A assessment of the PLS method, besides statistical arguments given by H.Wold[19], Dijkstra[102], stems from the observation that the vectors u_1, t_1, w_1 and q_1 are eigenvectors associated with the maximum eigenvalues of respectively YY'XX', XX'YY', X'YY'X, and Y'XX'Y. The same goes for u_2 to q_2 but then X and Y must be replaced by their residuals E and F and so on[105]. It is, however, hard to realize what e.g. XX'YY' means. Manne[106] shows that PLS1 is equivalent to the bidiagonalization algorithm of Golub and Kahan[109], which is used to invert a (probably singular) matrix.

Predictions at a new value of *x* can be obtained in the same way as in the PLS1 case. When the model dimensionality is established, say g, and $T=(t_1,\dots,t_g)$, $P=(p_1,\dots,p_g)$, $Q=(q_1,\dots,q_g)$ then[42]:

$$X = TP' + E \qquad (I.59)$$
$$Y = TQ' + F$$

The matrices X and Y are decomposed. Note that the decomposition of X given by the NIPALS algorithm yields another decomposition of X than PLS2. The P matrix from NIPALS is orthogonal, whereas the P matrix from PLS2 is not.

3.5 Unification of methods

A unification of the methods partial least squares (PLS), ridge regression (RR), James-Stein Regression (JS), ordinary least squares (OLS), and principal component regression (PCR) can be developed by realizing that each method somehow approximates the X'X matrix (Hocking[70], Naes and Martens[110], Hoerl and Kennard[91], Mason[67], Jolliffe[3]).

In OLS no approximation of X'X is made (X'X is used as such) and the estimator for the linear model $y=X\beta+\epsilon$ is $b_{ols}=(X'X)^{-1}X'y$. Ridge regression approximates the X'X with X'X+kI and therefore the RR estimator becomes $b_{rr}=(X'X+kI)^{-1}X'y$. If the JS estimator (I.50) is transformed into $b_{js}=(1/c_2X'X)^{-1}X'Y$, the approximation of X'X with $(1/c_2)X'X$ is clear. In case of PCR, the $(X'X)^{-1}$ matrix is approximated by the first part of the spectral decomposition of $(X'X)^{-1}$ (see Section 1.2) which is $(1/\lambda_1)p_1p_1'+\dots+(1/\lambda_g)p_gp_g'$ if the first g eigenvalues and eigenvectors are used, corresponding to the g largest eigenvalues of X'X. Especially if rank(X'X)<m, then g can be chosen as rank(X'X) and the inverse of X'X does not exist. The PCR estimator becomes then $b_{pcr}=((1/\lambda_1)p_1p_1'+\ldots+(1/\lambda_g)p_gp_g')X'y$ or $b_{pcr}=P_gD_g^{-1}P_g'X'y$ where the subscript g means that only the first g columns of the associated matrix are used. The matrix $P_gD_g^{-1}P_g'$ is called the generalized inverse of X'X. Marquardt[84] generalizes the idea of the rank g approximation by letting g be a continuous variable $0<g\leq m$ and defining the generalized inverse of X'X subsequently. On overview of the theory of generalized inverses is given by Rao and Mitra[10].

For PLS the derivation of its unified form is a bit more difficult (Naes and Martens[110]). It can be derived if applying the PLS variant with orthogonal P and a single y-variable. Let $U=(u_1,\ldots,u_g)$ be the (m×g) matrix consisting of the restricted (orthogonal) eigenvectors of X'X relative to the space spanned by P, then P=UC where C is an g×g orthogonal matrix. The restricted eigenvectors have the property that $U'X'XU = \mathrm{diag}(\Phi_1,\ldots,\Phi_g)$. The PLS predictor is then $b_{pls} = ((1/\Phi_1)u_1u_1'+\ldots+(1/\Phi_g)u_gu_g')X'y$.

Another estimation method for the linear model which can be utilized in case of multicollinearity is total least squares[111] (TLS). An application in the field of QSAR is given by S.Wold *et al.*[83]. With the use of the pseudo-inverse of X (which is a generalization of the usual inverse to singular and non-square matrices) a class of estimators can be defined which can also handle multicollinearity.

Chapter 4 Experimental design considerations

4.1 The classical situation

The ideas behind experimental design are described very well in textbooks (Box *et al.*[112], Deming and Morgan[113], Davies[114], Box and Draper[115]). One of these ideas is the theory of factorial experiments which results in a design matrix X with orthogonal columns. A factorial experiment can be augmented by some design points to obtain a central composite design. The notion of orthogonality in designs is important for the estimatibility of the effects (coefficients).

A special topic in the field of experimental design is the theory of optimal experimental design (Fedorov[116], Silvey[117]). If a linear model is postulated, $y=X\beta+\epsilon$, the variance-covariance matrix of b_{ols} is $\sigma^2(X'X)^{-1}$ and the purpose can be to choose X so that $\det(X'X)^{-1}$ is minimal (D-optimality) or to choose X so that $tr(X'X)^{-1}$ is minimal (A-optimality). Both A- and D-optimality aim at a regular spread of the design points in the design space, but the criteria are slightly different. It must be stressed that the specification of the model is assumed known throughout (whether it does contain quadratic or interaction terms etc.). The optimised design is, therefore, only optimal for that particular model. Obviously, the idea of optimal design can be broadened by choosing a design that performs reasonably with different (but equally probable) models. Another approach is outlined by Box and Draper[115]: the pay-off in using the wrong model and the variance in the estimation of b_{ols} or the prediction of y can be calculated by using an MSE criterion. Choosing a design which minimizes this MSE gives rise to the best compromise between model (mis)specification and variance.

The ideas of (optimal) experimental design are also introduced in the field of mixture experiments (Weyland[118,119], Debets[120], Snee[121], Cornell[65]). In a mixture experiment, the sum of the scores on the x variables always equals one and special models are developed to handle this situation[65]. A linear model is formulated in the mixture experiments and design considerations are also important in this area. The idea of a regular spread of the design points in the design space are followed almost everywhere (except by Drouen[122]). An application of D-optimality is reported by Bianchini *et al.*[123], for the use in the optimization of a chromatographic separation.

In the case of calibration, treated in this thesis, the problem becomes slightly more complicated. In the initial training set the stationary phases are given entities and the mobile phase compositions can be chosen only to the restricted area where meaningful retention times are obtained. In Figure I.11, this situation is depicted. A factorial arrangement of the experiment requires that each mobile phase composition is used on each stationary phase. The design used in Part III is shown in Figure I.12. In this design the factor "elution-strength" is varied at two levels (in Part IV: three levels) because each modifier is mixed with water in two different proportions. The factor "kind of modifier" is varied at three levels: two binary mixtures each comprising a different organic modifier and a ternary mixture made up of the two binary ones. The two ternary mixtures are incorporated for each stationary phase to measure the influence of mixing the binary ones. The factor "stationary phase" is varied at six levels. Note that the ideas of "regular spread" and "orthogonality" are used in this design.

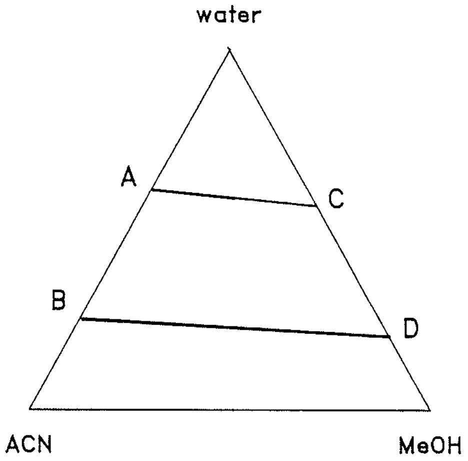


Figure I.11. Factor space of mobile phase compositions. ACN and MeOH are the abbreviations of acetonitrile and methanol, respectively. A to D are the limits of the factor space in which meaningful experiments are to be performed.

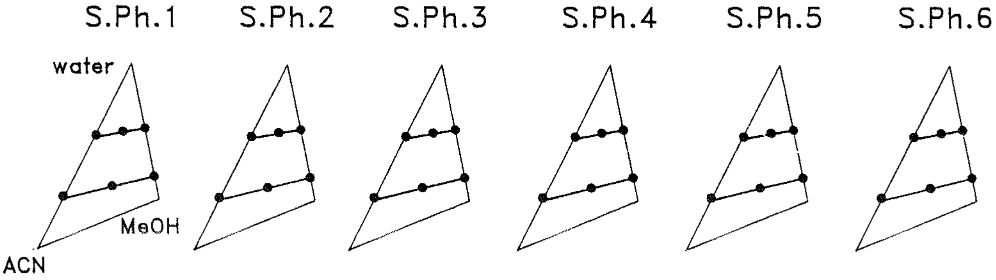


Figure I.12. Orthogonal design.

For practical reasons, the idea of orthogonality must be left sometimes because it is not possible to measure the retention of all solutes at the same mobile phase compositions if the stationary phases differ too much. The resulting design is shown in Figure I.13. The consequences of such a design with regard to the choice of the markers needs further research.

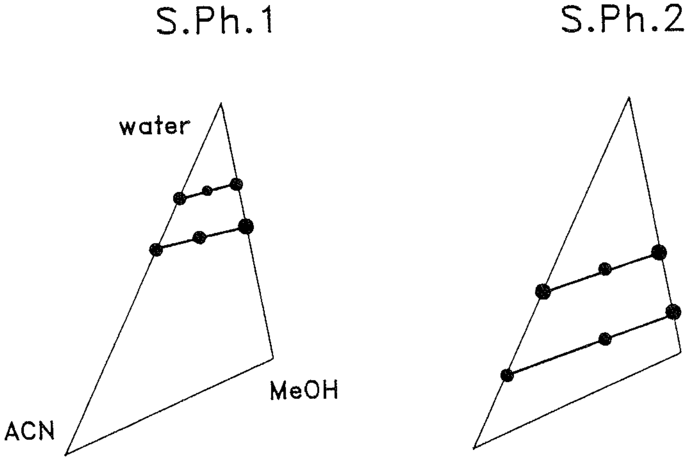


Figure I.13. Non-orthogonal design.

Anticipating the models which are used to calibrate with in Parts III and IV, generally some stationary phases are used in the training set. Markers are chosen with these training data. Then a model is build which relates the markers (and mobile phase composition measured as the fractions of the different modifiers) with the non-markers. The design matrix X associated with one of the models is shown in Figure I.14. The goal of designing experiments is to obtain a well conditioned X matrix (see Section 2.4). The first part of this design matrix is formed by the mobile phase modifier fractions. The second part of X is formed by the markers which are not known beforehand. The most sensible way to choose the mobile phase components is to isolate the mobile phase part from the total X matrix and optimize that part with the known theory.

This can be done on the assumption that the correlation between the mobile phase part and the marker part of X is low compared with the correlations within either the mobile phase- and/or the marker part of X. Calculations, see Part III, show that this is a reasonable assumption. The marker part of X depends on the kind of criterion used for the marker selection. The determinant criterion yields markers with a lower inter-correlation than the induced-variance criterion, as calculations will show (Part III). In terms of the condition of X, it might therefore be profitable to use the determinant criterion, which is also profitable for James-Stein regression (see Section 3.3). But the induced-variance criterion is more suited for our purpose: predicting the discarded solutes. So a conflict arises between both marker choice criteria. A yardstick is needed to validate the performance of both criteria. Cross-validation is such a yardstick and its merit will be assessed.

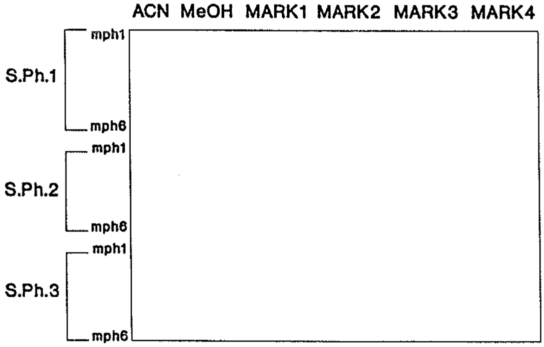


Figure I.14: Design matrix X of a model (see text).

A more sophisticated choice of the stationary phases can be made if they are characterized by physico-chemical measurements. A multivariate design approach as shown by S.Wold[9], can then be used. These measurements are, however, tedious and expensive.

4.2 The influence of the choice of the trainingset

If the stationary phases are chosen without design considerations then a random sample is the result. The result of the estimation process and the validation of that process will depend on the quality of that sample. It can be expected that CV is a good, or reasonable, validation criterion in cases of a representative sample. The robustness of the CV criterion with regard to the representativeness of the training sample must be established. This can be done by measuring the appropriateness of the random sample and confront the CV results with the results obtained in the test set, if such a test set is available.

Evidently, the appropriateness of a training set is difficult to assess. It is possible, however, to measure the difference between training set and test set and draw conclusions on that basis. Univariate measures of differences between training set and test set are the moments of first and second order for each variable, hence means and variances of the variables. A more sophisticated approach, pointed out by Picard and Cook[77] and Snee[124], is called matched split. The consequences of the split of the data in training- and test set is qualitatively checked by

$$V'V/n_v \sim X'X/n \tag{I.60}$$

where X is the matrix of the predictor variables of the total sample and V the matrix of predictor variables in the validation sample, the test set (test sample size n_v). The second order moments are matched, in order to obtain the same dispersion in the training set and the test set (note that $X'X/n$ and $V'V/n_v$ are dispersion matrices). It is important to emphasize that the unscaled (and uncentered) X matrix is used in (I.60). A constant in the models and therefore a column of ones is added to X.

Another measure of difference between the test set and the training set is the ratio of the mean Mahalanobis distances for the two sets, see Berk[78]. The Mahanalobis distance of a point x to the centroid (origin) is

$$ds = x(X'X)^{-1}x \tag{I.61}$$

assuming that X is mean-centered. It can be shown[78] that

$$E(MSEP_v/MSEP) = \frac{(n-m-1)}{(n-m-3)} (1 + 1/n + \frac{mDS}{nDS_0}) \tag{I.62}$$

where $MSEP_v$ is the mean squared error of prediction in the test set (validation set) and MSEP the mean squared error of prediction in the training set (both MSEP and $MSEP_v$ are MSEP(b,X) values, see formula (I.34)), DS is the mean Mahalanobis distance in the test set and DS_0 is the mean Mahalanobis distance in the training set. It is advantageous to keep the number of predictor variables, m, low, especially if DS/DS_0 is high. A danger of extrapolation is shown: MSEP in the training set is not a good estimator of MSEP in the test set with severe extrapolation (DS becomes large).

4.3 A summary of the preceding chapters

One of the keywords in this thesis is "model". The calibration of new stationary phases is done with the aid of a model. The relationship between predictors (markers) and dependent variables (nonmarkers) have to be modelled. The exact form of such a model is not known *a priori* (the number of predictors, which predictors, transformations of predictors etc.). Diagnostic tools are, therefore, necessary to judge the quality of predictors and of the model (see Section 2.2). Obviously, prior to the model-building process a set of predictors has to be chosen (see Chapter 1).

One of the first decisions to be made in the model-building process is the purpose of the model. The models used in this thesis have one explicit purpose: prediction. Therefore, the models must be judged with criteria measuring the predictive performance (see Section 2.3).

A difficult problem arises if the performance of a model has to be established. Incorporating too much variables in the model bears the risk of overfitting: irrelevant variables incorporated in the model introduce noise and obscure the effect of really important variables because of multicollinearity between the predictors (see Section 2.4). The need for variable selection is clear. If this variable selection is done using the data, uncertainty is introduced. The data is used to select the variables and, in a second step, to evaluate the model and the selected variables as if they were predetermined. Consequently, an assessment of the performance of the model tends to be optimistically biased.

Another issue related to model-building is the existence of influential observations. This subject is not treated extensively in this thesis, but some remarks are given (see Chapter 1).

After establishing the modelspecification, the estimation method with which the parameters in the model are to be estimated must be chosen (see Chapter 3). Ridge regression, James-Stein regression and partial least squares are designed to meet specific goals and should be chosen with these goals in mind.

All the above sketched topics and problems are intertwined. This makes the model-building process complex. For example: influential observations can induce multicollinearity; ridge regression and partial least squares can perhaps (partly) solve the problem of multicollinearity; jack-knife gives an idea on chance results when selecting markers, but depends on influential observations.

References Part I

1 T.W. Anderson; *An introduction to Multivariate Statistical Analysis (2nd edition)*, John Wiley & Sons, 1984

2 K.V. Mardia, J.T. Kent and J.M. Bibby; *Multivariate Analysis*, Academic Press, 1979

3 I.T. Jolliffe; *Principal Component Analysis*, Springer-Verlag, 1986

4 F.A. Graybill; *Matrices with Applications in Statistics (2nd edition)*, Wadsworth Statistics/ Probability Series, 1983, page 299

5 S. Wold, K. Esbensen and P. Geladi; *Principal Component Analysis*, Chemom.and Int.Lab.Systems, 2 (1987) 37

6 J.C. Davis, *Statistics and Data Analysis in Geology, John Wiley & Sons, 1973*

7 *I.T. Jolliffe; Discarding Variables in a Principal Component Analysis. I: Artificial Data*, Appl.Statist., 21 (1972) 160-173

8 I.T. Jolliffe; *Discarding Variables in a Principal Component Analysis. II: Real Data*, Appl.Statist., 22 (1973) 21-31

9 S. Wold, M. Sjöström, R. Carlson, T. Lundstedt, S. Hellberg, B. Skagerberg, C. Wikström and J. Öhman; *Multivariate Design*, Anal.Chim.Acta, 191 (1986) 17-32

10 C.R. Rao and S.K. Mitra; *Generalized Inverse of Matrices and its Applications*, John Wiley & Sons, 1971

11 G.H. Golub and C.F. van Loan; *Matrix Computations*, John Hopkins University Press, Baltimore, 1985

12 S. Wold; *Cross-Validatory estimation of the number of Components in Factor and Principal Component Analysis*, Technometrics, 20(4) (1978) 397-405

13 W.J. Krzanowski; *Cross-Validatory choice in Principal Component Analysis: some sampling results*, J.Statist.Computat.Simul., 18 (1983) 299-314

14 D.W. Osten; *Selection of Optimal Regression models via Cross-Validation*, J.Chemometrics, 2 (1988) 39-48

15 J.L. Horn; *A rationale and test for the number of factors in factor analysis*, Psychometrika 30 (1965) 179-186

16 E.R. Malinowski and D.G. Howery; *Factor Analysis in Chemistry*, John Wiley, New York, 1980

17 E.R. Malinowski; *Statistical F-Tests for Abstract Factor Analysis and Target Testing*, J.Chemometrics 3(1) (1988) 49-60

18 D.J. Aigner, C. Hsiao, A. Kapteyn and T.J. Wansbeek; *Latent Variables in Econometrics*, Handbook of Econometrics (eds. Z.Griliches and M.D.Intiligator), vol.2, North-Holland Publishing Company, Amsterdam, 1984

19. K.G. Jöreskog and H. Wold (eds); *Systems under Indirect Observation, Part I and Part II*, North-Holland Publishing Company, Amsterdam, 1982

20 S. Wold; *A Theoretical Foundation of Extrathermodynamic Relationships (Linear Free Energy Relationships)*, Chemica Scripta 5 (1974) 97-106

21 S. Wold and M. Sjöström; *Statistical Analysis of the Hammett Equation*, Chemica Scripta 2 (1972) 49-55

22 J.J. Daudin, C. Duby and P. Trecourt; *Stability of Principal Component Analysis studied by the Bootstrap method*, Statistics 19(2) (1988) 241-258

23 M.H. Ramsey and M. Thompson; *A caution of Principal Component Analysis: an example from inductively-coupled plasma/atomic emission spectrometry*, Anal.Chim.Acta, 206 (1988) 203-214

24 F.L. Ramsey; *A Fable of PCA*, Amer.Stat., 40(4) (1986) 323-324

25 T. Naes; *Leverage and Influence measures for Principal Component Analysis*, Chem.and Intell.Lab.Systems, 5 (1989) 155-168

26 C.J. Skinner, D.J. Holmes and T.M.F. Smith; *The Effect of Sample Design on Principal Component Analysis*, J.Amer.Stat.Ass., 81(395), Theory & Methods (1986) 789-797

27 G.P. McCabe; *Principal Variables*, Technometrics 26 (1984) 137-144

28 N.R. Draper and H. Smith; *Applied Regression Analysis*, John Wiley & Sons, New York, 1966

29 M. Okamoto; *Optimality of principal components*, Multivariate Analysis II (ed. P.R.Krishnaiah), 673-685, New York, Academic Press

30 H. Steigstra, A.P. Jansen and G. Kateman; *Multi-Inductive Component Analysis, a new approach in Pattern Recognition*, Anal.Chim.Acta, 186 (1986) 175-183

31 T. Steerneman; *Prediction Performance and the number of variables in Multivariate Linear Regression*, In: *Misspecification Analysis; Lecture notes in Economics and Math.Systems* (ed. T.K.Dijkstra), 118-129, Springer-Verlag, 1984

32 A.G.M. Steerneman; *On the Choice of Variables in Discriminant and Regression Analysis*, Ph.D.Thesis, Department of Econometrics, University of Groningen, 1987

33 W. Schaafsma and T. Steerneman; *Discriminant Analysis when the number of Features is Unbounded*, IEEE Transactions on Systems, Man and Cybernetics, 11(2) (1981) 144-151

34 J.G. Topliss and R.J. Costello; *Chance Correlations in Structure-Activity studies using Multiple Regression Analysis*, J.Med.Chem., 15(10) (1972)

35 J.G. Topliss and R.P. Edwards; *Chance Factors in studies of Quantitative Structure-Activity Relationships*, J.Med.Chem., 22(10) (1979)

36 W.J. Krzanowski; *Selection of Variables to preserve Multivariate Data Structure using Principal Components*, Appl.Stat., 36(1) (1987) 22-33

37 W.J. Krzanowski; *Principles of multivariate analysis: a user's perspective*, Oxford: Clarendon, 1988

38 J.C. Gower; *Statistical methods of comparing different multivariate analyses of the same data*, In: *Mathematics in the Archaeological and Historical Sciences* (eds. F.R.Hodson, D.G.Kendall and P.Tautu), pp. 138-149, Edinburgh, University Press, 1971

39 W.S. DeSarbo; *Canonical/Redundancy Factoring Analysis*, Psychometrika, 46(3) (1981) 307-329

40 P. Robert and Y. Escoufier; *A Unifying tool for linear multivariate statistical methods: the RV-coëfficiënt*, Appl.Statist., 25(3) (1976) 257-265

41 S. Wold; *Pattern Recognition by means of disjoint Principal Component Models*, Pattern Recogn., 8 (1976) 127

42 S. Wold, C. Albano, W.J. Dunn III, U. Edlund, K. Esbensen, P. Geladi, S. Hellberg, E. Johansson, W. Lindberg and M. Sjöström; *Multivariate Data Analysis in chemistry*. In: *Chemometrics: Mathematics and statistics in chemistry* (ed. B.R.Kowalski), Reidel, Dordrecht, 1984

43 R.A. Fisher; *The use of multiple measurements in taxonomic problems*, Annals of Eugenics 7 (1936) 179-188

44 G.P. McCabe; *Computations for variable selection in Discriminant Analysis*, Technometrics 17(1) (1975) 103-109

45 H. van der Voet and J.B. Hemel; *Multivariate Classification Methods and their Evaluation in Applications*, Ph.D.Thesis, University of Groningen, 1988

46 D. Coomans and I. Broeckaert; *Potential Pattern Recognition in Chemical and Medical Decision Making*, Wiley, New York, 1986

47 H. Steigstra, A.P. Jansen and G. Kateman; *SOLOMON, a classification program based on a statistical multivariate disjoint model*, Anal.Chim.Acta, 193 (1987) 269-276

48 J.D. Carroll, S. Pruzansky and J.B. Kruskal; *CANDELINC: A general approach to multidimensional analysis of many-arrays with linear constraints on parameters*, Psychometrika, 45(1) (1980) 3-24

49 P.M. Kroonenberg and J. de Leeuw; *Principal Component Analysis of Three-Mode data by means of Alternating Least Squares algorthms*, Psychometrika, 45(1) (1980) 69-97

50 R.A. Harshman; *Foundations of the PARAFAC procedure: Models and conditions for an "explanatory" multi-modal factor analysis*, UCLA Working Papers in Phonetics 16 (1970) 1-84

51 L.R. Tucker; *Implications of factor analysis of three-way matrices for measurement of change*, In: *Problems in measuring change* (ed. C.W.Harris), University of Wisconsin Press, Madison (Wisconsin), 1963

52 H.G. Law, C.W. Snyder, J.A. Hattie and R.P. McDonald (eds.); *Research Methods for Multimode Data Analysis*, Praeger Publ. New York, 1984

53 S. Wold, P. Geladi, K. Esbensen and J.Öhman; *Multi-Way Principal Components- and PLS-Analysis*, J.Chemometrics, 1 (1987) 41-56

54 C.L. de Ligny, M.C. Spanjer, J.C. van Houwelingen and H.M.Weesie; *Three-mode Factor Analysis of data on the retention in Normal-Phase LC*, J.Chromatogr., 301 (1984) 311

55 E. Sanchez and B.R. Kowalski; *Tensorial Calibration I. First-order Calibration*, J.Chemometrics, 2 (1988) 247

56 E. Sanchez and B.R.K owalski; *Tensorial Calibration II. Second-order Calibration*, J.Chemometrics, 2 (1988) 265

57 A. Lorber, *personal communication*, 1988

58 H. Wijnne; *Multivariate Analyse van selectieve interacties*, Ph.D.Thesis, University of Amsterdam, 1983

59 M.F. Delaney, A.N. Papas and M.J. Walters, *Chemometric classification of Reversed-phase HPLC columns*, J.Chromatogr., 410 (1987) 31-41

60 H.D. Vinod and A. Ullah; *Recent Advances in Regression Methods*, Marcel Dekker, New York, 1981

61 O.M. Kvalheim; *Scaling of Analytical Data*, Anal.Chim.Acta 177 (1985) 71-79

62 P.J. Brown; *Centering and Scaling in Ridge Regression*, Technometrics 19(1) (1977) 35-36

63 J. Johnston; *Econometric Methods (2nd ed)*, McGraw-Hill, Tokyo, 1972

64 G.G. Judge, W.E. Griffiths, R.C. Hill, H. Lütkepohl and T.C. Lee; *The Theory and Practice of Econometrics (2nd ed)*, John Wiley, New York, 1985

65 J.A. Cornell; *Experiments with Mixtures*, John Wiley, New York, 1981

66 D.A. Belsley, E. Kuh and R.E. Welsch; *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*, John Wiley, New York, 1980

67 R.L. Mason; *Latent Root Regression: A Biased Regression methodology for use with collinear predictor variables*, Commun.Statist.-Theor.Meth., 15(9) (1986) 2651-2678

68 D.M. Hawkins; *On the investigation of alternative regressions by principal component analysis*, Appl.Stat., 22 (1973) 275-286

69 J.T. Webster, R.F. Gunst and R.L. Mason; *Latent Root regression analysis*, Technometrics, 16 (1974) 513-522

70 R.R. Hocking; *The analysis and selection of variables in linear regression*, Biometrics 32 (1976) 1-49

71 D.M. Allen; *The relationship between variable selection and data augmentation and a method for prediction*, Technometrics, 16(1) (1974) 125-127

72 L. Breiman and D. Freedman; *How many variables should be entered in a regression equation?*, J.Amer.Stat.Ass., 78(381) (1983) 131-136

73 T. Amemiya; *Advanced Econometrics*, Basil Blackwell Ltd., 1985

74 O. Bunke and B. Droge; *Bootstrap and Cross-validation estimates of the prediction error for linear regression models*, Ann.Statist., 12(4) (1984) 1400-1424

75 O. Bunke and B. Droge; *Estimators of the Mean Squared Error of Prediction in Linear Regression;* Technometrics 26(2) (1984) 145-155

76 D.M. Allen; *Mean Squared Error of Prediction as a criterion for selecting variables*, Technometrics 13(3) (1971) 469-475

77 R.R. Picard and R.D. Cook; *Cross-Validation of Regression Models*, J.Amer.Stat.Ass., 79(387) (1984) 575-583

78 K.N. Berk; *Validating Regression procedures with new data*, Technometrics, 26(4) (1984) 331-338

79 C.L. Mallows; *Some comments on C_p*, Technometrics 15 (1973) 661-676

80 G.H. Golub, M. Heath and G. Wahba; *Generalized Cross-validation as a method for choosing a Good Ridge parameter*, Technometrics, 21(2) (1979) 215-223

81 B. Efron and G. Gong; *A Leisurely Look at the Bootstrap, the Jackknife, and Cross-Validation*, Amer.Statist., 37(1) (1983) 36-48

82 M. Stone; *Cross-validatory choice of Statistical Predictions*, J.Royal Stat.Soc.,Ser.B, 36 (1974) 111-147

83 S. Wold, A. Ruhe, H. Wold and W.J. Dunn III; *The collinearity problem in linear regression. The Partial Least Squares (PLS) approach to generalized inverses*, SIAM J.Sci.Stat.Comput., 5(3) (1984) 735-743

84 D.W. Marquardt; *Generalized Inverses, Ridge Regression, Biased Linear Estimation, and Nonlinear Estimation*, Technometrics, 12(3) (1970) 591-612

85 R.W. Hoerl, J.H. Schuenemeyer and A.E. Hoerl; *A Simulation of Biased Estimation and Subset Selection Regression Techniques*, Technometrics, 28(4) (1986) 369-380

86 T.K. Dijkstra (ed.); *Misspecification Analyses, Lecture Notes in Economics and Math.Systems*, Springer Verlag, 1984

87 P.P. Mager; *Multivariate Chemometrics in QSAR: A Dialogue*, John Wiley, New York, 1988

88 S.D. Silvey; *Multicollinearity and Imprecise Estimation*, J.Royal Stat.Soc.Ser.B, 35 (1969) 67-75

89 G.G. Judge and M.E. Bock; *Biased Estimation*, In: *Handbook of Econometrics* (Z.Griliches and M.D.Intriligator, eds), North-Holland, 1983

90 L.S. Mayer and Th.A. Willke; *On Biased estimation in Linear Models*, Technometrics, 15(3) (1973) 497-508

91 A.E. Hoerl and R.W. Kennard; *Ridge Regression: Biased Estimation for Nonorthogonal Problems*, Technometrics, 12(1) (1970) 55-67

92 A.E. Hoerl and R.W. Kennard; *Ridge Regression: Applications to Nonorthogonal Problems*, Technometrics, 12(1) (1970) 69-82

93 N.R. Draper and R.C. van Nostrand; *Ridge Regression and James-Stein Estimation: Review and Comments*, Technometrics, 21(4) (1979) 451-466

94 J. Riley; *Solving systems of linear equations with a positive definite, symmetric but possibly ill-conditioned matrix*, Mathematical Tables and Other Aids to Computation, 9 (1955) 96-101

95 D.W. Marquardt and R. Snee; *Ridge Regression*, Proc.of Univ. of Kentucky Conference on regression with a large number of predictor variables, W.O.Thompson and F.B.Cady (eds), Dept. of Stat. Univ. of Kentucky, Lexington Kentucky, 1973

96 G.C. McDonald and D.I. Galarneau; *A Monte Carlo Evaluation of Some Ridge Type Estimators*, J.Amer.Statis.Ass., 70 (1975) 407-416

97 A.P. Dempster, M. Schatzoff and N. Wermuth; *A Simulation study of alternatives to Ordinary Least Squares*, J.Amer.Statist.Ass., 72 (1977) 77-106

98 B.F. Swindel; *Geometry of Ridge Regression Illustrated*, Amer.Statist., 35(1) (1985) 12-15

99 W. James and C. Stein; *Estimation with Quadratic Loss*, In: *Proc. of the fourth Berkely Symposium Mathematical Statistics and Probability*, vol.1, Univ. of California Press, Berkely, 1961

100 R.I. Hennrich and S.D. Oman; *How much does Stein estimation help in Multiple Linear Regression?*, Technometrics, 28(2) (1986) 113-121

101 T.K. Dijkstra; *Some comments on Maximum Likelihood and Partial Least Squares Methods*, J.Econometrics 22 (1983) 67-90

102 T.K. Dijkstra; *Latent Variables in Linear Stochastic Models (2nd ed.)*, Sociometric Research Foundation, Amsterdam, 1985

103 P. Geladi and B.R. Kowalski; *Partial Least Squares: a Tutorial*, Anal.Chim.Acta, 185 (1986) 1-17

104 P. Geladi and B.R. Kowalski; *An example of 2-block predictive Partial Least Squares regression with simulated data*, Anal.Chim.Acta, 185 (1986) 19-32

105 A. Höskuldsson; *PLS Regression Methods*, J.Chemometrics 2 (1988) 211-228

106 R. Manne; *Analysis of Two Partial Least Squares algorithms for Multivariate Calibration*, Chemo.and Intell.Lab.Systems, 2 (1987) 187-197

107 H.A. Martens; *Multivariate Calibration*, Dr.Techn.Thesis, Technical University of Norway, Trondheim, 1985

108 K.R. Gabriël and S. Zamir; *Lower Rank Approximation of Matrices by Least Squares with any choice of weights*, Technometrics 21(4) (1979) 489-498

109 G.H. Golub and W. Kahan; *Calculating the singular values and pseudo-inverse of a matrix*, SIAM J.Numerical Analysis, Series B 2 (1965) 205-224

110 T. Naes and H. Martens; *Comparison of Prediction methods for Multicollinear data*, Commun.Statist.-Simul.Comput., 14(3) (1985) 545-576

111 G.H. Golub and C.H. van Loan; *An Analysis of the Total Least Squares Problem*, SIAM J.Numer.Anal., 17(6) (1980) 883-893

112 G.E.P. Box, W.G. Hunter and J.S. Hunter; *Statistics for Experimenters*, John Wiley, New York, 1978

113 S.N. Deming and S.L. Morgan; *Experimental Design: a chemometric approach*, Elsevier, Amsterdam, 1987

114 O.L. Davies (ed.); *The Design and Analysis of Industrial Experiments (2^{nd} ed)*, Longman Group Ltd., London, 1978

115 G.E.P. Box and N.R. Draper; *Emperical Model-Building and Response Surfaces*, John Wiley, New York, 1987

116 V.V. Fedorov; *Theory of Optimal Experiments*, Acad.Press, New York, 1972

117 S.D. Silvey; *Optimal Design*, Chapman and Hall, London, 1980

118 J.W. Weyland, C.H.P. Bruins and D.A. Doornbos; *Use of Three-Dimensional Minimum Alpha plots for optimization of mobile phase compositions for RP-HPLC separation of sulfonamides*, J.Chrom.Sci., 22 (1984) 32-39

119 J.W. Weyland; *Strategies for Mobile Phase Optimization in Chromatography*, Ph.D.Thesis, University Centre for Pharmacy, Groningen, 1986

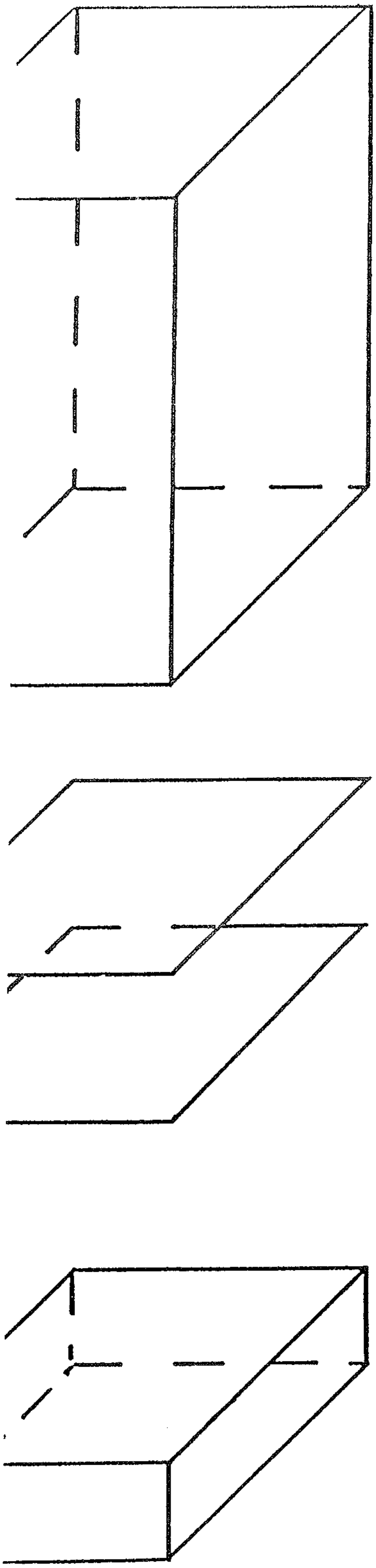
120 H.J.G. Debets; *The automatic optimization of reversed-phase hplc-separations (mobile phase composition)*, Ph.D.Thesis, University Centre for Pharmacy, Groningen, 1986

121 R. Snee; *Experimenting with Mixtures*, Chemtech (1979) 702-710

122 A.C.J.H. Drouen, H.A.H. Billiet, P.J. Schoenmakers and L. de Galan; *An improved optimization procedure for the selection of mixed mobile phases in reversed-phase liquid chromatography*, Chromatographia 16 (1982) 48

123 J.P. Bianchini, E.M. Gaydou, A.M. Siouffi, G. Mazerolles, D. Mathieu and R. Phan-Tan-Luu; *Optimization of the separation of Polymethoxylated Flavones in Reversed-Phase Liquid Chromatography*, Chromatographia 23(1) (1987) 15-20

124 R. Snee; *Validation of Regression Models: Methods and Examples*, Technometrics 19 (1977) 415-428



PART II
Reversed-Phase Chromatographic Introduction

Chapter 5 The mobile phase

5.1 Recent developments in mobile phase optimization

A lot of research has been done in the area of mobile phase optimization. Good reviews are given by Berridge[1] and Schoenmakers[2]. Recent developments in the area comprise the use of multi-criteria decision making[3], comparisons between different mobile phase systems[4], investigation of the selectivity of multisolvent systems[5], the use of optimal design theory[6] to select the initial mobile phase compositions[7], a statistical basis to select appropriate starting eluents[8], the use of diode array detection[9,10], new criteria to characterise a chromatographic separation[11], the use of retention indices[12,13], and an alternative optimization strategy[14].

5.2 Prediction of retention dependent on the mobile phase composition

In a series of papers[15-18], Jandera describes a method of retention prediction in binary eluentia. The key of this approach is the formula:

$$\log k = (a_0 + a_1 n_c)(1-px) - qx \tag{II.1}$$

where a_1 and p are assumed to depend only on the type of organic modifier, x is the fraction of organic modifier in the mobile phase, a_0 and q are solute/stationary phase specific parameters. Once the system is calibrated for a suitable homologous series, e.g. n-alkylbenzenes, the values of a_0, a_1 and p are assumed fixed. If prediction of the retention of a member of a homologous series is wanted, n_c is the number of carbon atoms of that member. Only one measurement of that member is necessary to calculate q, which is solute/stationary phase specific. If the prediction of an arbitrary solute is at hand, two measurements of that solute are needed to calculate q and n_c, which is then defined as the "carbon equivalent" of the solute. Modifications to incorporate ternary mixtures are reported. The formula is adapted to predict selectivity values (α values) of a solute relative to toluene. This allows the prediction of capacity factors at other stationary phases with the use of the measured capacity factor of toluene. The mean relative deviation of the predicted k from the experimental values was between 7 and 10% for a number of six test compounds on nine stationary phases.

Several comments are appropriate. The above mentioned formula is obtained after a series of linear approximations of a complicated function $F(n_c, x)$ describing the dependence of log k on n_c and x. This linear approximation holds for x values greater than 0.4. For the organic modifier tetrahydrofuran such a linearisation fails[16]. If the prediction of the k of a solute is wanted, two measurements are needed in Jandera's approach. It is questionable whether his approach is better than a simple linear approximation between the measured log k values as is used in the area of mobile phase optimization[2]. Closer examination of the formula used by Jandera, shows that this formula resembles a bilinear model: $a_0+a_1n_c$ and q depend only on the solute/stationary phase whereas 1-px and x depend only on the mobile phase. The influence of the mobile phase composition and the solute/stationary phase combination on log k is separated and combined in linear parts. As such the background of the formula becomes clear: it is a bilinear approximation of the complicated function F. Whether Jandera's approach is better than a direct bilinear approximation of this function F with the use of principal component- or factor analysis[19,20] is questionable.

The use of factor analysis for the prediction of retention has shown its utility in gas-chromatography[21,22] and reversed-phase chromatography[23]. This latter study uses a small (3x4) matrix of retention values as a training set. With the use of three retention values of the compound of interest, the retention values of that compound at the other mobile phase compositions can be predicted. An average relative prediction error in k of 5% is reported. Note that the data used to validate the procedure are chosen from the same set as the training set. A real independent test set is, therefore, not available and the results may be too optimistic.

An analogous prediction scheme as Jandera is provided by Jinno *et al.*[24] to predict the retention of PTH amino-acids. The retention of the amino-acids is predicted with linear models based on a solubility parameter R, the column temperature and the fraction of organic modifier in the mobile phase. The use of a parameter R is based on ideas current in the area of quantitative structure retention relationships[25]. The developed models can be used to optimize a separation.

Chapter 6 The stationary phase

6.1 Properties of reversed-phase stationary phases

Most reversed-phase materials are based on silica. The structure of this silica substrate and its influence on retention characteristics has received much attention[26-30]. The modification of the silica substrate in order to obtain the reversed-phase material is performed by silanisation of the silica. This is achieved by a reaction of the silanol groups at the surface of the bare silica with chlorosilanes or alkoxysilanes[31]. Groups of different functionalities can be bonded covalently to the silica surface, resulting in different types of stationary phases. Jones[32] has reported a study in which 21 variables are screened with respect to their influence on the silanisation process, using a Plackett-Burman design. A three level Plackett-Burman design was used thereafter to investigate in detail six variables[33]. This illustrates the complexity of the silanisation process, especially because Jones only investigated the dependence of carbon weight percentage on the 21 variables.

Apart from differences between the silica substrate material, the complicated silanisation process causes non-reproducible differences between stationary phases of the same type. Not only stationary phases of different brands differ[34-36], but even different batches of stationary phases from the same make[37-40] differ non-reproducibly. The bad reproducibility of retention values, due to changing stationary phases, is especially troublesome in the area of mobile phase optimization[41].

Due to sterical hindrance not every silanol group at the surface of the silica substrate reacts with the silanisation reagent, so that about 50% of the silanol groups is left unreacted[30]. One of the differences between stationary phases of the same type (i.e. with the same functional group attached to the silica surface) is caused by differences in amount and distribution of the free silanol groups[30,42-45]. This leads to a dual retention mechanism, not only hydrophobic interactions are important for the retention of a solute, but also the silanophilic (specific) interactions[46]. Obviously, this dual retention mechanism can be used to enhance selectivity[47,48]. On the other hand, however, the retention behaviour of basic solutes is hampered by the free silanol groups[49,50].

The elimination of the undesirable influence of free silanol groups is partly achieved by endcapping. After the silanisation process the stationary phase is treated with, e.g., trimethylchlorosilane that reacts with the free silanol groups[30,51]. Not all free silanol groups, however, are bonded to the endcapping reagent. Two new types of silane modified silica have been manufactured recently[52], which attempt to shield the free silanol groups by either attaching a monofunctional silane with a bulky group to the silica surface or bonding of bidendate silanes to the silica surface.

One of the problems encountered in reversed-phase chromatographic separations is the deterioration of the stationary phase material[53-55]. Retention values change due to ageing of the stationary phase material. A mobile phase optimization procedure performed on a fresh stationary phase is not easily updated if the column deteriorates.

The influence of the stationary phase on the retention of a solute depends on the mobile phase composition[56-59]. Therefore, the influence of the stationary phase and the mobile phase on the retention of a solute is difficult to entangle. Various types of stationary phases (i.e. modified with groups of different functionalities) show different behaviour with respect to the retention of a solute[60-62]. This leads to different selectivities which can be used to achieve a separation between a set of solutes. The selectivity of the stationary phase can also be changed by mixing different silylating reagents[63], so that the stationary phase consists of structure elements of different types. In fact, the endcapping procedure is a special example of such a mixed stationary phase[64].

6.2 Classification and characterization of reversed-phase material

Cluster analysis and principal component analysis (PCA) were used by Delaney *et al.*[65] to classify nine octadecyl stationary phases. Out of a set of ten benzene-derivatives, the solutes methylparaben, phenol and benzoic acid were found to contain the most information with regard to the differences between the stationary phases. Only one eluent composition (a binary water/acetonitrile mixture) was involved in the PCA calculations. Conclusions with regard to either other mobile phase systems or specific contributions of stationary/mobile phase combinations are, therefore, not available. With the use of linear discriminant analysis, Welsch *et al.*[66] found that the log k value of aniline, the log k value of butyrophenone and the asymmetry factor (a measure of peak asymmetry) of benzylalcohol were able to reproduce clusters of different kinds of stationary phase materials which were previously discovered with the use of cluster analysis. All measurements were performed with n-heptane as eluent, in order to find solutes sensitive to free silanol groups.

Antle *et al.*[67,68] define the concepts "strength" and "selectivity" of columns analogous to mobile phase solvents[69]. The column strength parameter or effective phase ratio, J, differs between reversed-phase columns and leads to corresponding shifts in retention for all solutes on a given column (at the same mobile phase composition). Column polarity can be measured by a parameter P, which further characterizes column selectivity. With the use of a set of benzene derivatives, estimates of J and P are found for six stationary phases of different types.

In a series of papers[70-73,48,74-75], multivariate analysis is used to study the selectivity of chalcones and other test solutes. In the first study[70], correspondence factor analysis (CFA, a version of PCA designed for nominal scaled variables[76]) is used to analyze a data set consisting of capacity factors of 53 chalcones, measured on four stationary phases of different types at the same mobile phase composition. The first axis of the CFA can be associated with the hydrophobicity of the compounds and stationary phases, the second axis describes specific interactions. In the fifth study[48], the selectivity of 22 chalcone configuration isomers on 23 reversed-phase packings are analysed with PCA. The scores on the first two principal components show clearly three different groups of stationary phases: octadecylsilyl endcapped phases, octadecylsilyl uncapped or partially capped phases and trimethyl, C_6, C_8, Phenyl, CN stationary phases. An interesting conclusion is that the selectivity seems to increase with increasing silanol group accessibility. In the sixth and seventh study[74,75], a distance metric (based on CFA) is defined that measures the differences in selectivity of fourteen octadecyl stationary phases, with respect to 63 compounds (chalcones and benzene derivatives). An attempt is made to estimate the hydrophobic- and non-hydrophobic part of these differences with the purpose to design a hydrophobicity scale for RP-HPLC packings. Packing characteristics governing selectivity are the carbon loading, the nature of the organic ligand and the accessibility of the silanol groups. The source of the silica gel, the shape of the silica and the type of organic layer do not have a significant influence on selectivity.

6.3 Simultaneous optimization of stationary- and mobile phase

A very straightforward strategy to optimize both stationary- and mobile phase is given by Lin[77]. Initially, the solutes were eluted on four stationary phases, a methyl- octyl- phenyl- and a cyano modified stationary phase, at varying binary mobile phase mixtures (water/methanol). Two stationary phases showed promising selectivity towards the test solutes and were tested further with water/acetonitrile and water/tetrahydrofuran mixtures, where the content of the organic modifier was varied from 10 to 50%. The best combination was chosen.

A strategy for the simultaneous optimization of stationary- and mobile phase is presented by Glajch *et al.*[78,79]. Three stationary phases of different types - C_8, CN, Phenyl - were incorporated in an earlier developed strategy for mobile phase optimization[80]. A complete mobile phase optimization scheme was performed for each type of stationary phase. Models were made to describe the influence of the mobile phase on the k values and resolution. The three stationary phases were treated as mixture variables, the k values of the solutes were assumed to depend linearly on the amount of stationary phase of a specific type in the column. With the use of this assumption and the models mentioned above, a grid search can be performed to scan all possible stationary/mobile phase combinations. Several comments are appropriate. First, the assumption of linearity is questionable[64]. This is also experimentally verified by Glajch and non-linear effects are detected. Second, not all solutes have to be measured on each stationary phase, this will be shown in Part III of this thesis.

Chapter 7 Calibration of reversed-phase chromatographic systems

7.1 Definition of calibration of a chromatographic system

The first step in calibration of a chromatographic system is the measurement of retention values of specific compounds (standards) on that system, with a specific purpose. The second step depends on the goal of calibration.

The goal of calibration can be to obtain a "measurement-system independent" retention value of a solute, measured on a new system. The second step then comprises the correction of the retention value of that solute, using the standards measured on the new system. This correction is particularly useful in the area of identification of unknown compounds with their retention values.

Another goal of calibration is the transfer of the retention value of a solute on one system to another system. The second step then comprises the prediction of that retention value on the new system, with the use of measured standards on that new system. This goal is particularly valuable in the area of mobile phase optimization where an optimal eluent composition must be updated if columns[41] are changed or if a column is deteriorated.

7.2 Calibration in gas-liquid-, paper- and thin-layer chromatography

A widespread used calibration system in gas-liquid chromatography (GLC) is the retention index system developed by Kovats[81]. These indices are based on n-alkanes as reference substances. The purpose of the correction of the retention value of a specific compound (by calculating its index) is to make that retention value less dependent on measurement conditions. This system is based on the linear relationship between the logarithm of the net retention times of the n-alkanes and the number of carbon atoms in the molecules. By definition, each homolog in the n-alkane series C_nH_{2n+2} receives the retention index RI=100n, so that fixed reference points are obtained. The retention index RI(A) of a compound A can then be obtained from a simple graph as shown in Figure II.1.

For use in paper-chromatography (PC) Galanos and Kapoulas[82] developed a method based on two reference compounds and the calculation of the corrected R_f values ($R_f(c)$) by a linear regression:

$$R_f(c) = aR_f + b \qquad (II.2)$$

where a and b are constants obtained from the calibration of the two measured R_f values of the reference compounds against tabulated values. These tabulated values are averages of repeated measurements of the reference compounds under controlled conditions.

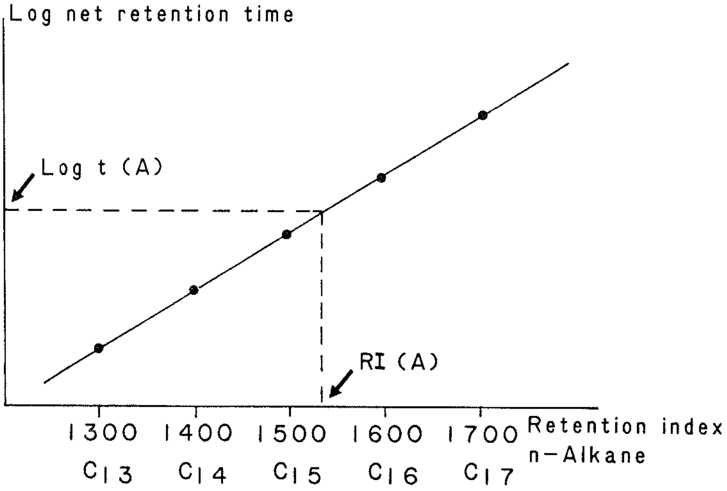


Figure II.1. Kovats indices. Log t(A) is the log net retention time of solute A. RI(A) is the retention index of solute A.

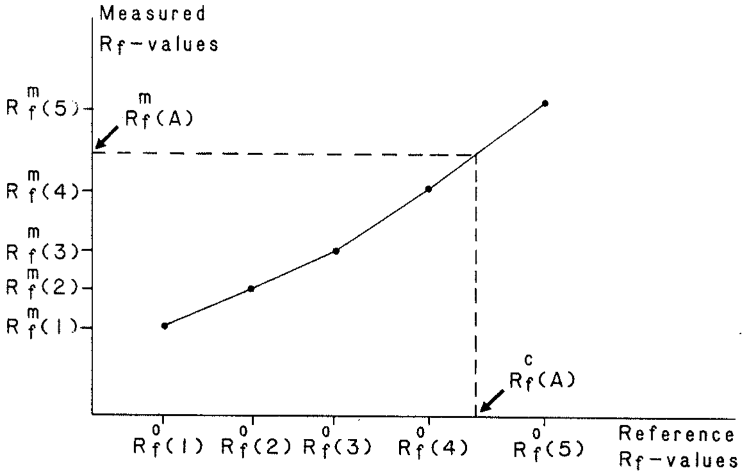


Figure II.2. Corrected R_f in TLC. $R_f^0(1)$ to $R_f^0(5)$ are the reference R_f values of the standards. $R_f^m(1)$ to $R_f^m(5)$ are the measured R_f values of the standards. The measured R_f value of solute A is $R_f^m(A)$, whereas the corrected R_f value of A is $R_f^c(A)$.

The same ideas used in PC are applied to thin-layer chromatography (TLC). Moffat[83] describes a procedure for TLC and advises four or five standards. The correction method is graphically depicted in Figure II.2. The reference R_f values of the standards (on the X-axis) are plotted against the measured R_f values (on the Y-axis). The reference values of the standards can be measurements made in one particular laboratory[84] or average values of several laboratories[83]. The corrected R_f value of a compound A is found by linear interpolation, as showed in Figure II.2. More than two standards are needed, because curvature may be present in the reference versus measured R_f values of the standards.

7.3 Calibration in reversed-phase chromatography (RP-HPLC) with retention indices

A review of the use of retention indices (RI) in RP-HPLC until 1987 is given by Smith[85]. Some results are summarized and extensions are given. The idea of using retention indices in RP-HPLC was introduced by Baker and Ma[86]. They proposed to use a homologous series of 2-keto alkanes. The method of calculating the RI of a compound A is depicted in Figure II.3. By definition, the retention index of a given 2-keto alkane standard is equal to 100 times the number of carbon atoms in that compound. Note that a non-linear relationship between the log k- and RI-values of the homologs is approximated by linear parts. The question is whether this approximation should be used or a curve fit procedure to estimate the relationship between log k and RI. The stationary phases tested were a octadecyl (ODS)- and a cyano (CN) bonded phase. The mobile phase comprised a mixture of 0.025 M NaH_2PO_4 aqueous solution and varying fractions of methanol (MeOH) or acetonitrile (ACN). Especially, the log k versus RI plot of the homologs with the combination of the CN phase and ACN containing eluent was curved. A number of eight different drugs was chosen as test compounds. From the combination of the ODS phase and the MeOH/NaH_2PO_4 mobile phase with varying methanol content (from 20-90% MeOH), the following was concluded: the RI values of the drugs decreased on average 18 units for each 10% increase in the methanol content of the mobile phase (The RI values of the drugs ranged from 230 to 870 RI units, approximately). This indicates that the RI values cannot be considered independent of the percentage MeOH in the mixture. The stability of a RI, under varying conditions, depends on the test compound, the stationary phase, the mobile phase composition and the particular combinations of these three. If the percentage of organic modifier is changed in the four measurement conditions - combinations of the ODS and CN phase with the two kinds of mobile phases - it appears that the RI values of acetophenone remains stable whereas the RI values of androsterone vary. Especially the combination of androsterone with the CN phase and ACN is unstable. This points to specific effects -interactions- between modifier, stationary phase and solute which makes calibration with retention indices based on a homologous series troublesome. Baker and Ma concluded that the retention index of a given drug does not markedly change with changes in solvent composition, solvent type and column type. No clear definition of "markedly" is given, so conclusions are not easy to draw.

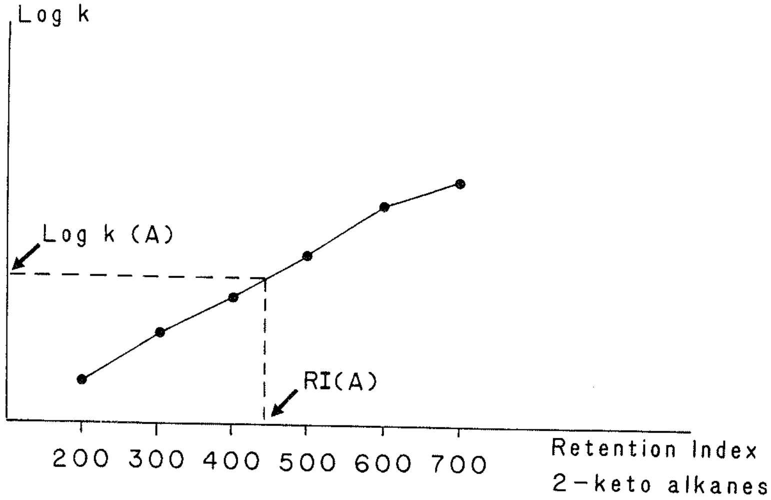


Figure II.3. Retention indices in RP-HPLC. Log k(A) is the logarithm of the capacity factor of a solute A, RI(A) is the calculated retention index of A.

Smith[87] proposed a retention index scale using alkylarylketones. He criticised the use of the 2-keto alkanes, because they have a limited absorbance at 254 nm and are not widely available. As with the 2-keto alkanes, the RI of the alkylarylketones are by definition 100 times the carbon number of that standard. Three to seven alkylarylketones are used to calculate the parameters a and b in log k = aRI + b. The RI of a specific test compound follows from its k value and application of the above mentioned equation with known a and b. Good linear relationships between carbon number (times 100) and log k of the alkylarylketones were found by Smith, on an octadecyl stationary phase at different binary water/methanol mixtures. Eight benzene derivatives were used as test compounds and eluted with mobile phases containing varying percentages of MeOH (on an ODS column). The RI values showed a small but general increase with increasing fractions of MeOH in the solvent. The average deviation of the RI values of these test compounds with a 10% change of the percentage of MeOH, was 13 RI units (the RI values ranged from 500 to 1100, approximately). Again it is not easy to derive conclusions. Yet the alkylarylketones seem to perform slightly better than the 2-keto alkanes. Experiments performed on three different stationary phases (SAS-Hypersil, C_{22}-Magnusil and Spherisorb-Phenyl) combined with the experiments on the ODS phase, showed that the RI values of the eight test compounds, measured at a constant mobile phase composition, varied considerably.

The effect of the eluent on the retention of barbiturates, using alkylarylketones as standards, was investigated by Smith *et al.*[88] The stationary phase is always the same ODS column. He concluded the following. The repeatability (measurements taken on different days) is better in terms of RI than k values. In a binary system - methanol/buffer pH 8.5 - the RI values of the barbiturates showed much smaller changes with varying fractions of methanol content than the capacity factors. Yet the average deviation of the RI values of the barbiturates with a 10% change of the percentage of methanol, was 26 RI units, whereas the analog value of the column test compounds-nitrobenzene, 2-phenylethanol, p-cresol, toluene, n-methylaniline-was 12 RI units. This last number is comparable with the earlier reported value[8,6] which was 13. For the RI values of the barbiturates the average deviation was much higher. According to Smith the reason for these higher deviations are twofold. First, the barbiturates are partly ionised under the changing apparent pH values of the eluent composition. The second reason is the difference in polarity of the barbiturates compared to the ketones. The RI values of the barbiturates showed significant changes with varying temperature, smaller changes however than the k values. The effect of the eluent pH on the RI values of the barbiturates is very strong, which prompts Smith to say that "for comparison purposes the pH of the eluent would probably be the principal factor that would need to be reproducible controlled".

The effect of the stationary phase on the RI values of barbiturates was also studied by Smith *et al.*[37]. The variation in retention values of the barbiturates on ODS columns of different batches was compared with the variation of the retention values on ODS columns of the same batch. It appeared that the k values had a slightly lower variation on the ODS columns of different batches than on the ODS columns of the same batch. The reverse was true for the RI values. This suggests small differences in selectivity between the batches. The variations in RI, on the ODS columns of different batches, were generally greater for the barbiturates than for the column test compounds (the same test compounds were used as in (88)). Different brands of ODS material showed less variation in RI values than k values of the barbiturates, but still the RI values are not as reproducible as on a single batch. The general conclusion is that RI values are more reliable than k values when comparing results obtained from different column packing materials and/or laboratories. The RI values are, however, still sensitive to column selectivity differences.

In another study[89], Smith *et al.* investigated the effect of different factors on the RI values of local anaesthetic drugs, again using the alkylarylketone scale. On an ODS column with an eluent containing methanol, the average deviation of the RI values of the local anaesthetic drugs with a 10% change of MeOH was 24. For the column test compounds - toluene, nitrobenzene, 2-phenylethanol, p-cresol - this number was 17. Although Smith concluded that the RI is virtually unaffected by changes in percentage of methanol, a complete independence is not present. The effect of the eluent pH is slightly visible in the RI values of the local anaesthetics, but not in the RI values of the column test compounds. The temperature of the column has influence and this prompts Smith to say that "running the analysis at a specified temperature will be an important requirement to obtain reproducible laboratory results". Comparing three different column packing materials showed large variation in k values of the local anaesthetics. If the results were expressed in RI units the variation was less, but still significant.

The use of the alkylarylketone scale was investigated in acetonitrile- or tetrahydrofuran containing binary eluents[90]. Good linear relationships were found between carbon number (times 100) and log k value of the homologs when changing the percentage acetonitrile in the binary eluent (acetonitrile-phosphate buffer pH 7.0). When changing the percentage of acetonitrile, the largest deviations in RI were found for the more polar test compounds, p-cresol and 2-phenylethanol. The average deviation, with a 10% percent change in acetonitrile, was 10 in RI units. Good linear relationships were also found for the log k values of the homologs and their carbon number in THF containing eluents (water-THF). The RI values, however, were not stable: the RI values of the relatively non-polar compound (toluene) increased steadily with increasing fractions of THF whereas for polar analytes, such as p-cresol, the values decreased markedly. The values of the retention indices in the THF containing eluents, particularly for p-cresol, were significant different from the acetonitrile containing eluent. In comparison with an earlier study[91], it appears that much smaller differences were observed between different makes of column packing material with THF containing eluents than with acetonitrile- or methanol containing eluents. Thus for THF containing eluents the RI values depend on the percentage modifier, but are more independent of the make of the ODS material.

In a study[34] on retention reproducibility of thiazide diuretics and related drugs, the influence of mobile phase properties (percentage acetonitrile, pH, proportion of acetic acid), temperature and stationary phase on RI values was established. The conclusion is that the proportion of acetonitrile in the mobile phase, the column temperature and the brand of column packing material are important. A common brand of packing material should be adopted for interlaboratory comparisons.

Bogusz and Aderjan[92] proposed a retention index scale based on 1-nitroalkanes. They criticised the use of alkylarylketones because the first reference compound on this scale (acetophenone) elutes comparatively late. The calculation of index values for fast eluting compounds is difficult, due to severe extrapolation on the index scale. The relationship between log k of the 1-nitroalkanes and the carbon number of those standards is linear from the second homologue. The RI values of some of the tested drugs - caffeine, barbital, paracetamol, theophylline - decrease sharply with increasing acetonitrile percentage in the binary mobile phase. The RI values of the column test compounds (toluene, methylbenzoate, methylaniline) remained virtually constant when changing the acetonitrile percentage. The same column test compounds were used by Smith in testing the alkylarylketones. Bogusz and Aderjan conclude that the behaviour of these column test compounds is virtually identical on both scales (alkylarylketones and 1-nitroalkanes). The first two homologs of the 1-nitroalkane scale, however, elute so early that it is possible to calculate the RI values of early eluting drugs.

A full comparison between the different scales has to be made, investigating the specific advantages of the particular homologous series. To make such a comparison a yardstick is necessary to evaluate the power of the different homologous series to correct retention values. Such a yardstick will be presented later on.

7.4 Calibration with corrected- and relative capacity factors or corrected retention index values in RP-HPLC

The notions of relative- and corrected capacity factors are discussed firstly. The relative capacity factor of a solute A is defined as k_A/k_S, where S is the standard compound used to calibrate. The idea of corrected capacity factors is borrowed from TLC. Reference k values are assigned to, say, four standards. These values may be mean capacity factors under controlled conditions. These reference values are plotted against the measured values and a least squares routine is used to estimate the best fitting straight line. A measured capacity factor of solute A is corrected with this equation to obtain a corrected k value, see Figure II.4.

Gill[93] investigated the merits of reporting retention values in RI units, relative k values and corrected k values for interlaboratory comparison. On a study with barbiturates, measured under controlled conditions in ten laboratories, he used the discrimination number (DN) as yardstick to compare the different options to report retention. The discrimination number is defined as the number of retention windows, each two standard deviations wide, which can be fitted into a defined chromatographic range. A large DN means that the chromatographic system, with the adopted option to report retention, has a large discriminating power towards the identification of a class of compounds. Gill assumed a predefined range of k values from 1 to 25 to calculate the discrimination numbers of the different options for reporting the retention of five barbiturates, and concluded the following. The corrected k values, based on four barbiturates as standards, gave the highest DN, which was 64. The relative k values, based on one barbiturate as a standard, were second best with a DN of 55. Relative retention times, analogously defined as relative k, were third best (DN=44). The use of RI values, based on the alkylarylketones, gave a DN of 34. The use of either k values or retention times for reporting retention gave a DN of, respectively, 16 and 10.

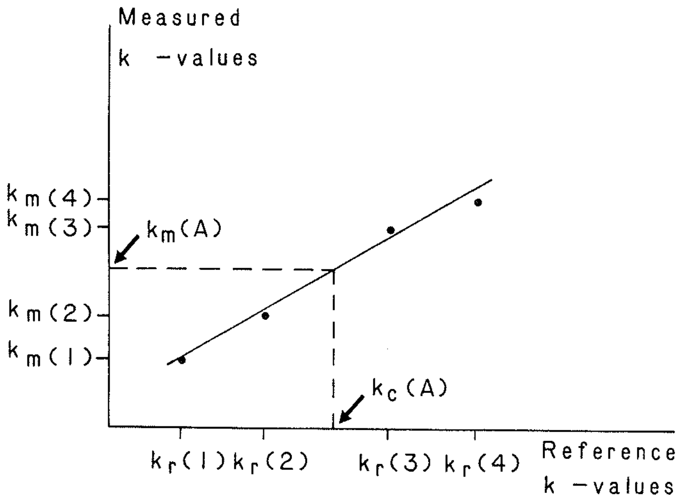


Figure II.4. Corrected k in RP-HPLC. The k_r(1) to k_r(4) values are reference k values of the standards, k_m(1) to k_m(4) are measured values of the standards. The measured capacity factor of solute A, k_m(A), is corrected with outcome k_c(A).

It is clear that corrections using standards of a similar nature as the test solutes, the barbiturates, give the best results. According to Gill, this is because of the fact that the alkylarylketones and barbiturates respond differently to small variations in chromatographic conditions. The use of more than one of such a standard is advantageous.

In the study on retention reproducibility of local anaesthetic drugs[89], Smith calculated relative capacity factors, using one of the local anaesthetics as a standard. His conclusion was that this correction cannot compensate for mobile phase properties (percentage methanol in a binary mobile phase, pH), temperature and different column packing materials.

In the study[34] on retention reproducibility of thiazide diuretics, Smith evaluated the use of relative capacity factors, using one of the thiazide diuretics as standard. He concluded that the proportion of acetonitrile in the (binary) mobile phase, the column temperature and the brand of column packing material are important factors influencing the reproducibility. It is important to use a relative method for reporting retention, such as relative capacity factors or retention indices, but there is no clear best method. For correcting differences due to different column materials the relative capacity factors have some advantage, but still significant differences are present. If relative capacity factors are used, those solutes with a capacity factor in the same order of magnitude as the standard, are corrected best.

Bogusz[94] proposed the use of corrected retention indices. As a first step the RI values of all solutes (barbiturates) relative to the alkylarylketone scale are calculated. Second, three of the barbiturates (located in the low, middle and high RI range) are chosen as correction standards. A plot is made of listed versus measured RI values of those standards. The RI values of the other barbiturates are corrected with linear interpolation, see Figure II.2. He concluded that the corrected RI values show less variation between six different ODS-silica columns than the ordinary RI values. Bogusz stated that the retention behaviour of chemically similar substances on different ODS columns is highly correlated. The method of corrected RI values may, therefore, enable the comparison of retention data of chemically similar compounds obtained on different ODS columns in different laboratories.

7.5 Some conclusions on calibration in RP-HPLC

The difference between the use of retention indices based on homologs in GC and RP-HPLC, finds its root in different retention mechanisms. While the retention mechanism in GC can be understood, to a high degree, as a partition mechanism, this is not true for RP-HPLC. The retention process in RP-HPLC is much less defined, and difficult to describe. The RI values cannot correct completely for the specific interactions between solute, stationary phase and modifier. The RI values are still sensitive to column selectivity differences, to the kind of organic modifier and to other properties of the mobile phase.

A disadvantage of the RI values, the inability to correct retention values of compounds which are chemically not related to the used homologs, is partly solved by using relative capacity factors. A standard is chosen with a chemical structure similar to the compounds which have to be corrected. In this case some improvement is reported compared to the use of RI values.

The use of more than one standard compound is more promising. On the notion that there are three principal sample-solvent interactions[95] (electron donating, electron withdrawing and dipole) Smith[96] argues that three test compounds, together with the RI standards to test polarity effects, should be sufficient to characterise any reversed-phase chromatographic system. Although the use of corrected capacity factors based on more than one standard compound seems to be advantageous, the question which standards should be used is still open. The RI values of the column test compounds (toluene, 2-phenylethanol, p-cresol and nitrobenzene) showed less variation, due to changing stationary/mobile phase conditions, than the RI values of barbiturates[88] and local anaesthetics[89]. The optimality of the choice of these test compounds is questionable in reflecting the selective differences between the stationary/mobile phase combinations with respect to the compounds of interest: the barbiturates and local anaesthetics. Note that the calibration along the lines suggested by Smith[96] and performed by Bogusz[94] requires the measurement of at least seven calibration standards (four homologs and three barbiturates). Besides, some criticism on Snyders solvent selectivity concept is reported[12,13]. Perhaps a more sound classification scheme for solvents is given by Chastrette[97].

Chapter 8 New calibration strategies aimed at prediction of retention on new chromatographic systems

8.1 Two-way approaches

8.1.1 Strategies based on one stationary phase in the training set

Suppose capacity factors are available of r solutes at m mobile phase compositions on one stationary phase. This data set can be arranged conveniently in a data matrix X with m rows and r columns. As was stressed in Section 2.1, homoscedasticity (the measurement error of a variable is constant) is a convenient property of the variables if linear models are applied. Assuming that the relative error of a measured capacity factor is independent of the mobile phase composition, the specific solute and the stationary phase, a logarithmic transformation of k gives ln k values with a constant variance[98-100]. This assumption will be verified in forthcoming cases (Section 9.5 and 14.5). It is assumed that X consists of ln k values of the solutes.

Two new approaches are introduced. The first approach takes the solutes as variables, whereas the second approach takes the mobile phase compositions as variables.

The first approach is depicted in Figure II.5. The measurements of the markers are gathered in X[M], the measurements of the non-markers are gathered in X[NM]. The markers are selected with one of the techniques of Chapter 1. Note that the markers in this case only represent differences due to the mobile phase, because measurements on other stationary phases are not available. For illustrative purposes a number of four markers is assumed to be sufficient. Different modelspecifications relating the X[M]-block to the X[NM]-block are possible. Ln k values of each non-marker (in the X[NM]-block) can be related to the measurements of the markers (in the X[M]-block) with a linear equation (for the sake of simplicity, an integer indexing the non-marker is left out)

$$\ln k_j = \beta_0 + \beta_1 M1_j + \beta_2 M2_j + \beta_3 M3_j + \beta_4 M4_j + \epsilon_j \qquad \text{(II.3)}$$

where j indicates the mobile phase composition, $M1_j$ to $M4_j$ are the ln k values of the markers measured at the jth mobile phase composition, ϵ_j is an error term. The parameters β_0 to β_4 have to be estimated with one of the methods mentioned in Chapters 2 and 3. If measurements of the markers on a new stationary phase at one specific mobile phase composition (not necessarily one of the mobile phase compositions used on the first stationary phase) are available, the ln k value of the non-marker can be predicted at that mobile phase composition on the new stationary phase. For each non-marker a separate model is build and can be used to predict on the new stationary phase. This can be done for every mobile phase composition, provided that the measurements of the markers at this mobile phase composition are available. It is also possible to relate all ln k values in X[NM] directly to X[M], with the use of, e.g., partial least squares (see Section 3.4).

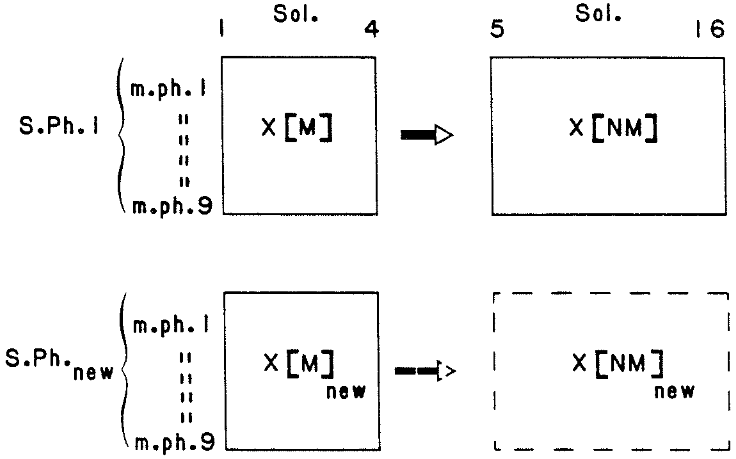


Figure II.5. One stationary phase in the training set, solutes treated as variables. In X[M], the ln k values of the four markers on the stationary phase (S.Ph.) at the nine mobile phases (m.ph.) are gathered. The ln k values of the non-markers, at the same measurement conditions as the markers, are gathered in X[NM].

Summarizing this approach, the relationship between the retention of the markers and non-markers on the first stationary phase is modelled. The same relationship is assumed to hold between the markers and non-markers on the new stationary phase.

The second approach is depicted in Figure II.6. The matrix X (r×m) is the transpose of the above mentioned X, for the sake of simplicity the same name is used. The mobile phase compositions are treated as variables. When applying this strategy, more markers have to be chosen, because the solutes are treated as objects. A number of five is minimally needed. These are chosen with the techniques of Chapter 1 on the basis of the measurements made on the only stationary phase. The number of variables in X[M] exceeds the number of objects, so a projection method like PLS is needed to relate the X[M]-block to the X[NM]-block. The corresponding model is:

$$X[M] = TP' + E \qquad (II.4)$$
$$X[M]_{new} = TQ' + F \qquad (II.5)$$

where T is a (5×g) matrix of scores on the g latent variables (g<5). P and Q are (m×g) matrices containing the loadings of X[M] and $X[M]_{new}$ on the latent variables, respectively. E and F are error matrices. In order to build the model, measurements of the ln k values of the markers have to be available on the new stationary phase at some mobile phase compositions (e.g. 7) not necessarily the same as used on the initial stationary phase. Prediction of ln k values of the non-markers on the new stationary phase are obtained at the (seven) mobile phase compositions chosen to calibrate that new stationary phase. This prediction is accomplished with the model and the measured ln k values of the non-markers on the original stationary phase. Summarizing this approach, the relationship between the behaviour of the markers on the original and on the new stationary phase is modelled and used to predict the behaviour of the non-markers on the new stationary phase.

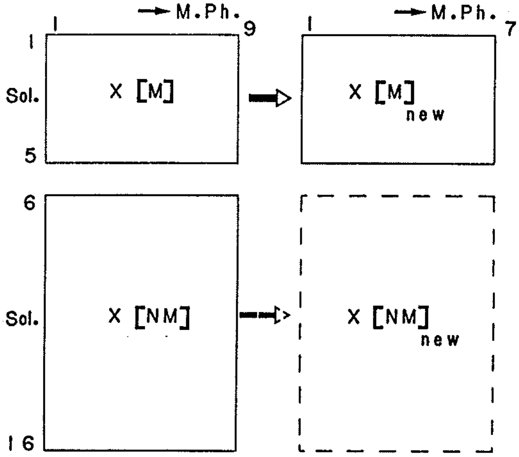


Figure II.6. One stationary phase in the training set, mobile phase compositions treated as variables. For abbreviations, see Figure II.5.

The relative merits of both approaches have to be established in future research. It is, e.g., not clear how the kind of scaling of the respective matrices effect predicting performance.

It seems that the above mentioned strategies are particularly of value if the problem of updating the optimal mobile phase composition on a deteriorating stationary phase is at hand. Only one - the original - stationary phase is available in the training set, after using this phase, it has to calibrated again with the use of the markers.

8.1.2 Strategies if two to four stationary phases are available in the training set

For the sake of illustration, suppose that measurements on two stationary phases are available. This data set can be depicted as in Figure II.7. The unfolding of this data set in the direction so that the mode of the solutes remains intact is shown in Figure II.8. The solutes are conceived as variables. This was also the case in the first strategy presented in Subsection 8.1.1, which can be thought of as a degenerate case of this unfolding. Markers can be selected with the help of the techniques described in Chapter 1.

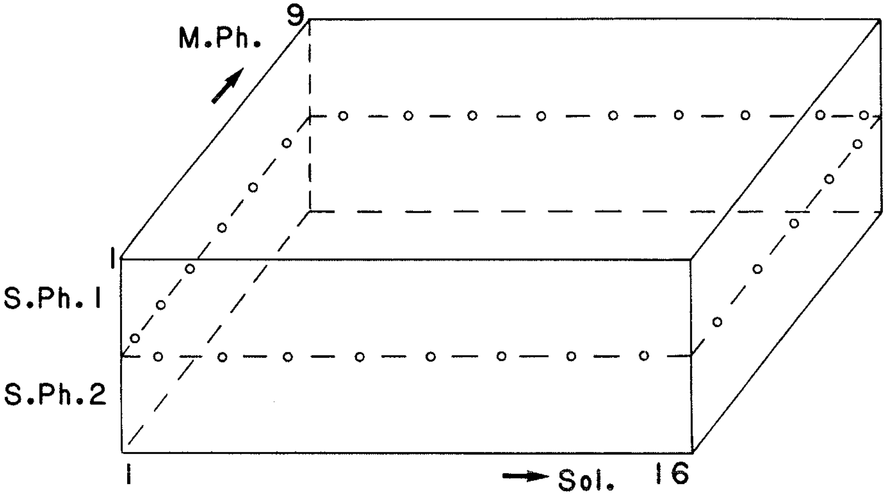


Figure II.7. Training set if measurements on two stationary phases are available. For abbreviations, see Figure II.5, Sol. is the abbreviation of solute.

The next step is the model-building step and is depicted in Figure II.9. The relationship between the ln k values of the markers and the non-markers can be modelled. This model can be used later to predict the retention of the non-markers on a new stationary phase, at an arbitrary mobile phase composition, provided that the retention values of the markers at that mobile phase have been measured, see Figure II.10. The specific form of the model is already explained in Subsection 8.1.1, see formula (II.3). An example in the area of the prediction of retention variability due to batch differences will be given in Chapter 15.

If measurements on three or four stationary phases are available, the matrix X can be augmented. The whole procedure remains the same. An example with three stationary phases in the training set is given in Chapters 10-12, where an attempt is made to incorporate explicitly the influence of the mobile phase composition in the model.

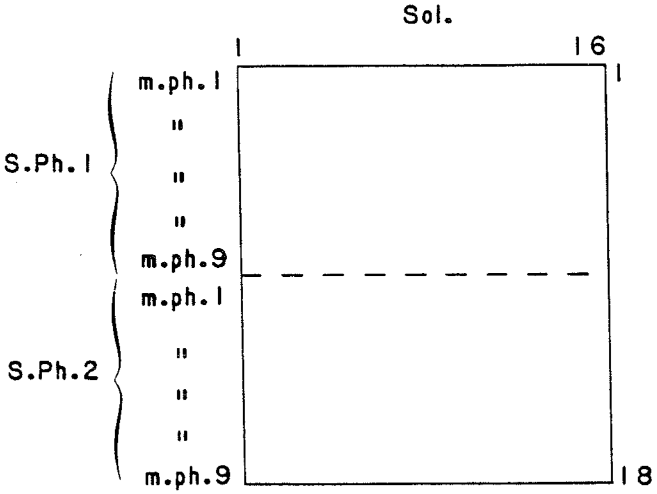


Figure II.8. Unfolded training set of Figure II.7. For abbreviations, see Figure II.7.

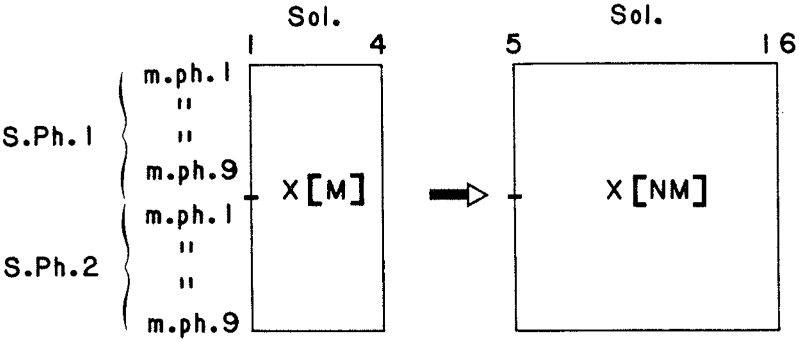


Figure II.9. Model if measurements on two stationary phases are available. For abbreviations, see Figures II.5 and II.7.

Analogous to the second strategy in Subsection 8.1.1, is the following. The first step is the choice of the markers, this can be done on a similar way as above, with the use of the matrix in Figure II.8. Next, the data cube is unfolded in such a way that the direction of the solutes remains intact. Suppose the ln k values of the markers are measured on a new stationary phase, not necessarily at the same mobile phase compositions as used on the original sta-

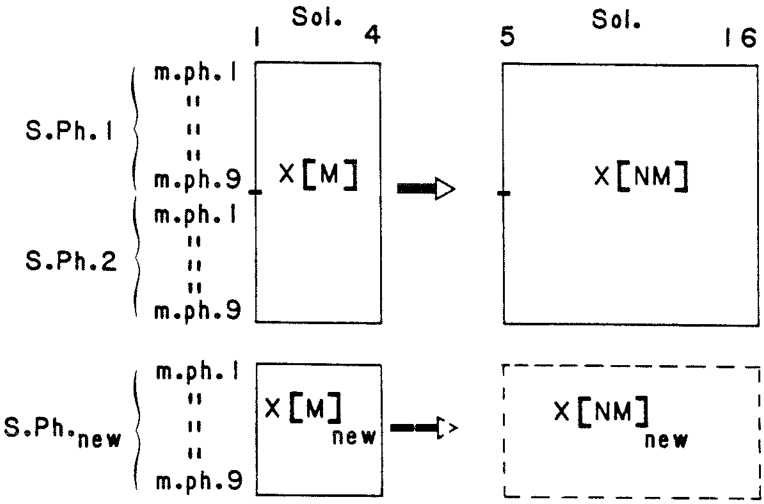


Figure II.10. Predictions for a new stationary phase if measurements are available on two stationary phases. For abbreviations, see Figure II.9. The subscript "new" refers to the stationary phase for which predictions are needed.

tionary phases. The data can be arranged as shown in Figure II.11. A model is made relating X[M] and $X[M]_{new}$, similar as in Section 8.1.1 (see formulas (II.4) and (II.5)).

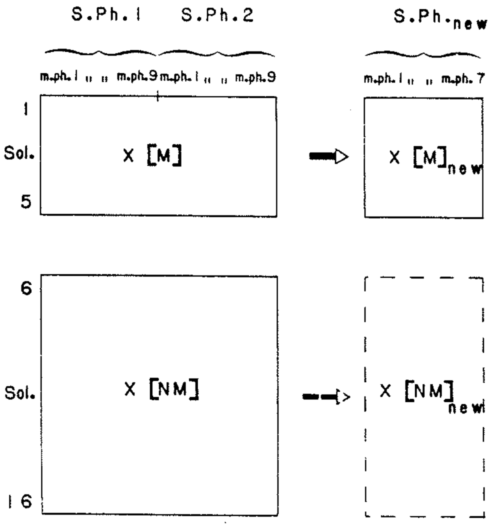


Figure II.11. Alternative model if measurements on two stationary phases are available. For abbreviations, see Figure II.10.

8.2 Three-way approaches: strategies if more than four stationary phases are available in the training set

Suppose ln k values of sixteen solutes are available on five stationary phases, at nine mobile phase compositions. This data set can be arranged in a data cube, as shown in Figure II.12. If the number of stationary phases is large enough, the way is open to use three-way decompositions to analyse this data cube. Note that this is impossible if too few stationary phases are available, because the rank of the data cube would become too low. Two possible alternatives to decompose this data cube are discussed: the unfold-PCA (and -PLS) and the PARAFAC solution (see Section 1.6).

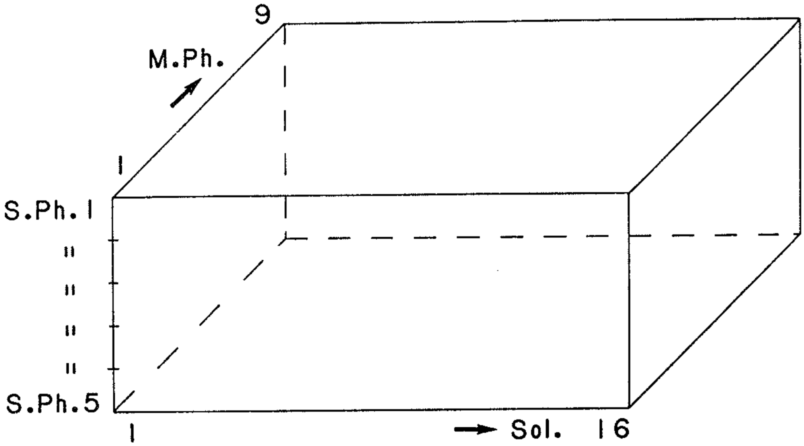


Figure II.12. Training set if measurements on five stationary phases are available. For abbreviations, see Figure II.7.

8.2.1 The unfold-PCA (and -PLS) solution

For the sake of simplicity, assume that four markers are chosen which have to be measured at four mobile phase compositions. The methods to select such combinations are presented in Section 1.6. A partition of the data cube of Figure II.12 can be made in such a way that the selected solute/mobile phase combinations are gathered in X[**MaMPh**] and the non-selected solute/mobile phase combinations constitute X[**NMaMPh**]. Note that the ln k values of the markers at the non-selected mobile phase compositions are also present in X[**NMaMPh**], as well as the ln k values of the non-markers at the selected mobile phase compositions, because only the ln k values of the markers at the selected mobile phase compositions are present in X[**MaMPh**]. The specific arrangement of all 128 (9x16-4x4) combinations in X[**NMaMPh**] is not important, as long as the arrangement is kept constant during all calculations. If PLS is used to model the relationship between **X**[MaMPh] and **X**[NMaMPh], the respective data cubes are decomposed in a summation of products of vectors times matrices, see Figure II.13.

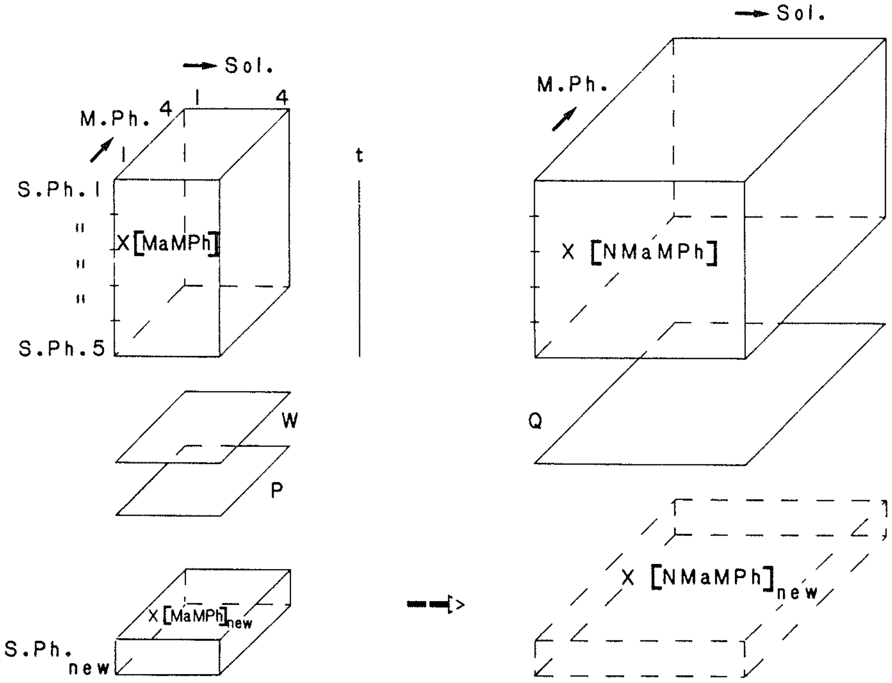


*Figure II.13. Unfold-PLS model. In **X[MaMPh]** the sixteen selected combinations of four markers, at four mobile phase compositions are gathered on all stationary phases. Similarly, the not selected marker/mobile phase combinations are gathered in **X[NMaMPh]**. For abbreviations, see Figure II.7 and the text.*

Considering only one dimension in the PLS model involves the calculation of t and W: scores and loadings on the first latent variable of the **X**[MaMPh]-block and u, Q: scores and loadings on the first latent variable of the X[NMaMPh]-block (see Section 3.4).

Let the typical element of W be w_{kj} ($k=1,...,4;j=1,...,4$), where k is the index of the markers and j the index of the mobile phase compositions. Measurements (ln k values) made on the new stationary phase, the sixth, are noted as x_{6kj}, where k and j are the same as in w_{kj}. Predictions of the retention of other mobile phase/solute combinations are obtained as follows. First the score of the new stationary phase on the latent variable, t_6, is calculated as $t_6=\mathbf{X}..\mathrm{W}"/\|\mathrm{W}\|^2$, where the notation for three-way arrays is used[101]. In this particular situation this is $t_6=\Sigma_{kj}(x_{6kj}w_{kj})/\Sigma_{kj}w_{kj}^2$. This score can be understood as the least squares solution of

$$x_{6kj} = w_{kj}t_6 + \epsilon_{6kj} \qquad \text{(II.6)}$$

because $t_6=(w'w)^{-1}w'x$, with $w=(w_{11},\ldots,w_{44})'$ and $x=(x_{11},\ldots,x_{44})'$ arranged in the same order, and ϵ_{6kj} is an error term. The **X**[NMaMPh] values of the new stationary phase are predicted by calculating t_6Q.

If the matrices **X**[MaMPh] and **X**[NMaMPh] are unfolded in such a way that the direction of the stationary phases is left intact, the result is shown in Figure II.14. Note that the arrangement of the solutes/mobile phase combinations in **X**[NMaMPh] is arbitrary as long as this arrangement is kept the same during the PLS calculations. If these two matrices are subjected to a normal PLS procedure (see Section 3.4), the result is exactly the same as above[101]. The loading matrix W becomes a loading vector w, the same w as defined below formula (II.6). The predictions are obtained according to the above mentioned scheme. This solution is called, therefore, the unfold-PLS solution.

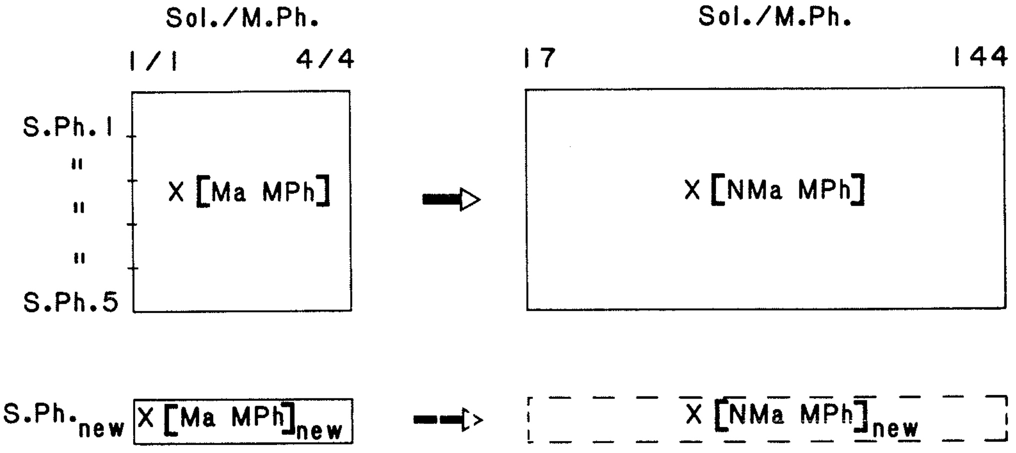


Figure II.14. Unfolding of the data cubes in Figure II.13. For abbreviations, see Figure II.13.

If more than one dimension in the PLS model is appropriate, the calculations are straightforward and comparable with the above sketched ones, see Section 3.4.

8.2.2 The PARAFAC solution

Explanation starts from the same figure as in Section 8.2.1: Figure II.12. The data cube is decomposed in a summation of products of vectors, see Section 1.6. For the sake of simplicity, it is assumed that one dimension is capable of describing sufficient variation in X. Let a,b and c be the modes describing, respectively, the stationary phases, the solutes and the mobile phase compositions. Suppose that, after rearranging, the selected markers and mobile phase compositions have loadings b_1 to b_4 and c_1 to c_4, respectively, on the first component. A new stationary phase has to be calibrated by measuring the markers at the selected mobile phase compositions (the "S.Ph.$_{new}$" block of Fig. II.15). Then a_6 can be calculated as the least squares solution of

$$x_{6kj} = b_k c_j a_6 + \epsilon_{6kj} \qquad \text{(II.7)}$$

where k=1,...,4 indexes the markers; j=1,...,4 indexes the selected mobile phase compositions and ϵ_{6kj} is an error term. The x_{6kj} values are the ln k outcomes of the measured markers at the selected mobile phase compositions. Predictions of the not selected solute/mobile phase combinations can be obtained by $x_{6kj}=b_k c_j a_6$, with k,j not equal to 1,1 ... 4,4 (see Fig. II.15). These b_k and c_j values are available from the primary decomposition. Note the resemblance between formulas (II.6) and (II.7), in both cases the x values are regressed in order to obtain a new score but the design matrices of the regressions differ.

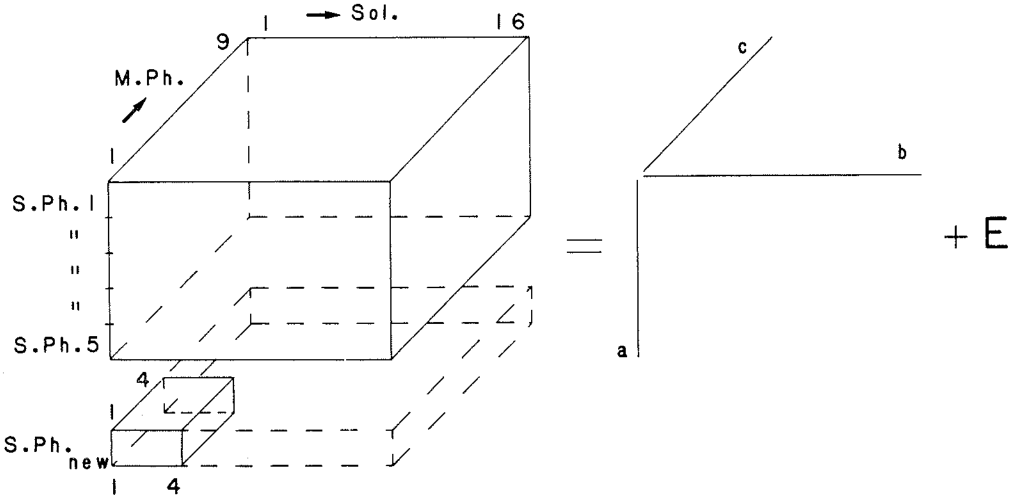


Figure II.15. The PARAFAC model with one component. For abbreviations, see Figure II.7 and the text.

If a decomposition of X with two components is wanted, then $a_1=(a_{1,1},...,a_{1,5})$, $a_2=(a_{2,1},...,a_{2,5})$, $b_1=(b_{1,1},...,b_{1,16})$, $b_2=(b_{2,1},...,b_{2,16})$, $c_1=(c_{1,1},...,c_{1,9})$ and $c_2=(c_{2,1},...,c_{2,9})$, with the first index indicating the component. Analogous to formula (II.7) is:

$$x_{6kj} = b_{1k}c_{1j}a_{16} + b_{2k}c_{2j}a_{26} + \epsilon_{6kj} \qquad \text{(II.8)}$$

where k, j, x_{6kj} and ϵ_{6kj} as in formula (II.7). The coefficients $a_{1,6}$ and $a_{2,6}$ are determined with multiple regression. Predictions of the not selected solute/mobile phase combinations can be done in a similar way as above.

For both the unfold-PLS and PARAFAC solution holds that if scaling the data is performed prior to the decomposition, rescaling of the predicted values is necessary to obtain predictions in the original scale.

References Part II

1 J.C. Berridge; *Techniques for the automated optimization of HPLC separations*, John Wiley, New York, 1985

2 P.J. Schoenmakers; *Optimization of chromatographic selectivity, Elsevier*, Amsterdam, 1986

3 A.K. Smilde, A. Knevelman and P.M.J. Coenegracht; *Introduction of multi-criteria decision making in optimization procedures for high performance liquid chromatography*, J.Chromatogr., 369 (1986) 1-10

4 P.M.J. Coenegracht, A.K. Smilde and A. Knevelman; *Performance characterization of multi-solvent mobile phase systems in RP-HPLC by multi-criteria decision making illustrated by the comparison of ternary and quaternary solvent systems*, J.Liquid Chromatogr., 12(1) (1989) 77-94

5 P.M.J. Coenegracht, H.J. Metting, A.K. Smilde and P.J.M. Coenegracht-Lamers; *A chemometric investigation of the selectivity of multisolvent mobile phase systems in RP-HPLC*, Chromatographia, 27(2/3) (1989) 135-141

6 V.V. Fedorov; *Theory of Optimal Experiments*, Ac.Press, New York, 1972

7 J.P. Bianchini, E.M. Gaydou, A.M. Siouffi, G. Mazerolles, D. Mathieu and R. Phan Tan Luu; *Optimization of the separation of polymethoxylated flavones in reversed phase liquid chromatography*, Chromatographia, 23(1) (1987) 15-20

8 D.P. Herman, H.A.H. Billiet and L. de Galan; *Statistical basis to select appropriate starting eluent compositions for solvent optimization in isocratic reversed phase liquid chromatography*, Anal.Chem., 58 (1986) 2999-3006

9 A.G. Wright, A.F. Fell and J.C. Berridge; *Strategies for automated Optimisation of high-performance liquid chromatographic separations incorporating Diode Array detection*, J.Chromatogr., 458 (1988) 335-353

10 J.K. Strasters, F. Coolsaet, A. Bartha, H.A.H. Billiet and L. de Galan; *Peak-tracking and subsequent choice of optimization parameters for the HPLC-separation of a mixture of local anaesthetics*, J.Chromatogr., submitted for publication

11 P.J. Schoenmakers; *Criteria for comparing the quality of chromatograms with great variations in capacity factors*, J.Liquid Chromatogr., 10 (1987) 163

12 S.D. West; *The prediction of reversed phase HPLC retention indices and resolution as a function of solvent strength and selectivity*, J.Chrom.Sci., 25 (1987) 122-129

13 S.D. West; *Correlation of retention indices with resolution and selectivity in reversed phase HPLC and GC*, J.Chrom.Sci., 27 (1989) 2-12

14 M.A. Quarry, R.L. Grob, L.R. Snyder, J.W. Dolan and M.P. Rigney; *Band spacing in reversed phase high performance liquid chromatography as a function of solvent strength. A simple and fast alternative to solvent optimization for method development*, J.Chromatogr., 384 (1987) 163-180

15 P. Jandera; *A method for characterization and optimization of RP-LC separations based on retention behaviour in homologous series*, Chromatographia, 19 (1984) 101

16 P. Jandera; *Reversed-phase liquid chromatography of homologous series. A general method for prediction of retention*, J.Chromatogr., 314 (1984) 13-36

17 P. Jandera; *Method for characterization of selectivity in reversed phase liquid chromatography I Derivation of the method and verification of the assumptions*, J.Chromatogr., 352 (1986) 91-110

18 P. Jandera; *Method for characterization of selectivity in reversed phase liquid chromatography II Possibilities for the prediction of retention data*, J.Chromatogr., 111-126

19 S. Wold and K. Andersson; *Major components influencing retention indices in gas chromatography*, J.Chromatogr., 80 (1973) 43-59

20 S. Wold; *A theoretical foundation of extrathermodynamic relationships (linear free energy relationships)*, Chemica Scripta, 5 (1974) 97-106

21 D.G. Howery, G.D. Williams and N. Ayala; *Predicting retention data by target factor analysis and multiple regression analysis*, Anal.Chim.Acta., 189 (1986) 339-351

22 R. Fellous, L. Lizzani-Cuvelier and R. Luft; *Predicting retention data in Gas-Liquid chromatography by multivariate analysis*, Anal.Chim.Acta, 174 (1985) 53-60

23 C.H. Lochmuller, S.J. Breiner, C.H. Reese and M.N. Koel; *Characterization and prediction of retention behaviour in reversed-phase liquid chromatography using factor analytical modeling*, Anal.Chem.,61 (1989) 367-375

24 K. Jinno, M. Yamagami and M. Kuwajima; *Retention prediction and computer assisted optimization for the separation of PTH-amino acids in isocratic reversed phase liquid chromatography*, Chromatographia, 25(11) (1988) 974-982

25 R. Kaliszan; *Quantitative Structure-Chromatographic Retention Relationships*, John Wiley, New York, 1987

26 R.K. Iler; *The Chemistry of Silica*, John Wiley, 1979

27 L.C. Sander and S.A. Wise; *Influence of substrate parameters on column selectivity with alkyl bonded-phase sorbents*, J.Chromatogr., 316 (1984) 163-181

28 J. Kohler, D.B. Chase, R.D. Farlee, A.J. Vega and J.J. Kirkland; *Comprehensive characterization of some silica-based stationary-phases for high-performance liquid chromatography*, J.Chromatogr., 352 (1986) 275-305

29 J. Kohler and J.J. Kirkland; *Improved silica-based column packings for high-performance liquid chromatography*, J.Chromatogr., 385 (1987) 125-150

30 J. Nawrocki and B. Buszewski; *Influence of silica surface chemistry and structure on the properties, structure and coverage of alkyl-onded phases for high-performance liquid chromatography*, J.Chromatogr., 449 (1988) 1-24

31 G.E. Berendsen; *Preparation and characterization of well-defined chemically bonded stationary phases for high pressure liquid chromatography*, Thesis, Delft, 1980

32 K. Jones; *Optimisation procedure for the silanisation of silicas for reversed-phase high-performance liquid chromatography. I Elimination of non-significant variables*, J.Chromatogr., 392 (1987) 1-10

33 K. Jones; *Optimisation procedure for the silanisation of silicas for reversed-phase high-performance liquid chromatography II Detailed examination of significant variables*, J.Chromatogr., 392 (1987) 11-16

34 R.M. Smith, G.A. Murrila, T.G. Hurdley, R. Gill and A.C. Moffat; *Retention Reproducibility of thiazide diuretics and related drugs in reversed-phase high-performance liquid chromatography*, J.Chromatogr., 384 (1987) 259-278

35 S.A. Wise and W.E. May; *Effect of C18 surface coverage on selectivity in reversed-phase liquid chromatography of polycyclic aromatic hydrocarbons*, Anal.Chem., 55 (1983) 1479-1485

36 L. Hanson and L. Troyer; *Characterization of Reversed-phase high performance liquid chromatographic stationary phases by means of pyrolysis gas chromatography*, J.Chromatogr., 207 (1981) 1-11

37 R.M. Smith, T.G. Hurdley, R. Gill and A.C. Moffat; *The application of retention indices using the alkylarylketone scale to the separation of the barbiturates by High performance liquid chromatography II The effect of the stationary phase*, Chromatographia, 19 (1984) 407-410

38 J.G. Atwood and J. Goldstein; *Testing and quality control of a reversed-phase column packing*, J.Chrom.Sci., 18 (1980) 650-655

39 A.P. Goldberg; *Comparison of columns for reversed-phase liquid chromatography*, Anal.Chem., 54 (1982) 342-345

40 L.C. Sander and S.A. Wise; *Synthesis and characterization of polymeric C18 stationary phases for liquid chromatography*, Anal.Chem., 56(3) (1984) 504-510

41 A.K. Smilde, C.H.P. Bruins, D.A. Doornbos and J. Vink; *Optimization of the reversed-phase high performance liquid chromatographic separation of synthetic estrogenic and progestogenic steroids using the multi-criteria decision making method*, J.Chromatogr., 410 (1987) 1-12

42 K.K. Unger; *Porous Silica*, Elsevier, Amsterdam, 1979

43 E. Bayer and A. Paulus; *Silanophilic interactions in reversed phase high performance liquid chromatography*, J.Chromatogr., 400 (1987) 1-4

44 E. Bayer, K. Albert, J. Reiners, M. Nieder and D. Muller; *Characterization of chemically modified silica gels by* ^{29}Si and ^{13}C cross-polarization and magic angle spinning nuclear magnetic resonance, J.Chromatogr., 264 (1983) 197-213

45 H.A. Claessens, L.J.M. van de Ven, J.W. de Haan and C.A. Cramers; *Correlations between HPLC and NMR properties of some selected alkyl bonded phases*, J.High Res.Chrom. & Chrom.Comm., 6 (1983) 433-435

46 A. Nahum and C. Horvath; *Surface silanols in silica-bonded hydrocarbonaceous stationary phases I Dual retention mechanism in reversed phase chromatography*, J.Chromatogr., 203 (1981) 53-63

47 K.E. Bij, C. Horvath, W.R. Melander and A. Nahum; *Surface silanols in silica-bonded hydrocarbonaceous stationary phases II Irregular retention behaviour and effect of silanol masking*, J.Chromatogr., 203 (1981) 65-84

48 B. Walczak, M. Lafosse, J.R. Chretien, M. Dreux and L. Morin-Allory; *Factor analyis and experiment design in hplc V. Selectivity of chalcone configuration isomers on 23 reversed phase packings*, J.Chromatogr., 369 (1986) 27-37

49 D.C. Leach, M.A. Stadalius, J.S. Berus and L.R. Snyder; *Reversed-phase HPLC of basic samples*, LC-GC 1(5) (1988) 22-30

50 P.C. Sadek, P.W. Carr and L.W. Bowers; *The significance of metallophilic and silanophilic interactions in Reversed Phase HPLC*, J.Liquid Chrom., 8(13) (1985) 2369-2386

51 G.E. Berendsen and L. de Galan; *Preparation and chromatographic properties of some chemically bonded phases for reversed phase liquid chromatography*, J.Liquid Chrom., 1 (1978) 561

52 J.J. Kirkland, J.L. Glajch and R.D. Farlee; *Synthesis and characterization of highly stable bonded phases for high performance liquid chromatography column packings*, Anal.Chem., 61 (1989) 2-11

53 H.A. Claessens, C.A. Cramers, J.W. de Haan, F.A.H. den Otter, L.J.M. van de Ven, P.J. Andree, G.J. de Jong, N. Lammers, J. Wijma and J. Zeeman; *Ageing processes of alkyl bonded phases in HPLC; a chromatographic and spectroscopic approach*, Chromatographia, 20(10) (1985) 582-586

54 N. Sagliano, R.A. Hartwick, R.E. Patterson, B.A. Woods, J.L. Bass and N.T. Miller; *Stabilization of reversed phases for liquid chromatography. Application of infrared spectroscopy for the study of bonded-phase stability*, J.Chromatogr., 458 (1988) 225-240

55 H.A. Claessens, J.W. de Haan, L.J.M. van de Ven, P.C. de Bruyn and C.A. Cramers; *Chromatographic and solid state nuclear magnetic resonance study of the changes in reversed phase packings for hplc at different eluent compositions*, J.Chromatogr., 436 (1988) 345-365

56 R.P.W. Scott and C.F. Simpson; *Solute-solvent interactions on the surface of reversed phases I Stationary phase interactions and their dependence on bonding characteristics*, J.Chromatogr., 197 (1980) 11-20

57 J. Stahlberg and M. Almgren; *Polarity of chemically modified silica surfaces and its dependence on mobile-phase composition by fluorescence spectrometry*, Anal.Chem., 57 (1985) 817-821

58 E. Bayer, A. Paulus, B. Peters, G. Laupp, J. Reiners and K. Albert; *Conformational behaviour of alkyl chains of reversed phases in high performance liquid chromatography*, J.Chromatogr., 364 (1986) 25-37

59 M.E. McNally and L.B. Rogers; *Examination of the effect of solvent composition on bonded phase liquid chromatography packings by ^{13}C fourier transform nuclear magnetic resonance spectroscopy*, J.Chromatogr., 331 (1985) 23-32

60 K. Jinno, T. Nagoshi, N. Tanaka, M. Okamoto, J.C. Fetzer and W.R. Briggs; *Elution behaviour of peropyrene-type polycyclic aromatic hydrocarbons in various chemically bonded stationary phases in reversed phase liquid chromatography*, J.Chromatogr., 386 (123-135)

61 W.T. Cooper and L.Y. Lin; *Effects of stationary phase polarity on retention in reversed bonded phase HPLC column*, Chromatographia, 21(6) (1986) 335-341

62 C.H. Lochmuller, H.H. Hangac and D.R. Wilder; *The effect of bonded ligand structure on solute retention in reversed phase high performance liquid chromatography*, J.Chrom.Sci., 19 (1981) 130-136

63 A.L. Colmsjo and M.W. Ericsson; *Synthesis and evaluation of selective reversed phase packing materials for HPLC*, Chromatographia, 21(7) (1986) 392-396

64 S.A. Tomellini, S.H. Hsu, S.D. Fazio and R.A. Hartwick; *Non-linear effects of TMS capping on dimethyloctyl HPLC stationary phases*, J.High Res.Chrom. & Chrom.Comm., 8 (1985) 337-340

65 M.F. Delaney, A.N. Papas and M.J. Walters; *Chemometric Classification of reversed phase hplc columns*, J.Chromatogr., 410 (1987) 31-41

66 T. Welsch, H. Frank, H. Zwanziger, S. Liebisch and W. Engewald; *Classification of reversed phase materials by means of selected chromatographic data*, Chromatographia, 19 (1984) 457-461

67 P.E. Antle, A.P. Goldberg and L.R. Snyder; *Characterization of silica-based reversed phase columns with respect to retention slectivity. Solvophobic effects*, J.Chromatogr., 321 (1985) 1-32

68 P.E. Antle and L.R. Snyder; *Selecting columns for reversed phase HPLC Part I: Column selectivity*, LC mag.,2 (1984) 840

69 L.R. Snyder and J.J. Kirkland; *Introduction to Modern Liquid Chromatography*, Wiley Interscience, New York, 2nd ed., 1979

70 B. Walczak, M. Dreux, J.R. Chretien, K. Szymoniak, M. Lafosse, L. Morin-Allory and J.P. Doucet; *Factor Analysis and experiment design in hplc I. Trends in selectivity of 53 chalcones in reversed phase hplc on alkyl- or phenyl-bonded stationary phases*, J.Chromatogr., 353 (1986) 109-121

71 B. Walczak, J.R. Chretien, M. Dreux, L. Morin-Allory, M. Lafosse, K. Szymoniak and F. Membrey; *Factor Analysis and experiment design in hplc II Normal phase hplc on columns of different polarities*, J.Chromatogr., 353 (1986) 123-137

72 B. Walczak, L. Morin-Allory, J.R. Chretien, M. Lafosse and M. Dreux; *Factor Analysis and experiment design in HPLC III Influence of mobile phase modifications on the selectivity of chalcones on a diol stationary phase*, Chem.and Intel.Lab. Systems, 1 (1986) 79-90

73 B. Walczak, J.R. Chretien, M. Dreux, L. Morin-Allory and M. Lafosse; *Factor Analysis and experiment design in hplc IV Influence of mobile phase modifications on the selectivity of chalcones on an ODS phase*, Chem.and Intel.Lab.Systems., 1 (1987) 177-189

74 J.R. Chretien, B. Walczak, L. Morin-Allory, M. Dreux and M. Lafosse; *Factor Analysis and experiment design in hplc VI Comparison of the retention mechanism for fourteen octadecyl-silica packings*, J.Chromatogr., 371 (1986) 253-267

75 B. Walczak, L. Morin-Allory, M. Lafosse, M. Dreux and J.R. Chretien; *Factor Analysis and experiment design VII Classification of 23 reversed phase high-performance liquid chromatographic packings and identification of factors governing selectivity*, J.Chromatogr., 395 (1987) 183-202

76 K.V. Mardia, J.T. Kent and J.M. Bibby; *Multivariate Analysis*, Ac.Press, 1979, pages 237-239

77 L.Y. Lin and W.T. Cooper; *Isocratic high-performance liquid chromatographic separation and multiple-wavelength ultraviolet detection of aldicarb and its soil degradation products. Optimization of stationary phase selectivity*, J.Chromatogr., 390 (1987) 285-295

78 J.L. Glajch and J.J. Kirkland; *Optimization of selectivity in liquid chromatography*, Anal.Chem., 55(2) (1983) 319A

79 J.L. Glajch, J.C. Gluckman, J.G. Charikofski, J.M. Minor and J.J. Kirkland; *Simultaneous selectivity optimization of mobile and stationary phases in reversed phase liquid chromatography for isocratic separations of phenylthiohydantoin amino acid derivatives*, J.Chromatogr., 318 (1985) 23-39

80 J.L. Glajch, J.J. Kirkland, K.M. Squire and J.M. Minor; *Optimization of solvent strength and selectivity for RP-LC using an interactive mixture-design statistical technique*, J.Chromatogr., 199 (1980) 57

81 E. Kovats; Helv.Chim.Acta, 41 (1958) 1915

82 D.S. Galanos and V.M. Kapoulas; J.Chromatogr., 13 (1964) 128

83 A.C. Moffat; *The standardisation of Thin layer chromatographic systems for the identification of basic drugs*, J.Chromatogr., 110 (1975) 341

84 J.H. Dhont, C. Vinkenborg, H. Compaan, F.J. Ritter, R.P. Labadie, A. Verweij and R.A. de Zeeuw; *Application of R_f correction in Thin-layer chromatography by means of two reference R_f values*, J.Chromatogr., 47 (1970) 376-381

85 R.M. Smith; *Retention Indices in Reversed-Phase HPLC*, In: Advances in Chromatography (eds J.C. Giddings, E. Grushka and Ph.R. Brown), vol 26, Marcel Dekker, New York, pp 277-319

86 J.K. Baker and C.Y. Ma; *Retention index scale for liquid-liquid chromatography*, J.Chromatogr., 169 (1979) 107

87 R.M. Smith; *Alkylarylketones as a retention index scale in liquid chromatography*, J.Chromatogr., 236 (1982) 313

88 R.M. Smith, T.G. Hurdley, R. Gill and A.C. Moffat; *The application of retention indices using the alklyarlyketone scale to the separation of the barbiturates by HPLC I. The effect of the eluent*, Chromatographia, 19 (1984) 401

89 R.M. Smith, T.G. Hurdley, R. Gill and A.C. Moffat; *Application of retention indices based on the alkylarylketone scale to the separation of the local anaesthetic drugs by hplc*, J.Chromatogr., 355 (1986) 75

90 R.M. Smith, G.A. Murilla and C.M. Burr; *Alkylarylketone index scale with acetonitrile or tetrahydrofuran containing eluents in reversed phase hplc*, J.Chromatogr., 388 (1987) 37

91 R.M. Smith; *Characterization of reversed phase liquid chromatography columns with retention indexes of standards based on an alkyl aryl ketone scale*, Anal.Chem., 56 (1984) 256

92 M. Bogusz and R. Aderjan; *Improved standardization in reversed phase hplc using 1-nitroalkanes as a retention index scale*, J.Chromatogr., 435 (1988) 43

93 R. Gill, A.C. Moffat, R.M. Smith and T.G. Hurdley; *A collaborative study to investigate the retention reproducibility of barbiturates in HPLC with a view to establishing retention databases for drug identification*, J.Chrom.Sci., 24 (1986) 153

94 M. Bogusz; *Correction of retention index values in hplc as a tool for comparison of results obtained with different octadecyl silica phases*, J.Chromatogr., 387 (1987) 404

95 L.R. Snyder; *Classification of the solvent properties of commom liquids*, J.Chrom.Sci., 16 (1978) 223

96 R.M. Smith; *the use of retention indices to measure the solvent selectivity of ternary eluents in reversed phase liquid chromatography*, J.Chromatogr., 324 (1985) 243

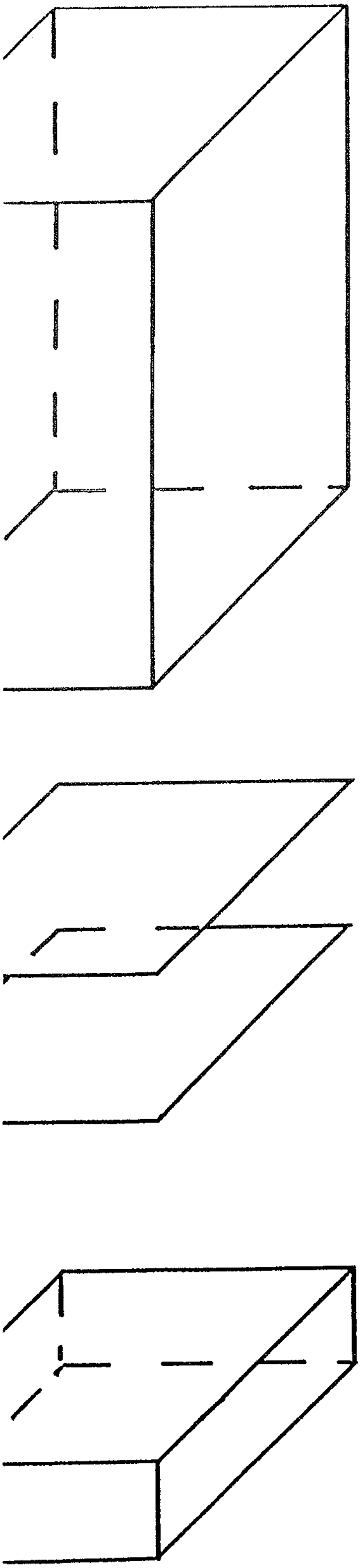
97 M. Chastrette, M. Rajzmann, M. Channon and K.F. Purcell; *Approach to a general classification of solvents using a multivariate statistical treatment of quantitative solvent parameters*, J.Amer.Chem.Soc., 107(1) (1985) 1-11

98 J.W. Weyland; *Strategies for mobile phase optimization in chromatography. A chemometrical approach*, Thesis, Groningen, 1986, page 61

99 S.T. Balke; *Quantitative column liquid chromatography*, Elsevier, Amsterdam, 1984, page 43

100 G.E.P. Box and N.R. Draper; *Emperical model building and response surfaces*, John Wiley, New York, 1987, page 284

101 S. Wold, P. Geladi, K. Esbensen and J. Öhman; *Multi-way Principal Components - and PLS analysis*, J.Chemometrics, 1 (1987) 41-56



PART III
Calibration of Various Types of Stationary Phases

Chapter 9 Experimental design

9.1 The choice of the stationary phases

Six stationary phases with varying ligands were chosen to build up a data set. In order to obtain stationary phases with a similar base silica structure, all stationary phases consisted of Spherisorb materials (Phase-Sep). These Spherisorb stationary phases were based on spherical particles with a mean particle size of 5 μm. Obviously, if different batches of base silica materials are used prior to the modification of the phases, the idea of a similar base silica has to be released to some extent.

The six stationary phases were modified with trimethylsilyl bonded to the silica (C1); hexylsilyl bonded to the silica, fully capped (C6); octylsilyl bonded to the silica, fully capped (C8); octadecylsilyl bonded to the silica, fully capped (C18); cyanopropylsilyl bonded to the silica (CN) and phenylsilyl bonded to the silica, partially capped (PHE).

These six stationary phases were considered diverse enough to represent different selectivities and to test the strategies mentioned in Section 8.2.

9.2 The choice of the mobile phase compositions

In order to develop calibration schemes for reversed-phase high performance liquid chromatography, relatively simple mobile phase compositions were selected. The exact compositions are given in Table III.1. The mobile phase compositions consist of binary and ternary mixtures of water, methanol and acetonitrile, regularly spread in the mixture triangle (see Figure I.11). The compositions were chosen in such a way that the capacity factors of the test solutes (see Section 9.3) ranged between one and thirty roughly. Similar mobile phase compositions were used on all stationary phases. This was done in order to avoid the problem of non-orthogonality in the design. This problem is indicated in Chapter 4.

Note that two aspects of a mobile phase composition were incorporated in this design. First, moving from wm1 to wm2 the eluent becomes weaker. The same holds for wa1 to wa2 and am1 to am2. Second, the kind of organic modifier changes when moving from a water/ methanol to a water/acetonitrile mixture.

9.3 The choice of the test solutes

Nine test solutes were chosen. These solutes were selected because of their relatively easy detectability with an UV variable wavelength detector; their expected reasonable retention behaviour with the selected mobile phase compositions (Section 9.2) and their frequent use in other studies on calibration of reversed-phase chromatographic systems (Chapter 6).

The nine selected testsolutes were acetophenone (ACP), n-butyl-4-aminobenzoate (BAB), ethyl-4-hydroxybenzoate (EHB), paracetamol (PAR), 2-phenylethanol (PE), toluene (TOL), n-propyl-4-hydroxybenzoate (PHB), ethyl-4-aminobenzoate (EAB) and methyl-4-hydroxybenzoate (MHB).

9.4 Experimental details

Methanol (MeOH) was of analytical grade and acetonitrile (ACN) was of chromatographic quality (both from Merck). Distilled water was used throughout. All test solutes were of pharmaceutical quality and obtained from various firms. The column dimensions were 150mm x 4.6mm (length x i.d.).

The instrument used was a Waters 6000A HPLC-pump fitted with a Kratos Spectroflow 757 variable wavelength detector (operating at a wavelength of 205 nm), an injection valve (Valco) fitted with a 20 μ-liter loop and an Omniscribe Recorder (Housten Instrument).

The dead time was measured as the retention time of a 5 μ-molar sodiumnitrate solution (in water). It ranged from 60 seconds for the C6 column to 68 seconds for the C1 and CN columns. The flow rate was 1.0 ml/min. The concentrations of the injected solutes were 0.02 mg/ml, except for toluene (0.04 mg/ml) and paracetamol (0.01 mg/ml).

Data acquisition and integration was performed by an Autolab System IVb Chromatography Data Analyzer (Spectra-Physics). The PLS and OLS calculations were performed on an Olivetti M-24 Personal Computer using the programs SIMCA (Sepanova, Sweden) and CLAS (University Centre for Pharmacy, The Netherlands). The selection of the markers was done with software written in Fortran V on the Cyber 170/760 computer of CDC at the Groningen University Computing Centre and on an Olivetti M-24 computer. The ridge and Stein calculations were performed on an Olivetti computer with software written in Fortran V. The PARAFAC calculations were performed on a Cyber with existing software developed by Harshman (see Section 1.6).

9.5 Reproducibility and Repeatability

The measurements were performed consecutively on each stationary phase so that, for practical reasons, no randomized design was used. On each stationary phase measurements started with the test mixture wm2, which consisted of a binary water/methanol (0.55/0.45 v/v) mixture. Then the other mobile phases were used. At the end of the series of measurements on a specific stationary phase, each solute was chromatographed again at the wm2 test mixture. The retention measurement of each solute was immediately repeated at each mobile phase mixture for three times. The repeatability of the retention of a solute at a specific stationary/mobile phase combination is defined as the standard deviation of those three repeated measurements. This repeatability can be defined in terms of k values or ln k values. The reproducibility of the retention of a solute on a stationary phase at the wm2 mixture is defined as the standard deviation of the two mean k values (or mean ln k values) that result from the three repeated measurements at mixture wm2 at the beginning and at the end of the series of measurements. It should be emphasized that this method for estimating the reproducibility is biased upwards. The change of a stationary phase is at its maximum between the beginning and the end of the series of measurements thereby yielding an estimate of the reproducibility which is too high. Actually, reproducibility defined in the way above is probably rather a yardstick of drift in the measurements than an indication of differences due to random causes. Another aspect that should be pointed out in this context is that the dependence of the reproducibility on mobile phase composition is not known and cannot be estimated with this experimental set-up.

In Table III.1 the capacity factors of the test solutes at the mobile phase compositions are given. Table III.2 states the reproducibility results in terms of k values. For every solute on each stationary phase, the mean k (or mean ln k) was higher at the start of the series than at the end of it, which indicates drift. It can be observed that reproducibility depends on the solute (a slow eluting compound has a worse reproducibility than a fast eluting one) and on the stationary phase.

The repeatability of the measurements (Table III.3) is much better than the reproducibility. This can be seen by comparing the standard deviations in the wm2 rows in Table III.3 with the corresponding rows in Table III.2. A pattern emerges from Table III.3. When paired mobile phase mixtures, that differ only in elution strength (wm1 and wm2, wa1 and wa2, am1 and am2), are compared, the strongest mixture (the "1" mixtures) generally has the best repeatability. This pattern breaks down for the CN phase and is somewhat less clear for the PHE phase, so that these phases can be considered as exceptions. A second, though vague, pattern is the dependence of the repeatability on the solute: a slower eluting compound has a worse repeatability.

If the capacity factor of a specific solute measured at different stationary/mobile phase combinations participates in the model-building proces, a constant absolute error is assumed. The error in the capacity factor depends on the mobile phase as was demonstrated earlier. This dependence can be removed by assuming a constant relative error in the capacity factor. By using a logarithmic transformation the constant relative error is transformed in an (almost) constant absolute error, see Balke[1], Weyland[2] and Box and Draper[3]. In Table III.4 the logarithms (base e logarithms) of the k values of the solutes are given. The corresponding reproducibilities and repeatabilities are given in the Tables III.5 and III.6. The reproducibilities (Table III.5) differ less than in Table III.2. The repeatabilities (Table III.6) show less dependence on the elution strength than in Table III.3, although complete dependence has not been removed.

Assuming that repeatability, in terms of ln k values, does not depend on the elution strength, the pooled standard deviations can be calculated for each solute on each stationary phase. These values are shown in Table III.6 and indicate no clear dependence on the stationary phase or the solute. The overall pooled repeatability is 0.013. It can, therefore, be concluded that by applying a logarithmic transformation a more homogeneous error structure will be obtained.

Another argument for using a logarithmic transformation can be found in Table III.7. In that table the moments of the retention measurements of the solutes measured in k values and in ln k values are compared. A normal distribution, for each solute, is approximated more when ln k values instead of k values are used, because a smaller skewness is obtained for the ln k values. Ideally, in case of perfect normality, the skewness is zero. The kurtosis, ideally zero for the normal case, is slightly worse however for the ln k values. In spite of the slightly worse kurtosis for the ln k values, a logarithmic transformation is performed because it renders a better skewness and because of the reason mentioned above regarding the constant absolute error.

A very serious problem was encountered when the PHE stationary phase was used. The measurement procedure as described above was applied to the PHE stationary phase. However, when the solutes were chromatographed at test mixture wm2 for the second time, a serious decrease of retention for every solute was observed. Table III.8 and III.9 summarize this problem. The whole procedure was repeated for the PHE stationary phase, but now at more instances in the measurement of the retention values on the PHE phase more wm2 test mixtures were used to evaluate that stationary phase. The results were much better. This last series was used and the results are given in Tables III.1 to III.6. By comparing Table III.8 with Table III.3 it can be inferred that the repeatability is in the same order of magnitude for the first (bad) series (Table III.8) and the good series (Table III.3). The reproducibility however, was much worse for the bad series (Table III.8) than for the good one (Table III.2); it differs by a factor four or five. The same can be said when comparing Table III.9 with Tables III.5 and III.6, although in that case the differences are less pronounced. Evidently, the PHE stationary phase was quite deteriorated which is alarming because the treatment of the phase was very gentle ie. no buffers with a high or low pH were used. The results for the second series, however, were good enough to use.

This extraordinary behaviour of the PHE phase was observed in a milder form on the other stationary phases. As mentioned earlier in the discussion on the reproducibility of the measurements, retention measured at the test mixture wm2, always decreases after measureming the whole series. This has a direct consequence for the consistency of the data set. It is therefore of utmost importance to minimize the number of measurements used to build the training set and to characterize a new stationary phase. A simultaneous decrease in the number of solutes and mobile phase compositions to process is necessary.

9.6 Univariate description of the data set

The univariate description of the data set is divided into three parts. The first part describes the dependence of the retention of the solutes on varying mobile phase compositions while the stationary phases are fixed. The second part describes the reverse: the dependence of the retention of the solutes on the different stationary phases at fixed mobile phase compositions. The last part is dedicated to the special role of the solute paracetamol.

Figures III.1a to III.1f give an impression of the dependence of some solutes (representative for the whole data set) on varying mobile phases with fixed stationary phases. Two different comparisons can be made by examining these figures. Comparison of the corresponding mobile phase mixtures indicated with "1" and "2" shows that retention usually increases when changing from the "series 1" mixtures to the "series 2" mixtures. This is due to a decrease of elution strength in that direction and is in accordance with reversed-phase theory[4]. This pattern breaks down for the solute PAR which is discussed later on. Inferences regarding the selectivity of a varying mobile phase in relation to a particular stationary phase, can be made by realizing that the distance of two lines in Figures III.1a to III.1f, measured in the ln k direction, is the logarithm of the selectivity factor α. The changes in selectivity due to mobile phase variations are small when all lines in a figure, corresponding to a particular stationary phase, are parallel to each other. The C1 phase shows hardly any changes in selectivity with the exception of PAR. The C6, C8 and C18 phases show some selectivity changes for ACP. The CN phase shows considerable changes in selectivity (including peak cross-over). The PHE phase is more regular.

Figures III.2a to III.2f give an impression of the differences between the stationary phases at fixed mobile phase compositions. When examining these figures some patterns emerge. It should be kept in mind that differences between stationary phases (at a specific mobile phase composition) must be examined reckoning with the reproducibility of the retention times. In Figure III.2b the 1 sigma bar of the ln k values of EHB are incorporated to give an impression of the significance of the differences. Again, a number of inferences can be made. When the mean retention (a measure of hydrophobicity of the stationary phase) of all the solutes is considered, the phases can be arranged in decreasing hydrophobic order for all mobile phases, namely C18, C8, C6, C1, PHE and CN. The differences between C6 and C8 are not very large. This pattern breaks down for PAR. For the water/methanol and the ternary mixtures, PAR shows hardly any difference between C1, C6, C8 and C18.

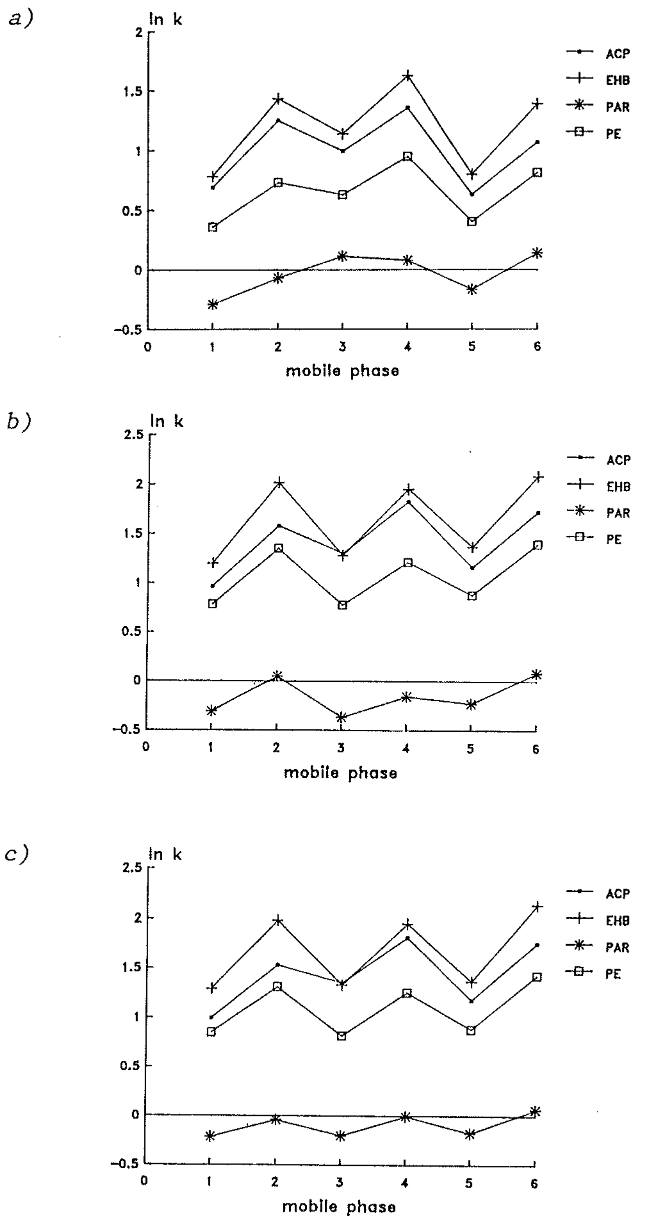


Figure III.1. Measured ln k values of some of the solutes on the C1 phase (III.1a), the C6 phase (III.1b) and the C8 phase (III.1c). For the abbreviations, see the text. The numbers 1 to 6 correspond to, respectively, wm1, wm2, wa1, wa2, am1, and am2.

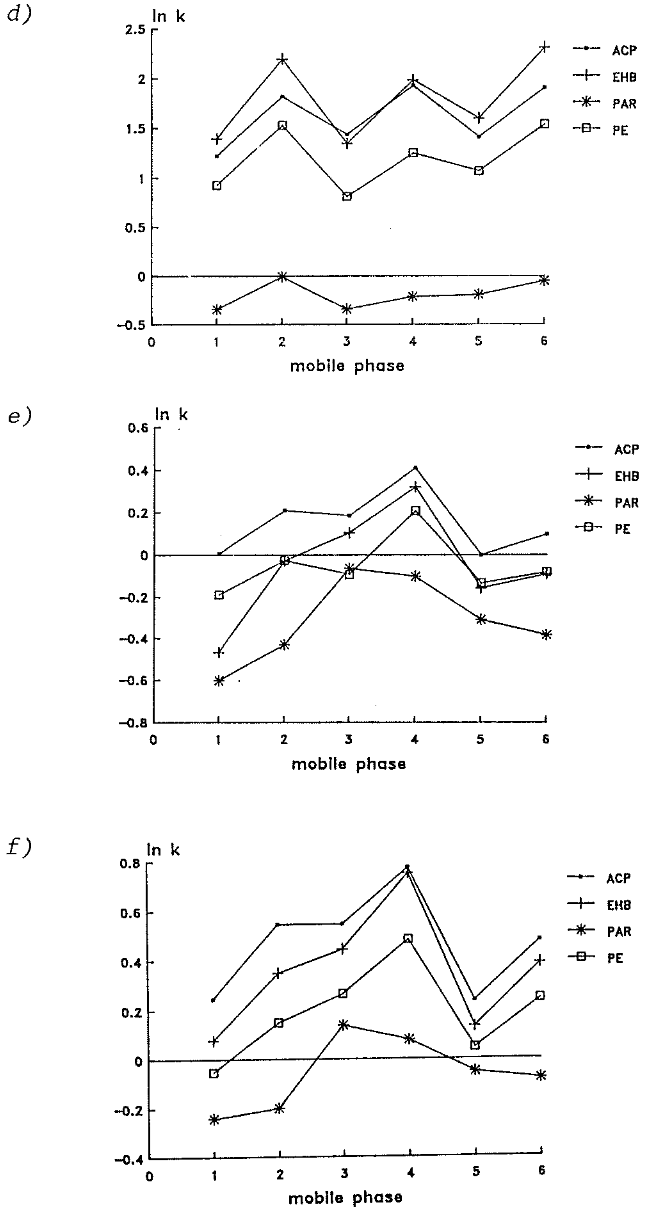


Figure III.1 (continued). Measured ln k values of some of the solutes on the C18 phase (III.1d), the CN phase (III.1e) and the PHE phase (III.1f). For the legend, see Fig. III.1a.

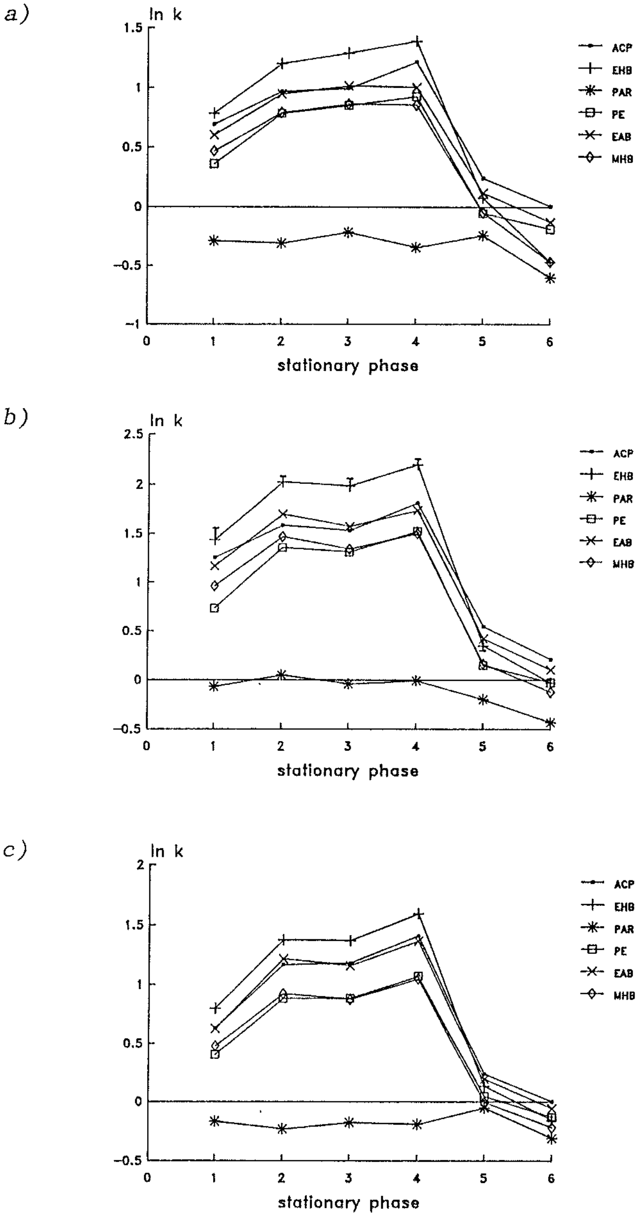


Figure III.2. Measured ln k values of some solutes on the stationary phases at mobile phase wm1 (III.2a), mobile phase wm2 (III.2b) and mobile phase am1 (III.2c). The numbers 1 to 6 correspond to, respectively, C1, C6, C8, C18, PHE, and CN. For the other abbreviations, see the text.

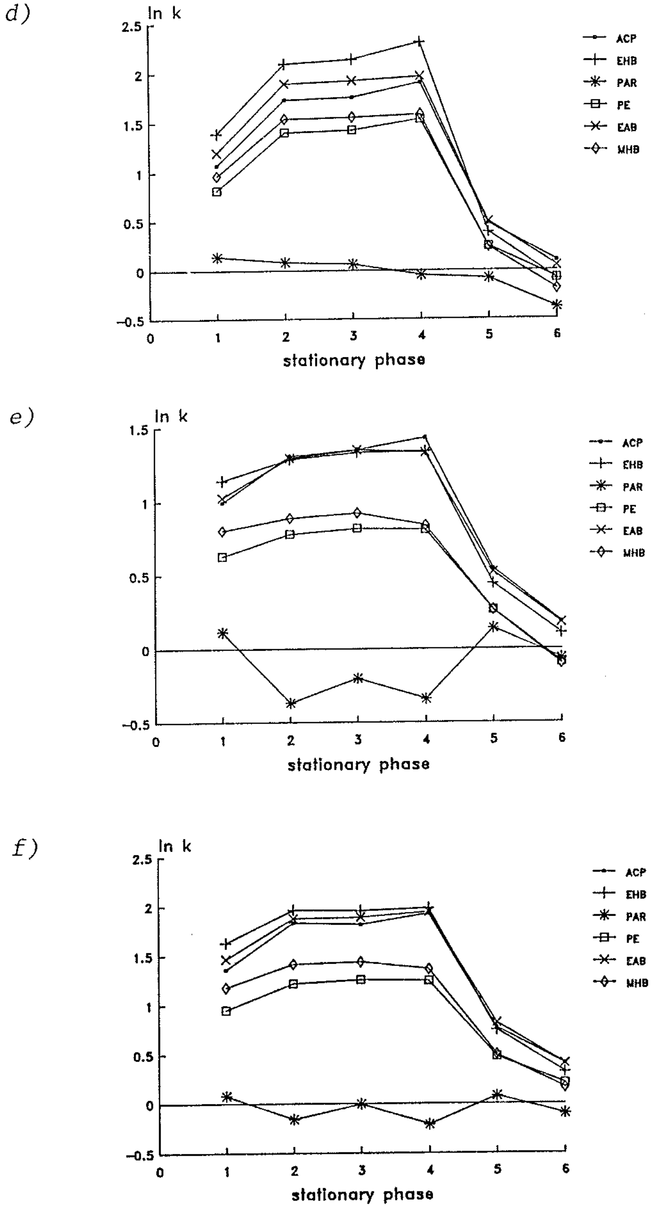


Figure III.2 (continued). Measured ln k values of some solutes on the stationary phases at mobile phase am2 (III.2d), mobile phase wal (III.2e) and mobile phase wa2 (III.2f). For the legend, see Fig.III.2a.

Especially water/ acetonitrile mixtures show more retention of PAR on C1 than on C18,but this will be commented on later. Changes in selectivity, due to varying stationary phases at a fixed mobile phase composition, can be observed for the CN and PHE phases. In particular for water/ acetonitrile mixtures these phases show peak cross-over. Incidently, C18 (Figure III.2a and III.2b) and C1 (Figures III.2b) show cross-over. In accordance with Berendsen[5] the alkyl-bonded phases show a flattening behaviour when the retention of a solute is plotted against the number of carbon atoms in the alkyl chain. The ternary mixtures show the least selective differences between the alkyl- bonded phases.

The special role of the solute paracetamol is illustrated in Figures III.3a to III.3f. In Figure III.3a and III.3b the ln k values of PAR at the water/methanol mixtures are plotted for the stationary phases (some extra measurements were available). When only a solvophobic mechanism is responsible for retention[6], all ln k values should increase in the order mobile phase 1 to 9 (decreasing elution strength). This is, however, not the case for the C18 phase at mobile phases 1 to 3 which indicates a dual retention mechanism[7]. A dual retention mechanism means in this case that, besides the solvophobic interactions, a second kind of interaction is present. This second kind of interaction might be a polar (specific) one of a solute with the free silanol at the surface of the modified silica material[8]. More indications of a dual retention mechanism can be obtained from Figure III.3e, the water/acetonitrile mixtures. The retention behaviour of paracetamol on the CN and PHE phases shows deviations from the behaviour based on solvophobic theory, which is in agreement with the observation that the CN and PHE phases can be regarded as moderately polar[9]. The irregularity of the C1 phase (Figure III.3f) with respect to the above mentioned solvophobic theory, might be due to measurement error and is hardly evidence of a dual mechanism. The ternary mixtures (Figures III.3c and III.3d) show also deviations from solvophobic theory for the CN and PHE phases. The irregular behaviour of the C1 phase is again doubtfull, although slightly more pronounced than in Figure III.3f. A reason for a more pronounced dual mechanism in water/acetonitrile mixtures can be that free silanol groups are more shielded in water/methanol mixtures because of hydrogen bonds between methanol and silanol[10]. In water/methanol mixtures, the number of free silanol sites accessible for specific interactions with a solute, is less. The observation that the solute paracetamol shows more retention on the C1 than on the C18 stationary phase, at water/acetonitrile mixtures, is may also be due to the free silanol sites. These sites are more accessible on the C1 phase than the C18 phase and cause specific interactions, especially in combination with water/acetonitrile mixtures[11].

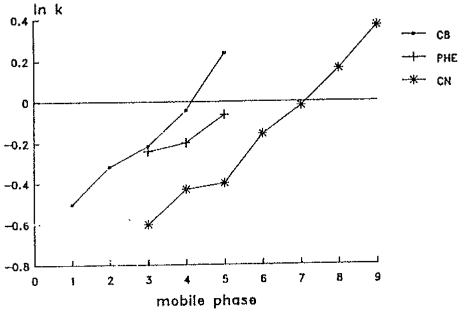


Figure III.3a. Measured ln k values of paracetamol on the C8, PHE and CN stationary phases. The mobile phases 1 to 9 are binary mixtures of water/methanol, the fraction of methanol is, respectively, 0.70; 0.62; 0.53 (wm1); 0.45 (wm2); 0.37; 0.28; 0.20; 0.12 and 0.03.

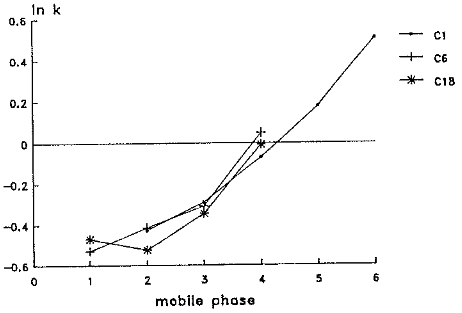


Figure III.3b. Measured ln k values of paracetamol on the C1, C6 and C18 stationary phases. For the mobile phases 1 to 6, see Fig.III.3a.

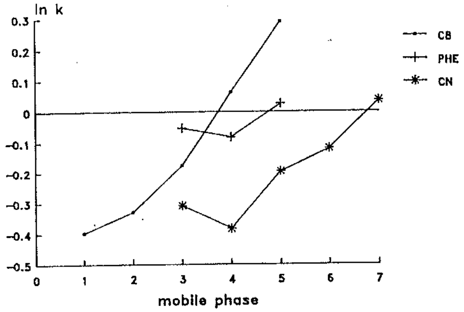


Figure III.3c. Measured ln k values of paracetamol on the C8, PHE and CN stationary phases. The mobile phases 1 to 7 are ternary mixtures with the following compositions (water/acetonitrile/methanol): 0.37/0.28/0.35; 0.45/0.23/0.31; 0.54/0.19/0.27 (am1); 0.63/0.15/0.22 (am2); 0.71/0.11/0.18; 0.80/0.06/0.14; 0.88/0.02/0.10.

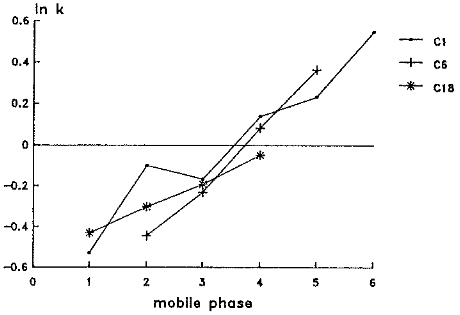


Figure III.3d. Measured ln k values of paracetamol on the C1, C6 and C18 stationary phase. For the mobile phase compositions 1 to 6, see Fig.III.3c.

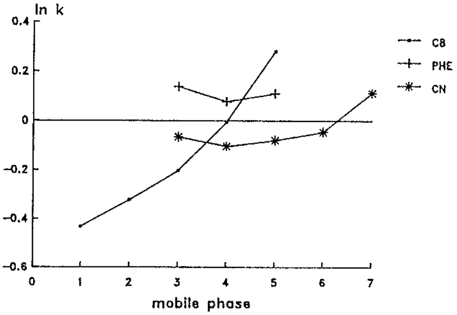


Figure III.3e. Measured ln k values of paracetamol on the C8, PHE and CN stationary phase. The mobile phases 1 to 7 are binary mixtures of water/acetonitrile, with fractions of acetonitrile of, respectively, 0.55; 0.47; 0.38 (wa1); 0.30 (wa2); 0.22; 0.13 and 0.05.

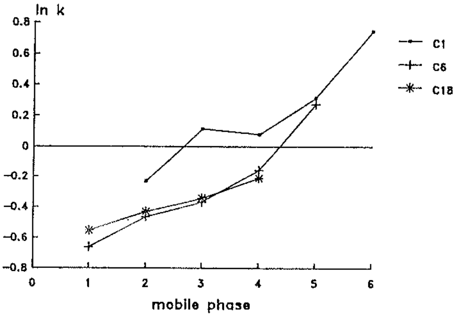


Figure III.3f. Measured ln k values of paracetamol on the C1, C6 and C18 stationary phases. Legend, see Fig.III.3e.

The repeatability and the reproducibility of the retention values of paracetamol on the six stationary phases were not extremely poor (see Tables III.2, III.3, III.5 and III.6) and there was no reason to assume, by examining the original measurements, that systematic errors were made. This is not the only peculiarity of the behaviour of paracetamol. The solute has almost everywhere the lowest retention. Differences between the stationary phases hardly show up for this solute because there is almost no interaction between paracetamol and the stationary phases. Yet the hypothesis that the ln k values of paracetamol are not influenced by the different stationary phases, at mobile phase composition wm2, can be checked by performing an analysis of variance. The associated F-ratio is 22.34 leading to a p-value smaller than 0.001. Altogether, there is no hard evidence to discard the measurements of the solute paracetamol.

The low retention of paracetamol is, of course, a calibration design problem. The mobile phase compositions in this particular experiment are chosen orthogonally (see Sections 4.1 and 9.2). This does not work out properly for paracetamol and weaker mobile phase compositions are, therefore, more appropriate in order to obtain more retention for this solute. This conflicts, however, with the very high retention of other solutes in the same calibration set. The only solution is a non-orthogonal design, calibrating the less polar compounds with a stronger eluent than the polar ones.

9.7 The choice of the training- and test set

If the stationary phases are conceived as objects, the solutes as the first category of variables and the mobile phase compositions as the second category of variables, a three-way data table is the result. This data table is presented in Figure III.4a. As shown in Section 1.6, one of the possible generalizations of PCA can be obtained by performing an ordinary PCA on the unfolded data cube, see Figure III.4b. Two principal component analyses were done, the first on the unscaled, the second on the scaled data. The data were always column-mean-centered and the scaling was done in such a way that the columns obtained variance one. The purpose of these principal component analyses is to make a proper division of the stationary phases into the training set and the test set.

With the first PCA (unscaled) three principal components accounted for 99.8% of the variation (respectively 98.6%, 1.1% and 0.1%). The score plot of the first two PC's are shown in Figure III.5a. From Figure III.5a, which represents 99.7% of the variation, it can be inferred that the C6 and C8 stationary phases are very similar, while the other stationary phases are dissimilar. The most extreme ones are the C18, CN and C1 phases. This is in accordance with the conclusions drawn in Section 9.6. As already mentioned in Section 1.2, it is usually difficult to interpret the principal components. Combining the observed patterns in Table III.4, which were already discussed in

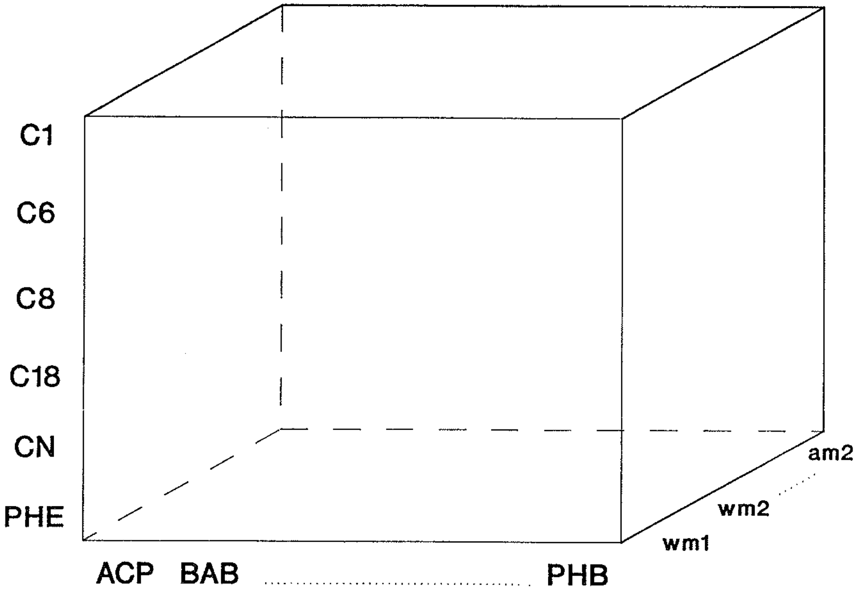


Figure III.4a. The complete data set arranged in a data cube. All entries are ln k values. For abbreviations, see the text.

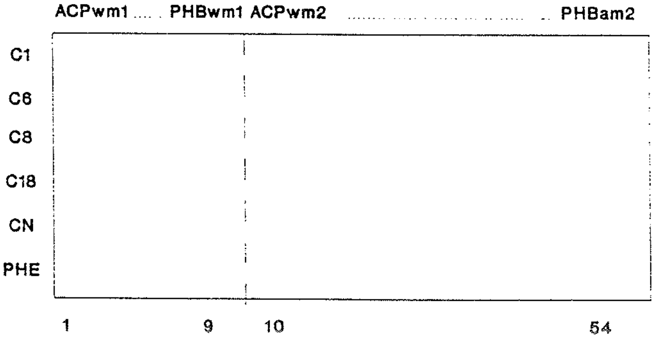


Figure III.4b. The unfolded data cube of Figure III.4a. The front layer of Fig.III.4a is placed most left in the unfolded matrix and so on. For abbreviations, see the text.

Section 9.6, with Figure III.5a, it might be concluded that the first principal component describes the hydrophobicity of the stationary phases. The CN and PHE phases are the least hydrophobic ones, the C18 phase the most hydrophobic, whereas the C6 and C8 phases have a similar hydrophobicity though somewhat less than C18. The C1 phase takes an intermediate position between C18 and CN. It is tempting to discuss differences in selectivity between the stationary phases in terms of scores on the second principal component but no attempts are made here because of the speculative nature. The loading plots, that belong to the first PCA, are complicated, see Figure III.5b, but a pattern emerges. The loadings on the first PC are mostly positive (except for the variables 22 and 31, which remain to be discussed later). This can be understood by observing, in Table III.4, that for every solute/mobile phase combination the ln k values become higher in the range CN, PHE, C1, C6, C8 and C18. This is, of course, a restatement of the earlier remark on the hydrophobicity of the stationary phases. The loadings on the first PC reflect the influence of the hydrophobicity of the stationary phases on every solute/ mobile phase combination. The variables 22 and 31 are exceptional cases (negative loadings on the first PC). In both cases the solute paracetamol is involved with the water/acetonitrile mixtures. The reason for this behaviour of paracetamol has already been discussed in Section 9.6.

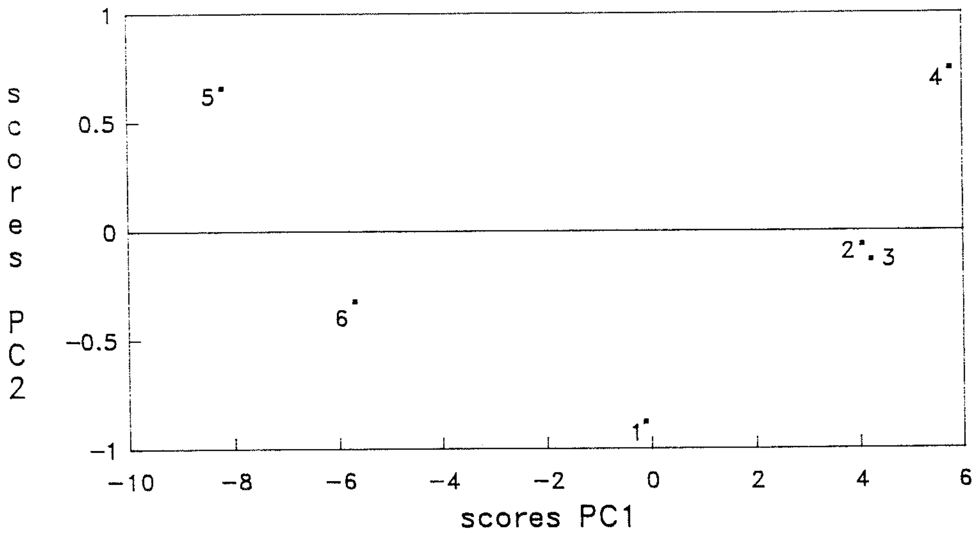


Figure III.5a. Score plot of an unscaled three-way PCA on the data cube of Fig.III.4a. Legend: 1=C1, 2=C6, 3=C8, 4=C18, 5=CN and 6=PHE.

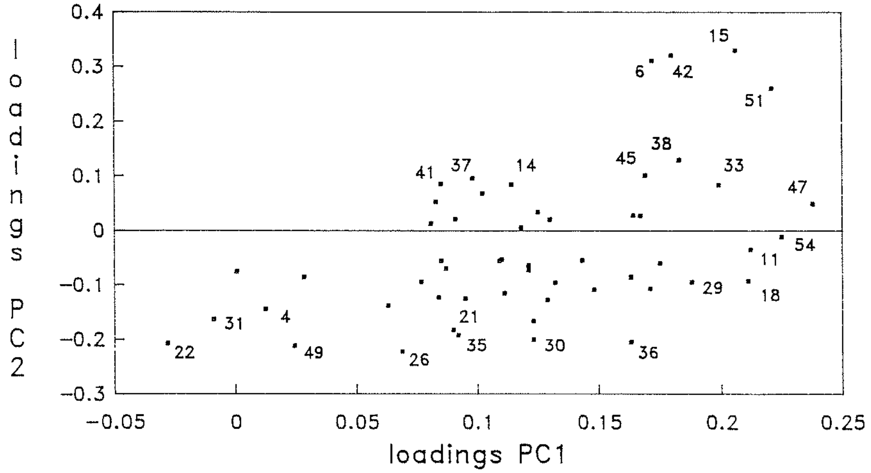


Figure III.5b. Loading plot of an unscaled three-way PCA on the data cube of Fig.III.4a. Legend: 4=ACP at wml, 6=ACP at wml, 11=BAB at wm2, 14=PE at wm2, 15=TOL at wm2, 18=PHB at wm2, 21=EHB at wal, 22=PAR at wal, 26=MHB at wal, 29=BAB at wa2, 30=EHB at wa2, 31=PAR at wa2, 33=TOL at wa2, 35=MHB at wa2, 36=PHB at wa2, 37=ACP at aml, 38=BAB at aml, 41=PE at aml, 42=TOL at aml, 45=PHB at aml, 47=BAB at am2, 49=PAR at am2, 51=TOL at am2 and 54=PHB at am2.

In order to perform the generalized Jolliffe-approach of variable selection (see Section 1.6) simultaneously in two directions, the loading plots must be examined carefully. It is very difficult to make a proper choice. One of the possible choices might be the solutes PAR, TOL and PHB at mobile phase compositions wm2, wal and am2. Variable selection was not the reason to perform three-way PCA but it illustrates how difficult it is to make a selection without procedures to evaluate the selection quantatively and to recognize "outlying" variables.

The second three-way PCA was performed on the auto-scaled data. The first three components explained 99.5% of the total variation (respectively 92.1%, 6.3% and 1.1%). The score plot, see Figure III.5c, shows the same pattern with regard to the hydrophobicity because the order of the scores on the first PC is the same as in the previous case. As can be seen, PCA is not scale invariant. The first PC in the scaled version explains less than the first PC in the unscaled version. This can be understood by realizing that in the unscaled PCA version variables with the highest variance contribute the most to the first principal component. This can be checked by comparing the loadings (see Figure III.5b) of the TOL related variables (6,15,33,42 and 51 are shown) on the first PC with the PE related ones (14 and 41 are shown). All the TOL related loadings are higher due to the higher variability of TOL. When the same comparison is made for the scaled PCA variant (Fig.III.5d), it can be observed that the loadings for the TOL and PE related variables on the first PC are almost equal (6 and 42 are shown, all other TOL and PE related loadings on the first PC are of the same order). Assuming that hydrophobicity of the stationary phases is the major cause of variability of the variables, the influence of hydrophobicity is blown up by using the unscaled variant which results in an extremely high percentage of variation explained by this phenomenon (98.6%, see also Section 1.2 and 1.7). This indicates that the scaled version may be more valuable.

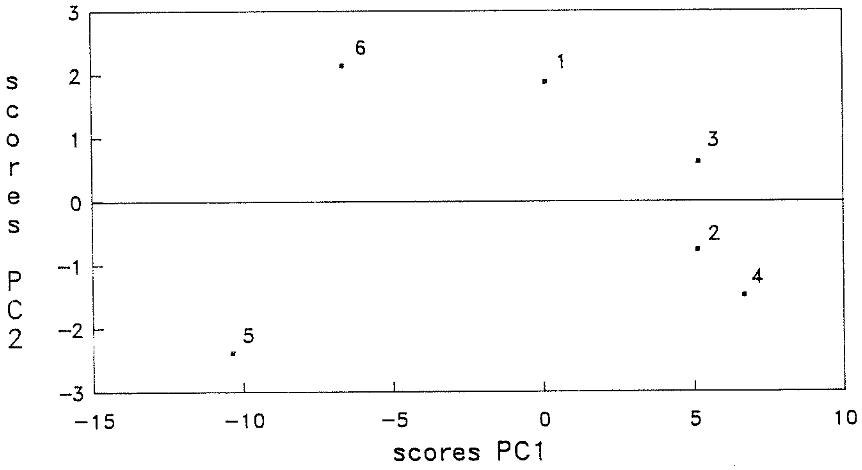


Figure III.5c. Score plot of a scaled three-way PCA on the data cube of Fig.III.4a. Legend: see Fig.III.5a.

Comparing the scores on the second component for the scaled and unscaled versions, differences show up (note that, due to sign arbitraryness, the reflection of Figure III.5a in the second axis should be compared with Figure III.5c). In the scaled version the PHE stationary phase is more similar to C1, the C6 and C8 phases are more dissimilar.

The stationary phases C1, C18 and CN are chosen for the training set, the other stationary phases constitute the test set. Looking at the score plot in Figure III.5c (scaled version) the phases in the training set span the three PC dimensions, with the exception of the PHE phase. The expectation is that when this choice of the training set is made, predictions on the PHE stationary phase might be a problem so that the performance of the prediction procedure can be tested.

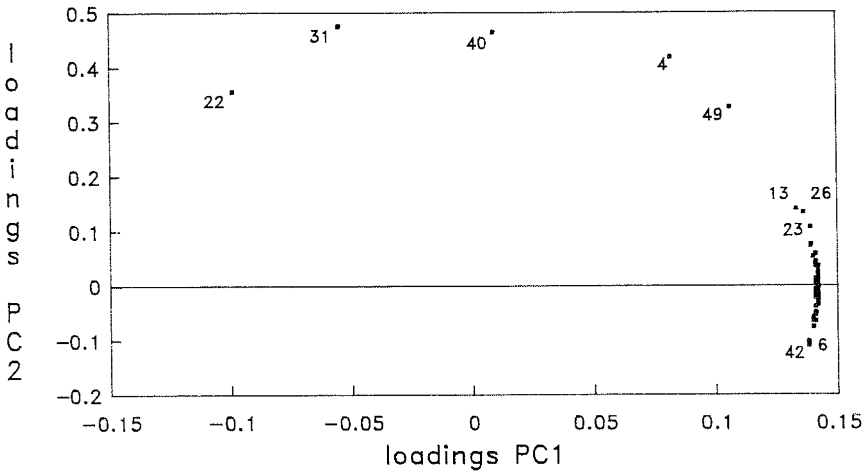


Figure III.5d. Loading plot of a scaled three-way PCA on the data cube of Fig.III.4a. Legend, see Fig.III.5c and 13=PAR at wm2, 23=PE at wa1 and 40=PAR at am1.

Chapter 10 Two-way approach: the induced-variance criterion

10.1 The choice of the markers

A first step in the marker choice procedure is the performance of a PCA on the training set, the stationary/mobile phase combinations as objects and the nine solutes as variables. The columns of this data matrix were again autoscaled. The cumulative percentages of explained variation were respectively 90.88, 99.22 and 99.60 for the first three dimensions. The score plot associated with the first two dimensions is shown in Figure III.6. In Figure III.6 a grouping of the observations can be seen according to the stationary phases. The data are spread regularly and no clear outliers are detected. The loading plot, Figures III.7, displays the contributions of the variables to the first two dimensions. The solutes PAR and TOL span the loading space together with one of the solutes MHB, EHB, PHB or PE (for an explication of the notion "span the loading space", see Sections 1.2 and 1.6). The reason for the extreme loadings of PAR is clear from the discussion in Section 9.6. The solute paracetamol is perhaps an outlying variable, and the Jolliffe-Wold approach in selecting variables from the loading plot which span the loading space is not robust against the selection of outlying- instead of representative variables.

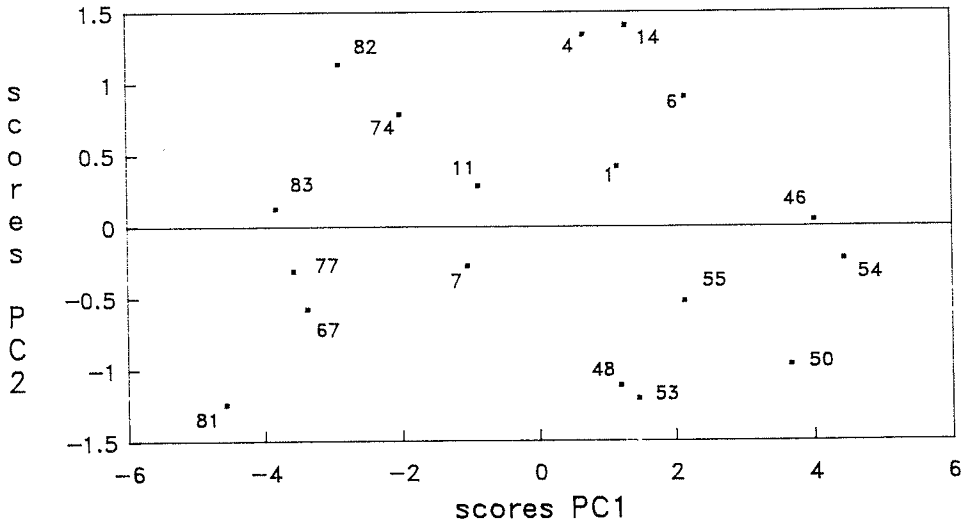


Figure III.6. Score plot of a PCA on the training set. Legend: 1=C1wm2, 4=C1wa1, 6=C1wa2, 7=C1wm1, 11=C1am1, 14=C1am2, 46=C18wm2, 48=C18wm1, 50=C18wa2, 53=C18wa1, 54=C18am2, 55=C18am1, 67=CNwm2, 74=CNwa2, 77=CNam2, 81=CNwm1, 82=CNwa1 and 83=CNam1.

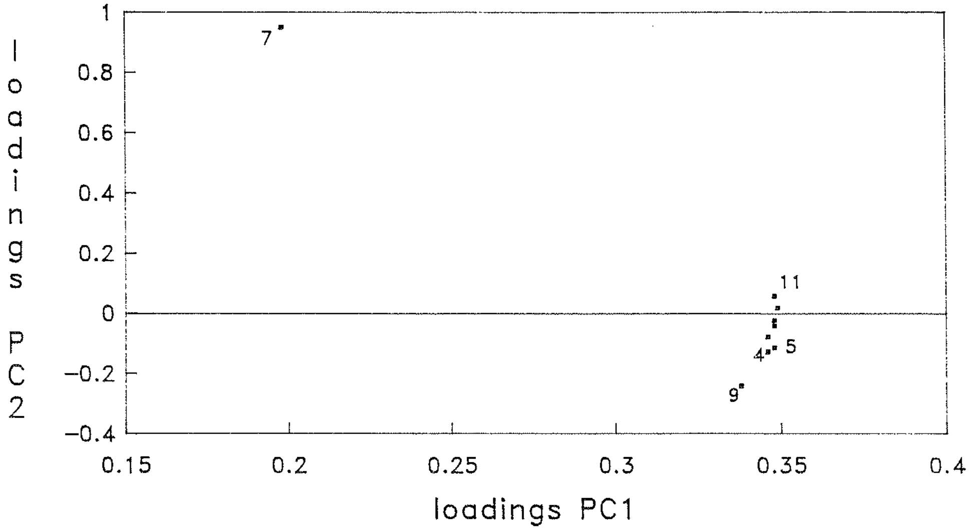


Figure III.7. Loading plot of a PCA on the training set. Legend: 4=ACP, 5=BAB, 7=PAR, 9=TOL and 11=MHB.

The induced-variance criterion, introduced in Section 1.3, is used to select the markers. For convenience, the training set from which the markers are chosen is depicted in Figure III.8 (this is also the matrix of which the PCA was calculated). The induced-variance criterion aims at the prediction of the omitted solutes (the non-markers), using the markers. The solutes that are selected perform well in this sense. The columns in the training set matrix are mean-centered and scaled-to-length-one in order to obtain the correlation form of the induced-variance criterion. This is a reasonable choice because the modelspecifications in the next stage of the prediction procedure comprise, besides the logarithms of the capacity factors of the markers, the fractions acetonitrile and methanol in the mobile phase. Different measurement units are therefore used in one model and autoscaling (or centering and scaling to length one which only differs a constant from autoscaling) seems appropriate. The results of the induced-variance calculations are shown in Table III.10.

From Table III.10b it can be inferred that the solutes BAB, EHB and PAR are the best choice. The first four subsets of markers, as shown in Table III.10b, explain nearly the same amount of variation and are therefore exchangeable.

The jack-knife procedure to check chance results (note that, in order to obtain the output in Table III.10b, 84 different combinations of solutes have to be tested) is summarized in Table III.10a. The marker selection is done six times, each time the three rows (cases) in the training set are omitted corresponding to C1, C18 and

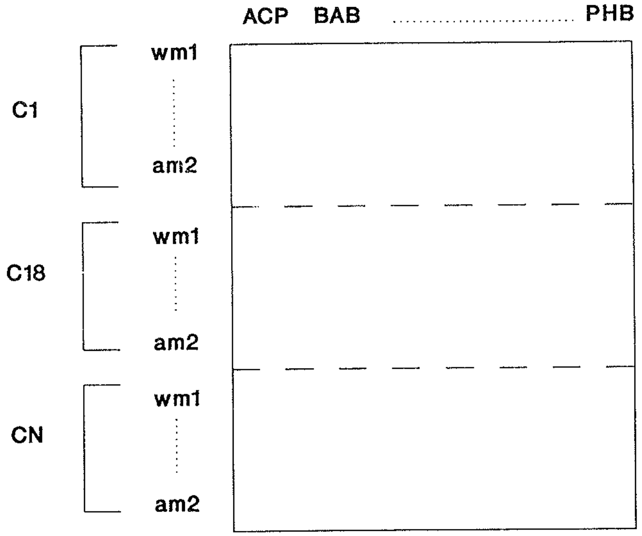


Figure III.8. The training set.

CN at the same mobile phase composition (see the legend of Table III.10). A full leave-one-out procedure was not performed in order to limit the calculations. The five best subsets of markers associated with each omitted mobile phase mixture are given. The combination of BAB, EHB and PAR is always one of the five best subsets. The combination of PAR, TOL and MHB is present in all but one of the five best subsets. In order to get a quantitative idea of the similarities between Table III.10a (the leave-more-out results) and Table III.10b (the optimal subsets), scores can be attributed to the subsets of Table III.10a which reflect the occurrence of the specific combination in the wml to am2 leave-more-out results. The subset BAB, EHB, PAR obtains score 3 from the wml block and score 5 from the am2 block (a subset obtains the highest score when it is the first in a particular leave-more-out block). The results of this score system, when these scores of the six omitted blocks are summed up, are: BAB,EHB,PAR 21; PAR,TOL,MHB 21; EHB,PAR,TOL 13; BAB,PAR,PHB 13 and BAB,PAR,MHB 4. Obviously, the outcome of the calculations as reported in Table III.10b, is stable. The difference between the combinations of BAB, EHB, PAR and PAR, TOL, MHB is very small. The combination of BAB, EHB, PAR is selected because it has the best performance.

A careful examination of the leave-more-out results show that the solute paracetamol is present in all subsets. This can be explained by the peculiarities in the behaviour of paracetamol, discussed in Section 9.6, and the dual character of the induced-variance criterion, see Section 1.3. The solute PAR is not incorporated in each subset because of its predictive quality, but due to its unpredictable behaviour. All other solutes, except PE, are present in one or more subsets. All solutes except PE and PAR are exchangeable as predictors for each other, although some combinations have a better performance than others. It is not possible to explain why the solute PE does not show up in any of the subsets. It may be a bad predictor or good predictable. It can be argued that PE is never selected because the predictability of PE is not so bad as the often selected solute PAR. A cautious conclusion might be that the non-robustness of the induced-variance criterion towards unpredictable- or outlying compounds can be overcome by careful examination of the leave-more-out results.

The choice of the numbers of markers can be made by examining the percentages of explained variation of the markers and the principal components. Table III.10c reflects this comparison and shows that a number of three markers is reasonable. Table III.10d gives calculations done on subsets of the training set. By choosing these subsets, the variation in ln k values of a solute due to the stationary phases is slightly more pronounced. The separate "series 1" and "series 2" mixtures contain less variation due to elution strength and, consequently, the stationary phase influence is enhanced. The selected combination performs well regarding the "series 1" mixtures. The performance with regard to the "series 2" mixtures is less good: the selected combination occupies a nineth place with 99.59% explained variation. Whether this constitutes a problem can be assessed by examining the prediction results. The combination of PAR, TOL, MHB performs well in both cases.

10.2 Modelspecifications and an evaluation of the design matrices

Two different modelspecifications (see Subsection 8.1.2) for the prediction of the non-markers are tested, which can be fine-tuned in a later stage (see Section 10.5). The first modelspecification, briefly called model 1, accounts explicitly for the mobile phase variation and describes the ln k value of each non-marker as dependent on markers and mobile phase compositions (for the sake of simplicity, an integer indexing the non-markers is left out):

$$\ln k_{non,t}=\beta_0 +\beta_1 A_t +\beta_2 M_t +\beta_3 BAB_t +\beta_4 EHB_t +\beta_5 PAR_t +\epsilon_t \qquad (III.1)$$

where t (=1,..,n) indexes the object number in the training set (the stationary/mobile phase combination), A_t and M_t are the fractions of acetonitrile and methanol respectively in the mobile phase, and BAB_t, EHB_t, PAR_t are the ln k values of the solutes BAB, EHB and PAR. The error terms ϵ_t are assumed to be distributed around zero with the covariance matrix $\sigma^2 I$. A justification of this constant variance is given in Section 9.5. This variance, however, may differ between the non-markers. After column-centering and omitting the column of zero's (see Section 2.1), the full rank design matrix X is the result. The matrix X is depicted in Figure III.9a. Figure III.9b depicts the resulting test set. The fraction of water is not taken into account because the fractions of A, M and water sum to one so that two variables are enough to describe the variation of the mobile phase composition. Moreover, when the fraction of water is incorporated in the model (the constant term has to be omitted in that case) a linear combination in X is introduced after autoscaling of the X block, see Section 1.7.

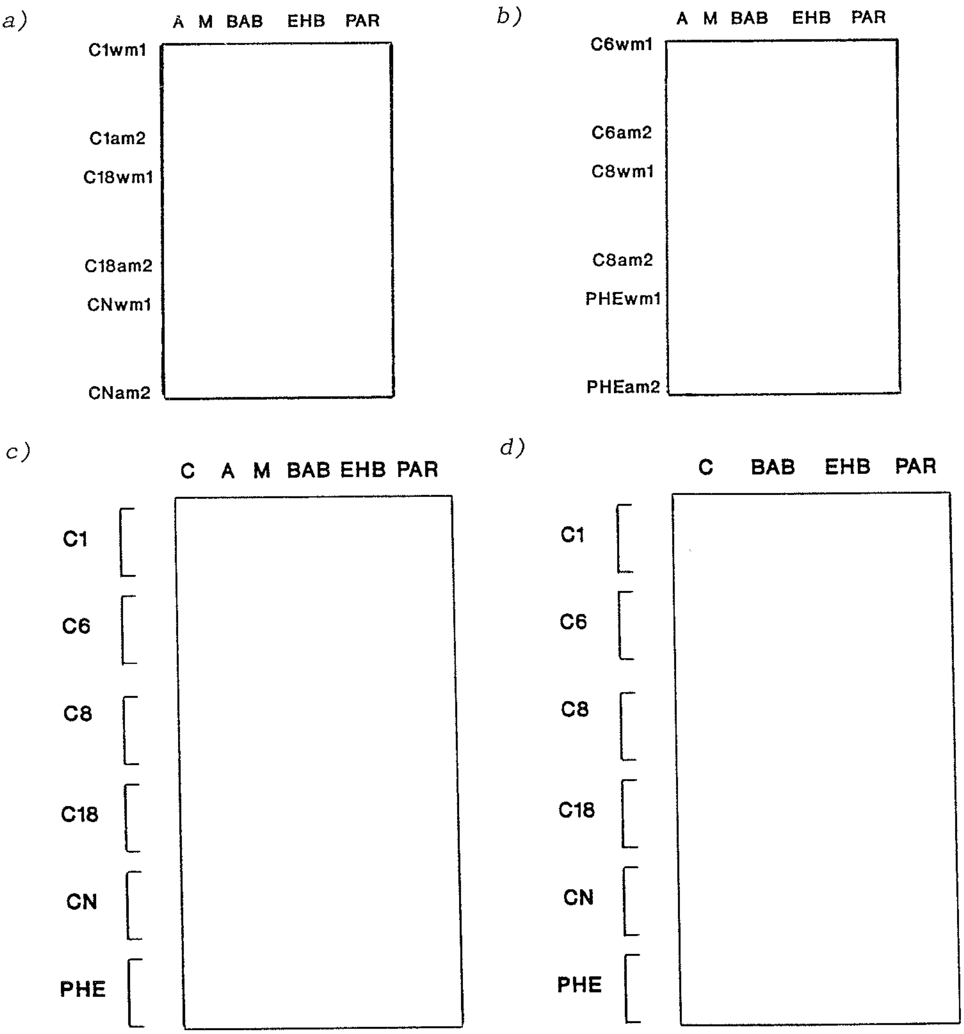


Figure III.9. Design matrix X (a) and the test set (b). Matrices Z associated with model 1 (c) and model 2 (d), see text.

The second modelspecification, model 2, does not take the mobile phase variation into account explicitly, but relies upon the predictive power of the ln k values of the markers, which are also influenced by the mobile phase composition. This model describes the ln k values of each non-marker as follows (again the integer indexing the non-marker is left out):

$$\ln k_{non,t} = \beta_o + \beta_1 BAB_t + \beta_2 EHB_t + \beta_3 PAR_t + \epsilon_t \qquad \text{(III.2)}$$

where t, BAB_t, EHB_t, PAR_t and ϵ_t stand for the same values as before. The design matrix associated with model 2 (again after the column-centering operation) is the same as for model 1 except for the first two columns.

A model relating the ln k values of the non-markers only to the mobile phase variables A and M can also be formulated. Preliminary calculations (not shown) indicated that the performance of these models was bad.

The most important characteristics of the design matrices are summarized in Table III.11. Model 1 is discussed firstly. The design matrix X of model 1 is scaled in such a way that X'X is in correlation form. Very strong correlation exists between the ln k values of BAB and EHB (Table III.11a). Multicollinearity is not difficult to discover in this case, the relationship between BAB and EHB being very clear. The drawback of using the correlation matrix as a diagnostic tool (see Section 2.3) does not hold in this simple case. There is also a strong correlation between A and M. The correlation between A and M can be influenced by using a proper design (see Section 4.1). In this case, the demand of orthogonality (the same mobile phase compositions on each stationary phase) results in correlation between the mobile phase variables. The correlation between the marker variables can be influenced by the marker choice, see Section 4.1. High correlations between the two mobile phase variables and the marker variables do not occur, the highest correlation is -0.469 between M and PAR. The multiple correlation coefficient (not shown in the Tables) between BAB and A,M is 0.31, for EHB and PAR these numbers are respectivily 0.32 and 0.53. These relatively low values give rise to some doubt about the predictive relevance of the mobile phase variables A an M. There are hardly serious connections between the mobile phase variables and the marker variables. Within the block of marker variables the solute PAR is again an exception: moderate correlations between PAR and BAB, EHB (0.484 and 0.577 respectively).

The variance inflation factors, Table III.11b, must be lower than five (see Section 2.4). Values above this threshold indicate multicollinearity. As already described in Section 2.2, a high variance inflation factor of a specific predictor variable indicates a strong dependence of that variable on the other predictor variables and, therefore, strong multicollinearity. Especially BAB and EHB have very high variance inflation factors, partly because of their strong correlation.

The relationship between BAB and EHB is also present in Table III.11c, presenting the variance decomposition proportions. Simulation results indicate that weak dependence shows itselves with condition indices around 10. Condition indices between 15-30 indicate relationships with associated correlations of about 0.9 and are considered serious. Condition indices above 100 are of great potential harm for the regression (see Section 2.4). It can be inferred from Table III.11c that a strong relationship exists between BAB, EHB and (to some extend) PAR. Note that the connection between BAB, EHB and PAR, as present in the variance decomposition proportions, cannot be inferred from the correlation matrix (Table III.11a). The relationship between A and M is associated with the fourth singular value and does not contribute much to the multicollinearity.

In order to give an overall picture of the design matrix of model 1 a principal component analysis is performed on this matrix. The columns of the design matrix are autoscaled. Figures III.10a to III.10c give the score plots of the first three dimensions. When the scores on PC1 and PC2 are considered, the CN measurements form a separate class, which is not surprising considering the remarks made in Section 9.6. Yet these measurements cannot be considered, on the whole, as outliers. When the scores on the second and third PC are plotted (Figure III.10c) a clustering of the measurements on the different stationary phases emerges. No clear outliers are visible so that all measurements can be retained. Examining Table III.11d, some patterns become visible. The only solute that loads high on the third PC is PAR which proves ones more that this solute has a deviating behaviour. The relationship between A and M is present in the fourth PC and between BAB and EHB in the fifth one. The variance decomposition proportions show these relationships more drastically because these proportions combine two aspects of a (near) linear combination of variables: the loadings of the variables on that combination and the strongness of that relation. Stated otherwise, the loadings of the variables on the last PC are combined (see Section 2.2) with the percentage of explained variation (eigenvalue) of that last PC. The variance decomposition proportions are therefore preferred as a diagnostic tool.

The design matrix of model 2 is closely related to the one of model 1. The correlation matrix of the variables in model 2 is the 3x3 lower-right submatrix of the correlation matrix of model 1, and the same remarks regarding the correlation of BAB and EHB are valid to this case. The variance inflation factors are lower than in the previous case (see Table III.11e). These variance inflation factors are the reciprocal of 1-R^2, where R^2 is the multiple correlation coefficient between the predictor variable and the other x variables. These R^2 values are a non-decreasing function of the number of the other x variables involved and are therefore not lower (and usually

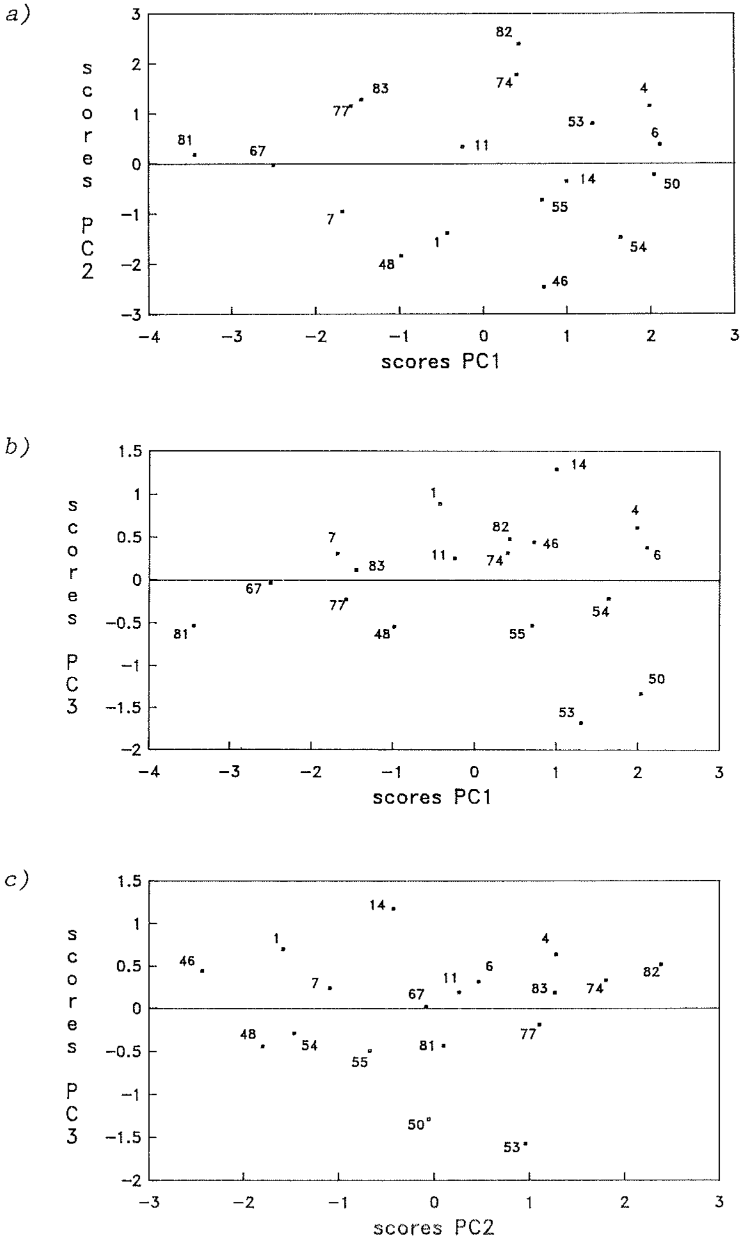


Figure III.10 a,b,c. Score plots of a PCA on the design matrix of model 1. Legend, see Fig.III.6.

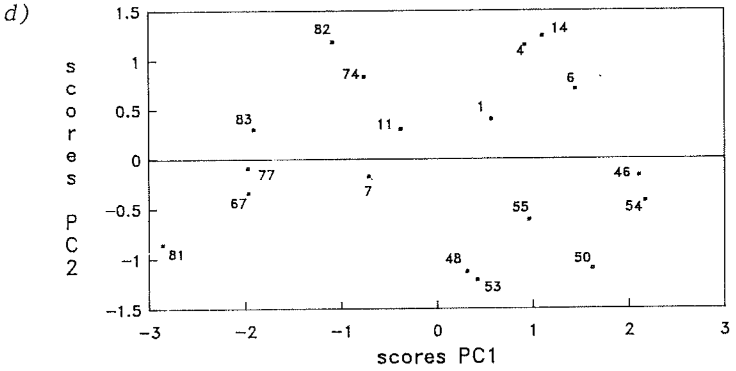


Figure III.10d. Score plot of a PCA on the design matrix of model 2. Legend, see Fig.III.6.

higher) for model 1 than for model 2. Yet the variance inflation factors of BAB and EHB are still very high. The variance decomposition proportions for model 2 show again a relationship between BAB, EHB and PAR although a slightly weaker than for model 1 (Table III.11f). Still, the condition index associated with this relationship is high, pointing to serious multicollinearity. A principal component analysis of this design matrix clearly reveals the relationship between BAB and EHB (see Table III.11g). The score plot (Figure III.10d) shows clustering of the stationary phases, but again no serious outliers are detected.

In relation to the marker choice and the choice of the modelspecification(s), the quality of the test set must be assessed. This is important because the ultimate yardstick in the evaluation of the estimation methods, of cross-validation and of the model specifications, is the quality of the predictions in the test set. The importance of a matched split is indicated in Section 4.2. If model 1 is assumed, the total data set, with regard to the predictor variables, is depicted in Figure III.9c and labelled Z, where Z is an 36x6 matrix. The test set values can be arranged in matrix V (18x6), in which the column entries are the same as those of Z. The rows of V consist of the stationary/mobile phase combinations which were chosen in the test set. Note that the elements of V are a subset of the elements of Z. The split of the data in a training set and a test (validation) set is a matched split when $Z'Z/n \approx V'V/n_v$, where n and n_v are the total number of observations (36) and the number of observations in the test (validation) set (18), respectively. When the two dispersion matrices $Z'Z/n$ and $V'V/n_v$ are compared, see Table III.11h, the problem arises of how close they resemble. No clear guidelines are given by Picard and Cook[12], but their numerical example show that the two dispersion matrices in Table III.11h are close enough to indicate a matched split.

The Z matrix, that corresponds with model 2, is depicted in Figure III.9d. The split can be judged by inspection of the two dispersion matrices in Table III.11h, the A and M entries being omitted. Again the tentative conclusion is that the test set is a reasonable reflection of the whole data set.

10.3 The results of the different estimation methods

10.3.1 Ordinary least squares (OLS)

The results of the ordinary least squares (OLS) calculations are summarized in Table III.12 (model 1) and Table III.13 (model 2). All y variables and columns of X are mean-centered and scaled-to-length-one and X is of full rank (see Section 2.1). Although this is not necessary for the OLS calculations, the estimated coefficients are now standardized and their influence on y can easily be compared. Model 2 is discussed firstly.

For each solute a high F value is obtained, indicating a significant decrease of unexplained variance due to the application of the model. Even for the solute with the lowest F value, TOL, the associated p value is still smaller than 0.01. The s values (standard deviations of the residuals) can be compared with the pooled standard deviations of the reproduced measurements in the training set (see Table III.5). They are in the same order of magnitude, so that there is no serious lack-of-fit (it should be kept in mind that these pooled standard deviations are biased estimates of the actual reproducibility of the measurements but these are the only ones available, see Section 9.5). The variances of the estimated coefficients b_{BAB} and b_{EHB} are regarded as high in relation to the other estimated coefficients in the same equation. This is due to the multicollinearity between the ln k values of BAB and EHB, and was already indicated by the variance inflation factors (discussed in Section 10.2). Despite this fact, the variable BAB has a significant influence in the regression of ACP, PE, EAB and TOL (the null hypothesis that $b_{BAB}=0$ is rejected at $\alpha=0.05$). The variable EHB has a significant ($\alpha=0.05$) influence in the regression of MHB and PHB, which is not surprising because the solutes MHB, EHB and PHB are quite similar (homologs). The variable PAR has some predictive power for PE and to a smaller extent for PHB, its role is not completely clear. The mobile phase variables A and M have predicting relevance for EAB, PHB and to a smaller extent for TOL, MHB. It is tempting to discard the variables without any significant coefficients in an equation. This is one of the remedies against multicollinearity but it should be done with care (see Section 2.4). One of the pitfalls of such a procedure is the strong influence multicollinearity has on the magnitude (and sign!) of an estimated coefficient, as will be illustrated in Subsection 10.3.2. The fine-tuning of the model specifications is postponed till Section 10.5.

In Table III.12b the root mean squared error of predictions (RMSEP) are presented. For the training set these values are calculated by the leave-one-out method to evaluate the predictive power of the estimation method and to check on influential observations[13,14]. The percentage of variation in the test set explained by the model has a minimum of 98.0 for PE and a maximum of 99.9 for PHB. This indicates that the predictions are good, despite the strong multicollinearity. Theoretically, this is possible if the linear combination(s) of the variables which cause the multicollinearity problem is (are) also present in the test set (see Section 2.4). To be more specific: when a new ln k value has to be predicted with the use of a (measured) predictor vector *x* (which consists of A, M and the ln k values of the markers) and when the *x* vector is subjected to the same linear combination(s) which determine the multicollinearity in the training set, and the outcome is near zero, then the multicollinearity will not harm the predictions. 'Harm' in this context means that the multicollinearity will cause high variances of the predicted ln k values[15]. Not only the scores of a test set predictor vector on the first PC's are of interest (in order to perform an outlier test) but also the score of a test set vector on the last PC is worthwhile to notice.

In Section 10.2 the design matrix of model 1 is discussed. There is one serious source of multicollinearity, given by the linear combination associated with the smallest eigenvalue of the design matrix, see Table III.11d. When the 18 test vectors x (see Section 9.7) are subjected to this linear combination, the outcomes range from -0.0353 to 0.0298. When these values are squared and averaged the result is $3.4\text{x}10^{-4}$. This value will be called mean sum of squared deviations (MSSD). For the individual stationary phases C6, C8 and PHE this mean sum of squared deviations was respectively $2.5\text{x}10^{-4}$, $0.68\text{x}10^{-4}$ and $6.9\text{x}10^{-4}$. The question rises whether these outcomes are near zero. For comparison: when the test vectors are subjected to the linear combination associated with the largest eigenvalue of the design matrix (see Table III.11d), the outcomes range from -0.5304 to 0.5751 with MSSD of $1079\text{x}10^{-4}$. When the training set objects are subjected to the linear combination associated with the smallest PC, the outcomes (which are proportional to the training set scores on the fifth PC) range from -0.0199 to 0.0142 with MSSD $0.75\text{x}10^{-4}$. The tentative conclusion may be that the positive predictive performance of the OLS model is (at least) partly due to a similar multicollinearity pattern in the training- and test set. Another reason for this positive predictive performance is mentioned in Section 2.4. A low s^2 decreases the effect of multicollinearity on the variance of the predictions. Hoerl *et al.*[16] showed that in combination with high signal-to-noise ratios (F-ratios > 150), the predictions are not necessarily bad with OLS. Whether the coefficients belonging to the predictor variables are stable under the multicollinearity is discussed in the subsection on ridge regression (Subsection 10.3.2).

The variable TOL has large prediction errors in the training set and on the test set phases. This can be explained by examining the loading plots of the PCA performed with all solutes as variables and the C1, C18 and CN phases as objects (at the six eluent mixtures), see Figure III.7. TOL represents one of the extremes in this loading plot, thereby indicating a deviant behaviour.

A closer look at the RMSEP values of the separate test stationary phases reveals no clear differences, although the predictions on the C6 stationary phase tend to be less than average, the RMSEP in the whole test set. The prediction of retention on the PHE stationary phase, which was expected to be difficult (see Section 9.7), is good, with a slight exception of PE. This variable is badly predicted, over-estimated, at mixtures wm1 and wm2. This might be caused by a deviating behaviour of the variable PE in relation to its predictors A, M, BAB, EHB and PAR, in the training set and on the PHE phase at the water/methanol mixtures. The variable EAB is badly predicted on C6 at eluent mixtures wa1, wa2 and am2; all ln k values are underestimated.

Note that the above mentioned mean sum of squared scores of the test predictors on the final PC are higher than the training set ones for the C6 and PHE column. The C8 column behaves very well in this respect and the predictions on this column are good.

The leave-one-out RMSEP (training set) can be regarded as predictors of the RMSEP in the test set (Section 2.3). As such, they perform reasonably, except in the case of TOL, which was already partly discussed above. A closer examination of the leave-one-out predictions of TOL reveals that especially the retention on the stationary/mobile phase combinations C1wm1, C18wm1 and C18wm2 is badly predicted. The same phenomenon as described above with regard to the possible harm of multicollinearity on predictions holds, of course, for the leave-one-out predictions. If the omitted predictor vector x (C1wm1, C18wm1 and C18wm2 consecutively) does not yield a near-zero value when subjected to the troublesome linear combination a bad prediction of the omitted ln k value of TOL is the result. The three mentioned combinations may be influential points in the regression. Careful examination with the use of robust techniques[17,18] and leverage diagnostics[13,14,19] is needed. If leverage is defined with regard to the dispersion of the X matrix, then C1wm1, C18wm1 and C18wm2 are no leverage points. This can be concluded from the score plots of the PCA made of the design matrix associated with model 1. Moreover, if these combinations were leverage points, bad predictions would also be present for the other non-markers at those points but that is not the case. This subject is not pursued any further, but it can be concluded that the reason for the bad predictions of TOL at C1wm1, C18wm1 and C18wm2 is not clear.

The results of model 2 are shown in Table III.13. The F values are again very high, indicating a significant regression. The s values are in the same order of magnitude as the standard deviations of the reproduced measurements, so there is no serious lack-of-fit present. The variances of the estimated coefficients b_{BAB} and b_{EHB} are high in comparison with b_{PAR}, as was already inferred from the high variance inflation factors of these specific coefficients. The variable BAB has a significant (α=0.05) influence on ACP, EAB and TOL. The variable EHB has a significant influence on EAB, TOL, MHB and PHB. The influence of PAR is significantly present in the regression of EAB and PHB. The danger of using the significance of an estimated coefficient for variable selection purposes has already been pointed out.

The RMSEP values of model 2 are summarized in Table III.13b. The predictions in the test set are good, except for TOL. If the predictor vectors *x* in the test set are subjected to the linear combination causing the multicollinearity in the training set (see Table III.11g), the outcomes range from -0.0376 to 0.0301, with an MSSD of $3.5\text{x}10^{-4}$ (for the individual stationary phases of C6, C8 and PHE these numbers are respectively $2.1\text{x}10^{-4}$, $2.7\text{x}10^{-4}$ and $5.8\text{x}10^{-4}$). The outcomes are small in comparison with the outcomes of the *x* vectors subjected to the linear combination associated with the highest eigenvalue, which range from -0.3608 to 0.5415 (MSSD of $908\text{x}10^{-4}$). Moreover, the training set outcomes of the linear combination associated with the smallest PC range from -0.0252 to 0.0307 (MSSD of $2.2\text{x}10^{-4}$). The phenomenon described above when discussing model 1 regarding the cancelling of the damage of multicollinearity, seems to hold for model 2 as well. For the test set predictions the variable TOL is an exception. Relatively bad predictions are obtained, particularly for the PHE phase. The RMSEP of TOL at the individual mobile phase compositions on the PHE phase are 0.119 at wm1, 0.389 at wm2, 0.041 at wa1, 0.219 at wa2, 0.092 at am1 and 0.253 at am2. For the "serie 2" mixtures, the predictions are bad, the predicted value is always much too high. This is in agreement with the results in Section 10.1 where it was shown that the combination of the markers BAB, EHB and PAR are not the optimal choice for the "series 2" mixtures. The influence of elution strength is not reflected adequately by the markers on the PHE column to become good predictors of TOL on the PHE phase. Incorporation of A and M improves the prediction of TOL on PHE, see model 1.

Again the predictions on the C6 phase are slightly worse than average. Especially EAB (and TOL) are badly predicted. The largest prediction errors with regard to EAB on C6 are made at the mixtures wa1 and wa2. The retention of EAB is underestimated at those mixtures. A similar line of reasoning to explain these bad predictions can be followed in the case of PE on PHE with model 1. The relationship between EAB and its predictors BAB, EHB, PAR differs between the training set and the C6 phase. Note that incorporation of the mobile phase variables improves the prediction of EAB on the C6 phase.

The leave-one-out RMSEP is a reasonable predictor of the test set RMSEP except for PE, as the leave-one-out RMSEP for PE is high. The leave-one-out RMSEP of PE on the different phases are 0.0759 on C1, 0.1226 on C18 and 0.0747 on CN, so the leave-one-out predictions are particularly bad on C18. This subject shall be treated in Subsection 10.4.4.1.

A discussion of the differences between model 1 and model 2 should include two aspects. The first aspect concerns the differences between the models with respect to their predictive performance. The second aspect concerns the performance of different validation criteria which are used to choose between the different models. The second aspect is discussed in Subsection 10.4.4. With regard to the first aspect, the RMSEP in the test set as a whole is treated firstly. Model 1 surpasses model 2 for the solutes EAB, TOL, MHB and PHB, although the difference for MHB is not very large. PE is better predicted by model 2, and for ACP there is no difference between the models.

When comparing the RMSEP of the models for the individual phases in the test set (Tables III.12b and III.13), the reproducibility of the separate solutes should be taken into account. The differences between the predictive performances of the models is especially manifest for the PHE stationary phase. The variable PE has a lower RMSEP on the PHE phase if model 2 instead of model 1 is used: 0.0517 versus 0.0935, which is a clear difference compared with the reproducibility of PE on the PHE phase (0.079). The incorporation of the variables A and M in the model for PE only increases the variability of the predictions on the PHE phase, and do not contribute much to the influence of the other variables, the markers (see Subsection 10.4.4.1). With the variables EAB and TOL the reverse is true. The RMSEP for EAB is slightly lower if model 1 is used, compared with the reproducibility of EAB on the PHE phase, which is 0.105. For TOL, a large difference between the RMSEP on PHE for model 1, 0.1287, and for model 2, 0.1758, shows up compared with the reproducibility of TOL on the PHE stationary phase, which is 0.1098. It can be concluded that model 2 shows a lack-of-fit in predicting TOL.

In order to get an impression of the predictive performance of the models, some predicted versus observed values are given in Table III.14

10.3.2 Ridge regression (RR)

The results of ridge regression (RR) applied to model 1 (see Subsection 10.3.1) is discussed firstly. The y variables and the columns of the X matrix are all mean-centered and scaled-to-length-one (see Section 2.1). In contrast with the OLS technique this scaling has influence on the RR results[20]. The question whether or not to scale is debated in the literature (see Section 1.7). The data are scaled in such a way that X'X is in correlation form[21]. By centering y and the columns of X, a constant is implicitly assumed, which is not subjected to the shrinkage operation of the ridge estimation[22].

The influence of the k parameter on the estimated coefficients can be seen in Figures III.11a to III.11d, the so-called ridge traces. Even within a very small range of the k parameter considerable changes in the coefficients of BAB and EHB are observed. The estimated coefficient for the marker EHB changes sign in the models of PE, EAB and TOL: applying RR with a k parameter of 0.01, instead of k=0.00 (OLS), changes the coefficient of EHB in the model of PE from -1.368 to 0.808. These changes in estimated coefficients show clearly that the OLS coefficients, as presented in Subsection 10.3.1, should be interpreted carefully. The instability of the coefficients of the markers EHB and BAB is the result of their strong correlation. Note that the sum of the estimated coefficients of BAB and EHB remains almost constant if the ridge results are compared with the OLS results. Clearly, the sum of β_{BAB} and β_{EHB} can be estimated by OLS and is not affected by the multicollinearity[20]. This may indicate another reason for the good predictive performance of OLS: the markers BAB and EHB are exchangeable to a large extent (correlation of 0.99) and their simultaneous influence can be is estimated despite the multicollinearity.

For almost every non-marker, the estimated coefficients of the mobile phase variables A and M level off and decrease. According to Hoerl and Kennard[23], such behaviour points to the loss of predictive power of these variables. Note that as a result of the scaling, the coefficients are standardized and the impact of a variable is reflected in its estimated coefficient. For TOL the variable A does not decrease very much. Ridge traces can be used for the purpose of variable selection[23,24]. This subject will be discussed later (Section 10.5). However, it should be pointed out that discarding the mobile phase variables in the way suggested by the ridge traces is not advantageous for all non-markers, especially not for TOL, with respect to the test set results.

The role of the marker PAR is discussed separately. For the variables PE and EAB the coefficient of PAR decreases rapidly for increasing k values. The PAR coefficients for the other variables (ACP, TOL, MHB and PHB), however, do not diminish when k increases. Conclusions regarding the predictive power of PAR are difficult to draw.

The choice of the k parameter is, of course, essential. A full discussion on this topic is postponed till Subsection 10.4.2. The method used here employs the leave-one-out method (see Section 3.2).

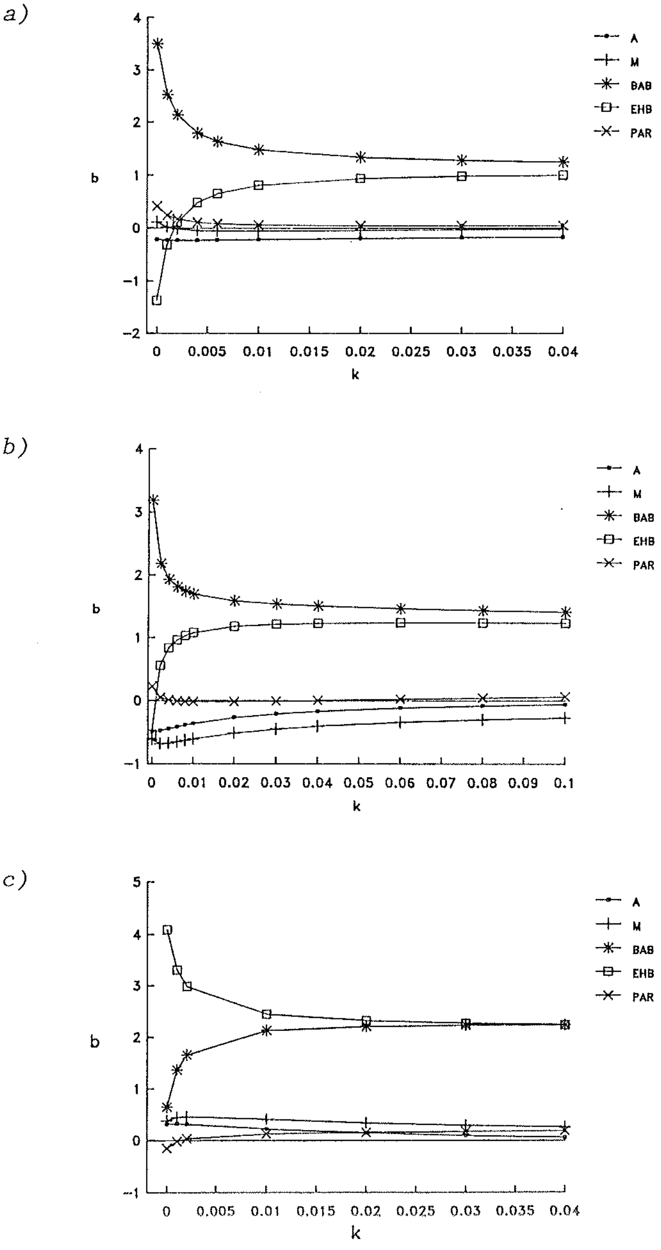


Figure III.11. Ridge traces of model 1 for the variables PE (a), EAB (b) and TOL (c). For the abbreviations, see the text.

d)

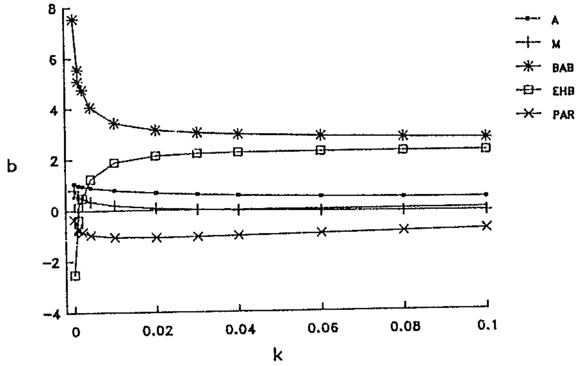


Figure III.11 (continued). Ridge traces of model 1 for the variable PHB (d). For the legend, see Fig. III.10a.

The validation of this leave-one-out method for the choice of k isalso postponed till Subsection 10.4.2. For a grid of k values between 0.00 and 0.20 the leave-one-out prediction error sum of squares (LOOPRESS) outcomes are compared. Assuming that the LOOPRESS is a unimodal function of k, the value of k can be established by approximating the minimum of the LOOPRESS. Some typical LOOPRESS versus k functions are shown in Figures III.12a to III.12d.

Table III.15 summarizes the results of the ridge regression when k is chosen with the leave-one-out method. All k parameters are low. The standard deviations of the estimated coefficients can be calculated (see Section 3.2). For the variables involved in the multicollinearity, EHB and BAB, the standard deviations are much lower than the OLS results. The standard deviations of the residuals, the s values, are always higher than the corresponding OLS results. This is due to optimum properties of OLS, see Section 3.2. Only for the variable TOL the s value exceeds the standardized reproducibility, but no serious lack-of-fit is present.

A comparison of the Tables III.12b and III.15b shows the effects of ridge regression in terms of predictions (see also Figures III.12a to III.12d). In case of PE, EAB and TOL it is advantageous to use ridge regression, the opposite is the case for ACP and MHB. The solute PHB is an exception because OLS and RR coincide. The advantage of using RR in case of EAB and TOL is a minor one. This is in agreement with the observation of Hoerl *et al.*[16]. They show in their study that the improvement of RR over OLS is not large when F ratios of 150 and more

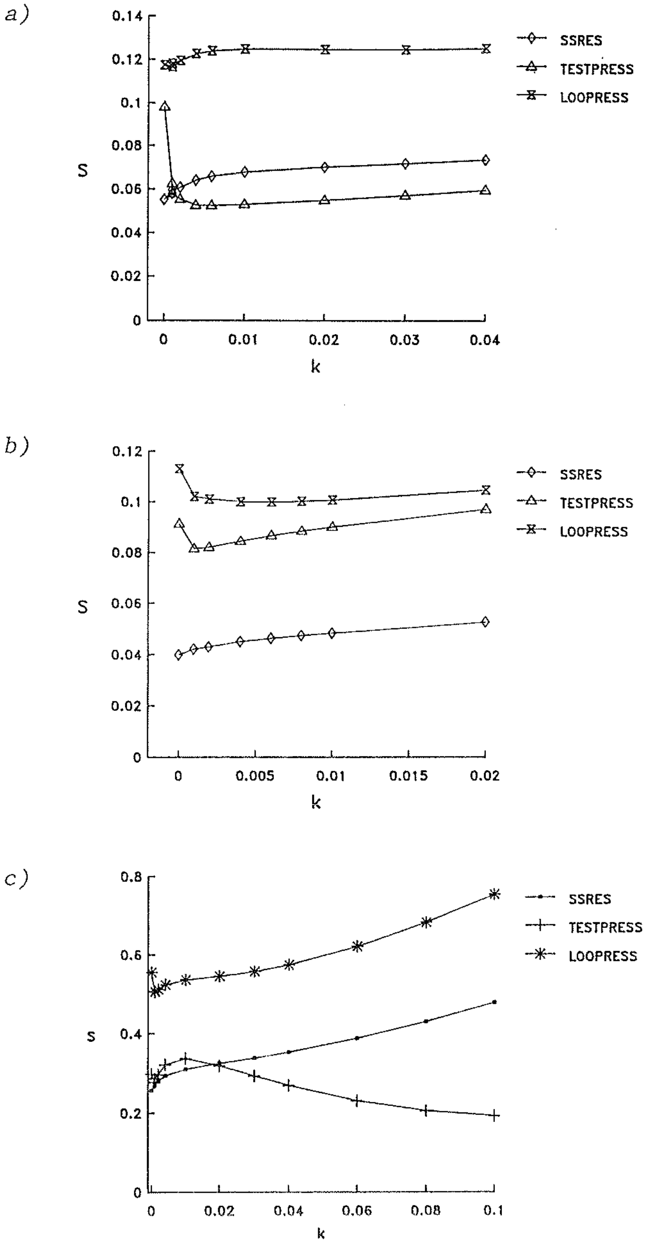


Figure III.12. Plots of sum of squares (S) against k for model 1, PE (a), EAB (b) and TOL (c). SSRES is the residual sum of squares, TESTPRESS is the sum of squared prediction errors in the test set. For LOOPRESS, see text.

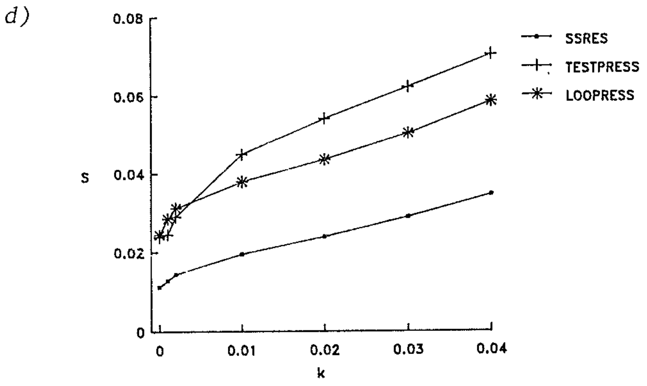


Figure III.12 (continued). Plots of sum of squares (S) against k for model 1, PHB (d). Legend, see Fig. III.12a.

are obtained in OLS. This is the case for all non markers. The improvement in the predictions for PE and TOL are mainly caused by the predictions on the PHE stationary phase. This can be checked by comparing the mean RMSEP values of OLS and RR with the leave-one-out k parameter. The reason for the specific improvement of the prediction of PE and TOL on PHE is not clear.

An assessment of the use of RR is obscured by the necessity to choose k from the training data. This choice may not be the optimum one. If the test set results only are considered, an optimal k value can be calculated: i.e. the k value that gives rise to the lowest RMSEP in the whole test set. The results are presented in Table III.17. Ridge regression is only useful in case of the prediction of PE, EAB and TOL. The solutes PE and TOL profit the most from RR; PE especially when predicting on the PHE column, TOL in the predictions on the C6 and C8 columns.

The profit from using ridge regression for the predictions on the individual phases can be expressed by comparing the mean RMSEP values of these phases for the OLS- and the ridge case. All stationary phases in the test set profit from the ridge operation. From the MSSD values it could be expected, that the multicollinearity harms the predictions on the C6 and PHE phase most, although not serious, as argued in Subsection 10.3.1. Yet, also C8 profits from the Ridge regression.

A modified t-test is proposed by Hoerl *et al.*[16]. In this test the ratios $b_i(k)/s_{bi(k)}$ are used to test the hypotheses H_0: $\beta_i(k)=0$, where $\beta_i(k)=E(b_i(k))$. The denominators of the ratios are the coefficient's standard deviations (see Section 3.2) assuming a fixed k, they ignore the bias. The $b_i(k)$ and $s_{bi(k)}$ values are reported in Tables III.15 and III.17. The threshold value of a ratio to indicate a significant coefficient is roughly 2. For the solute TOL a considerable change in important predictor variables is observed when applying RR, reflected in Table III.17. The variables PAR and EHB become important in predicting TOL, which was obscured by the multicollinearity in the OLS results. The reason why ACP, MHB and PHB do not profit from the RR may be that a given vector of random errors in the y variable, causes the non-existence of a suitable k in a particular application[24].

The results of the ridge regression on model 2 are given in Table III.16. For ACP, EAB and PHB the leave-one-out method recommends OLS (k=0.0). Some typical ridge traces are shown in Figures III.13a to III.13c. Except for PE, the estimated coefficients for EHB and BAB change rapidly. The marker PAR has only some relevance for EAB and MHB. For ACP, EAB and TOL the OLS estimates of β_{EHB} and β_{BAB} have opposite signs. This is suspect because the markers EHB and BAB have a highly positive correlation. They describe, therefore, the same aspect of the training data and should influence the non-markers similarly. According to Hoerl and Kennard[23], the reason for the opposite signs in the OLS estimates might be the negative covariance between the estimates (one of them is estimated too high and the other one therefore too low, becoming negative). This is indeed the case: the covariance between b_{BAB}(ols) and b_{EHB}(ols) is $-360\sigma^2$. In the ridge analysis they do not keep their opposite signs. This can be concluded from Figures III.13b and III.13c. The reason why the ridge trace for PE is stable in comparison with the other non markers is not clear. For those solutes where the leave-one-out method advises a k value >0, only TOL profits from the RR, which can be seen in Table III.16b: the test set results. If the individual test set phases are observed, it becomes clear that only the PHE phase profits to some extent from the ridge regression with the leave-one-out k parameter.

The usefulness of RR can be assessed with Table III.18. Again the solute TOL profits the most from the RR, especially in the predictions on the PHE column. A comparison of the mean RMSEP values of RR, if the best ridge estimator is applied, and OLS reveals that only the RR predictions on the PHE phase are better than the OLS predictions on that phase. This conclusion is in agreement with the high MSSD value of PHE on the last PC of the X matrix of model 2. The stationary phase of the test set that profits most from RR, shows a pattern in the ln k values of the markers which resemble the least the "multicollinear pattern" in the training set.

A comparison between model 1 and model 2 can be made with the test set results given in the Tables III.17 and III.18. Model 1 has always the best performance, although the differences for ACP and MHB are small. Of special interest is PE because this solute is predicted best with model 2 when OLS is used and with model 1 when RR is used.

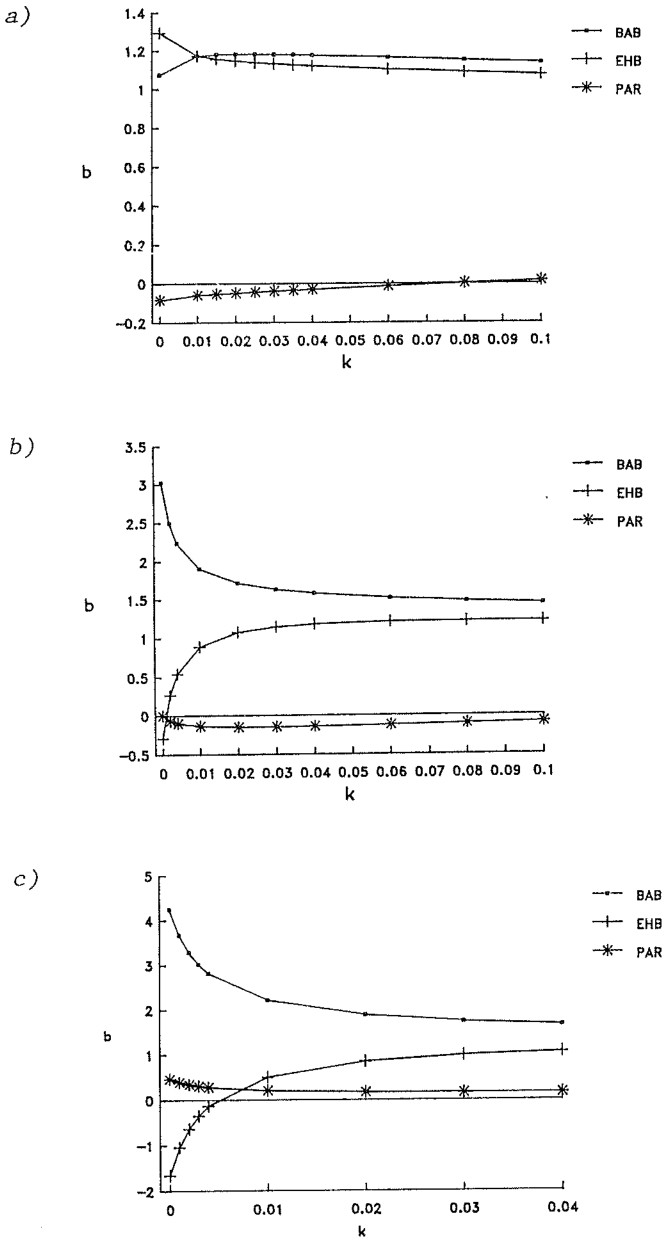


Figure III.13. Ridge traces of model 2 for the variables PE (a), ACP (b) and EAB (c). For the abbreviations, see text.

This will be commented on in Subsection 10.4.4.1. A large difference is observed with the RMSEP on the PHE phase for TOL, model 1 predicts TOL much better on the PHE phase.

10.3.3 James-Stein regression (JSR)

As already indicated in Section 3.3, a necessary condition for the JS estimator to be better than the OLS estimator is $d=1_m\Sigma\lambda_i^{-1}>2$ where $\lambda_1\geq\ldots\geq\lambda_m\geq0$ are the ordered eigenvalues of X'X (m is the number of predictors in X). In case of model 1, d=1.05 and for model 2, d=1.04. Note that another necessary condition in this context, namely that the number of predictors should be greater than two, is exactly fullfilled in model 2. There is no guarantee that the JS estimator will give a smaller PMSE than the OLS estimator for all values of β. Yet, there might be a value of β for which the PMSE of the JS estimator is smaller than the PMSE of the OLS estimator. It is, therefore, still worthwhile to calculate the JS estimator. This holds both for model 1 and model 2.

For both models the y and columns of X are centered and scaled-to-length-one (see Section 2.1). Scaling influences Stein regression and the traditional scaling procedure[21] is followed. Note that, by adopting this kind of scaling, a constant is implicitly assumed, which is not subjected to the shrinkage operation of the Stein regression[20]. The results of model 1 are presented firstly.

The value of c is chosen with the leave-one-out method in the same way as the ridge parameter in the previous chapter. The results of the calculations are reported in Table III.19. The first part of Table III.19a shows that for TOL and PHB no shrinkage is recommended by the leave-one-out method. For the other variables mild shrinkage is advised and the s values increase therefore slightly. In Table III.19b the results of the predictions are shown. When these are compared with the OLS results of model 1 (Table III.12), it appears that only the predictions of ACP improve to some extent, the other variables for which shrinkage is proposed are predicted worse. The mean RMSEP values of the individual phase show that the Stein regression with the leave-one-out c choice does not improve the predictions.

Whether or not these conclusions are influenced by the leave-one-out choice of c can be checked by calculating the best c value with the use of the test set results (the assessment of the leave-one-out criterion for choosing c is postponed till Subsection 10.4.3). The choice of the best c value is carried out similarly to the leave-one-out choice of c, only in this case the PRESS values in the test set are used, which cannot be done in practice, of course. The best c values are given in the second part of Table III.19a. Only for ACP, TOL and PHB shrinkage is advantageous. The prediction results are shown in Table III.19c. When applying Stein regression, the mean RMSEP values of the individual phases show a slight improvement of the predictions on C6 and C8 and a worsening on the PHE phase. There is no connection between this conclusion and the MSSD values (the multicollinearity patterns) of the test set phases. The variable ACP gains a little from the shrinkage, especially on the PHE column. For TOL the gain is in the predictions on the C6 and C8 column, while a considerable worsening is seen in the predictions on the PHE column. The same holds for PHB. The reason for this phenomenon is not clear. The conclusion is that the shrinkage does not provide considerably better estimators in the MSEP sense for model 1, probably because of the severe multicollinearity in X.

By examining Table III.20a, it is clear that, in case of model 2, the leave-one-out procedure only advises shrunken estimators for PE and TOL. The predictions of these solutes (Table III.20b) improve little for PE and become worse for TOL. No improvement of Stein estimation (with the leave-one-out c choice) is seen with regard to the mean RMSEP values.

The best choice of the c value (Table III.20a, second part) suggests shrinkage for ACP, TOL and PHB, again using the test set predictions. For ACP this results in only slightly better predictions (Table III.20c), for PHB the RMSEP in the test set decreases clearly. It is, however, difficult to assess this decrease in relation to the estimated reproducibility of PHB in the test set. The profit TOL gains from the shrinkage is minor. The mean RMSEP values for C6, C8 and PHE when applying Stein regression (with the best choice of c) show hardly any difference with the OLS ones, perhaps a slight improvement occurs on the C8 phase. Again, as in model 1, the predictions on the PHE column become worse. The conclusion with regard to the use of JS regression in estimating the parameters in model 2 is that only for PHB a small advantage is obtained in terms of PMSE. A comparison of the different modelspecifications 1 and 2, when subjected to the JS regression, can be done with the use of Table III.19c and Table III.20c. The same conclusions can be drawn as in Chapter 10.3.1 with regard to the OLS results: only for the solute PE model 2 performs better. Again the success of model 2 is in the prediction of PE on the PHE column.

10.3.4 Partial least squares (PLS)

The PLS calculations are performed with the X-block variables A, M, BAB, EHB, PAR and the Y-block variables ACP, PE, EAB, TOL, MHB, PHB. Modelspecification 2 is not estimated with PLS because in that case X will have only three columns which will make the choice of PLS less obvious (note that if in the PLS calculations g=m, the number of dimensions is the number of columns in X, then PLS becomes OLS[25]). All variables are column-autoscaled. Figure III.14 depicts this set up. The number of PLS components is determined to be three by cross-validation, see Subsection 10.4.1. In Table III.21 some diagnostics and the results of the predictions are shown. The loading of a variable on the PLS component of a specific dimension (Table III.21a) measures the contribution of that variable to that PLS component. For variables within the X-block the contribution, measured with the absolute value of the loading, can be interpreted as the amount of variation in the variable that is used to explain the Y-block variation. For variables within the Y-block, the contribution (absolute value of the loading) can be interpreted as the amount of explained variation of the specific variable.

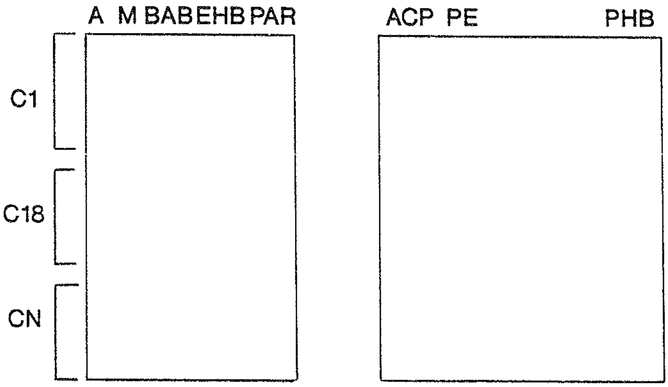


Figure III.14. Set-up of the PLS calculations.

The value s^2_{unexpl} is calculated as the variance of the variable after the contribution of the variable to the three PLS dimensions. Although the exact behaviour of these values is not yet clear[26], they can be used for indicative purposes. In case of X-block variables, these s^2_{unexpl} values can be interpreted as the unused variance in the modelling process. For the Y-block variables, s^2_{unexpl} can be interpreted as the variance unexplained by the model. Note that, at the beginning of the estimation process each variable has variance one, so that each s^2_{unexpl} equals one. The square root of the s^2_{unexpl} values are reported in Table III.21a and can be compared with the s values obtained in the OLS estimation. After applying three dimensions, 99.42% of the variation in the X-block is used to explain 99.21% of the Y-block variation. The markers BAB and EHB contribute mainly in the first dimension, the mobile phase variables mainly in the second one. The marker PAR plays a particular role and contributes to a large extent in the third dimension, which is in agreement with the principal component analysis on the X-block (see Section 10.2). Especially the non-marker TOL profits from this contribution of PAR because a high loading of TOL on this PLS component is observed. This is in accordance with the ridge regression results, see Table III.15 and Table III.17, where PAR contributes to the explanation of the variation of TOL. A warning with regard to the interpretation of these loadings is appropriate: it might be concluded that ACP also profits, to a lesser extent than TOL, from the contribution of PAR, but this is contradicted in Table III.17a.

The results of the predictions in the test set are given in Table III.21b. When these results are compared with the OLS results for model 1 (Table III.12b), it appears that only PE is better predicted by PLS, especially on the PHE and C6 column. The retention of the other non-markers are worse predicted by PLS than OLS, sometimes considerably worse e.g. TOL on C6. The mean RMSEP values for C6, C8 and PHE when applying PLS show, in comparison with the OLS ones, no improvement. However, the predictions on the C6 and PHE phases are worse. It should be kept in mind that PLS is a biased estimation procedure, which does not give always better predictions than the unbiased one: OLS. Another reason for the weak performance of PLS in relation to OLS might be the influence of autoscaling (both X- and Y-block) on the predictive performance of PLS[27]. This is still to be investigated.

10.4 The performance of cross-validation.

10.4.1 Determining PLS model-complexity

The use of cross-validation (CV) in estimating the optimal model complexity in PLS is discussed firstly. In the PLS calculations, as described in Subsection 10.3.4, the PLS algorithm was used as implemented in the SIMCA-3B program. All non-markers were collected in one Y-block and simultaneously modelled. Whether it is appropriate to model each y variable separately depends on the postulated latent variable structure of the Y-block. In the case at hand a latent structure within the Y-block measurements is supposed. The hydrophobicity of the stationary phase, the polarity of the mobile phase and the polarity of the solutes are ingredients of this latent structure. All y variables are the manifest variables of that structure and should be incorporated into the model simultaneously. An advantage of this ste-up is that it creates the possibility to reckon with the measurement error in the x variables[28].

The CV as implemented in the SIMCA-3B program is not a leave-one-out method, but splits the Y-block data in four groups and leaves these groups out successively. For each dimension in the PLS model, the sum of squares in the Y-block is calculated before applying the next dimension (SSBEF). In calculating the next dimension of the model the ln k values of the omitted group are predicted and the quadratic differences with the actual values are summed in the sum of squares due to Cross-validation (SSCSV). This is done four times to ensure that each data point in the Y-block is predicted once by the next dimension. The SSBEF can be interpreted as the error sum of squares if, instead of applying the next dimension, each Y-block data point is predicted with the value of zero. If the ratio SSCSV/SSBEF is smaller than 1 the next dimension in the model will predict better than the zero values.

As a threshold for the significance of the next dimension the SIMCA-3B manual advises that the square root of SSCSV/SSBEF (CSV/SD) is smaller than 0.95. These CSV/SD values can also be calculated also for each of the y variables separately. These values should be taken into account when a decision is made on the significance of a new dimension, i.e. when the overall CSV/SD is greater than 0.95 but a considerable number of individual CSV/SD is smaller than 0.95, this dimension (component) can still be considered significant.

The above sketched CV procedure showed the following characteristics when applied to model 1. The first three dimensions gave overall CSV/SD values of below 0.95. The fourth dimension gave an overall CSV/SD value of 1.00, indicating non-significance. Two of the six individual CSV/SD values were below 0.95 and the increase of explained variation in the Y-block was 0.09%, which is very low. The fifth dimension gave an overall CSV/SD value of 0.94, which is on the borderline, and explained only an additional 0.14% of the variation in the Y-block (see also Table III.21a). It seems therefore reasonable to choose three dimensions.

The results of the leave-one-out strategy are given in Table III.22. Concentrating on the rows with entry "TOTAL" it appears that 5 dimensions give the lowest leave-one-out prediction error sum of squares (LOOPRESS) summed over all non-markers. Actually, PLS with five dimensions is OLS because X is a 18x5 matrix. Osten[29] describes how to use PRESS values in the decision on the number of PLS components: choosing the number of dimensions which corresponds to a local minimum in the PRESS versus number of dimensions plot is a reasonable procedure. An alternative method to decide on the significance of the difference between two succesive PRESS values is based on an F test. This F test, however, is outlined only for the case of a single y variable, which is the situation at hand. Besides, the F test lacks a rigorous statistical treatment. The local minimum in the PRESS values is attained at five PLS dimensions.

Selecting five PLS components gives rise to the lowest PRESS in the test set, so that, obviously, five components is the best choice. This TESTPRESS value, associated with five PLS dimensions, is evidently smaller than the value associated with three PLS components, which is chosen on the basis of the PLS results obtained from the SIMCA-3B program. Several reasons for the discrepancy between the use of CV from SIMCA and the leave-one-out procedure with a PRESS value can be given. First, the SIMCA program splits the data in four groups which are subsequently omitted while the CV procedure reported in Table III.22 leaves out one data point at a time. Osten[29] shows that this might influence the situation when the data are not an adequate sample of the population. Another reason for the above mentioned discrepancy is the "soft" way by which the significance of a new calculated component must be judged in the SIMCA program. This leaves some margin for the investigator to decide. The leave-one-out method is preferred because its statistical basis is to retain as much as possible of the original data in the training set to evaluate properly the predictive power of the model (see Section 2.3). A leave-one-out method might be more robust against non-adequate sampling.

In Table III.22 the LOOPRESS and TESTPRESS values are also given for the separate non-markers. It should be kept in mind that the modelling process is done in such a way that all non-markers constitute the Y-block. Discrepancies arise between the number of PLS components advised by the leave-one-out method and the best test results for the individual non-markers. Whether the non-markers should be modelled individually to obtain a better fine tuning of the number of components, by means of an leave-one-out method, is not known.

10.4.2 Determining the ridge parameter k.

Several procedures for choosing the ridge parameter k are described in Section 3.2. The (ordinary) cross-validatory estimate of k is compared with reasonable alternatives for the situation at hand: Hoerl's recommended value $k=ms^2/(b'_{ols}b_{ols})$ and the k value as obtained from the ridge trace. The McDonald and Galarneau alternative seems also reasonable, but is not implemented yet. The method using the variance inflation factors of b_{rr} has the disadvantage of not taking the y variable[30] into account. The generalized cross-validation (GCV) was designed to meet the problem of near-orthogonal design matrices. This GCV is not necessary in our case because the design matrices are not near orthogonal.

The results for model 1 are shown in Table III.23. For the sake of convenience this table partly duplicates Table III.17. When all three alternatives are considered, only for the solutes MHB and PHB the leave-one-out strategy yields a k value which gives rise to a TESTRMSEP as close as possible to the best TESTRMSEP. For the other non-markers Hoerl's choice (four times) and the trace choice (two times) give rise to the "closest" TESTRMSEP. There is no clear preference for the leave-one-out strategy. The other strategies do not perform any worse and are computationally much more convenient. A greater variety of modelled non-markers is needed to evaluate properly the success of CV in establishing k. For this moment Hoerl's choice seems to be reasonable.

The results of model 2 are presented in Table III.24. A comparison of the different choices of k is skewed by the fact that Hoerl's method and the leave-one-out method give rise to the same k in three cases. The performance of the ridge trace criterion is worse than in model 1. Hoerl's method, however, performs well and so does the leave-one-out method. The only non-marker for which the leave-one-out method is better than Hoerl's method is ACP. The conclusion is that also for model 2 Hoerl's method for choosing k is reasonable. The only improvement in choosing k might consist of applying the algorithm given by McDonald and Galarneau (see Section 3.2), because this algorithm seems to be particularly appropriate for large signal-to-noise ratios[16] which is the case with the models described above.

10.4.3 Determining the James-Stein shrinkage parameter c.

The leave-one-out procedure applied to calculate c is compared with the c value as proposed by Stein (see Section 3.3 formula I.10): the positive Stein rule estimator.

The results of this comparison are summarized in Table III.25. When the c choice procedure, which gives rise to a TESTRMSEP as close as possible to the best TESTRMSEP value, for model 1 is counted, it appears that the STEIN choice beats the leave-one-out choice. For model 2 such a comparison remains undecided.

There is no clear preference for the leave-one-out method. The STEIN choice has the advantage of being simple to calculate.

10.4.4 Choosing between the modelspecifications.

Two different models (modelspecifications) are calculated throughout: model 1 with the inclusion of the mobile phase fractions and model 2 with these fractions omitted. Cross-validation can be used to make a choice between these models. For the OLS estimation method different criteria are available to judge the predictive performance of a model, see Section 2.3. Generalizations of such criteria for ridge- and Stein regression are also reported[31,32]. The discussion concentrates on the criteria for the OLS estimation. Specifically, the prediction criterion of Amemiya (PRC), the C_p value of Mallows and the R^2_{adj}, are compared with the cross-validation (PRESS), see Section 2.3.

Table III.26 contains the result of this comparison. The ultimate decision which model has the best performance can be made with the yardstick RMSEP in the test set, as noted down in Tables III.12 and III.13. The values of C_p, PRC and PRESS have to be minimized, whereas R^2_{adj} must be maximized.

Concentrating on PE, EAB, TOL and PHB, where clear differences between the predictive performance of the respective models are observed, the following can be concluded. For PE, all criteria give the wrong advice, i.e. the criteria do not recognize model 2 as the best model for predicting the ln k values of PE. For TOL, all criteria give the good advice, although C_p is not very decisive. In case of PHB three criteria give the right advise, but R^2_{adj} is indecisive. The PRESS criterion performs worse than the other criteria in case of EAB. A preliminary conclusion is that PRESS is not the best criterion for choosing between modelspecifications because the PRC criterion performs better. A more exhaustive evaluation of the criteria is needed with a wider range of non-markers and model specifications, e.g. by choosing a subset of the markers for the prediction of a particular non-marker (see Section 10.5).

10.4.4.1 A closer look at the solute PE

There are two reasons to take a closer look at the results for PE. First, PE is the only solute for which model 2 is the best and second, no validation criterion - PRESS, R^2_{adj}, PRC or C_p - indicates this. The behaviour of the PRESS criterion is treated first. A PCA on the augmented matrix [X,y], as described in Section 2.4, reveals some interesting features. In Table III.14b the loadings of this PCA of the training data for model 1 and 2 are given, together with the variation explained by the components. Model 2 is treated firstly.

The last component reveals a non-predictive multicollinearity between BAB and EHB (see Section 2.4). PE loads high on the third component but no clear pattern emerges on which predictors are connected with PE for this dimension. The leave-one-out RMSEP of the data points C1wm1, C1wm2, C18wm2, C18wa1 and C18wa2 are respectively 0.114, 0.135, 0.199, 0.146 and 0.140, the five highest values of all eighteen. These values can be compared with the leave-one-out RMSEP, Table III.13b, which is .0937. Deletion statistics are indeed capable of detecting influential observations but the PRESS value becomes high when there are influential observations. In order to illustrate the effect of masking[19], data points C18wa1 and C18wa2 are omitted simultaneously and predicted by the model on the basis of the sixteen other data points. The RMSEP values are respectively 0.253 and 0.262. This is considerably worse than the leave-one-out results because the influence of C18wa1 is no longer masked by C18wa2 and vice versa. The PRESS value is sensitive to leverage points and influential observations. To assess the meaning of a PRESS value in a specific situation, a careful analysis with regard to leverage- and influential points is needed.

How the observation of the high leave-one-out RMSEP value for model 2 is related to the conclusion drawn in Section 10.1 regarding the peculiar role of PE in the marker choice assessment is difficult to disentangle. Two possibilities for the absence of PE in one of the best subsets were given, either PE is a bad predictor or the retention of this solute can be predicted very well. The tentative conclusion can be drawn that PE is a bad predictor. Note the close connection between the results of model 2 for each non-marker and the induced-variance criterion: the mean R^2 (averaged over all non-markers) is equal to the induced-variance.

The same analysis can be performed in case of model 1. The loadings on the sixth component of the PCA on the augmented model 1 matrix, Table III.14b, reveal a non-predictive multicollinearity between BAB and EHB. The variable PE is mainly present in the fifth dimension and is related, to some extent, to the predictor BAB. The data points

Clwm2, C18wm2 and C18wa2 have a high leave-one-out RMSEP, respectively 0.173, 0.142 and 0.139. No related observations are seen in Figure III.9d and no further attempts were made to investigate masking effects. Note that examining Figure III.9d only reveals masking leverage points, not masking influential points.

A conclusion is that the high leave-one-out RMSEP for model 2 is (partly) due to some influential observations, which are less pronounced for model 1. It should be stressed that a full analysis of the prediction of PE, and indeed of all non-markers, with the use of models like model 1 and 2 should encompass robust estimation methods[17,18,19,33]. A rigorous treatment of the impact of influential observations on the behaviour of R^2_{adj}, PRC and C_p is not readily available, so that an assessment of the behaviour of these criteria in deciding whether to use model 1 or model 2 is difficult.

The question remains why model 2 predicts PE better on the PHE column than model 1. A LR-PCA on the PHE data set was performed to reveal this difference. The results for model 1 and model 2 are recorded in Table III.14b. In order to make an honest comparison between training set and test set results, the PHE data are, prior to the LR-PCA, subjected to the scaling constants of the training set. The loadings on the second and sixth component in the LR-PCA of model 1 are important to PE, especially the sixth. In neither of these dimensions the mobile phase variables are present, which indicates that these variables possess no clear predictive power. This conclusion is supported by the results in Table III.17, where the best ridge results are presented of model 1. The mobile phase variables have hardly any influence, see Table III.17a. The mobile phase variables, therefore, only attribute to the variance in the prediction of PE if incorporated in the model.

A particular phenomenon is observed when comparing Table III.15b with Table III.16b and Table III.17b with Table III.18b. For both the leave-one-out and the best choice of the ridge parameter, model 1 predicts better than model 2 in case of PE. This is in agreement with the conclusions and warning of Hoerl and Kennard[22] and Hoerl *et al.*[16] that variables that lose their predictive power and obtain a low coefficient estimate in the equation should not be discarded in the ultimate prediction equation.

10.4.5 Choosing between estimation methods.

A particular advantage of cross-validation is the potential it offers to choose between estimation methods. The leave-one-out method is utilized to obtain PRESS values for different estimation methods, for each non-marker separately. The method with the lowest PRESS is advised as the best prediction method. Table III.27 summarizes some results. Concentrating first on model 1, Table III.27a, it is clear that cross-validation gives a good advice for the solutes EAB, TOL and PHB. For ACP cross-validation advises the worst method, PLS (the highest TESTPRESS value) and for PE and MHB the second- respectively the third best choice. Cross-validation performs well in advising the estimation method, except for ACP.

For model 2, Table III.27b, the assessment of the potential of CV to choose between estimation methods is obscured by the fact that for three solutes no alternative estimation methods are available. When only observing the non-markers PE, TOL and MHB it appears that CV advises two times the best method, and one time the second best. Also for model 2 CV behaves reasonably.

It can be concluded from Table III.27 that there is no best estimation method, although there may be a slight advantage in using ridge regression. For diagnostic purposes, however, ridge regression is useful, probably in combination with influence measures in ridge regression[14,34].

10.5 Fine tuning the modelspecification of EAB

As an example model 1 is taken as starting point for the non-markers to illustrate modifications of the model. Eliminating variables, which are of no predictive importance, gives rise to a decrease of variance in the predictions of the non-markers. An example is given by fine tuning the modelspecification of EAB.

There are two possible modifications for EAB, starting from model 1. The first modification emerges from the observation that the marker PAR is, perhaps, obsolete in the prediction of EAB. Indications of this conclusion can be found in the low significance of PAR in the OLS calculation of EAB, see Table III.12a, in the behaviour of the estimated coefficient of PAR in the ridge trace, Figure III.11b, and in the high loading of PAR on the third component of the LR-PCA on EAB, Table III.28a, which is hardly associated with EAB. If the marker PAR is omitted, the resulting model is called model 1a. The results are as shown in Table III.28b and Table III.29a. The design matrix, associated with model 1a, has a better condition: a condition index of 15 instead of 45 for the full model, so that there is less multicollinearity. This indicates that the deletion of variables is one of the remedies against multicollinearity (see Section 2.4). The coefficient of EHB in the model equation changes considerably. The variance inflation factors decrease considerably, resulting in a much lower standard error of the estimated coefficients for BAB and EHB. A LR-PCA on the augmented correlation matrix of model 1a reveals no non-predictive multicollinearity between BAB and EHB, which was present in the full model. This illustrates the tentative nature of the LR-PCA considerations; the elimination of the variable PAR changes the pattern of the LR-PCA considerably, even though the correlation between PAR and BAB, EHB is low (see Table III.11). The leave-one-out RMSEP of model 1a is lower than the analogous value of model 1. The PRESS criterion, therefore, advises model 1a, which is a correct advice because model 1a predicts better than model 1, see the

TESTRMSEP values. Ridge and Stein calculations on model 1a did not improve the predictions (results not shown).

It might be argued that a further step in the fine tuning of model 1a can be made by eliminating variable EHB, because this marker is not significant (Table III.29a). The resulting model, EAB= F(A,M,BAB), has a lower leave-one-out RMSEP (0.0694) but a higher TESTRMSEP (0.0690) than model 1a. This fine tuning of model 1a is, therefore, not appropriate although the PRESS criterion does not reveal this.

The second possible modification, starting from model 1, may consist of deleting EHB. Because EHB and BAB are highly collinear, EHB is not significant in model 1, see Table III.12, and a non-predictive multicollinearity exists between EHB and BAB, see Table III.28a. Note, however, that this conclusion is contradicted by the behaviour of EHB in the ridge trace, but for the sake of illustration model 1b is postulated. Model 1b is model 1 with EHB omitted. The results are shown in Tables III.28 and III.29. The condition number of the design matrix associated with model 1b, becomes 9 and again the multicollinearity is much decreased, which was expected because one of the two correlated predictors has been left out. The leave-one-out RMSEP as well as the TESTRMSEP are lower for model 1b than for model 1, see Table III.29b, which indicates a better predictive performance of model 1b. This is noticed by the PRESS criterion.

The standard error of the estimated coefficients decrease considerably for BAB and PAR, because of a lower multicollinearity, and the marker PAR obtains a significant coefficient. Note that, if the variable PAR is deleted from model 1b, the same model as above EAB=F(A,M,BAB) is obtained. Model 1b has both a lower leave-one-out RMSEP and a lower TESTRMSEP than the model EAB=F(A,M,BAB), so model 1b surpasses this model. This is indicated by the PRESS criterion. Ridge and Stein calculations, starting from model 1b, do not yield improvement (results not shown).

If a decision has to be taken whether to chose model 1a or 1b, on the basis of the PRESS criterion model 1b is preferred. Both models, 1a and 1b, have the same predictive performance.

Some final remarks are appropriate in this context. First, note that the best ridge solution, see Table III.17, of model 1 gives a TESTRMSEP of 0.0674, which is lower than the best TESTRMSEP after fine tuning model 1. This illustrates the notion that elimination of variables is not always the preferred remedie against multicollinearity (see Section 2.4). Even the ridge results of model 1 with the leave-one-out choice of k, see Table XV, gives a TESTRMSEP close to the one obtained after fine tuning. Second, an evaluation of the PRESS criterion, when using it for fine tuning, is difficult. On the occasions that it failed, it was in distinguishing between model 1a and the model EAB=F(A,M,BAB) and in recognizing the equal predictive performance of model 1a and model 1b. An explanation of this failure might be that the PRESS value (and hence the leave-one-out RMSEP value) is too high for model 1a. This is possibly due to the high correlation between EHB and BAB in model 1a, which indicates a relation between the performance of the PRESS criterion and the degree of multicollinearity in the design matrix.

The precision of the predictions of the retention of the non-markers must be very high. Calculations show[35] that a gain in RMSEP from 0.100 to 0.0600 seems little but results in a reduction of the relative prediction error in k values from 10% to 6%. This reduction is of interest when the relative measurement precision of the k values is about 3-5%. The purpose of the calibration should be kept in mind: predicting simultaneously the retention values of a set of solutes, or stated otherwise, predicting a chromatogram. Small prediction errors result in chromatograms with overlapping peaks[36]. It is therefore of utmost importance to obtain very precise predictions.

Chapter 11 Two-way approach: the determinant criterion

11.1 The choice of the markers

The data set from which the markers must be chosen comprises the logarithms of the capacity factors of all solutes measured on the C1, C18 and CN stationary phase at the six mobile phase compositions. A principal component analysis on this data set has already been carried out and discussed in Section 10.1. Prior to the marker selection the data were centered and scaled-to-length-one column wise, for the same reasons as outlined in Section 10.1.

If the markers are chosen according to the determinant criterion the five best subsets are PAR,TOL,MHB; PAR,PE,TOL; PAR,TOL,PHB; EHB,PAR,TOL and PAR,PE,EAB. The determinants of these subsets were respectively 0.0331; 0.0306; 0.0302; 0.0299; 0.0288. The first subset determinant is slightly higher than the one for the other subsets, the last four subsets have almost equal determinants. The best subset, PAR,TOL,MHB, induces 99.47% of the variance in the whole data set (see Table III.10).

If the markers are calculated with the determinant criterion in combination with a leave-more-out evaluation, in the same way as in Section 10.1, it appears that PAR,TOL,MHB is a stable solution. Note that this subset spans the loading space of first principal components, see Figure III.7. The solute PAR is always part of the five best subsets for each omitted group, again pointing to an outlying behaviour of PAR. All other solutes were present in one or more of the previous subsets. The deviating behaviour of PE was not present in these leave-more-out results.

If the markers are chosen on the basis of the "series 1" mixtures only, the subset PAR,TOL,MHB is the second best choice, with a determinant of 0.0609. Compared with the determinant of the optimal subset, in this case (PAR,PE,EAB) 0.0610, there is hardly a difference. This does not hold for the "series 2" mixtures. The best choice is ACP,PAR,PE with a determinant of 0.0273 which is clearly higher than 0.0214, the determinant of PAR,TOL,MHB, the sixth best choice. Whether this will results in worse predictions at the "series 2" mixtures if PAR,TOL,MHB are used as markers is an open question.

The best subset of size 2 is PAR,TOL with an induced-variance of 97.59%, and the best subset of size 4 is PAR,PE,TOL,EAB with an induced-variance of 99.57%. When these induced variances are compared with the explained variances of the principal components, Table III.10c, a subset size of 3 seems reasonable.

11.2 Model specifications and evaluation of the design matrices

Two different model specifications are distinguished for reasons already discussed in Section 10.2. Model 1 is defined as (an integer indexing the non-marker is left out):

$$\ln k_{non,t}=\beta_0+\beta_1 A_t+\beta_2 M_t+\beta_3 PAR_t+\beta_4 TOL_t+\beta_5 MHB_t+\epsilon \quad (III.3)$$

where t is an index indicating the measurement conditions (stationary/mobile phase combinations); A_t and M_t are the fractions of acetonitrile and methanol, respectively, in the mobile phase and PAR_t, TOL_t, MHB_t are the ln k values of the markers PAR, TOL and MHB. The error terms ϵ_t are assumed to be distributed around zero with covariance matrix $\sigma^2 I$.

The second model specification is the same as model 1 except for the mobile phase fractions which are omitted. For details see Section 10.2.

Table III.30 reports the diagnostics for these models. Model 1 is discussed firstly. High correlations are present between A and M and between TOL and MHB. The variance inflation factors for A, M, TOL and MHB are greater than 10 and point to multicollinearity. The variance decomposition proportions, Table III.30c, show that a multicollinearity is present and caused by a relationship between the markers. Note that the degree of multicollinearity is much lower than that of the markers BAB, EHB and PAR, see Section 10.2. A weak dependence is present between A and M.

A principal component analysis on the design matrix associated with model 1 was performed (all columns in X were autoscaled). The loadings are given in Table III.30d and the scores are plotted in Figures III.15a to III.15c. When these figures are examined, it appears that the CN measurements form a separate group in the scores on the first two PC's. A clear clustering can be observed in Figure III.15c, the three stationary phases being separated. In the plot of PC1 versus PC3 two observations are outlying: 50 (C18wa2) and 53 (C18wa1). This could be related to the relative low retention of PAR on C18 at the water/acetonitrile mixtures (note that PAR loads high on the third PC), see Section 9.6. This outlying behaviour does not show up in Figures III.15a and III.15c and these data points are therefore not considered to be leverage points. On the whole, these score plots are comparable with the corresponding ones from the two-way approach with the induced-variance criterion (Chapter 10), see Figures III.10a to III.10c. The loadings of the variables on the PC's behave also similar as in case of Chapter 10. The solute PAR is almost the only one present in the third dimension. The solutes TOL and MHB constitute the final dimension.

The design matrix associated with model 2 will not be discussed extensively, reference is given to Table III.30e to Table III.30g and Figure III.16. An overall conclusion is that the corresponding design matrices for the determinant markers and the induced-variance markers bear the same characteristics, whereas the relationships are (much) more pronounced for the induced-variance case.

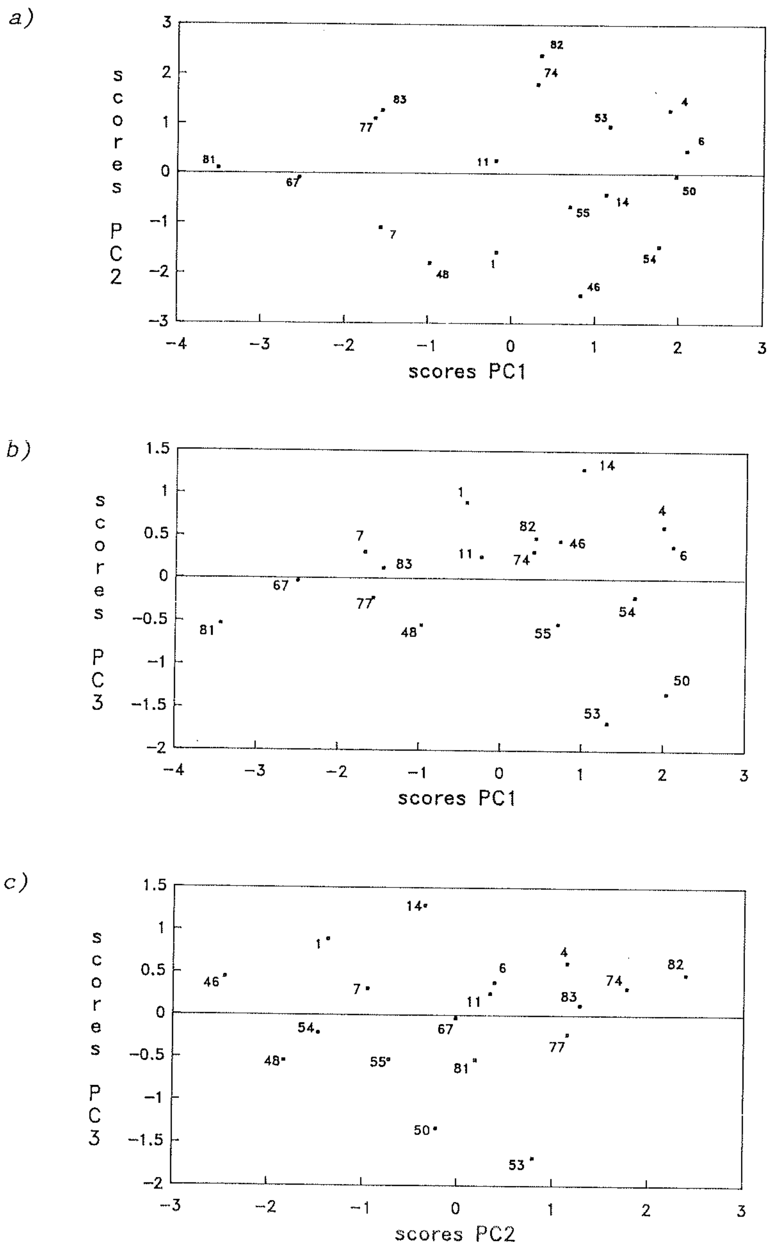


Figure III.15a,b,c. Score plots of a PCA on the design matrix of model 1. Legend, see Fig.III.6.

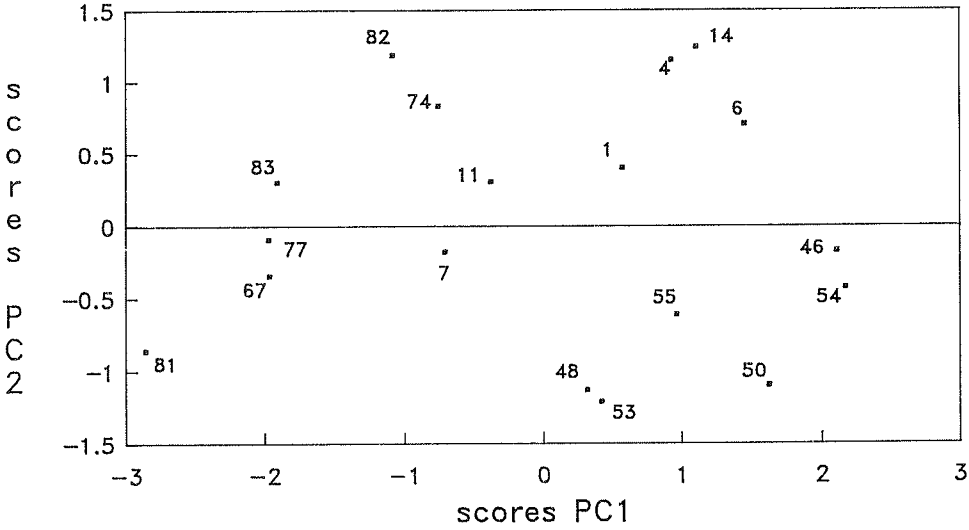


Figure III.16. Score plot of a PCA on the design matrix of model 2. Legend, see Fig.III.6.

An assessment of the quality of the test set in terms of model 1 and model 2 can be done with the use of Table III.30h. In case of both models all entries of the dispersion matrices in the whole data set and the test set are in the same order of magnitude. A matched split (see Sections 4.1 and 10.2) is therefore also present in the two-way approach using the determinant criterion.

11.3 The results of the different estimation methods

11.3.1 Ordinary least squares

The results of the OLS calculations on model 1 are summarized in Table III.31. All non-markers are modelled very well with high F-ratios and low s values. When these s values are compared with the corresponding standard deviations due to reproducibility, s_{repro}, (Table III.5) no lack of fit shows up. The standard errors of the estimates of the marker coefficients are lower than in Table III.12, where the induced-variance markers were used. This is due to the lower degree of multicollinearity in the X matrix. The marker MHB is a valuable predictor for every non-marker, except for PE. The variable TOL is a relevant predictor for BAB, EHB, PE and PHB. Again PAR plays a dubious role: it has relevance for PE. The mobile phase variables, A and M, are only relevant for PE and EAB. An assessment of the stability of the estimated coefficients under the multicollinearity is postponed till Subsection 11.3.2.

None of the RMSEP values of the whole test set (Table III.31b) is higher than the average standard deviation due to reproducibility. The predictions are good. As in Subsection 10.3.1, three reasons for this good performance can be given. First, the F-ratios are high, indicating a high signal-to-noise ratio which makes the need for biased estimation less[16]. Second, low s values make the variance in the predictions low, see Section 2.4, formula I.42. Third, the influence of the multicollinearity on the predictive performance is not severe. As already explained in Subsection 10.3.1, this can be checked by subjecting the x vectors used for prediction to the linear combination which causes the multicollinearity, with loadings given by the last PC on X, Table III.30d. The mean sum of squared deviations (MSSD) for the eighteen predictor vectors (six mobile phases times three stationary phases), when subjected to this combination is 12.3×10^{-4}. The analogous value in the training set is 4.8×10^{-4}. If the test set vectors are subjected to the loadings associated with the first PC, Table III.30d, the resulting MSSD is 1115×10^{-4}. The analogous value of the training set is 1506×10^{-4}. A tentative conclusion is that the predictor vectors have the same multicollinear pattern as the X vectors in the training set and that therefore the influence of the multicollinearity on the predictive performance is damped.

The predictions for the individual phases can be judged by the individual RMSEP values in Table III.31b. The RMSEP values for the C6 phase are higher than the corresponding s_{repro}, see Table III.5, for each non-marker. The MSSD values associated with the last PC for the C6, C8 and PHE phase are respectively, 25.6×10^{-4}, 4.7×10^{-4} and 6.4×10^{-4}. The MSSD values associated with the first PC are, respectively, 992×10^{-4}, 1197×10^{-4} and 1158×10^{-4}. These values support the relatively bad predictions on the C6 phase. Only for PE and PHB the RMSEP values of the C6 phase are high compared with the RMSEP values for the whole test set. PE is discussed firstly.

The measurements of the solute PE are underestimated at the ternary mobile phase compositions on the C6 column. Inspection of the ln k values of PE on C6 (Table III.4) shows a slight non-linear effect of the mixing of the binary mobile phase "series 1" compositions. The difference between the average of the measurements at wal and wml (0.783) and the measurement at aml (0.881) is larger than the s_{repro} of PE on C6. The same holds for the "series 2" mixtures: (1.357+1.219)/2=1.288 is smaller than 1.405, compared with s_{repro}. The most influential predictor of PE is TOL. If the non-linear behaviour of TOL, as a result of mixing binary eluentia, is investigated in the training set, it appears that especially on the C1 and CN stationary phase, TOL shows a pattern similar to PE on these phases. This is not the case if the measurements on the C6 phase of TOL and PE are compared. There is an interaction between non-linear mixing behaviour of the solutes TOL, PE and the stationary phases. This makes the prediction of PE at the ternary mobile phase compositions on the C6 phase with the use of TOL, quite troublesome.

For PHB, bad predictions are observed at the "series 2" mixtures, especially the wm2 and am2 mixtures on the C6 phase. All the predictions of PHB on C6 are too high, which indicates that the mean retention of the markers (especially TOL and MHB) estimates the mean retention of PHB too high. Stated otherwise, in relation to the retention of TOL and MHB, there is a difference in the behaviour of PHB in the training- and test set. This can be verified roughly by calculating the weighed average, with weighing-constant the corresponding b_{ols}, of the mean ln k values of TOL and MHB in the training set and compare this with the mean ln k value of PHB. Since TOL and MHB are the most important predictors of PHB, only the ln k values of these markers are needed to predict PHB on the C1, CN and C18 phases at the six mobile phase compositions. The ratio between this weighed average and mean ln k of PHB is 1.8. The corresponding value on the C6 phase is 1.5, which means that the mean retention of PHB on C6 (averaged over the six mobile phase compositions) is overestimated. The particularly bad predictions at the wm2 and am2 mixtures are probably related to the remark in Section 11.1, where it was pointed out that the particular combination PAR, TOL, MHB was not the best one with respect to the "series 2" mixtures. An explanation of the bad predictions at the "series 2" mixtures might be that the dependence of PHB on elution strength, in relation to the dependence of the markers (especially TOL and MHB) on the same elution strength, differentiates between the training set and the C6 phase. The influence of elution strength is overestimated by the markers on the C6 phase, resulting in predictions which are too high for PHB at the "series 2" mixtures. Stated otherwise, the interaction between elution strength and stationary phase should be represented well by the markers. In the case of PHB, this interaction is not represented well enough by the markers.

The variable ACP is badly predicted on the PHE phase. The retention of ACP on PHE at mixtures wm2 and wal is severely underestimated. The reason of this is not clear.

Summarizing the comments on the bad predictions of PE and PHB, three aspects are to be distinguished. First, the particular position (hydrophobicity) of a new stationary phase in relation to the stationary phases in the training set. This particular position of the new stationary phase must be adequately assessed by the markers. Second, non-linear mixing behaviour must follow the same patterns in the training set as on the new phase, as well as for the markers and for the non-markers. Third, the simultaneous influence of elution strength on the markers and non-markers on the new phase must reflect the patterns in the training set. Unfortunately none of these aspects can be tested beforehand, a careful (theoretical) examination of the prediction problem at hand is, therefore, necessary.

The variable PHB has a high RMSEP value in the training set, calculated with the leave-one-out procedure, higher than s_{repro}.

Particularly bad predictions are obtained for PHB at the wal mixtures on both Cl and CN. The score plots, Figures III.15a to III.15c, do not reveal these points as leverage points and they might be considered influential. The leave-one-out RMSEP without the contribution of these points is 0.066, which is a much better result. The results of model 2 are discussed in the following. The estimated coefficients and diagnostic values are reported in Table III.32b. Again high F-ratios are obtained. Only the s values for PE and PHB are higher than the corresponding s_{repro}, indicating a slight lack-of-fit (note that in model 1 the variable A was significant in the model equation of PE and M nearly significant for PHB). The marker MHB is a relevant predictor for all non-markers. For ACP, BAB and EAB the marker TOL is also relevant, whereas PAR contributes only to BAB.

The predictions performed with the use of model 2 in the test set are good, there is not one RMSEP in the test set larger than the corresponding s_{repro}. As for model 1, this is due to high F-ratios, low s values and a similar multicollinear pattern in the training and test set. To illustrate this last point the MSSD should be calculated when the training- and test set vectors are subjected to the linear combination associated with the last PC of the design matrix of model 2, see Table III.30. The outcomes are respectively, $12\text{x}10^{-4}$ and $14\text{x}10^{-4}$. The analogous values with the use of the loadings on the first PC are $1288\text{x}10^{-4}$ and $891\text{x}10^{-4}$.

The RMSEP of the individual phases, shows that the C6 phase has RMSEP values which are always higher than the s_{repro} values. The MSSD values of C6, C8 and PHE with respect to the last PC are, respectively, $19.9\text{x}10^{-4}$, $9.3\text{x}10^{-4}$ and $13.8\text{x}10^{-4}$. For the first PC these values are, respectivily, $1045\text{x}10^{-4}$, $1128\text{x}10^{-4}$ and $500\text{x}10^{-4}$. The high MSSD value of C6, associated with the last PC, supports the view that the multicollinear pattern in the training set is not completely present on the C6 phase, so the multicollinearity may cause harm on C6. Only for the solutes EAB and PHB, the RMSEP on the C6 phase is higher than on the other phases. EAB is discussed firstly. Particularly the retention at the mixtures wa2 and am2 is badly predicted. The predictions are too low, which indicates a different influence of elution strength on the C6 phase and on the phases of the training set with regard to the solutes TOL, MHB (the important predictors of EAB) and EAB. If the mobile phase variables are incorporated (model 1) better results are obtained, see Table III.31.

For PHB holds that the predictions obtained on the C6 phase at mixtures wal and wa2 are too high. MHB is the most important predictor of PHB in model 2. It is possible that the relative sensitivity of MHB and PHB for the water/acetonitrile mixtures in the training set differs from this sensitivity on the C6 phase. PHB profits from the incorporation of the variables A and M in the model.

The bad prediction of ACP on PHE is caused by a severe underestimation of the observed values at mixtures wm2 and wal.

The leave-one-out RMSEP are high for EAB and PHB. In case of EAB the ln k values at combinations C18wm1 and C18wa2 are considerably overestimated. If the contribution of these prediction errors is subtracted from RMSEP the new RMSEP becomes 0.0695. For PHB the ln k value at the combination CNwa1 is underestimated. With this prediction error omitted the RMSEP becomes 0.0836. The need for outlier detection and regression diagnostics is again clear, but this is not pursued any further.

A discussion on the differences between model 1 and 2 is divided in two parts. The part describing the power of validation criteria to choose between the models is postponed till Subsection 11.4.3. The differences between the models with regard to their predictive performance are discussed. At this moment it is very hard to derive solid conclusions by comparing the predictive performance of the models. A comparison of the RMSEP values should be done with the corresponding s_{repro} kept in mind. All RMSEP in the test set are lower than the associated s_{repro} values, for both models. The only tentative conclusion is that model 1 predicts EHB, EAB and PHB slightly better. For the other solutes no real differences between the models are observed. The profit of using model 1 in predicting EAB and PHB is obtained at the C6 phase, which has already been discussed.

A comparison between the predictive performance of the marker set chosen with the induced-variance criterion (Chapter 10) and the set chosen with the determinant criterion (Chapter 11), in terms of OLS results, is postponed till Section 11.5.

11.3.2 Ridge regression

Prior to the performance of ridge regression the data are centered and scaled-to-length-one. The evaluation of cross-validation with regard to the choice of the ridge parameter is postponed till Subsection 11.4.1. The results of model 1 are discussed firstly.

The ridge estimates for these models, with a leave-one-out choice of k, are presented in Table III.33. Some representative ridge traces are shown (Figures III.17a to III.17d) and are much more stable than the analogous ones in the case of the induced-variance markers (Figures III.11a to III.11d). Comparing, e.g., the ridge traces of PE in both cases, very small changes in k have a large effect on the parameter estimates of the induced-variance markers (Figure III.11a), whereas the influence on the parameter estimates of the determinant markers is much milder (Figure III.17b). This is due to its milder multicollinearity. Changes in signs of the estimated coefficients of the markers, at low values of k as in Figure III.11a, are not observed. For ACP the marker TOL becomes more pronounced than in the OLS estimation. The variable MHB obtains some predictive relevance for PE. The mobile phase variables do not diminish rapidly in the ridge trace of EAB, which is in agreement with the better predictive performance of model 1 in the case of EAB. The differences between the OLS estimates (Table III.31a) and the ridge estimates (Table III.33a) are not as large as the analogous difference with the induced-variance markers (Chapter 10).

With respect to the predictions with the ridge estimates of model 1, Table III.33b, the conclusion is that slightly better predictions are obtained, in the whole test set, with ridge regression. As a warning against drawing conclusions too quickly, note that the RMSEP values of the whole test set for all non-markers, both for the OLS and ridge estimates, are below the corresponding s_{repro} values.

The predictions on the C6 phase, especially the solutes BAB, PE and PHB, profit considerably from applying the ridge estimates. This is also reflected in the mean RMSEP values. Only PE has still a relative high prediction error on C6, the ternary mobile phase compositions are still troublesome. The leave-one-out RMSEP is rather high for PHB, which was also the case for OLS. The solute ACP is badly predicted on the PHE phase: the same eluent mixtures as in the OLS case are troublesome.

The real power of ridge regression can be assessed if the ridge parameter is chosen to minimize the prediction error in the test set, Table III.35. An evaluation of the leave-one-out choice of k is postponed till Subsection 11.4.1. The estimated coefficients given in Table III.33 do not differ much from the ones given in Table III.35. The predictions show improvement in relation to the OLS predictions. Yet, the remark regarding the s_{repro} is still valid.

An impression of the power of ridge regression for each test set stationary phase separately, can be obtained by comparing the mean RMSEP for the particular phases in the OLS case and the ridge case. The stationary phase C8 does not profit from the ridge regression, this is in agreement with the low MSSD value of C8 associated with the last PC in the training set, see Subsection 11.3.1. The high prediction error of ACP on the PHE phase has vanished.

It is interesting to note that all non-markers profit from ridge regression, which was certainly not the case for the two-way approach with the induced-variance markers, see Table III.17. A full comparison of both two-way approaches using the induced-variance- or the determinant markers is postponed till Section 11.5.

Some representative ridge traces of model 2 (Figure III.18) show a smooth and similar behaviour for all solutes. Only for EAB the variable PAR holds some predictive power, still the influence of PAR remains small. This can be checked in Table III.34. The markers MHB and TOL have predictive relevance for all non-markers. For the solutes PE and EAB the ridge regression is not advantageous, compare Tables III.34b and III.32b. Although these solutes profit from the ridge operation with regard to the predictions on the PHE phase, the predictions on the other phases, C6 and C8, are worse than with OLS. On the whole, only the C6 and PHE phase profit slightly from this ridge regression.

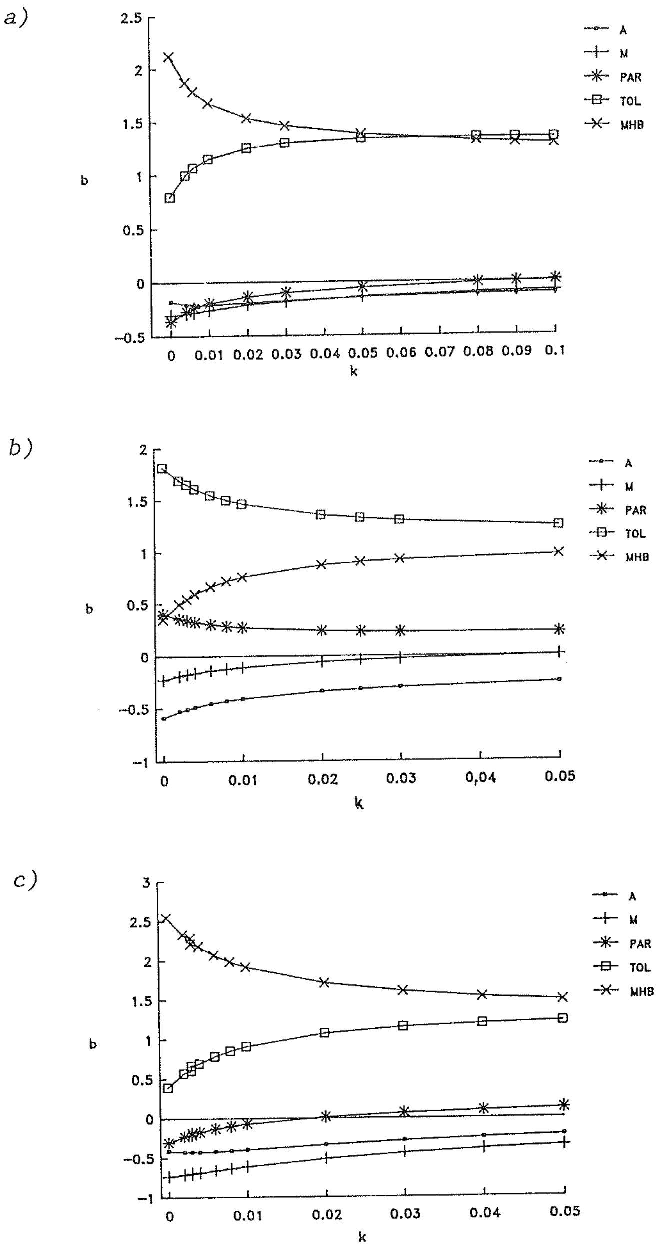


Figure III.17. Ridge traces of model 1 for the solutes ACP (a), PE (b) and EAB (c). For abbreviations, see the text.

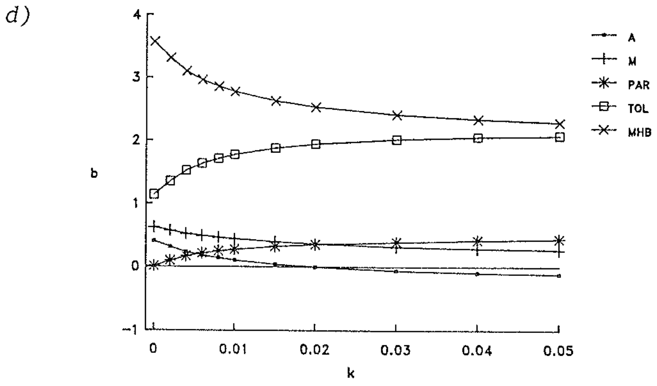


Figure III.17 (continued). Ridge traces of model 1 for the solute PHB (d). Legend, see Fig. III.17a.

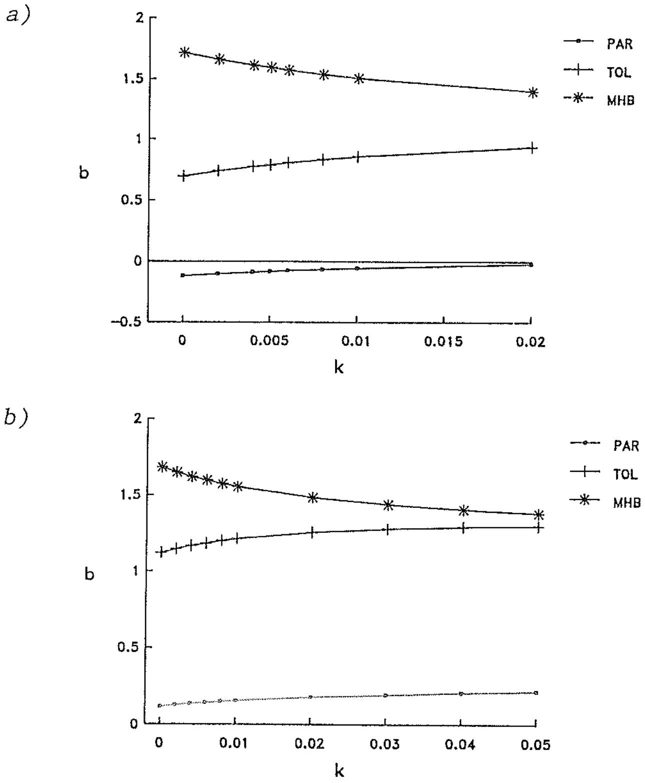


Figure III.18. Ridge traces of model 2 for the solutes PE (a) and EAB (b).

The best results which can be obtained with ridge regression are reported in Table III.36. The estimated coefficients are in the same order of magnitude as the leave-one-out choice k results. The mean RMSEP values (when applying the optimal ridge regression) show that only the predictions on C6 and PHE become better. This is in agreement with the MSSD considerations in Subsection 11.3.1, where it was shown that the ln k values of the markers on the C8 phase resemble the multicollinear pattern in the training set most. The solutes PE and EAB do not profit from applying ridge regression.

11.3.3 James-Stein regression

All y vectors and columns of the X matrix are centered and scaled-to-length-one prior to the James-Stein shrinkage calculations (see Section 2.1). One of the necessary conditions to obtain a lower MSEP for all values of β is not fulfilled. This condition is $d=l_p\Sigma\lambda_i^{-1}>2$, but in case of model 1 d=1.32 and for model 2 d=1.04. This implies that there is no guarantee for better predictions, but, as argued earlier in Subsection 10.3.3, there might be such a β value that better predictions can be obtained by applying the JS shrinkage. The results of model 1 are discussed firstly.

The leave-one-out method always advises shrinkage, all c values are smaller than 1.00, see Table III.37a. An evaluation of the leave-one-out choice of c is postponed till Subsection 11.4.2. By comparing the s values of Table III.31a and Table III.37a it can be observed that, in terms of sum of squared error, the leave-one-out c Stein solution does not differ much from the OLS solution. Only the non-markers ACP, BAB, EHB and PHB profit, to some extent, from the shrinkage. The variable PE is predicted worse. Comparing the mean RMSEP values for OLS and Stein regression with the leave-one-out c choice, hardly shows any improvement.

By comparison, the Stein results with the best choice of c are given in Table III.37. These results are compared with the OLS ones for model 1. The non-markers ACP, BAB, EHB and PHB profit from applying Stein regression. Note that the differences between RMSEP values used to decide whether or not Stein regression is useful, are small compared with the associated s_{repro} values. Too decisive conclusions are not allowed. The predictions of ACP on the PHE phase become much better but the opposite is true for PHB. The profit for PHB is on the C6 phase.

The mean RMSEP values for the C6, C8 and PHE phase, when applying the Stein regression with the optimal choice of c, show that only the C6 and PHE phase profit to some extent from the Stein regression. This is related to the "multicollinearity pattern" of the markers on these phases. Note that the performance of Stein regression, in relation to OLS, is better with the markers PAR, TOL and MHB than with the markers BAB, EHB ,PAR as in Chapter 10. This may be due to the lesser degree of multicollinearity between the markers chosen with the determinant criterion.

The leave-one-out method in case of model 2 advises always shrinkage, see Table III.38a. According to the RMSEP values in the whole test set, there is hardly any profit from this shrinkage. This is partly due to a bad choice of c, as can be seen in Table III.38c, where the results for the optimal c choice are given. The variable PHB profits most from the optimal shrinkage. On all test set phases there is some improvement after using shrinkage, especially on C6. Again the gain from shrinkage (apart from the problem of the choice of c) is greater for the two-way approach with the determinant markers than with the induced-variance markers.

11.3.4 Partial least squares

The PLS calculations were performed in exactly the same way as described in Subsection 10.3.4. The X-block constists of A, M, PAR, TOL and of MHB and the Y-block of all non-markers. All columns in X and Y were autoscaled prior to the calculations. Again only model 1 was calculated, otherwise too few predictors remain in the X-block.

Cross-validation, in the version of the SIMCA program (see Subsection 10.3.4), was used to establish the number of relevant dimensions, which was three. The loadings, as reported in Table III.39a, reveal the same pattern as in the case of the induced-variance markers (Table III.21). The markers load on the first dimension, together with all non-markers. The second dimension in the X-block consitutes the mobile phase variables. The relative low loading of EAB on this dimension does not point to irrelevance of the mobile phase variables in the prediction of EAB, see Tables III.38 and III.35, but, on the contrary, the ridge results show that the mobile phase variables have some relevance. The marker PAR loads high on the third dimension in the X-block. The related non-markers in the Y-block are ACP, BAB and EAB. The conclusion that PAR may have some relevance for the prediction of these non-markers is contradicted in Tables III.33 and III.35.

The summary statistics regarding the predictive performance of PLS, Table III.39b, can be compared with the OLS ones (Table III.31b). For the individual phases in the test set, differences are absent between OLS and PLS. The non-markers ACP, BAB and PE profit to some extent from applying PLS. The same was observed for the non-markers with the best ridge results.

11.4 The performance of cross-validation

11.4.1 Determining the ridge parameter k

The different ways a ridge parameter can be choosen (Sections 10.4.2 and 3.2) is already discussed. The same methods will be discussed in the context of the two-way approach with the determinant markers. Tables III.40 and III.41 state the results. The ultimate yardstick for the comparison of the different choices is the "best" choice: the k values that gives rise to the lowest RMSEP in the test set. Obviously, this value is never known in practice, but for the sake of comparison this yardstick is chosen. The results of model 1 are discussed firstly.

From the "k-choice methods" - leave-one-out method, Hoerl and trace - the number of times a method gives a TESTRMSEP value closest to the best one is counted. These counts are four, one and two for, respectively, the leave-one-out-, Hoerl- and trace method. A slight advantage of using the leave-one-out method is seen. This is in contrast with the conclusion in Subsection 10.4.2 (Table III.23), with the induced-variance markers, where Hoerl's method surpassed the leave-one-out method. Whether this observation is related to the lower degree of multicollinearity when the markers PAR, TOL and MHB are used is not clear.

A similar pattern is seen with regard to the results of model 2, Table III.41. Again the leave-one-out method has four times a TESTRMSEP value closest to the optimal one. For both Hoerl's- and the trace method this number is two. After comparing this with the results of Table III.24, where the induced-variance markers were used, the same conclusion as above can be drawn. The leave-one-out method seems to perform better in case of a lower degree of multicollinearity.

11.4.2 Determining the James-Stein shrinkage parameter c

The leave-one-out choice of c is compared with the c value as proposed by Stein, see Sections 10.4.3 and 3.3. The results of model 1, Table III.42, is discussed firstly. The best choice of c gives rise to the lowest TESTRMSEP value. The comparison between the c choice methods is done analogous to the comparison between the k choice methods in the previous chapter. The Stein c value always advises less shrinkage than the leave-one-out method, which results in a good advice regarding the shrinkage factor for PE. Yet, the number of times leave-one-out recommends a c value that gives rise to the TESTRMSEP value closest to the optimal one is four whereas for the Stein c choice this number is two. This is in contrast with the results in Table III.25, where the Stein method performed better than the leave-one-out method.

The same conclusions hold for model 2. Again the leave-one-out method is better than Steins choice, except in cases where no shrinkage is profitable (PE and EAB), because of the more conservative nature of Steins choice.

The same phenomenon with regard to the leave-one-out choice of the shrinkage parameter c is observed as in the leave-one-out choice of the ridge parameter k. There seems to be a relation between the performance of the leave-one-out method and degree of multicollinearity in the X matrix. A high degree of multicollinearity is problematic with the performance of the leave-one-out method. Further research is needed on this topic.

11.4.3 Choosing between modelspecifications

The performance of cross-validation, and more specifically the prediction error sum of squares (PRESS) as a result of the cross-validation, with respect to the choice between the models 1 and 2, will be discussed within the OLS context. In this perspective the performance of the PRESS criterion is compared with the C_p, PRC and R^2_{adj} criteria (Table III.43). The PRESS, C_p and PRC have to be minimized whereas the R^2_{adj} should be maximized when choosing between model 1 and 2.

The variables BAB, EHB, EAB and PHB are concentrated on because the largest difference in predictive performance between the models is observed with these solutes. C_p gives a wrong advice for EHB. The PRESS, PRC and $R^2{}_{adj}$ always give a good advice. Again, as in the case with the induced-variance markers (Table III.26), there is no clear preference for the PRESS criterion. Note that none of the validation criteria gives good advice in choosing between model 1 and 2 in case of ACP and PE, where the difference between the models is less pronounced.

If contrasted with the results given in Table III.26 (the markers BAB, EHB, PAR) a slight improvement of the performance of the PRESS criterion is observed, indicating a relation between the performance of PRESS and the degree of multicollinearity.

11.4.4 Choosing between estimation methods

Cross-validation can be used to choose between estimation methods. For a particular non-marker, the method that gives rise to the lowest leave-one-out PRESS value is advised by cross-validation. This advice can be judged by the test set results. The evaluation of cross-validation is done with Table III.44. Model 1 is discussed firstly.

Cross-validation gives a good advice in five of the six cases, only for PE the ridge estimation is advised, which is the second best solution. An assessment of the performance of cross-validation with respect to the choice of the estimation method is obscured by the fact that ridge is almost everywhere the best estimation method, in the training set as well as in the test set. If cross-validation has to choose from a wider variety of best estimation methods then a better assessment is possible. A comparison between Table III.34a with its analogon Table III.27, is skewed on behalf of the same reason: the variety of best estimation methods is greater in case of the two-way approach with the induced-variance markers. Note that PLS is the best estimation method for PE in both cases, the induced-variance markers and the determinant markers. The reason of this particular phenomenon is not clear.

In case of model 2, Table III.34b, the same line of reasoning as above can be applied. Except for PE and EAB, ridge is always the best estimation method. Cross-validation does always advise ridge, and gives a false advice twice. This wrong advice is related to the performance of CV in order to choose the ridge parameter, see Subsection 11.4.1. CV fails to advise the right ridge parameter twice, in case of PE and EAB.

A reasonable judgement regarding the performance of cross-validation with respect to the choice of the estimation method is only possible for a wider range of non-markers and, perhaps, estimation methods.

11.5 Comparing the induced-variance- and the determinant markers

Both two-way approaches can be compared with regard to the consequence they have for the predictive performance of the applied estimation methods for both model 1 and 2. The main differences between the markers BAB, EHB, PAR and PAR, TOL, MHB should be kept in mind. The first set of markers have a low determinant value and give rise to serious multicollinearity. The second set of markers have the highest determinant, moderate multicollinearity and an induced-variance which is not much lower than for the first set of markers. The key question when comparing the outcomes of the induced-variance- and the determinant approaches is: is the decrease in induced-variance, resulting from choosing with the determinant criterion instead of the induced-variance criterion, payed off by a decrease in multicollinearity, which at the end, results in better predictions with the markers chosen with the determinant criterion? This question is obscured by the fact that the markers, chosen with the determinant criterion, also have a high induced-variance. Therefore, the main difference between the marker sets lies in their multicollinear behaviour.

First, a comparison is made of the performance of all estimation methods separately. Such a comparison can be done in two ways. The first point of view is the evalution of the estimation methods with regard to the prediction of ACP, PE, EAB and PHB. These solutes are not contained in either of the markers sets. The second point of view is the judgement of estimation methods averaged over all non-markers. Ideally, a large set of solutes, which are not contained in either of the marker sets, should be available for the assessment of the marker choice.

The OLS results of model 1 are discussed firstly. Judging the predictive performance of both two-way approaches, it should be kept in mind that the multicollinear pattern on C6 deviates from the training set pattern in case of the induced-variance approach, see the MSSD values in Subsection 10.3.1. The same holds for the determinant approach, where the PHE phase has a relatively high MSSD value. The overall predictive performance of both approaches can be judged by comparing the mean RMSEP values. There is hardly any difference, especially if the bad predictions of TOL are observed which make the determinant approach less favourable. The predictions on C6 with the determinant approach are better than with the induced-variance approach, despite the high MSSD value. This might be due to the lower degree of multicollinearity with the determinant approach. If the individual solutes are examined, it appears that only EAB is better predicted with the determinant approach. PHB has a better predictor in EHB than MHB and ACP gains much from the induced-variance approach in the predictions on the PHE phase.

The variable PE is better predicted on the C6 and C8 phases with the use of the induced-variance approach, and on the PHE phase with the determinant approach. This indicates two properties. First, an inhomogeneity of the test set: the PHE phase plays a particular role. Second, PE has a specific predictor in the marker set PAR, TOL, MHB which improves the predictions on PHE, but these predictors worsen the prediction on the C6 and C8 phases. A specific combination of a non-marker and a test set stationary phase demands its own predictor(s). This stresses the notion of a representative training set. Not all non-markers reveal these differences with respect to the test set phases: the solute PHB is predicted better on all test set phases with the induced-variance markers, which shows that EHB is a better predictor than MHB in all cases. There is no interaction between stationary phase, non-marker and predictor. The results of model 2 are not discussed separately, the conclusions are similar as in the case of model 1.

The ridge results of model 1, with the leave-one-out choice of k, show the same pattern as the OLS results for model 1. The determinant approach performs slightly better. In case of these ridge results, the predictions of TOL are not as bad as in the OLS case with the determinant markers. The mean RMSEP value (averaged over the whole test set) is, therefore, a better yardstick for comparison in the ridge case than in the case of OLS. The ridge results of model 2, again with respect to the leave-one-out choice of k, show a slightly better performance of the determinant approach. This is, however, obscured by the moderately bad predictions of TOL. For the non-markers ACP, PE, EAB and PHB, the induced-variance approach performs better.

The ridge results for model 1, with the best choice of k, show no real advantage in using the determinant marker set. The solute EAB is better predicted with the determinant markers, only on the PHE phase the induced-variance marker set is better. This indicates again an interaction between non-marker, stationary phase and predictor(s). In this case the predictor BAB, a homolog of EAB might have predictive relevance for EAB on PHE, see Tables III.17 and III.18. This phenomenon, with regard to the prediction of EAB, is also visible in the ridge results of model 2, again with the best choice of the ridge parameter. When the ridge results of model 2 are examined, it is worth noticing that PE is now better predicted on PHE with the use of the determinant markers set, contrary to the ridge (and OLS) results of model 1. The mobile phase variables in combination with the induced-variance marker set are harmful for the prediction of PE on PHE.

The results of the Stein estimation of model 1, with the leave-one-out choice of c, show an advantage with the use of the determinant marker set. Only on the PHE phase the induced-variance marker set has a better performance. PE is predicted better by the induced-variance marker set on the C6 and C8 phase. On the PHE phase the determinant marker set is better. Exactly the opposite is true for EAB. These observations indicate again interactions between non-markers, stationary phases and predictors. The variables ACP and PHB are better predicted with the induced-variance markers. Roughly the same conclusions can be drawn regarding the results of the Stein estimation of model 1 with the best choice of c.

The Stein results of model 2 show, both for the leave-one-out- and best choice of c, a slight advantage in using the determinant marker set. The variable PE is now better predicted with the induced-variance markers on PHE. This is in agreement with the ridge results, see above.

The results for the PLS estimation show that the determinant approach performs better than the induced-variance approach, on all phases (particular C6). It is difficult to relate this to the degree of multicollinearity, because PLS decomposes the X matrix in orthogonal t vectors, so the multicollinearity is removed. Besides, the solute TOL is badly predicted with the induced-variance markers and PLS, especially on the C6 phase.

The influence of the marker choice (and the degree of multicollinearity) on the performance of ridge regression as a whole, can be established by comparing the mean RMSEP values for the whole test set, and for the individual test set phases, of OLS and ridge (both the leave-one-out- and best choice of k). Although ridge regression is more advantageous if the induced-variance markers are used instead of the determinant markers, when it is evaluated on the whole set, the general advantage of ridge is small. For model 2 such a comparison reveals a slight advantage in using ridge. These kinds of comparisons should be related to the reproducibility and are therefore tentative. A phenomenon which emerges is that on the phases with a deviating multicollinear pattern of the markers as compared with the training set (C6 and PHE), ridge regression is relativily more advantageous. Especially in case of the prediction on C6 with the induced-variance markers, ridge improves the predictions, whereas the MSSD value of C6 is high.

Theoretically Stein regression is not profitable when a high degree of multicollinearity is present. When the appropriate mean RMSEP values are compared it appears that, for both model 1 and 2, the determinant approach profits more from the Stein shrinkage than the induced-variance approach, which profits hardly from it. Again, as above, the test set phase with a higher deviating multicollinear behaviour profits the most, i.e. C6 and PHE with the determinant approach.

Some summarizing is convenient. First, there is no clear preference for the induced-variance- or determinant criterion to choose the markers. A slight advantage of the latter criterion is its ability to chose extreme solutes, such as TOL, which cannot be predicted easily. However, if the retention of a solute is very difficult to predict (e.g. PAR), this solute will automatically be a member of the induced-variance marker set. Both marker choice criteria perform well. Second, an ultimate yardstick to measure the difference between the multicollinear pattern in the training set phases and a specific test set phase is hard to derive. It can be concluded that a decrease in predictive performance of OLS, resulting from deviating multicollinear behaviour of a marker set on a specific test set phase as measured by MSSD, can be counteracted by a shrinkage operation. Either ridge or Stein estimation will improve the predictions to some extent. Finally, specific interactions between non-markers (e.g. PE and EAB), markers and test set stationary phases (e.g. PHE) cause trouble and unexpected (and unforeseen) prediction results.

A natural question to ask is whether it is possible to validate the marker choice and choose that marker set which is expected to perform best. The validation criteria in combination with OLS estimation are discussed firstly. Several validation criteria for choosing between rival models have been presented. Our purpose is to compare model 1 when using the induced-variance marker set and the determinant marker set, the two models will be labelled 1A and 1B, respectivily.

The C_p criterion is not suitable because it focusses on discriminating between a full model and a reduced one. The problem at hand is the comparison of models of the same size, but with different markers. With the comparison of model 1A and 1B, the values of n and m remain the same. This means that both PRC and R^2_{adj} will rely upon the residual sum of squares in their judgement which model to choose. The model with the smallest residual sum of squares will be advised as the best model, both criteria give the same advice. The only comparison made, therefore, is between PRESS and PRC.

In case of the solutes ACP and PHB, model 1A is the best choice which is clearly detected by the criteria PRC and PRESS. Solute EAB is slightly better predictable with model 1B. It is not detected with PRC and PRESS. Models 1A and 1B differ hardly in case of PE with respect to their predictive power. This is not detected by PRC and PRESS and these criteria give a convinced wrong advice. Judging the behaviour of both validation criteria with respect to PE and EAB, it should be kept in mind that differences between the training-and test set are clearly present for these solutes.

The same analysis can be done with regard to model 2. Two models are compared, model 2A and 2B, both contain only the markers, respectively the induced-variance markers and the determinant markers. Again the behaviour of PRC and PRESS with regard to ACP and PHB are good. The wrong advice is given for the solutes PE and EAB, by both criteria.

The conclusion is that clear differences between the predictive power of the models in the test set are detected by the PRC and PRESS criteria. The inverse is not true, when the criteria indicate a clear difference between either models 1A and 1B, or 2A and 2B, this is no guarantee for the same differences between the models with respect to their predictive performance.

The only criterion by which the performance of the marker sets in combination with PLS estimation can be assessed is PRESS. Concentrating again on the solutes ACP, PE, EAB and PHB the PRESS criterion only fails in recognizing that PE is better predicted with the determinant marker set. The particular behaviour of PE, as already discussed, has again its influence.

Chapter 12 Two-way approach: other marker choices

12.1 A homologous series as markers

The solutes MHB, EHB and PHB are members of a homologous series. Such a homologous series is often used to correct differences in retention times due to variation of stationary phases, see Chapter 7. They can be evaluated with regard to their predictive power. Only the results of the predictions of the capacity factors of the non-markers obtained with model 1 (each ln k of a non-marker is a linear function of a constant, the mobile phase fractions and the ln k values of the homologs), are discussed. The training- and test set are the same as in Chapters 10 and 11. Within the data matrix as pictured in Figure III.8, the homologs induce 97.2% of the variation in that matrix, which is slightly lower than the induced-variance markers (99.5%). The determinant of the dispersion matrix of the homologs is 0.6×10^{-5}, which is considerably lower than in case of the determinant markers (3310×10^{-5}). This means that from the point of view of the induced-variance criterion, the homologs form a reasonable set of markers. From the viewpoint of the determinant criterion, however, the homologs form a bad set of markers.

A principal component analysis on the column autoscaled training set - the variables are the mobile phase fractions and ln k values of MHB, EHB and PHB- shows that two PC's already explain 99.3% of the variation (the first PC accounts for 60.9%). The score plot of the two first PC's is shown in Figure III.19. The CN measurements form a different class, but no clear outliers are present.

Some diagnostics of the associated design matrix are presented in Table III.45. High correlations are present between the homologs, see Table III.45a. Very high variance inflation factors are observed, which indicates strong multicollinearity. This is also reflected in the very high condition number (105, see Table III.45c). Such a high condition number points to very serious multicollinearity[13]. From Table III.45c it can be inferred that strong multicollinearity is caused by the homologs, which, as expected, show a very similar retention behaviour in the training set. A moderate relation is seen between A, M, MHB and PHB with associated condition index of 27. The estimated coefficient of solute EHB has the highest variance inflation factor of the series, indicating that, of all homologs, this solute can be predicted best from the other X variables in the model. This is in agreement with the observation that EHB is the intermediate in the homologous series, the fastest (MHB) and slowest (PHB) eluting homologs are good predictors for EHB. Table III.45c shows that a very large proportion of the variance in the estimated coefficient of EHB is due to the main source of multicollinearity.

The results of the predictions, when ordinary least squares is used to estimate the coefficients, are presented in Table III.46. The variable TOL is predicted not very well, especially on the C8 and

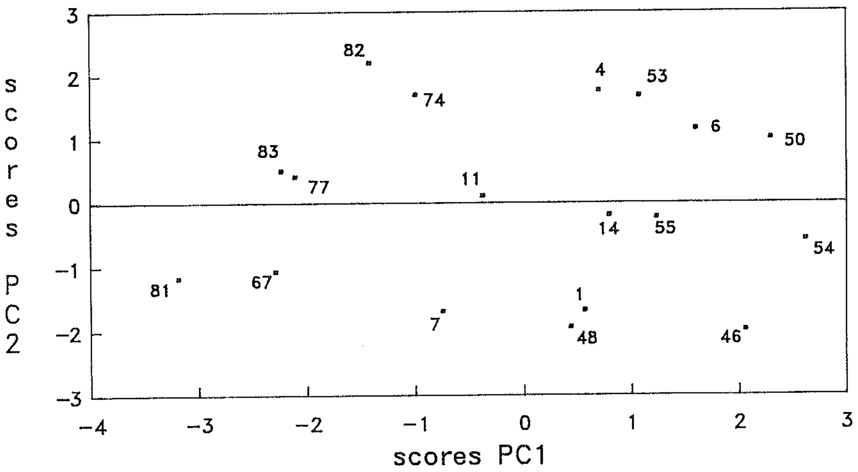


Figure III.19. Score plot of a PCA on the design matrix of model 1. Legend, see Fig.III.6.

PHE phase. PAR is badly predicted, especially on the PHE phase. This is in agreement with the remarks made in Sections 10.1 and 11.1. The results of the OLS calculations can be compared with their counterparts in Tables III.12 (induced-variance markers) and III.31 (determinant markers). Note that the marker MHB is also present in the determinant marker set and is a good predictor. The mean RMSEP in the whole test set is higher for the homolog markers than the other marker sets. If the solutes ACP, PE and EAB are examined- note that none of these solutes is contained in the three marker sets- it appears that PE is slightly better predicted with the homologs. The outcome of an evaluation of the predictive performance of the three different marker sets, depends on whether all non-markers are examined or only specific ones. This illustrates the "averaging" character of the induced-variance- and determinant marker choice criteria. With these criteria, markers are selected that give a good predictive performance on average. For the prediction of a specific solute on a specific stationary phase, a marker set, not chosen by one of the two criteria, might be more appropriate.

Whether the measurements of the markers (homologs) on the test set stationary phases bear the same multicollinear pattern as in the training set, can be assessed by calculating the mean sum of squared deviations (MSSD, see Subsection 10.3.1). For the C6, C8 and PHE phase the MSSD values are respectively, $3.26\text{x}10^{-5}$; $796.3\text{x}10^{-5}$ and $1.88\text{x}10^{-5}$. The training set MSSD is $2.49\text{x}10^{-5}$. The difference in predictive performance between C6 and C8 might, partly, be due to the deviating behaviour of C8 with respect to the MSSD values; the variance of the predictions on C8 is blown up by the multicollinearity. The predictions on the PHE phase are relatively bad. The variance in the predictions on the PHE phase is not severely blown up by the multicollinearity because this phase has a low MSSD value. As pointed out in Section 2.4, there is another aspect which may hamper predictions: the bias of the predictor $\hat{y}$. This bias may be caused by misspecification. Stated otherwise, another set of markers might be more appropriate for the prediction on the PHE phase.

Some of the ridge results are presented in Figures III.20a to III.20c and Table III.47. The ridge traces are highly unstable, because of the high degree of multicollinearity. *A priori*, the estimated coefficients of the homologs are expected to have the same sign, because the ln k values of the homologs are pairwise highly correlated. For all non-markers, the OLS estimate of β_{EHB} has always the opposite sign as those of β_{MHB} and β_{PHB}. The reason of this is, as pointed out in Subsection 10.3.1, a high negative covariance between the estimated coefficients of β_{EHB}, β_{MHB} ($-1130x\sigma^2$) and of β_{EHB}, β_{PHB} ($-1301x\sigma^2$). Only for TOL and PAR the sign difference remains with increasing k. The leave-one-out procedure only advises a non-zero k value in case of PE, see Table III.47a, the main advantage is in the predictions on C8. An evaluation of the leave-one-out choice of the ridge parameter is not pursued, but it should be noted that, contrary to the induced-variance and determinant two-way approaches, a non-zero k value is only advised in one case. This might indicate a relation between multicollinearity and the performance of the leave-one-out choice of k. The ridge results with the best k value, Table III.47b, show that improvement can be obtained in the predictions on the PHE phase. On the other test set phases the predictions become worse: the reduction of variance in the predictions does not compensate for the increase of bias.

The hypothesis that the relatively bad prediction on the PHE phase is due to specification error- the wrong set of markers is used to predict on the PHE phase- is contradicted with the ridge results. Note the considerable improvements (eg. TOL), when the best ridge estimates are used.

The results of the Stein estimation are presented in Table III.48. The Stein estimation with leave-one-out choice of the shrinkage parameter c, Table III.48a, hardly improves the predictions. The results of Stein estimation with the best choice of c, Table III.48b, does not show much improvement either. This is in agreement with the theoretical result that Stein estimation does not improve in case of severe multicollinearity, see Section 3.3.

The PLS calculations are performed with all non-markers gathered in the Y-block. Both X- and Y matrices were column-autoscaled prior to the PLS calculations. Calculations with PAR as a separate y variable (scaled or unscaled) did not yield any improvement (results not shown). Two dimensions in the PLS model were applied, explaining

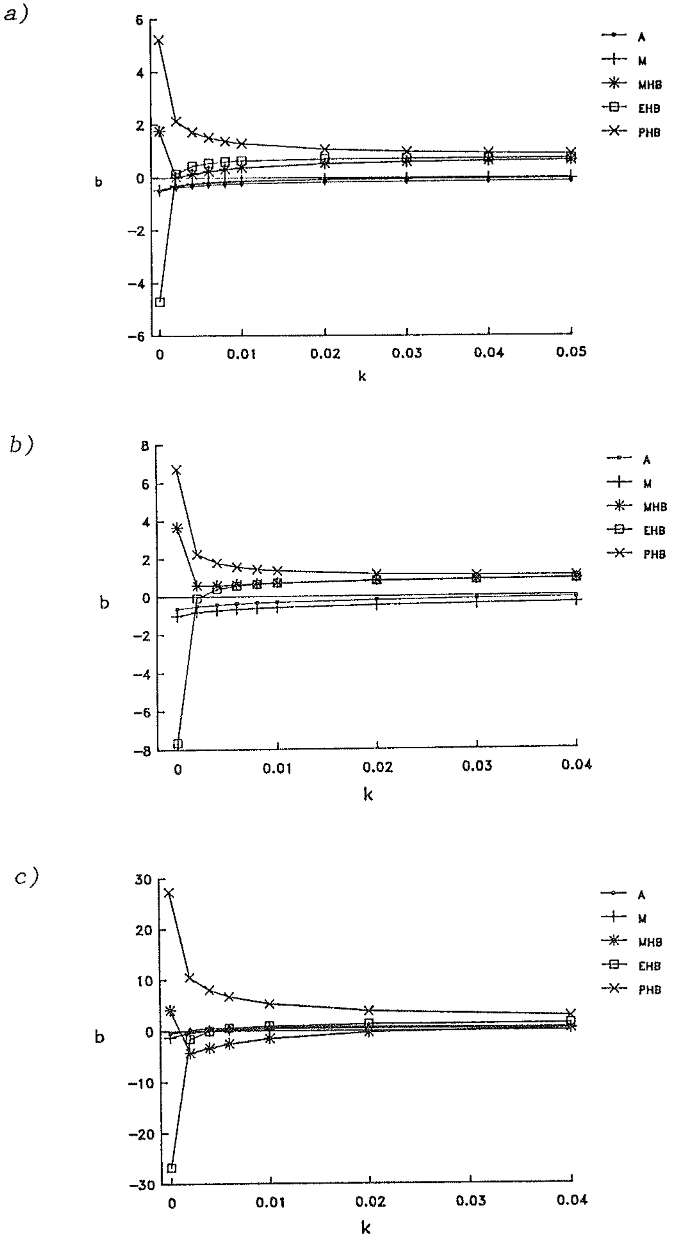


Figure III.20. Ridge trace of model 1 for the solutes PE (a), EAB (b) and TOL (C).

99.3% of the variation in the X matrix and 88.5% of the variation in the Y matrix. The results of the predictions with this PLS model are presented in Table III.49. Only the variables PE and BAB are better predicted with PLS than with OLS. PE profits from PLS prediction on C8, and BAB is better predicted with PLS on C8 and PHE. The reason for these improvements is not clear. On the whole, as measured by the mean RMSEP of all non-markers in the whole test set, PLS gives worse predictions than OLS. This holds also for each particular test set stationary phase. The conclusion is that PLS generally does not improve over OLS in this situation with severe multicollinearity. Whether or not this result depends on the kind of scaling applied, is not known (see Subsection 10.3.4).

The performance of cross-validation with respect to the choice of the estimation method is shown in Table III.50. Only in case of ACP, TOL and PAR the right advice is given. For the other three solutes the second best choice is given. Compared with the analogous results of the two-way approach with determinant markers, Table III.44a, the performance of CV is worse. Again, as in the case of the two-way approach with the induced-variance markers, this might be connected with the multicollinearity. The results with respect to the performance of CV for the induced-variance markers are comparable to those of the homologous markers. This contradicts the hypothesis that the use of CV as procedure to choose between estimation methods depends on the degree of multicollinearity.

There is no clear preference for a particular estimation method. It is again striking that PE is predicted best by PLS, as was the case when using the other sets of markers. A detailed analysis is necessary, but will not be pursued any further.

12.2 Using bad markers

Using the induced-variance criterion a set of bad markers can be chosen. The variables ACP, BAB and PE induce only 90.9% of the variation in the whole data table of Figure III.8 which is the lowest when three markers are used. A model which relates the non-markers linearly to the mobile phase fractions and the ln k values of ACP, BAB, PE will be tested.

The diagnostics of the resulting design matrix are presented in Table III.51. Again high correlations between the markers are visible. The variance inflation factors are high, but not as high as in case of the homologs. The multicollinearity is in the same order of magnitude as with the induced-variance markers. A plot of the scores on the first two components of a PCA (autoscaled) on the design matrix is shown in Figure III.21. No clear leverage points are present.

The results of the predictions with OLS, Table III.52, show a worse performance of this marker set than in case of the induced-variance- or the determinant markers, especially on the PHE phase. The

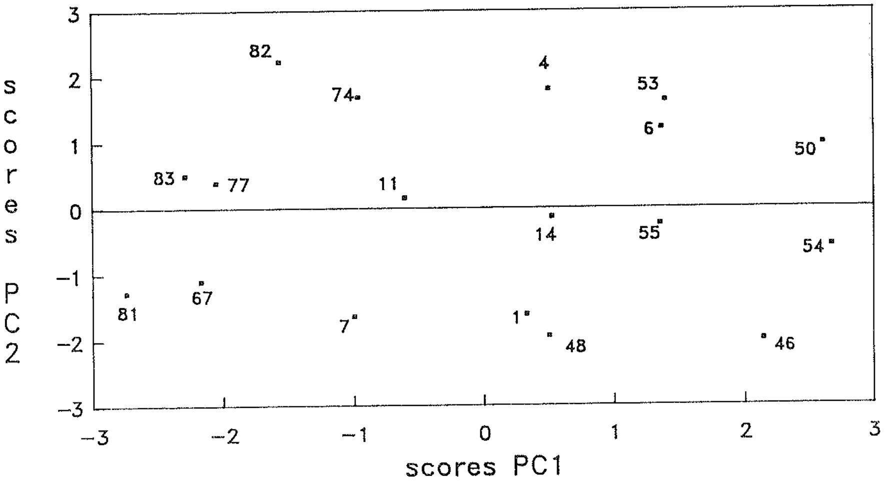


Figure III.21. Score plot of the design matrix of model 1. Legend, see Fig.III.6.

retention of the homologs- MHB, EHB, PHB- is well predicted with this marker set, but TOL and PAR are predicted poorly, again especially on the PHE phase. The MSSD value for the training set, as calculated with the loadings on the final PC (Table III.51d), is $8x10^{-5}$. These values are $8.2x10^{-5}$; $10.4x10^{-5}$ and $11.4x10^{-5}$ for the C6, C8 and PHE phase, respectively. This indicates that the multicollinearity has no serious effect, therefore, on the variance of the predictions in the test set.

Ridge regression, with leave-one-out choice of k (Table III.53a), is only to a small extent profitable for EAB, the other solutes are predicted worse with ridge than with OLS. If the best choice of k is made, Table III.53b, only EAB has a non-zero k value, and profits to some extent of the ridge operation. On the whole it can be concluded that ridge regression does not improve the predictions when compared with OLS.

Stein regression with the leave-one-out choice of c, Table III.54a, shows small improvements for PAR (especially on PHE) and PHB. If the best choice of c is made, Table III.54b, also small improvements are visible. But the bad predictions of PAR and TOL are not compensated for. These bad OLS predictions are probably due to bias caused by misspecification: the wrong marker set is used.

The PLS calculations are done with all non-markers gathered in the Y-block, both matrices X and Y were column autoscaled and two dimensions were used in the PLS model (explaining 99.1% and 88.3% of the variation in X and Y, respectively). PLS estimation (Table III.55) is only, and to some extent, profitable for EAB. All other non-markers are predicted worse with PLS than with OLS. Obviously, the bias introduced in the predicted values by using PLS is not compensated for by the reduction of the variance.

A particularly poor performance of CV is shown in Table III.56. Cross-validation never advises the best estimation method and hardly ever the second best choice. How this behaviour is related to the bad choice of the markers is not clear and still to be investigated.

12.3 Conclusions of the two-way approaches

Some overall conclusions can be drawn. It should be kept in mind that the data set is rather small, especially as far as the number of test solutes is concerned. The reported calculations should, therefore, be considered a pilot study and an example by which calibration calculations can be performed.

1. The tentative conclusion can be drawn that choosing markers with one of the two proposed criteria is better than using homologs or random solutes as markers. There are other criteria which are still to be evaluated (see Chapter 1), preferably on a larger data set. One of these criteria is designed to choose the solutes, given a subset-size, that maximize the minimal R^2 over all non-markers.

2. Both described marker choice criteria are sensitive to "outlying" variables (solutes). Prior to the marker selection a thorough analysis of the training data is needed with this respect. The solute PAR shows outlying behaviour.

3. There seems to be a slight advantage in the use of the determinant criterion for the marker selection. The markers chosen with the determinant criterion give a lower degree of multicollinearity in the ultimate models. This is to some extent advantageous for Stein regression and the performance of cross-validation in choosing k, c and the estimation method.

4. When an estimation method has to be chosen which has a reasonable performance on average, ridge regression seems appropriate. The choice of the ridge parameter k is a problem. The leave-one-out choice works only reasonable in case of low multicollinearity (determinant markers), while Hoerl's method is preferable in cases of moderate multicollinearity (induced-variance markers). The method of McDonald and Galarneau is worth trying.

5. *A priory* knowledge on the specific interactions between training- and test set phases, non-markers, markers, and mobile phase composition is needed. There is no guarantee that a set of markers performs well in predicting the retention of each non-marker on each test set phase (at each mobile phase composition), because of the above

mentioned interactions. Speaking in statistical terms, the availability of a representative sample of stationary phases is crucial.

6. Models incorporating explicitly the mobile phase variables seem to work slightly better, on average, than models without these variables. This has consequences for the choice of the markers. A marker choice criterion that reckons with the incorporation of the mobile phase variables in the final model is worth trying.

7. A good yardstick is needed to decide if the multicollinearity pattern of the markers (and mobile phase variables) in a test set are similar to the pattern in the training set. The performance of the yardstick used here, the mean sum of squared deviations (MSSD), should be assessed.

8. Some properties of the prediction procedure still have to be tested. The influence of scaling and the incorporation of interaction terms in the models (to account for non-linear mixing behaviour of the mobile phase) should be investigated.

Chapter 13 Three-way approaches

13.1 Three-way PCA and PARAFAC on the whole data cube

One of the conclusions in Chapters 10-12 was that the solute PAR showed deviating behaviour. This solute is therefore discarded from the following calculations. The three-way PCA and PARAFAC model is performed on the whole data cube firstly. This data cube, with the typical element x_{ijk}, consists of ln k values with the stationary phases as objects (index i=1,...,6; the first mode), the solutes (index j=1,...,8; the second mode) and the mobile phase compositions (index k=1,...,6; the third mode) as variables. This arrangement is visualized in Figure III.4a of Chapter 9 with the appropriate modification of the index values.

Data centering is performed in such a way that for each j,k it holds that $\Sigma x_{ijk} = 0$, where the summation runs from i=1,..,6. Because differences between stationary phases is of primary interest, this centering operation seems reasonable. Contrary to the arguments given in Section 9.7 and 10.1, scaling of the data cube is not performed, because all measurements are in the same units and scaling of data in three-way analyses is not straightforward[38].

After the centering operation, the mean sum of squares of each stationary phase i, before applying the model, can be calculated as $\Sigma x_{ijk}{}^2/(9x6)$, the summation running over all combinations of j and k. These MS_{bef} values can be regarded as the mean sum of squares that is to be explained by the model, or, stated otherwise, the <u>m</u>ean <u>s</u>um of squares <u>bef</u>ore the model is applied. After applying the model, values similar to MS_{bef} can be calculated for the residuals: the MS_{res} values. An important question in modelling a data cube is the number of components which should be retained. Because the experiment was designed in such a way that estimates of the variance of the measurements due to experimental error were available, these estimates can serve as a yardstick to choose the number of components, by comparing the MS_{res} values with the measurement error variances.

The results of the unfold-PCA will be discussed firstly. This is equivalent to an ordinary PCA on the unfolded data cube, where the data cube is unfolded so that the direction of the stationary phases (the objects) remains intact. (see Section 1.6). All parameters can then be estimated with the usual PCA algorithms. The results are presented in Table III.57.

The mean of all ln k values on a stationary phase, after the centering operation, can be calculated as the arithmetic mean of all values in the corresponding layer in the data cube. These mean values are reported in Table III.57a. This mean value is almost zero for stationary phase C1. This stationary phase takes an intermediate position in the range of the six stationary phases and its retention values hardly deviate from the mean values. Comparing the MS_{bef} values with the corresponding s^2_{repro}, it is clear that the retention values on the C1 phase hardly show any systematic variation after the centering operation. Both observations - near-zero mean and hardly systematic variation - will have its consequences for the calibration of the C1 phase, as will become apparent later on.

The variation of the retention values on all stationary phases, except C1, can be explained with one principal component: 98.8% of the variation in the whole data cube is explained by the first component, a percentage comparable with the previously found one in Section 9.7. Contrary to the results reported in Section 9.7, the loadings of all 48 (8x6) variables on the first PC are positive, because the solute PAR has been omitted. For the interpretation of the loadings and scores (Table III.57d) on the first PC, reference is given to Section 9.7.

Only the C18 phase has a MS_{res} value which is higher than the corresponding s^2_{repro}, but the s^2_{repro} of C18 is very low. A better approach might be to compare the MS_{res} values with the mean s^2_{repro}. In this context only the variation on C1 is not explained well, as already discussed, but this stationary phase can be regarded as an exception.

When the MS_{res} values of the individual solutes are considered (Table III.57b), it appears that only TOL is not explained completely by the first component. High positive residuals of this solute are observed on the C18 phase, at mobile phase compositions wm1, wm2, am1, and am2, indicating a systematic error of the model. The same pattern is present, although somewhat weaker, in the residuals of TOL on the CN phase. The stationary phases C18 and CN are the extreme stationary phases, see also Table III.57d, which may partly explain the situation. Yet, the fit of the retention values of the other solutes on these phases is reasonable even in case of the solute BAB, which is also a slow eluting compound. A three-way PCA with two components (not shown) decreases the residuals of TOL on C18 and CN to a reasonable size, but, as argued earlier, this second component cannot be regarded as significant in comparison with the reproducibility. The conclusion is that the solute TOL shows selective behaviour that is not accounted for by the first component. If more precise measurements are available this selective effect of TOL can perhaps be modelled with an extra component in the three-way PCA.

The MS_{res} values of the individual mobile phase compositions are reported in Table III.57c, and do not reveal deviating behaviour.

Starting with the same centered data cube as above, the PARAFAC model was applied. For an explanation of this model, reference is given to Section 1.6. The main difference between unfold-PCA and PARAFAC is that the latter decomposes the data cube as the summation of products of vectors, whereas unfold-PCA decomposes the data cube as a summation of products of vectors and two-way matrices.

In order to explain the consequences of the difference between the unfold-PCA and PARAFAC model, formulas of Section 1.6 are repeated. One component is assumed, in both cases. The unfold-PCA model is:

$$x_{ijk} = t_{i1} \cdot p_{1jk} + e_{ijk} \quad i=1,\ldots,6;\ j=1,\ldots,8;\ k=1,\ldots,6$$

where t_{i1} ($i=1,\ldots,6$) are the scores on the first principal component and p_{1jk} ($j=1,\ldots,8$; $k=1,\ldots,6$) are the loadings. The e_{ijk} values are residuals. The PARAFAC model is:

$$x_{ijk} = a_{i1} . b_{j1} . c_{k1} + e_{ijk} \quad i=1,\ldots,6;\ j=1,\ldots,8;\ k=1,\ldots,6$$

where a_{i1} ($i=1,..,6$) are the scores of the stationary phase on the first PARAFAC component, b_{j1} ($j=1,,.8$) and c_{k1} ($k=1,..,6$) are the loadings of the solutes and mobile phases, respectively, on the first PARAFAC component.

The PARAFAC model puts a constraint on the loadings in the second and third mode. Whereas unfold-PCA assumes the general form p_{1jk}, PARAFAC assumes that $p_{1jk} = b_{j1} . c_{k1}$. The meaning of this constraint was already explained in Section 1.6.

A consequence of the constraint in the PARAFAC model is that less parameters have to be estimated when the PARAFAC model is used instead of the unfold-PCA model. This means that, if the assumptions of the PARAFAC model are fulfilled a gain in degrees in freedom is obtained by applying this model.

The results of the PARAFAC model are reported in Table III.58. One component in this model explains 98.5% of the variation in the whole data cube. Considering the average s^2_{repro} and MS_{res}, one component is sufficient. The peculiarities of stationary phase C1 are already described. Again the C18 stationary phase is perhaps not completely described by this one component model, but compared with the average s^2_{repro} some doubt is present regarding this conclusion. Considering the individual solutes (Table III.58b), it appears that the variation in the retention of solute TOL is not completely described by the one component model. Closer examination reveals that this is especially the case for the retention of this solute on the C18 and CN phases. A pattern in the residuals of TOL on these stationary phases can be observed similar to the unfold-PCA case. The reasons for the relatively poor fit of the TOL retention values on the C18 and CN phase are already described above. Note that again the stationary phases C18 and CN are extremes (Table III.58d). A two component PARAFAC model takes away the high residuals, but the same warning as in the unfold-PCA case is appropriate here. The results for the individual mobile phase compositions (Table III.58c) show no deviating behaviour.

Which model is best: unfold-PCA or PARAFAC? It is clear that no decisive conclusion is possible on the basis of the above results, both models explain with one component about 98-99% of the variation in the data cube and the differences between the MS_{res} values of both methods are not significant compared to the s^2_{repro}. Moreover, a decision on which model is appropriate should not be done on the basis of statistics only. S.Wold *et al.*[39] propose a method of estimating the PARAFAC model by subjecting the loading matrix P_1 (with typical element p_{1jk}) to a PCA. Note that with this extra PCA step, the original data cube can be decomposed as a sum of products of vectors. In fact, the decomposition that is estimated by PARAFAC directly is then obtained. If the PCA on the matrix P_1 is performed, one principal component explains 99.7% of the variation in P_1. The conclusion is that the three-way PCA model degenerates to the trilinear PARAFAC model. Whether the unfolding solution is the best way to estimate the parameters in the PARAFAC model is questionable.

13.2 Choice of the markers/mobile phase compositions

The calibration procedures will be validated by successively leaving out one stationary phase, building the models with the five stationary phases left over, and using the omitted stationary phase as an independent test sample. The problem reduces, therefore, to choosing solute/mobile phase combinations, with the use of the retention measurements made on the five stationary phases, which explain enough variation of those retention measurements. Because no generalizations of the principal variables approach, as described in Section 1.3, are available, the following procedure is adopted.

The first step in the selection procedure is the unfolding of the 5x8x6 data cube in such a way that the direction of the solutes is left intact. The result is a 30x8 matrix, where the objects are stationary/mobile phase combinations and the variables are the 8 solutes. The three solutes that give the highest sum of all multiple correlation coefficients between the three solutes and the 5 non-selected ones are selected; the induced-variance criterion is applied on the non-scaled data gathered in the 30x8 matrix. The variation in retention of these selected solutes explain the variation of the non-selected ones at best. The outcome of this procedure is the set of solutes toluene, ethylaminobenzoate, and propylhydroxybenzoate. This outcome is found six times: for each omitted stationary phase the procedure was repeated and gave the same results. This places some confidence on the selection procedure. The variation of the retention of these three solutes explain on average 99.8% of the variation in the matrix from which they are chosen.

The second step in the selection procedure is the unfolding of the 5x8x6 data cube in such a way that the direction of the mobile phases is left intact. This results in a 40x6 data matrix, with the mobile phase compositions as variables. The same procedure as above was applied, which resulted in the choice of the wm1, wa1, and am2 mixtures. For each omitted stationary phase the procedure was repeated, but now the above mentioned mobile phase compositions were not always the best choice; they were the best compromise. On average the selected mobile phase compositions explain 99.7% of the variation in the matrix from which they are chosen.

The conclusion is that calibration of a new stationary phase can be performed by measuring the retention values of the markers TOL, EAB, and PHB at mobile phase compositions wm1, wa1, and am2.

13.3 Results of the unfold-PLS calibration

The calibration procedure is performed in such a way that each stationary phase is omitted once. This stationary phase is then used as an independent test set. The calibration is, therefore, done six times. The calibration procedure is explained in Subsection 8.2.1. Column-centering in the way described in Section 13.1 is always performed on the **X[MaMPh]** and **X[NMaMPh]** data cubes or, alternatively, the X[MaMPh] and X[NMaMPh] data matrices. The results of the unfold-PLS predictions are presented in Table III.59.

In order to explain Table III.59 the results of the C18 phase are examined extensively. The training set consists of all other stationary phases. In the block of explaining variables - the X[MaMPh] matrix - one component takes away 99.2% of the variation present in that block; only noise is left which can be checked by comparing the residual standard deviation of the X[MaMPh] block (not given) after applying the first PLS component with the mean s_{repro} of the training set. Analogous reasoning holds for the X[NMaMPh] block: 99.0% of the variation in that block is explained by the first PLS component and only noise is left. One PLS component is therefore sufficient to build the PLS model.

The next step in the calibration process is the measurement of the retention values of the three markers at the three mobile phase compositions on the C18 stationary phase. A yardstick of the information contained in these marker retention values is the MS_{mark}, which has the following meaning: the nine (3x3) marker retention values on the C18 stationary phase are subtracted from the corresponding training set means. These differences are squared and averaged. Compared with the s^2_{repro} of the retention values on the C18 phase-which has to be estimated by reproducing some of the marker measurements - this MS_{mark} gives an impression of the differences between the C18 phase and the training set (the reported s^2_{repro} in Table III.59 is calculated with the use of all measurements on C18). In practice, these s^2_{repro} have to be estimated or must be known, in order to judge the information content of the marker retention values. For the C18 phase this MS_{mark} has the value of 1.3605, which is high relative to the s^2_{repro} of C18 (0.0011).

With the use of the nine marker retention values the score of the new stationary phase - the C18 phase - can be calculated (we refer again to Subsection 8.2.1). The analogon of %X[MaMPh] for the nine selected marker/mobile phase combinations is %MARK: the percentage of variation of the nine marker retention values used to estimate the score of the new stationary phase. This value is 96.9% for the C18 phase.

Analogous to MS_{mark}, the MS_{toex} can be calculated, where toex is the abbreviation of "to explain", of the non-selected solute/mobile phase combinations on the C18 phase (the test set). The differences incorporated in this MS_{toex} have to be explained by the PLS model. Although in practice the MS_{toex} value is never known, the MS_{mark} can serve as a rough estimate. The MS_{toex} value of the C18 phase is 1.4814, this value represents the "signal" which is to be explained.

If the unfold-PLS model is applied to predict the retention values of all non-selected solute/mobile phase combinations, the resulting prediction errors are summarized in MS_{res}. This MS_{res} is the mean sum of squared differences between the observed and predicted values. The percentage of explained variation in the test set is simply $100((MS_{toex}-MS_{res})/MS_{toex})$ and takes on the value of 97.0% for the C18 phase.

The C1 stationary phase is a special case. The MS_{res} is higher than the MS_{toex}. The peculiar behaviour of the C1 phase is already indicated in Section 13.1. From Table III.59a it is clear that the MS_{mark} does not exceed the $s^2{}_{repro}$ of the C1 phase, therefore the signal-to-noise ratio of the markers (MS_{mark}/s^2_{repro}) is lower than one. This means that no systematic information is present in the marker retention values, the score of C1 in the PLS model is estimated as -0.34, which is very small in comparison with e.g the C8 score: 4.6. The signal-to-noise ratio of the test set (MS_{toex}/s^2_{repro}) is also lower than one indicating that there is no systematic variation to explain. The diagnostic value of the MS_{mark} and %MARK is stressed: both indicate the peculiar behaviour of the C1 phase. The opposite holds for %X[MaMPH] and %X[NMaMPh]: their diagnostic value is questionable. The reason for the special behaviour of the C1 phase is directly connected with the observation in Section 13.1 that the C1 phase takes an intermediate position in the range of the six studied stationary phases. The behaviour of the C1 phase can be elucidated by the following analogy. A single y vector of mean-centered measurements is regressed on an x vector, which is also mean-centered. A value of y in the neighborhood of zero is predicted with a near-zero x value. The MS_{toex} is very low in this case and will be approximately equal to MS_{res}. Yet, it cannot be said that the model is wrong.

Examining Table III.59a, the MS_{mark} values are found to be reasonable approximates of the MS_{toex} values, thereby fortifying their diagnostic power. Only in the predictions on the C18 and CN phases lack-of-fit is present. This can be concluded from the fact that the MS_{res} is more than twice the corresponding s^2_{repro} for both stationary phases. Note that both the MS_{mark} and MS_{toex} values of the C18 and CN phases are the highest ones. This illustrates that these stationary phases are the extreme ones, as was already concluded in Section 13.1. Extrapolation on the "stationary phase scale" is therefore, probably, one of the reasons for the lack-of-fit. Besides, a CN phase is known to be more polar than alkyl-bonded phases[40] and induces, therefore, selective interactions. This partly explains the anomalous behaviour of the CN phase.

When the root mean squared error of prediction values of the solutes (Table III.59b) are compared with the corresponding s_{repro} (Table III.5), it appears that the predictions of the retention values of all solutes on the C18 phase show lack-of-fit (RMSEP more than twice the s_{repro}). Closer examination of the prediction errors do not show systematic errors. The variables PE, TOL, MHB, and PHB show a lack-of-fit in the prediction on the CN phase, whereas ACP, EHB, and EAB are doubtful in this respect. Again no systematic prediction errors can be observed. The lack-of-fit in calibrating the C18 and CN phase can not be attributed to some of the solutes alone, although TOL is the one predicted worst. Note that TOL is a marker, therefore this variable is only predicted at the wm2, wa2, and aml mixtures. Even with the availability of the measurements of this solute at the selected mobile phase compositions, predictions at the other mobile phase compositions are troublesome. High residuals are observed for the prediction of TOL at the wm2 and aml mixtures on both the C18 and the CN phase. This poor predictability was already foreboded by the results of the unfold-PCA and PARAFAC model in Section 13.1 with respect to this solute. The variable MHB has a low percentage of explained variation on the C18 phase. This is due to a low MS_{toex} for this solute and poor predictions at the water/acetonitrile mixtures. Despite the presence of PHB - a related compound of MHB - as a marker, the unfold-PLS model is not able to predict MHB well at the water/acetonitrile mixtures on C18. The retention behaviour of PHB at the water/acetonitrile mixtures on C18 is not representative of the behaviour of MHB under these circumstances and/or the unfold-PLS model is inappropriate.

The RMSEP values of the individual mobile phase compositions show some high values for the water/acetonitrile mixtures on the C18 and CN phase indicating interaction effects between these mixtures and the C18 and CN phases not accounted for by the marker retention values and the model. As already discussed, MHB is one of the solutes which contributes to the high value on the C18 phase. The results of the C1 phase for the individual solutes and mobile phase compositions will not be discussed, because of the peculiar behaviour of the C1 phase as discussed above.

The results of the unfold-PLS calibration of the stationary phases and the calibration results in Chapters 10 and 11 (two-way approaches) can be compared. The differences between the two calibration strategies, however, should be kept in mind. In the two-way approach the training set consisted of three stationary phases, whereas five stationary phases constitute the training set in the three-way case. More effort is needed therefore, to build a training set in the three-way approach. When calibration of the new stationary phase at the specific mobile phase compositions is needed, the markers have to be measured at all these mobile phase composition in case of the two-way approach. In the three-way case, measurements of the markers at some mobile phase compositions are needed to predict at all other training set mobile phase compositions; less effort is needed compared with the two-way case. The pay-off between the two approaches is clear: less effort in the training state (two-way) or less effort in the calibration state (three-way). Note that the two-way approach leaves the way open for calibration at a mobile phase composition that is not incorporated in the training set; such a feature is not yet available in the three-way approach.

The RMSEP values of the C6, C8, and PHE phases are, respectively, 0.0762; 0.0686; and 0.0794 (the square roots of the MS_{res} values). Comparing these values with the corresponding ones from the two-way approach, as presented, e.g., in Table III.12 (respectively 0.0832, 0.0743, and 0.0703), no clear distinctions are visible. The two differences between the two-way and three-way approach work in opposite direction on the prediction errors and balance each other.

13.4 Results of the PARAFAC calibration

The application of the PARAFAC model for calibration purposes is already explained in Subsection 8.2.2. As in Section 13.3, one stationary phase is omitted successively, which is used later as an independent test set. Prior to applying the PARAFAC model, the data cube is centered as described in Section 13.3. The same marker/mobile phase combinations are used as earlier in Section 13.3. The results for the individual stationary phases are presented in Table III.60a which partly reproduces Table III.59a for the sake of convenience.

One component in the PARAFAC model explains 98-99% (the R^2_{train} values times 100) of the variation in the training cube. For each training cube the mean sum of squares of the residuals after applying one PARAFAC component was always lower than the average s^2_{repro} of that training cube. Therefore, PARAFAC models with one component were always appropriate.

The remarks made in Section 13.3 with respect to the signal-to-noise ratios and the special behaviour of the C1 phase are valid here. The MS_{mark} and %MARK values are again of diagnostic importance, whereas $R^2{}_{train}$ is not. The same pattern with respect to the quality of the predictions on the separate stationary phase is present as in the unfold-PLS calibration. Retention on the stationary phases C18 and CN are again predicted with a lack-of-fit, which can not be assigned to some of the solutes. No differences are visible in the predicting power of unfold-PLS and PARAFAC for the individual stationary phases.

The variable TOL is predicted on the C18 and CN phases with high prediction errors. Especially at the am1 mixture on the C18 phase, TOL has a large residual, probably because of non-linear mixing behaviour. The variable MHB has large prediction errors at the water/acetonitrile mixtures on the C18 phase, already discussed in Section 13.3.

Considering the RMSEP values of the individual mobile phase compositions, RMSEP of wal on C18 is high. The largest prediction errors are observed for MHB, EHB, and PHB. These homologs show specific behaviour in water/acetonitrile mixtures on the C18 phase which are not accounted for by the markers and the PARAFAC model.

13.5 Conclusions of the three-way approaches

1. There is no clear preference, on purely statistical grounds, for the unfold-PLS or PARAFAC calibration. The question rises how much the unfold-PLS calibration differs from the PARAFAC calibration, because it was shown that the unfold-PCA and PARAFAC models of the whole data cube do not differ very much with respect to the underlying factor structure.

2. If unfold-PLS and PARAFAC show differences between their predicting performance, statistical tools have to be developed in order to chose between both models. These tools should be evaluated with respect to their performance. Besides, a choice between unfold-PLS or PARAFAC should also be performed on chromatographic arguments. The data set at hand was too small to try a cross-validation within the training set: if one of the five stationary phases in the training set is omitted, only four stationary phases remain in that training set, a number which is probably too low.

3. Non-linear mixing behaviour of the retention of the solutes (see the remarks regarding TOL) may cause trouble when predicting on a new stationary phase, like it did in the two-way approaches (see Section 12.3). Modifications of the unfold-PLS and PARAFAC models which reckon with these phenomena are worth trying.

4. The fifth conclusion of Section 12.3 holds also for the three-way case (see the remarks regarding MHB, EHB, and PHB at the water/acetonitrile mixtures).

5. It is unknown how the drift in the measurements, as reported in Section 9.5, affects the performance of the unfold-PLS and PARAFAC predictions.

6. One of the draw-backs of the three-way approaches is that predictions are only available at the mobile phase compositions used in the training set. Explicit incorporation of the solvent fractions, present in a mobile phase mixture, in the three-way models might solve this problem. This would also serve another purpose: when predictions at every mobile phase composition on a new stationary phase are available, the way is open to predict an optimal mobile phase composition on that new stationary phase for a given separation problem.

References Part III

1 S.T. Balke; *Quantitative Column Liquid Chromatography*, J.Chrom.Lib. vol 29, Elsevier, Amsterdam, 1984

2 J.W. Weyland ; *Strategies for mobile phase optimization in chromatography*, PhD.Thesis, University of Groningen, 1986

3 G.E.P. Box and N.R. Draper; *Emperical model building and response surfaces*, John Wiley, New York, 1987

4 L.R. Snyder and J.J. Kirkland; *Introduction to Modern Liquid Chromatography (2nd ed)*, John Wiley, New York, 1979

5 G.E. Berendsen; *Preparation and characterization of well-defined chemically bonded stationary phases for high pressure liquid chromatography*, PhD.Thesis, Delft University, 1980

6 C. Horvath, W.R. Melander and I. Molnar; *Solvophobic interactions in Liquid Chromatography with non-polar stationary phases*, J.Chrom.,125 (1976) 129-156

7 A. Nahum and C. Horvath; *Surface Silanols in Silica-Bonded Hydrocarbonaceous stationary phases I. Dual retention mechanism in reversed-phase chromatography*; J.Chrom., 203 (1981) 53-63

8 K.E. Bij, C. Horvath, W.R. Melander and A. Nahum; *Surface silanols in Silica Bonded Hydrocarbonaceous stationary phases II. Irregular retention behaviour and effect of silanol masking*, J.Chrom., 203 (1981) 65-84

9 P.J. Schoenmakers; *Optimization of Chromatographic Selectivity*, J.Chrom.Lib. (vol 35), Elsevier, Amsterdam, 1986

10 J.W. Munson; *High-performance liquid chromatography: Theory, Instrumentation, and Pharmaceutical applications*, In: *Pharmaceutical Analysis* (ed J.W.Munson), Marcel Dekker, New York, 1984

11 W.R. Melander and C. Horvath; *Reversed-Phase Chromatography*, In: *High-Performance Liquid Chromatography: Advances and Perspectives (Vol.2)* (ed.C. Horvath), Ac.Press, New York, 1980

12 R.R. Picard and R.D. Cook; *Cross-validation of regression models*, J.Amer.Stat.Ass., 79(387) (1984) 575-583

13 D.A. Belsley, E. Kuh and R.E. Welsch; *Regression Diagnostics*, John Wiley, New York, 1980

14 R.D. Cook and S. Weisberg; *Residuals and Influence in Regression*, Chapman and Hall, 1982

15 J. Mandel; *Use of the Singular Value Decomposition in Regression Analysis*, Amer.Statist., 36(1) (1986) 369-380

16 R.W. Hoerl, J.H. Schuenemeyer and A.E. Hoerl, *A simulation of biased estimation and subset selection regression techniques*, Technometrics 28(4) (1986) 369-380

17 P.J. Rousseew and A.M. Leroy; *Robust Regression and Outlier Detection*, John Wiley, New York, 1987

18 P.J. Huber: *Robust Statistics*, John Wiley, New York,1981

19 A.C. Atkinson; *Plots, Transformations and Regression: an introduction to graphical methods of diagnostic regression analysis*, Clarendon, Oxford, 1985

20 H.D. Vinod and A. Ullah; *Recent advances in regression methods*, Marcel Dekker, New York, 1981

21 N.R. Draper and R.C. van Nostrand; *Ridge Regression and James-Stein Estimation: Review and Comments*, Technometrics 21(4) (1979) 451-466

22 P.J. Brown; *Centering and Scaling in Ridge Regression*, Technometrics 19(1) (1977) 35-36

23 A.E. Hoerl and R.W. Kennard; *Ridge Regression: Applications to Nonorthogonal Problems*, Technometrics 12(1) (1970) 69-82

24 R.R. Hocking; *The analysis and selection of variables in linear regression*, Biometrics, 32 (1976) 1-49

25 C. Albano, G. Blomqvist, D. Coomans, W.J. Dunn III, U. Edlund, B. Eliasson, S. Hellberg, E. Johansson, B. Norden, D. Johnels, M. Sjöström, B. Söderström, H. Wold and S. Wold; *Pattern Recognition by means of disjoint principal component models (SIMCA). Philosophy and Methods*, Proc.Symp. on Applied Statistics, Copenhagen, 1981

26 P. Geladi and B.R. Kowalski; *Partial Least-Squares Regression: A Tutorial*, Anal.Chim.Acta, 185 (1986) 1-17

27 S. Wold, personal communication, 1988

28 T.K. Dijkstra; *Latent Variables in Linear Stochastic Models (2nd ed)*, Sociometric Research Foundation, Amsterdam, 1985

29 D.W. Osten; *Selection of Optimal Regression models via Cross-Validation*, J.Chemometrics 2 (1988) 39-48

30 G.C. McDonald and D.I. Galarneau; *A Monte Carlo Evaluation of Some Ridge-type Estimators*, J.Amer.Stat.Ass., 70(350) (1975) 407-416

31 C.L. Mallows; *Some comments on C_p*, Technometrics, 15 (1973) 661-675

32 G.G. Judge and M.E. Bock; *Biased Estimation*, In: *Handbook of Econometrics vol.1* (eds. Z. Griliches and M.D. Intriligator), North Holland, 1983

33 D.L. Massart, L. Kaufmann, P.J. Rousseeuw and A. Leroy; *Least Median of Squares: A robust method for outlier and model error detection in regression and calibration*, Anal.Chim.Acta, 187 (1986) 171-179

34 E. Walker and J.B. Birch; *Influence Measures in Ridge Regression*, Technometrics, 30(2) (1988) 221-227

35 P.M.J. Coenegracht, personal communication, 1989

36 P.J. Schoenmakers and Th. Blaffert; *Effect of model inaccuracy on selectivity optimization procedures in reversed-phase liquid chromatography*, J.Chromat., 384 (1987) 117-133

37 I.T. Jolliffe; *Discarding variables in a principal component analysis, II: Real Data*, Appl.Statist., 22 (1973) 21-31

38 H.G. Law, C.W. Snyder, J.A. Hattie and R.P. McDonald (eds); *Research methods for Multimode Data Analysis*, Praeger Publ., New York, 1984

39 S. Wold, P. Geladi, K. Esbensen and J. Öhman; *Multi-way Principal Components- and PLS-Analysis*, J.Chemometrics, 1(1), (1987) 33-40

40 R.M. Smith and S.L. Miller; *Comparison of the selectivity of Cyano-bonded silica stationary phases in Reversed-phase Liquid Chromatography*, J.Chromatogr., 464 (1989) 297-306

Table III.1. The capacity factors of the solutes

		ACP	BAB	EHB	PAR	PE	TOL	EAB	MHB	PHB
C1	wm1	2.00	3.65	2.19	0.75	1.43	2.79	1.83	1.60	3.21
	wm2	3.53	8.78	4.23	0.93	2.09	4.82	3.22	2.62	7.00
	wa1	2.71	6.07	3.13	1.12	1.88	5.67	2.79	2.23	4.50
	wa2	3.89	11.53	5.10	1.08	2.59	9.02	4.33	3.25	8.44
	am1	1.88	3.92	2.22	0.85	1.50	3.01	1.87	1.61	3.06
	am2	2.92	8.18	4.02	1.15	2.26	5.25	3.31	2.62	6.22
C6	wm1	2.63	6.93	3.33	0.73	2.19	6.24	2.58	2.21	5.49
	wm2	4.89	20.07	7.56	1.05	3.89	12.78	5.48	4.35	14.59
	wa1	3.69	9.74	3.61	0.69	2.18	10.71	3.65	2.43	6.11
	wa2	6.25	24.48	7.10	0.85	3.38	20.31	6.53	4.12	13.45
	am1	3.21	9.72	3.95	0.79	2.41	8.12	3.37	2.51	6.77
	am2	5.65	24.29	8.14	1.09	4.08	15.63	6.65	4.67	15.70
C8	wm1	2.71	7.90	3.63	0.81	2.34	7.33	2.76	2.37	6.37
	wm2	4.65	19.73	7.30	0.96	3.72	12.94	4.85	3.85	14.10
	wa1	3.86	10.36	3.78	0.82	2.27	11.38	3.86	2.52	6.27
	wa2	6.15	25.33	7.06	0.99	3.51	22.05	6.61	4.24	13.30
	am1	3.25	9.85	3.92	0.84	2.42	8.17	3.18	2.39	6.52
	am2	5.79	27.48	8.50	1.06	4.17	17.47	6.85	4.74	17.61
C18	wm1	3.38	10.19	4.02	0.71	2.53	12.59	2.73	2.35	7.69
	wm2	6.15	27.54	8.98	0.99	4.60	25.71	5.69	4.50	19.02
	wa1	4.19	12.51	3.82	0.71	2.25	15.87	3.79	2.31	6.78
	wa2	6.87	30.62	7.27	0.81	3.48	33.22	6.98	3.92	15.13
	am1	4.09	14.97	4.93	0.83	2.91	15.32	3.91	2.85	9.80
	am2	6.74	37.24	10.10	0.95	4.66	29.75	7.16	4.88	23.35
CN	wm1	1.01	1.00	0.63	0.55	0.83	1.08	0.88	0.63	0.67
	wm2	1.24	1.46	0.98	0.65	0.97	1.45	1.11	0.89	1.06
	wa1	1.20	1.51	1.11	0.94	0.91	1.48	1.19	0.90	1.26
	wa2	1.51	2.22	1.38	0.90	1.23	1.99	1.50	1.17	1.60
	am1	1.00	1.09	0.85	0.73	0.88	1.11	0.94	0.80	0.90
	am2	1.10	1.35	0.91	0.68	0.92	1.32	1.04	0.83	1.01
PHE	wm1	1.28	1.50	1.08	0.79	0.95	1.19	1.13	0.95	1.22
	wm2	1.73	2.42	1.42	0.82	1.16	1.56	1.53	1.18	1.72
	wa1	1.72	2.56	1.56	1.15	1.30	2.20	1.68	1.30	1.86
	wa2	2.16	4.06	2.11	1.08	1.62	2.99	2.27	1.65	2.71
	am1	1.27	1.67	1.14	0.95	1.05	1.39	1.21	0.99	1.26
	am2	1.61	2.48	1.47	0.92	1.28	1.81	1.64	1.26	1.85

Legend: the mobile phase compositions are abbreviated to wm (water/methanol), wa (water/acetonitrile) and am (water/acetonitril/methanol). The compositions are (in volume fractions): wm1 0.47/0.00/0.53; wm2 0.55/0.00/0.45; wa1 0.62/0.38/0.00; wa2 0.70/0.30/0.00; am1 0.54/0.19/0.27; am2 0.63/0.15/0.22. The solutes were injected individually, so that no recognition problem arose.

Table III.2. Reproducibility (k-values)

	ACP	BAB	EHB	PAR	PE	TOL	EAB	MHB	PHB
C1 wm2									
mean k	3.527	8.776	4.233	0.933	2.089	4.824	3.220	2.624	7.002
stdev	0.552	2.267	0.742	0.018	0.219	0.993	0.447	0.347	1.516
CV	15.65	25.84	17.52	1.97	10.47	20.59	13.89	13.23	21.66
C6 wm2									
mean k	4.894	20.08	7.565	1.053	3.888	12.78	5.476	4.353	14.59
stdev	0.331	1.356	0.293	0.113	0.240	1.095	0.464	0.295	0.819
CV	6.76	6.75	3.87	10.71	6.18	8.57	8.46	6.78	5.61
C8 wm2									
mean k	4.646	19.73	7.297	0.962	3.722	12.94	4.85	3.85	14.10
stdev	0.556	2.589	0.817	0.101	0.432	1.878	0.708	0.595	1.803
CV	11.96	13.13	11.19	10.49	11.61	14.51	14.59	15.46	12.78
C18 wm2									
mean k	6.148	27.54	8.98	0.992	4.600	25.71	5.693	4.498	19.02
stdev	0.094	1.613	0.495	0.036	0.110	0.685	0.125	0.009	0.027
CV	1.52	5.86	5.52	3.59	2.40	2.66	2.19	0.19	0.14
CN wm2									
mean k	1.237	1.464	0.976	0.651	0.975	1.450	1.113	0.886	1.063
stdev	0.137	0.308	0.122	0.006	0.071	0.207	0.124	0.075	0.110
CV	11.05	21.03	12.47	0.95	7.33	14.29	11.11	8.44	10.35
PHE wm2									
mean k	1.730	2.419	1.420	0.822	1.163	1.565	1.525	1.182	1.716
stdev	0.183	0.376	0.132	0.080	0.077	0.159	0.140	0.081	0.202
CV	10.37	15.53	9.27	9.74	6.61	10.17	9.19	6.86	11.75

Legend: at the start of each series of measurements on a column, eluent mixture wm2 (W/A/M/ 0.55/0.00/0.45 v/v) was used. The same mixture was used at the end of each series of measurements. The result is a pair of k values for every solute (averaged over the repeated measurements, see text). The mean, the standard deviation and coefficient of variation (CV) of these two k values are given. Only with the Phenyl column the procedure differed with respect to the number of reproduced measurements, which was five for this column.

Table III.3. Repeatability (k-values)

		ACP	BAB	EHB	PAR	PE	TOL	EAB	MHB	PHB
C1	wm1	0.025	0.031	0.051	0.008	0.000	0.008	0.000	0.011	0.086
	wm2	0.023	0.068	0.046	0.013	0.009	0.072	0.029	0.051	0.079
	wa1	0.032	0.023	0.016	0.015	0.030	0.112	0.041	0.011	0.049
	wa2	0.094	0.106	0.078	0.024	0.031	0.100	0.082	0.033	0.184
	am1	0.000	0.068	0.024	0.008	0.011	0.043	0.014	0.003	0.029
	am2	0.032	0.015	0.009	0.019	0.015	0.060	0.020	0.016	0.060
C6	wm1	0.009	0.034	0.039	0.016	0.029	0.137	0.024	0.018	0.032
	wm2	0.167	0.410	0.132	0.025	0.060	0.390	0.100	0.087	0.174
	wa1	0.009	0.016	0.043	0.014	0.000	0.111	0.044	0.034	0.016
	wa2	0.075	0.124	0.038	0.017	0.030	0.184	0.192	0.047	0.059
	am1	0.020	0.060	0.018	0.000	0.000	0.018	0.000	0.009	0.009
	am2	0.047	0.163	0.122	0.010	0.062	0.212	0.124	0.038	0.219
C8	wm1	0.034	0.024	0.039	0.000	0.009	0.009	0.009	0.016	0.009
	wm2	0.043	0.315	0.110	0.000	0.032	0.089	0.009	0.040	0.390
	wa1	0.047	0.019	0.000	0.000	0.000	0.019	0.025	0.000	0.000
	wa2	0.118	0.329	0.162	0.050	0.029	0.088	0.033	0.599	0.243
	am1	0.009	0.032	0.009	0.000	0.000	0.009	0.044	0.037	0.086
	am2	0.045	0.150	0.106	0.000	0.025	0.627	0.079	0.040	0.104
C18	wm1	0.000	0.188	0.055	0.009	0.020	0.000	0.009	0.009	0.065
	wm2	0.067	0.599	0.084	0.012	0.026	0.622	0.066	0.106	0.145
	wa1	0.009	0.019	0.000	0.000	0.009	0.150	0.049	0.036	0.019
	wa2	0.074	0.050	0.078	0.049	0.042	0.613	0.074	0.048	0.140
	am1	0.053	0.316	0.018	0.009	0.047	0.058	0.009	0.000	0.028
	am2	0.167	0.689	0.311	0.015	0.009	0.128	0.046	0.044	0.629
CN	wm1	0.009	0.015	0.009	0.027	0.009	0.000	0.000	0.009	0.015
	wm2	0.018	0.020	0.005	0.022	0.006	0.006	0.006	0.010	0.020
	wa1	0.000	0.009	0.000	0.000	0.009	0.009	0.009	0.009	0.010
	wa2	0.000	0.000	0.009	0.009	0.000	0.009	0.009	0.008	0.000
	am1	0.000	0.009	0.009	0.000	0.000	0.019	0.009	0.016	0.017
	am2	0.000	0.008	0.000	0.000	0.016	0.000	0.000	0.000	0.000
PHE	wm1	0.009	0.000	0.016	0.022	0.009	0.016	0.000	0.000	0.029
	wm2	0.016	0.018	0.015	0.012	0.020	0.011	0.028	0.026	0.019
	wa1	0.009	0.009	0.000	0.034	0.000	0.009	0.000	0.000	0.009
	wa2	0.016	0.027	0.000	0.000	0.024	0.009	0.016	0.000	0.009
	am1	0.000	0.000	0.000	0.009	0.000	0.000	0.020	0.018	0.000
	am2	0.000	0.009	0.000	0.000	0.009	0.025	0.009	0.002	0.017

Legend: all numbers are standard deviations calculated on the basis of three repeated experiments. For the wm2 mobile phase compositions, the standard deviations are calculated as the square root of the pooled variances over the reproduced measurements (see Table III.2).

TABLE III.4. Logarithms of the capacity factors of the solutes

		ACP	BAB	EHB	PAR	PE	TOL	EAB	MHB	PHB
C1	wm1	0.693	1.294	0.784	-0.289	0.361	1.027	0.602	0.468	1.167
	wm2	1.254	2.155	1.435	-0.069	0.734	1.563	1.165	0.960	1.934
	wa1	0.995	1.804	1.141	0.114	0.630	1.734	1.027	0.804	1.503
	wa2	1.359	2.445	1.628	0.078	0.951	2.200	1.466	1.177	2.133
	am1	0.633	1.366	0.799	-0.167	0.404	1.103	0.627	0.478	1.118
	am2	1.071	2.102	1.392	0.139	0.816	1.659	1.198	0.965	1.828
C6	wm1	0.969	1.936	1.202	-0.309	0.785	1.831	0.947	0.792	1.704
	wm2	1.587	2.998	2.023	0.049	1.357	2.545	1.699	1.470	2.680
	wa1	1.307	2.276	1.285	-0.368	0.780	2.371	1.294	0.889	1.811
	wa2	1.832	3.198	1.960	-0.159	1.219	3.011	1.876	1.417	2.599
	am1	1.167	2.274	1.373	-0.231	0.881	2.095	1.214	0.922	1.912
	am2	1.732	3.190	2.097	0.083	1.405	2.749	1.895	1.541	2.754
C8	wm1	0.996	2.067	1.290	-0.215	0.852	1.992	1.017	0.863	1.851
	wm2	1.532	2.978	1.984	-0.042	1.311	2.555	1.573	1.342	2.642
	wa1	1.351	2.338	1.331	-0.203	0.818	2.432	1.350	0.923	1.835
	wa2	1.816	3.232	1.954	-0.006	1.256	3.094	1.889	1.439	2.588
	am1	1.180	2.288	1.367	-0.176	0.884	2.100	1.156	0.873	1.875
	am2	1.757	3.313	2.140	0.063	1.428	2.860	1.924	1.557	2.869
C18	wm1	1.218	2.321	1.390	-0.345	0.927	2.533	1.004	0.856	2.040
	wm2	1.816	3.315	2.194	-0.008	1.526	3.246	1.739	1.504	2.945
	wa1	1.432	2.526	1.341	-0.343	0.810	2.764	1.331	0.839	1.915
	wa2	1.926	3.422	1.983	-0.211	1.246	3.503	1.944	1.365	2.717
	am1	1.409	2.706	1.596	-0.191	1.069	2.729	1.364	1.048	2.283
	am2	1.908	3.617	2.312	-0.049	1.538	3.393	1.969	1.586	3.150
CN	wm1	0.005	0.000	-0.468	-0.603	-0.188	0.073	-0.129	-0.468	-0.406
	wm2	0.209	0.370	-0.029	-0.430	-0.027	0.366	0.104	-0.123	0.058
	wa1	0.185	0.409	0.104	-0.065	-0.093	0.392	0.176	-0.104	0.229
	wa2	0.411	0.795	0.322	-0.103	0.208	0.688	0.407	0.161	0.470
	am1	0.000	0.085	-0.158	-0.309	-0.134	0.108	-0.059	-0.219	-0.109
	am2	0.097	0.302	-0.091	-0.384	-0.081	0.277	0.043	-0.191	0.014
PHE	wm1	0.244	0.406	0.076	-0.242	-0.054	0.172	0.118	-0.048	0.199
	wm2	0.544	0.874	0.347	-0.200	0.149	0.444	0.419	0.165	0.535
	wa1	0.545	0.940	0.442	0.137	0.264	0.789	0.520	0.264	0.622
	wa2	0.770	1.402	0.747	0.076	0.481	1.097	0.820	0.501	0.997
	am1	0.236	0.514	0.132	-0.054	0.046	0.330	0.192	-0.005	0.232
	am2	0.476	0.908	0.384	-0.083	0.243	0.594	0.492	0.233	0.618

Legend: see Table III.1.

TABLE III.5. Reproducibility (ln k-values)

	ACP	BAB	EHB	PAR	PE	TOL	EAB	MHB	PHB
C1 wm2									
mean lnk	1.254	2.155	1.435	-0.069	0.734	1.563	1.165	0.960	1.934
stdev	0.157	0.261	0.176	0.020	0.105	0.208	0.139	0.133	0.218
CV	12.53	12.13	12.27	28.45	14.28	13.28	11.96	13.80	11.28
C6 wm2									
mean lnk	1.587	2.998	2.023	0.049	1.357	2.545	1.699	1.470	2.680
stdev	0.067	0.068	0.039	0.108	0.062	0.086	0.085	0.068	0.056
CV	4.24	2.25	1.91	220.5	4.55	3.36	4.99	4.61	2.10
C8 wm2									
mean lnk	1.532	2.978	1.984	-0.042	1.311	2.555	1.573	1.342	2.642
stdev	0.120	0.132	0.112	0.105	0.116	0.146	0.147	0.155	0.129
CV	7.83	4.42	5.65	253.1	8.88	5.67	9.31	11.57	4.86
C18 wm2									
mean lnk	1.816	3.315	2.194	-0.008	1.526	3.246	1.739	1.504	2.945
stdev	0.015	0.058	0.055	0.036	0.024	0.027	0.022	0.002	0.002
CV	0.837	1.763	2.517	429.1	1.573	0.828	1.258	0.119	0.050
CN wm2									
mean lnk	0.209	0.370	-0.03	-0.430	-0.03	0.366	0.104	-0.123	0.058
stdev	0.111	0.212	0.125	0.010	0.073	0.143	0.111	0.085	0.104
CV	52.93	57.26	435.6	2.32	272.2	39.17	107.4	68.91	178.1
PHE wm2									
mean lnk	0.544	0.874	0.347	-0.200	0.149	0.444	0.419	0.165	0.535
stdev	0.102	0.148	0.090	0.099	0.066	0.100	0.092	0.068	0.116
CV	18.75	16.95	25.98	49.69	43.91	22.58	21.90	41.48	21.64
stdev in training set	0.111	0.197	0.129	0.024	0.075	0.147	0.103	0.086	0.058
stdev in test set	0.100	0.131	0.087	0.102	0.079	0.110	0.105	0.095	0.109

Legend: see Table III.2. The last two rows give the mean standard deviations of the training- and test set. The training set consists of the ln k values of the phases C1, CN and C18. The other phases constitute the test set.

TABLE III.6. Repeatability (ln k-values)

		ACP	BAB	EHB	PAR	PE	TOL	EAB	MHB	PHB
C1	wm1	0.013	0.009	0.024	0.011	0.000	0.003	0.000	0.007	0.027
	wm2	0.006	0.009	0.011	0.014	0.004	0.026	0.008	0.018	0.010
	wa1	0.012	0.004	0.005	0.014	0.016	0.020	0.015	0.005	0.011
	wa2	0.024	0.009	0.015	0.022	0.012	0.011	0.019	0.010	0.022
	am1	0.000	0.017	0.011	0.009	0.007	0.014	0.008	0.002	0.010
	am2	0.011	0.002	0.002	0.017	0.007	0.011	0.006	0.006	0.010
	sp	0.012	0.010	0.013	0.015	0.009	0.018	0.011	0.011	0.016
C6	wm1	0.004	0.005	0.012	0.021	0.013	0.022	0.009	0.008	0.006
	wm2	0.033	0.020	0.018	0.025	0.015	0.029	0.018	0.020	0.012
	wa1	0.003	0.002	0.012	0.020	0.000	0.010	0.012	0.014	0.003
	wa2	0.012	0.005	0.005	0.020	0.009	0.009	0.029	0.011	0.004
	am1	0.006	0.006	0.005	0.000	0.000	0.002	0.000	0.004	0.001
	am2	0.008	0.007	0.015	0.009	0.015	0.014	0.019	0.008	0.014
	sp	0.019	0.012	0.013	0.019	0.012	0.019	0.017	0.013	0.009
C8	wm1	0.012	0.003	0.011	0.000	0.004	0.001	0.003	0.007	0.001
	wm2	0.009	0.015	0.015	0.000	0.008	0.006	0.002	0.010	0.029
	wa1	0.012	0.002	0.000	0.000	0.000	0.002	0.007	0.000	0.000
	wa2	0.019	0.013	0.023	0.050	0.008	0.004	0.005	0.136	0.018
	am1	0.003	0.003	0.002	0.000	0.000	0.001	0.014	0.015	0.013
	am2	0.008	0.005	0.012	0.000	0.006	0.036	0.012	0.008	0.006
	sp	0.012	0.010	0.013	0.019	0.006	0.014	0.008	0.052	0.018
C18	wm1	0.000	0.019	0.014	0.013	0.008	0.000	0.003	0.004	0.008
	wm2	0.011	0.021	0.009	0.012	0.006	0.024	0.011	0.024	0.008
	wa1	0.002	0.001	0.000	0.000	0.004	0.009	0.013	0.016	0.003
	wa2	0.011	0.002	0.011	0.062	0.012	0.018	0.011	0.012	0.009
	am1	0.013	0.021	0.004	0.011	0.016	0.004	0.002	0.000	0.003
	am2	0.025	0.018	0.031	0.016	0.002	0.004	0.006	0.009	0.027
	sp	0.013	0.017	0.014	0.026	0.009	0.015	0.009	0.015	0.012
CN	wm1	0.009	0.015	0.014	0.049	0.011	0.000	0.000	0.014	0.023
	wm2	0.016	0.015	0.006	0.034	0.006	0.004	0.006	0.011	0.018
	wa1	0.000	0.006	0.000	0.000	0.010	0.006	0.008	0.010	0.008
	wa2	0.000	0.000	0.006	0.010	0.000	0.004	0.006	0.008	0.000
	am1	0.000	0.008	0.011	0.000	0.000	0.017	0.010	0.020	0.019
	am2	0.000	0.006	0.003	0.000	0.018	0.000	0.000	0.000	0.000
	sp	0.009	0.011	0.008	0.026	0.009	0.007	0.006	0.012	0.015

TABLE III.6 (continued).

PHE	wm1	0.007	0.000	0.015	0.028	0.010	0.013	0.000	0.000	0.024
	wm2	0.010	0.008	0.011	0.015	0.018	0.007	0.019	0.023	0.012
	wa1	0.005	0.004	0.000	0.030	0.000	0.004	0.000	0.000	0.005
	wa2	0.007	0.007	0.000	0.000	0.015	0.003	0.007	0.000	0.003
	am1	0.000	0.000	0.000	0.010	0.000	0.000	0.016	0.018	0.000
	am2	0.000	0.004	0.000	0.000	0.007	0.014	0.005	0.002	0.009
	sp	0.008	0.006	0.009	0.017	0.014	0.008	0.015	0.017	0.012

Legend: all numbers are standard deviations of the three ln k values measured at each mobile phase composition. The sp's are pooled standard deviations over the mobile-phase compositions. The numbers in the wm2 rows are averaged over the reproduced measurements (two for all phases except Phenyl (five)). The number for the C8 phase wa2 and MHB is an outlier. Examination of the original measurements did not show the cause. If this number is deleted the corresponding sp becomes 0.010.

TABLE III.7. Statistics for the C1, C6, C8, C18, CN and PHE data set

Moments	ACP	BAB	EHB	PAR	PE	TOL	EAB	MHB	PHB
A									
mean	3.28	10.96	3.96	0.88	2.27	9.33	3.34	2.44	7.16
variance	3.36	100.6	7.35	0.02	1.32	75.11	4.06	1.74	0.95
skewness	0.53	1.02	0.66	0.04	0.57	1.16	0.62	0.45	0.95
kurtosis	-0.90	0.03	-0.66	-0.77	-0.74	0.75	-0.85	-1.00	0.07
B									
mean	1.02	1.89	1.11	-0.14	0.69	1.74	1.01	0.73	1.54
variance	0.37	1.21	0.63	0.03	0.28	1.19	0.43	0.36	1.05
skewness	-0.18	-0.20	-0.30	-0.32	-0.12	-0.12	-0.15	-0.29	-0.29
kurtosis	-1.21	-1.22	-1.08	-0.30	-1.18	-1.39	-1.17	-1.05	-1.15

Legend: the A-part are the moments (about the mean) of the k values. The B-part are the moments (about the mean) for the ln k values. In both cases the statistics are calculated for the values in Tables III.1 and III.4 respectively.

TABLE III.8. Reproducibility and repeatability of the old Phenyl stationary phase (k-values)

Reproducibility

	ACP	BAB	EHB	PAR	PE	TOL	EAB	MHB	PHB
wm2									
mean k	2.160	3.554	1.905	0.874	1.323	1.978	1.891	1.417	2.340
stdev	0.698	1.689	0.655	0.113	0.225	0.642	0.520	0.313	0.838
CV	32.31	47.52	34.38	12.93	17.01	32.46	27.50	22.09	35.81

Repeatability

	ACP	BAB	EHB	PAR	PE	TOL	EAB	MHB	PHB
wm1	0.009	0.009	0.020	0.020	0.018	0.000	0.020	0.000	0.021
wm2	0.035	0.039	0.038	0.021	0.017	0.042	0.008	0.030	0.019
wa1	0.032	0.054	0.018	0.000	0.009	0.009	0.009	0.009	0.000
wa2	0.016	0.009	0.000	0.000	0.000	0.009	0.000	0.025	0.009
am1	0.009	0.000	0.009	0.000	0.000	0.000	0.009	0.000	0.000
am2	0.009	0.000	0.024	0.009	0.000	0.024	0.009	0.000	0.034

Legend: see Table III.2 and Table III.3.

TABLE III.9. Reproducibility and repeatability of the old Phenyl stationary phase (ln k-values)

Reproducibility

	ACP	BAB	EHB	PAR	PE	TOL	EAB	MHB	PHB
wm2									
mean lnk	0.743	1.208	0.614	-0.139	0.273	0.655	0.618	0.336	0.817
stdev	0.329	0.494	0.351	0.130	0.171	0.330	0.278	0.222	0.366
CV	44.28	40.89	57.17	93.53	62.64	50.38	44.98	66.07	44.80

Repeatability

	ACP	BAB	EHB	PAR	PE	TOL	EAB	MHB	PHB
wm1	0.006	0.005	0.017	0.029	0.018	0.000	0.016	0.000	0.016
wm2	0.013	0.009	0.022	0.022	0.012	0.017	0.004	0.019	0.010
wa1	0.017	0.019	0.011	0.000	0.007	0.004	0.005	0.007	0.000
wa2	0.007	0.002	0.000	0.000	0.000	0.003	0.000	0.014	0.003
am1	0.007	0.000	0.007	0.000	0.004	0.000	0.007	0.000	0.000
am2	0.005	0.000	0.015	0.010	0.000	0.012	0.005	0.000	0.016
sp	0.011	0.009	0.015	0.017	0.010	0.010	0.008	0.011	0.010

Legend: see Table III.5 and Table III.6.

TABLE III.10. Markers selected with the induced-variance criterion

a) leave-more-out evaluation of the markers: the five best sets of markers

wm1

solutes	%
EHB,PAR,TOL	99.606
PAR,TOL,MHB	99.567
BAB,EHB,PAR	99.559
BAB,PAR,TOL	99.535
PAR,TOL,PHB	99.532

wm2

solutes	%
ACP,EHB,PAR	99.609
ACP,PAR,PHB	99.601
BAB,EHB,PAR	99.581
BAB,PAR,PHB	99.575
ACP,PAR,MHB	99.533

wa1

solutes	%
PAR,TOL,MHB	99.571
BAB,PAR,PHB	99.567
BAB,EHB,PAR	99.561
EHB,PAR,TOL	99.540
BAB,PAR,EAB	99.536

wa2

solutes	%
PAR,TOL,MHB	99.527
BAB,PAR,TOL	99.519
EHB,PAR,TOL	99.519
BAB,EHB,PAR	99.485
BAB,PAR,MHB	99.482

am1

solutes	%
BAB,EHB,PAR	99.457
PAR,TOL,MHB	99.430
BAB,PAR,PHB	99.429
EHB,PAR,TOL	99.421
BAB,PAR,MHB	99.406

am2

solutes	%
BAB,EHB,PAR	99.395
BAB,PAR,PHB	99.361
PAR,TOL,MHB	99.355
BAB,PAR,MHB	99.336
EHB,PAR,TOL	99.334

b) the five best sets of markers

solutes	%
BAB,EHB,PAR	99.483
PAR,TOL,MHB	99.470
EHB,PAR,TOL	99.460
BAB,PAR,PHB	99.460
BAB,PAR,MHB	99.437

c) the best markers if the subset size varies from 2 to 4

solutes	%	principal components %
BAB,PAR	99.121	99.220
BAB,EHB,PAR	99.483	99.601
EHB,PAR,TOL,EAB	99.767	99.850

Table III.10 (continued).

d) the best markers if only the (nearly) iso-elutropic eluent mixtures are used

only wm1/am1/wa1		only wm2/am2/wa2	
solutes	%	solutes	%
BAB,EHB,PAR	99.496	EHB,PAR,TOL	99.667
ACP,EHB,PAR	99.484	PAR,TOL,MHB	99.648
BAB,PAR,PHB	99.471	PAR,TOL,PHB	99.645
ACP,PAR,PHB	99.470	BAB,PAR,PE	99.625
PAR,TOL,MHB	99.439	PAR,PE,EAB	99.613

Legend: a) the heading wm1 means that all three wm1 mixtures are omitted; the % are the percentages of explained variation by the set of solutes; b, and c): calculations on the whole training set.

TABLE III.11. Diagnostics of the design matrices of the induced-variance markers

a) correlation matrix model 1

	A	M	BAB	EHB	PAR
A	1.000	-.968	.072	.057	.392
M	-.968	1.000	-.144	-.133	-.469
BAB	.072	-.144	1.000	.990	.484
EHB	.057	-.133	.990	1.000	.577
PAR	.392	-.469	.484	.577	1.000

b) variance inflation factors model 1

A	M	BAB	EHB	PAR
18.0	22.9	334.3	392.6	12.9

c) variance decomposition proportions model 1:

SV	A	M	BAB	EHB	PAR	CI
1	.0030	.0029	.0002	.0002	.0068	1
2	.0098	.0067	.0003	.0003	.0000	1.24
3	.0068	.0030	.0009	.0002	.1230	2.42
4	.9686	.8102	.0001	.0004	.0053	9.93
5	.0117	.1773	.9984	.9989	.8650	44.95

d) loadings of the variables on the PC's model 1

PC	A	M	BAB	EHB	PAR	%explained
1	.383	-.423	.460	.473	.489	54.5
2	.561	-.521	-.450	-.459	-.024	35.6
3	-.239	.179	-.367	-.180	.862	9.34
4	.694	.715	-.033	.065	.043	0.55
5	-.017	-.074	-.671	.727	-.123	0.03

e) variance inflation factors model 2

BAB	EHB	PAR
119.4	137.0	3.5

f) variance decomposition proportions model 2

SV	BAB	EHB	PAR	CI
1	.0013	.0012	.0267	1
2	.0023	.0009	.3614	1.99
3	.9964	.9979	.6119	24.84

g) loadings of the variables on the PC's model 2

PC	BAB	EHB	PAR	%
1	.613	.632	.474	79.81
2	-.403	-.266	.876	20.06
3	-.679	.728	-.092	0.13

TABLE III.11 (continued).

h) dispersion matrix model 1 total data set and test set

total data set

	C	A	M	BAB	EHB	PAR
C	1.00					
A	.171	.049				
M	.246	.014	.101			
BAB	1.893	.336	.430	4.760		
EHB	1.106	.193	.252	2.933	1.832	
PAR	-.142	-.017	-.048	-.195	-.095	.052

test set

	C	A	M	BAB	EHB	PAR
C	1.00					
A	.171	.049				
M	.246	.014	.101			
BAB	2.063	.365	.470	5.201		
EHB	1.230	.212	.285	3.213	2.000	
PAR	-.104	-.015	-.034	-.192	-.109	.032

Legend: abbreviations: singular value (SV), condition index (CI), principal component (PC), and C is an abbreviation of the constant term, see text.

TABLE III.12. OLS results model 1

a) estimated coefficients and diagnostic values

solute	b_A	b_M	b_{BAB}	b_{EHB}	b_{PAR}	F	s
ACP	-.163	-.216	2.588	.175	-.098	441.1	.0583
	.247	.278	1.064	1.154	.210		
PE	-.213	.112	3.500	-1.368	.412	234.7	.0678
	.288	.324	1.240	1.344	.244		
EAB	-.491	-.619	3.190	-.532	.227	485.9	.0574
	.244	.276	1.054	1.141	.207		
TOL	1.062	.796	7.539	-2.521	-.357	223.0	.1463
	.620	.699	2.674	2.896	.526		
MHB	-.234	-.244	-.110	2.687	.077	1079	.0361
	.153	.172	.660	.713	.129		
PHB	.321	.380	.644	4.088	-.154	4608	.0304
	.129	.145	.555	.602	.109		

b) results of predictions in training- and test set

root mean squared error of predictions (RMSEP)

solute	RMSEP in training set	RMSEP C6	RMSEP C8	RMSEP PHE	RMSEP in test set
ACP	.0703	.0502	.0693	.0660	.0624
	(98.8)	(99.3)	(98.6)	(98.2)	(98.8)
PE	.0808	.0753	.0444	.0935	.0739
	(97.8)	(98.1)	(99.4)	(95.5)	(98.0)
EAB	.0793	.1012	.0640	.0297	.0712
	(98.6)	(97.9)	(99.1)	(99.7)	(98.8)
TOL	.1756	.1313	.1335	.1208	.1286
	(97.7)	(97.9)	(98.1)	(98.8)	(98.3)
MHB	.0478	.0641	.0549	.0162	.0496
	(99.4)	(99.0)	(99.2)	(99.9)	(99.3)
PHB	.0369	.0413	.0378	.0290	.0364
	(99.9)	(99.8)	(99.9)	(99.9)	(99.9)
mean	.0933	.0832	.0743	.0703	.0761

Legend: a) the numbers in the rows with the solutes as entries are the estimated coefficients, the numbers below these estimates are their standard deviations; b) root mean squared error of prediction is calculated with the following formula: $\sqrt{[(1/z)\Sigma(y_i-\hat{y}_i)^2]}$, where z is the number of predicted values $\hat{y}$, and the summation is over all the z quadratic deviations. The numbers below the RMSEP values, in parenthesis, are the percentages of variation explained by the model. Note that y and $\hat{y}$ are in ln k units.

TABLE III.13. OLS results model 2

a) estimated coefficients and diagnostic values

solute	b_{BAB}	b_{EHB}	b_{PAR}	F	s
ACP	3.022	-.291	-.002	815.2	.0553
	.604	.647	.124		
PE	1.074	1.294	-.084	285.7	.0791
	.865	.927	.149		
EAB	4.246	-1.660	.464	663.2	.0636
	.696	.745	.119		
TOL	9.365	-4.571	-.017	322.5	.1567
	1.713	1.834	.305		
MHB	.009	2.569	.111	1753	.0364
	.383	.427	.068		
PHB	.144	4.618	-.270	5692	.0353
	.386	.413	.059		

b) results of predictions in training- and test set

root mean squared error of predictions (RMSEP)

solute	RMSEP in training set	RMSEP C6	RMSEP C8	RMSEP Phe	RMSEP in testset
ACP	.0606	.0597	.0740	.0509	.0623
	(99.2)	(99.0)	(98.4)	(98.9)	(98.8)
PE	.0937	.0800	.0621	.0517	.0657
	(97.4)	(97.8)	(98.7)	(98.6)	(98.4)
EAB	.0734	.1181	.0610	.0733	.0876
	(98.9)	(97.1)	(99.2)	(98.0)	(98.1)
TOL	.1820	.1300	.1671	.2189	.1758
	(97.8)	(98.0)	(96.9)	(96.1)	(96.9)
MHB	.0427	.0653	.0544	.0203	.0504
	(99.6)	(98.9)	(99.2)	(99.8)	(99.2)
PHB	.0400	.0511	.0372	.0381	.0426
	(99.9)	(99.7)	(99.9)	(99.8)	(99.8)
mean	.0952	.0892	.0869	.1003	.0923

Legend: see Table III.12.

TABLE III.14. Some extensions of the OLS results

a) predicted versus observed capacity factors

solute ACP on Phenyl	observed	predicted
wm1	1.28	1.18
wm2	1.73	1.56
wa1	1.72	1.57
wa2	2.16	2.10
am1	1.27	1.25
am2	1.61	1.59

solute PE on Phenyl	observed	predicted
wm1	0.95	1.08
wm2	1.16	1.37
wa1	1.30	1.34
wa2	1.62	1.66
am1	1.05	1.13
am2	1.28	1.35

solute PHB on Phenyl	observed	predicted
wm1	1.22	1.24
wm2	1.72	1.73
wa1	1.86	1.84
wa2	2.71	2.69
am1	1.26	1.28
am2	1.85	1.74

Legend: prediction results for model 1 (OLS): relatively bad predictions (PE on Phenyl), moderate good predictions (ACP on Phenyl) and very good predictions (PHB on Phenyl).

Table III.14 (continued).

b) latent root PCA (LR-PCA) of PE in the training set and on the Phenyl column

model 1: training set

dim	%	A	M	BAB	EHB	PAR	PE
1	57.56	.177	-.220	.504	.513	.393	.499
2	33.68	.647	-.629	-.207	-.206	.201	-.245
3	8.13	-.261	.196	-.267	-.082	.888	-.166
4	0.47	.684	.711	.017	.121	.041	-.104
5	0.15	.123	.092	-.278	-.523	.032	.790
6	0.02	-.003	.062	.744	-.632	.120	-.168

model 1: Phenyl test set

dim	%	A	M	BAB	EHB	PAR	PE
1	53.96	.608	-.617	.178	.176	.378	.209
2	45.09	-.107	.071	.517	.507	-.463	.495
3	0.85	.780	.418	-.133	-.031	-.428	-.122
4	0.10	.075	.615	.294	.313	.650	-.092
5	0.00	-.065	-.102	-.588	.772	-.031	-.205
6	0.00	.019	.224	-.501	-.130	.189	.804

model 2: training set

dim	%	BAB	EHB	PAR	PE
1	83.26	.534	.544	.362	.536
2	16.38	-.264	-.130	.927	-.232
3	0.27	-.458	-.364	.022	.811
4	0.10	-.660	.745	-.092	-.036

model 2: Phenyl test set

dim	%	BAB	EHB	PAR	PE
1	77.96	-.541	-.531	.387	-.525
2	22.00	.203	.199	.921	.268
3	0.02	-.274	-.543	-.053	.792
4	0.01	.769	-.620	.011	-.158

Legend: for the training set, the data are autoscaled. For the Phenyl set, the data are scaled with the scaling constants (mean, standard deviation) of the training set.

TABLE III.15. Ridge regression LOO-k parameter results model 1

a) estimated coefficients and related values

solute	b_A	b_M	b_{BAB}	b_{EHB}	b_{PAR}	k	s(k)
ACP	-.126	-.251	1.638	1.183	-.248	0.01	.0605
	.182	.187	.133	.141	.078		
PE	-.231	.012	2.528	-.316	.236	0.001	.0696
	.277	.294	.713	.772	.158		
EAB	-.414	-.652	1.807	.959	-.009	0.006	.0622
	.200	.206	.197	.211	.082		
TOL	.990	.557	5.557	-.381	-.716	0.001	.1496
	.597	.635	1.539	1.666	.341		
MHB	-.216	-.176	.472	2.055	.185	0.001	.0372
	.147	.157	.380	.411	.084		
PHB	.321	.380	.644	4.088	-.154	0.00	.0300
	.129	.145	.555	.602	.109		

b) results of predictions in training- and test set

root mean squared error of predictions (RMSEP)

	RMSEP in training set	RMSEP C6	RMSEP C8	RMSEP PHE	RMSEP in test set
solute					
ACP	.0674	.0534	.0716	.0917	.0739
PE	.0805	.0621	.0421	.0693	.0590
EAB	.0745	.0879	.0725	.0383	.0694
TOL	.1675	.1594	.1211	.0782	.1241
MHB	.0449	.0733	.0516	.0154	.0525
PHB	.0369	.0413	.0378	.0290	.0364
mean	.0894	.0884	.0718	.0604	.0744

Legend: a) the numbers in the rows with the solutes as entries are the estimated coefficients at the particular value of k and the numbers below these estimates are their standard deviations ($s_{bi(k)}$), calculated for the special k parameter

TABLE III.16. Ridge regression LOO-k parameter results model 2

a) estimated coefficients and related values

solute	b_{BAB}	b_{EHB}	b_{PAR}	k	s(k)
ACP	3.022	-.291	-.002	0.00	.0553
	.604	.647	.124		
PE	1.179	1.130	-.039	0.03	.0800
	.111	.114	.090		
EAB	4.246	-1.660	.464	0.00	.0636
	.696	.745	.119		
TOL	8.014	-3.127	-.195	0.001	.1602
	1.363	1.459	.260		
MHB	.267	2.292	.146	0.001	.0370
	.314	.336	.060		
PHB	.144	4.618	-.270	0.00	.0353
	.386	.413	.059		

b) results of predictions in training- and test set

	root mean squared error of predictions (RMSEP)				
solute	RMSEP in training set	RMSEP C6	RMSEP C8	RMSEP Phe	RMSEP in testset
ACP	.0606	.0597	.0740	.0509	.0623
PE	.0860	.0831	.0682	.0410	.0664
EAB	.0734	.1181	.0610	.0733	.0876
TOL	.1809	.1263	.1606	.1936	.1625
MHB	.0426	.0677	.0507	.0191	.0500
PHB	.0400	.0511	.0372	.0381	.0426
mean	.0936	.0890	.0852	.0903	.0882

Legend: see Table III.15

TABLE III.17. Ridge regression best k parameter results model 1

a) estimated coefficients and related values

solute	b_A	b_M	b_{BAB}	b_{EHB}	b_{PAR}	k	s(k)
ACP	-.163	-.216	2.588	.175	-.098	0.00	.0583
	.247	.278	1.064	1.154	.210		
PE	-.229	.060	1.631	.649	.079	0.006	.0740
	.236	.244	.232	.250	.097		
EAB	-.487	-.675	2.474	.243	.099	0.001	.0588
	.234	.249	.604	.654	.134		
TOL	.357	-.147	2.663	2.209	-.796	0.12	.2106
	.199	.198	.133	.111	.220		
MHB	-.234	-.244	-.110	2.687	.077	0.00	.0361
	.153	.172	.660	.713	.129		
PHB	.321	.380	.644	4.088	-.154	0.00	.0300
	.129	.145	.555	.602	.109		

b) results of predictions in training- and test set

	root mean squared error of predictions (RMSEP)				
	RMSEP in training set	RMSEP C6	RMSEP C8	RMSEP PHE	RMSEP in test set
solute					
ACP	.0703	.0502	.0693	.0660	.0624
PE	.0830	.0563	.0434	.0607	.0539
EAB	.0757	.0924	.0672	.0242	.0674
TOL	.2151	.0971	.0941	.1155	.1027
MHB	.0478	.0641	.0549	.0162	.0496
PHB	.0369	.0413	.0378	.0290	.0364
mean	.1060	.0696	.0639	.0620	.0654

Legend: see Table III.15.

TABLE III.18. Ridge regression best k parameter results model 2

a) estimated coefficients and related values

solute	b_{BAB}	b_{EHB}	b_{PAR}	k	s(k)
ACP	3.022	-.291	-.002	0.00	.0553
	.604	.647	.124		
PE	1.159	1.192	-.066	0.006	.0792
	.344	.367	.103		
EAB	3.288	-.638	.338	0.002	.0678
	.462	.495	.097		
TOL	6.014	-.997	-.454	0.004	.1769
	.848	.906	.215		
MHB	.267	2.292	.146	0.001	.0370
	.314	.336	.060		
PHB	.600	4.126	-.207	0.001	.0370
	.301	.322	.057		

b) results of predictions in training- and test set

	Root Mean Squared Error of Predictions (RMSEP)				
	RMSEP in training set	RMSEP C6	RMSEP C8	RMSEP Phe	RMSEP in testset
solute					
ACP	.0606	.0597	.0740	.0509	.0623
PE	.0879	.0811	.0640	.0473	.0656
EAB	.0770	.1140	.0771	.0481	.0842
TOL	.1913	.1315	.1628	.1734	.1569
MHB	.0426	.0677	.0507	.0191	.0500
PHB	.0424	.0497	.0435	.0269	.0412
mean	.0979	.0890	.0880	.0799	.0857

Legend: see Table III.15

TABLE III.19. Stein regression results model 1

a) estimated c and s values

	LOO choice of c		best choice of c	
solute	c	s(c)	c	s(c)
ACP	.997	.0583	.965	.0644
PE	.995	.0679	1.00	.0679
EAB	.990	.0582	1.00	.0577
TOL	1.00	.1463	.960	.1568
MHB	.995	.0362	1.00	.0361
PHB	1.00	.0304	.975	.0451

b) results of predictions in training- and test set with LOO choice of c

solute	RMSEP in training set	RMSEP C6	RMSEP C8	RMSEP PHE	RMSEP in test set
ACP	.0720	.0505	.0689	.0647	.0619
PE	.0807	.0768	.0455	.0950	.0753
EAB	.0788	.1065	.0684	.0310	.0753
TOL	.1756	.1313	.1334	.1208	.1286
MHB	.0477	.0667	.0568	.0148	.0513
PHB	.0369	.0413	.0378	.0290	.0364
mean	.0932	.0848	.0752	.0705	.0771

c) results of predictions in training- and test set with best choice of c

solute	RMSEP in training set	RMSEP C6	RMSEP C8	RMSEP PHE	RMSEP in test set
ACP	.0730	.0571	.0678	.0519	.0593
PE	.0808	.0753	.0444	.0935	.0739
EAB	.0793	.1012	.0640	.0297	.0712
TOL	.1826	.0991	.1017	.1545	.1211
MHB	.0478	.0641	.0549	.0162	.0496
PHB	.0462	.0203	.0293	.0351	.0289
mean	.0963	.0747	.0644	.0792	.0730

TABLE III.20. Stein regression results model 2

a) estimated c and s values

	LOO choice of c		best choice of c	
solute	c	s(c)	c	s(c)
ACP	1.00	.0553	.980	.0572
PE	.985	.0797	1.00	.0791
EAB	1.00	.0636	1.00	.0636
TOL	.995	.1569	.970	.1615
MHB	1.00	.0364	1.00	.0364
PHB	1.00	.0353	.970	.0511

b) results of predictions in training- and test set with LOO choice of c

solute	RMSEP in training set	RMSEP C6	RMSEP C8	RMSEP PHE	RMSEP in test set
ACP	.0606	.0597	.0740	.0509	.0623
PE	.0934	.0840	.0675	.0497	.0685
EAB	.0734	.1181	.0610	.0733	.0876
TOL	.1819	.1261	.1629	.2223	.1750
MHB	.0427	.0653	.0544	.0203	.0504
PHB	.0400	.0511	.0372	.0381	.0426
mean	.0951	.0888	.0862	.1014	.0924

c) results of predictions in training- and test set with best choice of c

solute	RMSEP in training set	RMSEP C6	RMSEP C8	RMSEP PHE	RMSEP in test set
ACP	.0621	.0631	.0720	.0446	.0610
PE	.0937	.0800	.0621	.0517	.0656
EAB	.0734	.1181	.0610	.0733	.0876
TOL	.1848	.1082	.1430	.2398	.1729
MHB	.0427	.0653	.0544	.0203	.0504
PHB	.0515	.0291	.0313	.0284	.0296
mean	.0972	.0828	.0787	.1070	.0904

TABLE III.21. PLS results model 1

a) diagnostics

	loadings			
variable	diml	dim2	dim3	s_{unexpl}
A	.146	-.648	.495	.1269
M	-.197	.643	-.431	.1311
BAB	.603	.169	.221	.0276
EHB	.615	.149	.049	.0316
PAR	.446	-.339	-.720	.0090
ACP	.406	.428	.459	.0872
PE	.406	.497	.206	.1269
EAB	.416	.269	.358	.1034
TOL	.391	.435	.764	.1315
MHB	.418	.335	.030	.0616
PHB	.412	.443	.185	.0447
explained variation				total
X-block	50.55	38.09	10.79	99.42
Y-block	91.76	4.95	2.50	99.21

b) results of predictions in training- and test set

solute	RMSEP in training set	RMSEP C6	RMSEP C8	RMSEP PHE	RMSEP in test set
ACP	.0653	.0614	.0751	.1014	.0810
PE	.0831	.0540	.0424	.0626	.0536
EAB	.0847	.0909	.0801	.0598	.0780
TOL	.1767	.2053	.1090	.1385	.1562
MHB	.0454	.0861	.0471	.0445	.0622
PHB	.0580	.0559	.0572	.0647	.0594
mean	.0957	.1062	.0721	.0848	.0888

Legend: diml to dim3 stand for the number of PLS components in the model, for an explanation of X-block and Y-block, see the text.

TABLE III.22. Cross-validatory choice of PLS model complexity

solute		dim2	dim3	dim4	dim5
ACP	LOOPRESS	.4145	.0767	.0834	.0890
	TESTPRESS	.3735	.1182	.1073	.0700
PE	LOOPRESS	.1636	.1242	.1347	.1175
	TESTPRESS	.0746	.0518	.0544	.0983
EAB	LOOPRESS	.3476	.1292	.1012	.1131
	TESTPRESS	.2894	.1095	.0889	.0914
TOL	LOOPRESS	3.344	.5619	.5745	.5549
	TESTPRESS	2.209	.4392	.4247	.2979
MHB	LOOPRESS	.0441	.0371	.0411	.0412
	TESTPRESS	.0837	.0696	.0666	.0444
PHB	LOOPRESS	.2689	.0605	.0401	.0244
	TESTPRESS	.3181	.0635	.0470	.0238
TOTAL	LOOPRESS	4.583	.9896	.9750	.9401
	TESTPRESS	3.348	.8518	.7889	.6258

Legend: the LOOPRESS values are prediction error sum of squares in the training set based on the leave-one-out results. The TESTPRESS values are the prediction error sum of squares in the test set. The row of totals correspond to the summation of the entities in the Table over the solutes. See also the legend of Table III.21.

TABLE III.23. Cross-validatory choice of the ridge parameter k, results for model 1.

solute		LOOM	HOERL	TRACE	BEST
ACP	k	.01	.002	.01	.00
	LOORMSEP	.0674	.0680	.0674	.0703
	TESTRMSEP	.0739	.0707	.0739	.0624
PE	k	.001	.002	.01	.006
	LOORMSEP	.0805	.0815	.0832	.0830
	TESTRMSEP	.0590	.0555	.0543	.0539
EAB	k	.006	.001	.01	.001
	LOORMSEP	.0745	.0753	.0748	.0757
	TESTRMSEP	.0694	.0674	.0707	.0674
TOL	k	.001	.002	.04	.12
	LOORMSEP	.1675	.1685	.1785	.2151
	TESTRMSEP	.1241	.1282	.1222	.1027
MHB	k	.001	.001	.01	.00
	LOORMSEP	.0450	.0450	.0463	.0478
	TESTRMSEP	.0525	.0525	.0620	.0496
PHB	k	.00	.00	.02	.00
	LOORMSEP	.0368	.0368	.0492	.0368
	TESTRMSEP	.0364	.0364	.0548	.0364

Legend: the LOORMSEP is the leave-one-out value of the RMSEP, see legend of Table III.12. Stated otherwise, LOORMSEP is the square root of the average LOOPRESS value. The TESTRMSEP values are defined analogously.

TABLE III.24. Cross-validatory choice of ridge parameter k. Results for model 2.

solute		LOOM	HOERL	TRACE	BEST
ACP	k	.00	.001	.06	.00
	LOORMSEP	.0606	.0608	.0740	.0606
	TESTRMSEP	.0623	.0670	.0839	.0623
PE	k	.03	.006	.02	.006
	LOORMSEP	.0860	.0879	.0861	.0879
	TESTRMSEP	.0664	.0656	.0660	.0656
EAB	k	.00	.001	.04	.002
	LOORMSEP	.0734	.0745	.0977	.0770
	TESTRMSEP	.0876	.0846	.1010	.0842
TOL	k	.001	.001	.06	.004
	LOORMSEP	.1809	.1809	.2280	.1913
	TESTRMSEP	.1625	.1625	.1683	.1569
MHB	k	.001	.001	.03	.001
	LOORMSEP	.0426	.0426	.0504	.0426
	TESTRMSEP	.0500	.0500	.0620	.0500
PHB	k	.00	.00	.03	.001
	LOORMSEP	.0400	.0400	.0669	.0424
	TESTRMSEP	.0426	.0426	.0627	.0412

Legend: see Table III.23.

TABLE III.25. Cross-validatory choice of Stein parameter c.

	model1			model2		
solute	LOOM	STEIN	BEST	LOOM	STEIN	BEST
ACP						
c	.997	.999	.965	1.00	1.00	.980
LOORMSEP	.0702	.0703	.0730	.0606	.0606	.0621
TESTRMSEP	.0619	.0622	.0593	.0623	.0623	.0610
PE						
c	.995	.998	1.00	.985	.999	1.00
LOORMSEP	.0807	.0807	.0808	.0934	.0937	.0937
TESTRMSEP	.0753	.0744	.0739	.0685	.0658	.0657
EAB						
c	.990	.999	1.00	1.00	1.00	1.00
LOORMSEP	.0787	.0792	.0793	.0734	.0734	.0734
TESTRMSEP	.0753	.0716	.0713	.0876	.0876	.0876
TOL						
c	1.00	.998	.960	.995	.999	.970
LOORMSEP	.1756	.1756	.1826	.1819	.1819	.1848
TESTRMSEP	.1286	.1279	.1211	.1750	.1757	.1729
MHB						
c	.995	1.00	1.00	1.00	1.00	1.00
LOORMSEP	.0477	.0478	.0478	.0427	.0427	.0427
TESTRMSEP	.0513	.0497	.0497	.0504	.0504	.0504
PHB						
c	1.00	1.00	.975	1.00	1.00	.970
LOORMSEP	.0368	.0368	.0462	.0400	.0400	.0515
TESTRMSEP	.0364	.0364	.0289	.0426	.0426	.0296

Legend: see Table III.23.

TABLE III.26. Cross-validatory choice of modelspecification.

solute	C_p	PRESS	PRC	R^2_{adj}
ACP				
model 1	6	.0889	.0045	.992
model 2	2.6	.0662	.0037	.993
PE				
model 1	6	.1175	.0061	.986
model 2	9.0	.1582	.0077	.981
EAB				
model 1	6	.1131	.0044	.993
model 2	7.1	.0970	.0050	.992
TOL				
model 1	6	.5548	.0285	.985
model 2	6.1	.5959	.0300	.983
MHB				
model 1	6	.0412	.0017	.997
model 2	4.3	.0328	.0016	.997
PHB				
model 1	6	.0244	.0012	.999
model 2	8.8	.0288	.0015	.999

Legend: C_p, PRESS, PRC, and R^2_{adj} are explained in the text.

TABLE III.27. Cross-validatory choice of estimation method.

a) model 1

solute		OLS	PLS	RIDGE	STEIN
ACP	LOOPRESS	.0889	.0767	.0818	.0887
	TESTPRESS	.0700	.1182	.0983	.0689
PE	LOOPRESS	.1175	.1242	.1167	.1172
	TESTPRESS	.0982	.0518	.0626	.1020
EAB	LOOPRESS	.1131	.1292	.0998	.1116
	TESTPRESS	.0914	.1095	.0867	.1020
TOL	LOOPRESS	.5548	.5619	.5047	.5548
	TESTPRESS	.2979	.4392	.2773	.2979
MHB	LOOPRESS	.0412	.0371	.0364	.0409
	TESTPRESS	.0444	.0696	.0496	.0473
PHB	LOOPRESS	.0244	.0605	.0244	.0244
	TESTPRESS	.0238	.0635	.0238	.0238

b) model 2

solute		OLS	RIDGE	STEIN
ACP	LOOPRESS	.0662	.0662	.0662
	TESTPRESS	.0698	.0698	.0698
PE	LOOPRESS	.1582	.1330	.1571
	TESTPRESS	.0776	.0794	.0845
EAB	LOOPRESS	.0970	.0970	.0970
	TESTPRESS	.1382	.1382	.1382
TOL	LOOPRESS	.5959	.5890	.5957
	TESTPRESS	.5565	.4753	.5512
MHB	LOOPRESS	.0328	.0327	.0328
	TESTPRESS	.0458	.0451	.0458
PHB	LOOPRESS	.0288	.0288	.0288
	TESTPRESS	.0327	.0327	.0327

Legend: all ridge and Stein estimation is performed with the LOO-choice of the k and c values. PLS is performed with the three significant dimensions, see text.

TABLE III.28. Fine tuning EAB

a) LR-PCA on EAB in the training set

dim	%	A	M	BAB	EHB	PAR	EAB
1	58.83	.21	-.25	.49	.50	.39	.50
2	32.32	.64	-.62	-.26	-.26	.17	-.20
3	8.27	-.25	.18	-.25	-.07	.89	-.18
4	0.50	.67	.67	.09	.18	.03	-.23
5	0.07	.17	.26	-.17	-.57	.06	.74
6	0.02	-.03	-.01	.77	-.57	.11	-.26

b) LR-PCA on EAB in training set with PAR omitted

dim	%	A	M	BAB	EHB	EAB
1	61.44	.17	-.22	.55	.55	.56
2	37.66	.69	-.67	-.17	-.18	-.12
3	0.61	.69	.68	.16	.09	-.20
4	0.22	-.03	-.06	-.49	.81	-.32
5	0.06	.16	.22	-.63	-.07	.72

c) LR-PCA on EAB in training set with EHB omitted

dim	%	A	M	BAB	PAR	EAB
1	55.62	.40	-.44	.45	.47	.48
2	33.91	.56	-.51	-.49	-.03	-.44
3	9.85	-.23	.18	-.29	.88	-.24
4	0.57	.68	.70	.17	.06	-.14
5	0.05	.14	.18	-.67	-.03	.71

Legend: dim is the abbreviation of dimension and refers to the number of the principal component. For the other abbreviations, see the text.

TABLE III.29. New models for EAB

a) model 1a: EAB=F(A,M,BAB,EHB)

b_A	b_M	b_{BAB}	b_{EHB}	s	LOORMSEP	TESTRMSEP
-.531	-.762	2.129	.633	.0581	.0734	.0678
.244	.245	.418	.419			

b) model 1b: EAB=F(A,M,BAB,PAR)

b_A	b_M	b_{BAB}	b_{PAR}	s	LOORMSEP	TESTRMSEP
-.501	-.671	2.700	.137	.0559	.0670	.0678
.236	.245	.065	.005			

Legend: see Table III.12.

TABLE III.30. Diagnostics of the design matrices of the determinant criterion markers

a) correlation matrix model 1

	A	M	PAR	TOL	MHB
A	1.000	-.968	.392	.112	.068
M	-.968	1.000	-.469	-.166	-.150
PAR	.392	-.469	1.000	.383	.603
TOL	.112	-.166	.383	1.000	.945
MHB	.068	-.150	.603	.945	1.000

b) variance inflation factors model 1

A	M	PAR	TOL	MHB
21.4	19.6	8.7	43.3	61.6

c) variance decomposition proportions model 1

SV	A	M	PAR	TOL	MHB	CI
1	.0029	.0038	.0097	.0016	.0013	1
2	.0080	.0075	.0001	.0029	.0023	1.26
3	.0048	.0027	.1488	.0091	.0001	2.24
4	.7074	.9089	.0535	.0183	.0134	9.73
5	.2769	.0772	.7880	.9681	.9829	17.79

d) loadings of the variables on PC's model 1

PC	A	M	PAR	TOL	MHB	%explained
1	.410	-.446	.478	.435	.465	54.2
2	.542	-.501	-.035	-.464	-.488	34.3
3	-.237	.169	.836	-.462	-.057	10.8
4	.657	.713	.115	.151	-.154	0.57
5	.225	.114	-.242	-.599	.720	0.17

e) variance inflation factors model 2

PAR	TOL	MHB
3.2	19.2	25.8

f) variance decomposition proportions model 2

SV	PAR	TOL	MHB	CI
1	.0296	.0081	.0070	1
2	.3438	.0185	.0018	1.9
3	.6266	.9734	.9912	10.4

g) Loadings of the variables on the PC's model 2

PC	PAR	TOL	MHB	%explained
1	.471	.601	.646	77.24
2	.857	-.485	-.174	22.04
3	-.209	-.636	.743	0.72

TABLE III.30 (continued).

h) dispersion matrix model 1

total data set

	C	A	M	PAR	TOL	MHB
C	1.00					
A	.171	.049				
M	.246	.014	.101			
PAR	-.142	-.017	-.048	.052		
TOL	1.737	.318	.383	-.194	4.168	
MHB	.729	.129	.161	-.054	1.871	.881

test set

	C	A	M	PAR	TOL	MHB
C	1.00					
A	.171	.049				
M	.246	.014	.101			
PAR	-.104	-.015	-.034	.032		
TOL	1.842	.339	.404	-.182	4.337	
MHB	.841	.147	.190	-.071	2.054	.993

Legend: see Table III.11.

TABLE III.31. OLS results model 1

a) estimated coefficients and diagnostic values

solute	b_A	b_M	b_{PAR}	b_{TOL}	b_{MHB}	F	s
ACP	-.182	-.306	-.363	.796	2.121	340.8	.0662
	.306	.293	.195	.436	.520		
BAB	-.169	-.239	-.352	1.863	3.274	881.4	.0741
	.342	.328	.218	.487	.581		
EHB	.152	.259	.096	.567	1.975	1359	.0430
	.199	.190	.127	.283	.338		
PE	-.590	-.227	.401	1.814	.349	434.0	.0500
	.231	.221	.147	.329	.393		
EAB	-.413	-.737	-.306	.393	2.543	348.6	.0680
	.314	.301	.200	.448	.534		
PHB	.403	.621	.008	1.141	3.567	760.0	.0747
	.345	.331	.220	.492	.587		

b) results of predictions in training- and test set

	Root Mean Squared Error of Predictions (RMSEP)				
	RMSEP in training set	RMSEP C6	RMSEP C8	RMSEP PHE	RMSEP in test set
solute					
ACP	.0842	.0737	.0649	.1233	.0910
	(98.3)	(98.5)	(98.8)	(93.7)	(97.4)
BAB	.0923	.0832	.0702	.1075	.0773
	(99.4)	(99.4)	(99.6)	(98.7)	(99.3)
EHB	.0576	.0517	.0617	.0373	.0512
	(99.5)	(99.6)	(99.4)	(99.7)	(99.5)
PE	.0617	.0947	.0780	.0440	.0752
	(98.7)	(97.0)	(98.0)	(99.0)	(97.9)
EAB	.0856	.0514	.0446	.0844	.0626
	(98.4)	(99.5)	(99.6)	(97.3)	(99.0)
PHB	.0988	.0901	.0849	.0388	.0749
	(99.2)	(99.1)	(99.3)	(99.8)	(99.4)
mean	.0815	.0761	.0686	.0803	.0731

Legend:
a) the numbers in the rows with the solutes as entries are the estimated coefficients, the numbers below these estimates are their standard deviations (see also legend Table III.12).
b) the numbers below the RMSEP values, in parenthesis, are the percentages of variation explained by the model.

TABLE III.32. OLS results model 2

a) estimated coefficients and diagnostic values

solute	b_{PAR}	b_{TOL}	b_{MHB}	F	s
ACP	-.201	1.069	1.800	558.2	.0667
	.120	.293	.339		
BAB	-.263	1.997	3.119	1606	.0708
	.127	.311	.359		
EHB	-.044	.332	3.254	1999	.0458
	.082	.201	.232		
PE	-.118	.695	1.714	306.3	.0765
	.137	.335	.388		
EAB	.119	1.120	1.683	318.4	.0915
	.164	.401	.465		
PHB	-.272	.688	4.098	972.5	.0852
	.153	.374	.432		

b) results of predictions in training- and test set

	Root Mean Squared Error of Predictions (RMSEP)				
solute	RMSEP in training set	RMSEP C6	RMSEP C8	RMSEP Phe	RMSEP in testset
ACP	.0814	.0813	.0829	.1079	.0915
	(98.4)	(98.1)	(98.1)	(95.2)	(97.4)
BAB	.0841	.0771	.0715	.1009	.0749
	(99.5)	(99.5)	(99.6)	(98.8)	(99.3)
EHB	.0596	.0695	.0741	.0306	.0613
	(99.5)	(99.2)	(99.1)	(99.8)	(99.3)
PE	.0887	.0832	.0837	.0519	.0744
	(97.4)	(97.7)	(97.7)	(98.6)	(97.9)
EAB	.1034	.0988	.0902	.0609	.0849
	(97.6)	(98.0)	(98.3)	(98.6)	(98.2)
PHB	.1080	.1302	.1142	.0274	.1012
	(99.0)	(98.2)	(98.7)	(99.9)	(98.9)
mean	.0890	.0922	.0872	.0706	.0824

Legend: see Table III.12.

TABLE III.33. Ridge regression LOO-k results model 1

a) estimated coefficients and related values

solute	b_A	b_M	b_{PAR}	b_{TOL}	b_{MHB}	k	s(k)
ACP	-.212	-.263	-.198	1.150	1.679	0.01	.0685
	.208	.213	.115	.209	.244		
BAB	-.230	-.215	-.155	2.271	2.751	0.01	.0770
	.233	.238	.129	.233	.273		
EHB	.077	.223	.181	.775	2.721	0.002	.0440
	.179	.176	.108	.231	.275		
PE	-.489	-.163	.323	1.607	.592	0.004	.0509
	.192	.191	.111	.227	.269		
EAB	-.427	-.715	-.227	.572	2.326	0.002	.0686
	.284	.279	.170	.365	.434		
PHB	.176	.483	.209	1.628	2.962	0.006	.0782
	.266	.267	.150	.295	.348		

b) results of predictions in training- and test set

	Root Mean Squared Error of Predictions (RMSEP)				
	RMSEP in training set	RMSEP C6	RMSEP C8	RMSEP PHE	RMSEP in test set
solute					
ACP	.0804	.0619	.0664	.1060	.0806
BAB	.0886	.0546	.0728	.0853	.0688
EHB	.0564	.0417	.0577	.0463	.0490
PE	.0609	.0804	.0730	.0450	.0679
EAB	.0849	.0524	.0505	.0768	.0611
PHB	.0953	.0655	.0729	.0594	.0662
mean	.0791	.0606	.0661	.0732	.0663

Legend: a) the numbers in the rows with the solutes as entries are the estimated coefficients at the particular value of k and the numbers below these estimates are their standard deviations, calculated for the special k parameter (see also legend Table III.15).

TABLE III.34. Ridge regression LOO results model 2

a) estimated coefficients and related values

solute	b_{PAR}	b_{TOL}	b_{MHB}	k	s(k)
ACP	-.089	1.293	1.476	0.03	.0698
	.081	.129	.145		
BAB	-.174	2.199	2.846	0.01	.0726
	.103	.215	.247		
EHB	.031	.547	2.993	0.004	.0477
	.074	.170	.196		
PE	-.075	.807	1.572	0.006	.0769
	.119	.265	.305		
EAB	.181	1.257	1.485	0.02	.0924
	.119	.215	.244		
PHB	-.145	1.046	3.661	0.006	.0883
	.133	.295	.339		

b) results of predictions in training- and test set

Root Mean Squared Error of Predictions (RMSEP)

solute	RMSEP in training set	RMSEP C6	RMSEP C8	RMSEP Phe	RMSEP in testset
ACP	.0757	.0724	.0753	.0935	.0809
BAB	.0817	.0632	.0764	.0910	.0711
EHB	.0580	.0562	.0748	.0393	.0586
PE	.0880	.0853	.0892	.0454	.0759
EAB	.1020	.1096	.0946	.0593	.0903
PHB	.1042	.1083	.1160	.0380	.0942
mean	.0864	.0851	.0889	.0653	.0794

Legend: see Table III.15.

TABLE III.35. Ridge regression best results model 1

a) estimated coefficients and related values

solute	b_A	b_M	b_{PAR}	b_{TOL}	b_{MHB}	k	s(k)
ACP	-.120	-.104	-.014	1.346	1.312	0.08	.0822
	.078	.081	.071	.064	.060		
BAB	-.199	-.128	-.004	2.452	2.430	0.04	.0869
	.131	.136	.091	.104	.111		
EHB	.077	.223	.181	.775	2.721	0.002	.0440
	.179	.176	.108	.231	.275		
PE	-.319	-.042	.230	1.326	.900	0.025	.0556
	.112	.116	.068	.094	.105		
EAB	-.428	-.703	-.200	.632	2.252	0.003	.0690
	.272	.269	.159	.335	.397		
PHB	.134	.456	.244	1.709	2.858	0.008	.0795
	.249	.254	.139	.262	.307		

b) results of predictions in training- and test set

	Root Mean Squared Error of Predictions (RMSEP)				
	RMSEP in training set	RMSEP C6	RMSEP C8	RMSEP PHE	RMSEP in test set
solute					
ACP	.0893	.0754	.0759	.0746	.0753
BAB	.0941	.0497	.0817	.0621	.0588
EHB	.0564	.0417	.0577	.0463	.0490
PE	.0620	.0666	.0695	.0485	.0622
EAB	.0849	.0541	.0529	.0739	.0611
PHB	.0953	.0633	.0710	.0636	.0661
mean	.0818	.0595	.0688	.0625	.0626

Legend:
a) the numbers in the rows with the solutes as entries are the estimated coefficients at the particular value of k and the numbers below these estimates are their standard deviations, calculated for the special k parameter.

TABLE III.36. Ridge regression best results model 2

a) estimated coefficients and related values

solute	b_{PAR}	b_{TOL}	b_{MHB}	k	s(k)
ACP	-.001	1.328	1.303	0.10	.0813
	.066	.067	.067		
BAB	-.053	2.373	2.536	0.04	.0821
	.082	.118	.130		
EHB	.060	.631	2.891	0.006	.0496
	.071	.158	.182		
PE	-.118	.695	1.714	0.00	.0765
	.137	.335	.388		
EAB	.119	1.120	1.683	0.00	.0915
	.164	.401	.465		
PHB	-.028	1.357	3.270	0.015	.0959
	.116	.226	.257		

b) results of predictions in training- and test set

	Root Mean Squared Error of Predictions (RMSEP)				
solute	RMSEP in training set	RMSEP C6	RMSEP C8	RMSEP Phe	RMSEP in testset
ACP	.0833	.0777	.0765	.0741	.0761
BAB	.0873	.0517	.0871	.0739	.0641
EHB	.0588	.0514	.0755	.0435	.0584
PE	.0887	.0832	.0837	.0519	.0744
EAB	.1034	.0988	.0902	.0609	.0849
PHB	.1077	.0905	.1196	.0559	.0924
mean	.0756	.0777	.0900	.0611	.0759

Legend: see Table III.15.

TABLE III.37. Stein regression results model 1

a) estimated c and s values

	LOO-choice of c		best choice of c	
solute	c	s(c)	c	s(c)
ACP	.995	.0663	.91	.0970
BAB	.995	.0744	.96	.0933
EHB	.995	.0433	.975	.0500
PE	.995	.0502	1.00	.0500
EAB	.990	.0685	.995	.0681
PHB	.990	.0759	.960	.0917

b) results of predictions in training- and test set with LOO-choice of c

solute	RMSEP in training set	RMSEP C6	RMSEP C8	RMSEP PHE	RMSEP in test set
ACP	.0841	.0723	.0643	.1206	.0892
BAB	.0919	.0801	.0712	.1028	.0756
EHB	.0574	.0478	.0596	.0398	.0498
PE	.0616	.0964	.0798	.0452	.0768
EAB	.0854	.0549	.0504	.0788	.0626
PHB	.0982	.0800	.0787	.0456	.0699
mean	.0812	.0738	.0681	.0786	.0717

c) results of predictions in training- and test set with best choice of c

solute	RMSEP in training set	RMSEP C6	RMSEP C8	RMSEP PHE	RMSEP in test set
ACP	.1021	.0686	.0740	.0783	.0737
BAB	.1003	.0670	.0881	.0721	.0633
EHB	.0601	.0328	.0535	.0507	.0466
PE	.0617	.0947	.0780	.0440	.0752
EAB	.0855	.0531	.0474	.0816	.0625
PHB	.1042	.0505	.0668	.0684	.0624
mean	.0876	.0640	.0694	.0673	.0646

Legend: see Table III.19.

TABLE III.38. Stein regression results model 2

a) estimated c and s values

	LOO-choice of c		best choice of c	
solute	c	s(c)	c	s(c)
ACP	.995	.0668	.940	.0798
BAB	.995	.0711	.970	.0811
EHB	.995	.0460	.960	.0594
PE	.990	.0768	1.00	.0765
EAB	.995	.0916	1.00	.0915
PHB	.990	.0861	.930	.1211

b) results of predictions in training- and test set with LOO-choice of c

solute	RMSEP in training set	RMSEP C6	RMSEP C8	RMSEP PHE	RMSEP in test set
ACP	.0813	.0807	.0826	.1055	.0903
BAB	.0840	.0745	.0729	.0966	.0733
EHB	.0595	.0657	.0719	.0319	.0592
PE	.0885	.0842	.0858	.0492	.0750
EAB	.1034	.1012	.0923	.0598	.0863
PHB	.1076	.1205	.1083	.0256	.0947
mean	.0888	.0897	.0865	.0685	.0807

c) results of predictions in training- and test set with best choice of c

solute	RMSEP in training set	RMSEP C6	RMSEP C8	RMSEP PHE	RMSEP in test set
ACP	.0894	.0822	.0864	.0812	.0833
BAB	.0902	.0668	.0847	.0767	.0653
EHB	.0673	.0411	.0625	.0464	.0508
PE	.0887	.0832	.0837	.0519	.0744
EAB	.1034	.0988	.0902	.0609	.0849
PHB	.1292	.0695	.0919	.0595	.0749
mean	.0965	.0757	.0838	.0640	.0732

Legend: see Table III.20.

TABLE III.39. PLS results model 1

a) diagnostics

	loadings			
variable	dim1	dim2	dim3	s_{unexpl}
A	.161	-.657	.407	.1281
M	.210	.651	-.338	.1308
PAR	.438	-.314	-.767	.0252
TOL	.592	.167	.363	.0707
MHB	.623	.137	-.028	.0787
ACP	.405	.416	.630	.1063
BAB	.407	.418	.565	.0656
EHB	.413	.379	.021	.0700
PE	.404	.496	.245	.0970
EAB	.413	.263	.436	.1288
PHB	.409	.440	.182	.0775

explained variation				total
X-block	49.56	38.34	11.45	99.25
Y-block	93.21	5.14	0.93	99.28

b) results of predictions in training- and test set

solute	RMSEP in training set	RMSEP C6	RMSEP C8	RMSEP PHE	RMSEP in test set
ACP	.0821	.0702	.0739	.1019	.0832
BAB	.0903	.0514	.0779	.0860	.0691
EHB	.0686	.0706	.0519	.0829	.0697
PE	.0650	.0523	.0618	.0523	.0557
EAB	.1058	.1018	.0843	.0721	.0869
PHB	.0987	.0785	.0699	.0774	.0754
mean	.0864	.0728	.0708	.0802	.0740

Legend: see Table III.21.

TABLE III.40. Cross-validatory choice of ridge parameter k. Results for model 1.

solute		LOOM	HOERL	TRACE	BEST
ACP	k	.01	.004	.02	.08
	LOORMSEP	.0804	.0812	.0809	.0893
	TESTRMSEP	.0806	.0845	.0780	.0753
BAB	k	.01	.002	.01	.04
	LOORMSEP	.0886	.0902	.0886	.0941
	TESTRMSEP	.0688	.0823	.0720	.0588
EHB	k	.002	.001	.01	.002
	LOORMSEP	.0564	.0566	.0598	.0564
	TESTRMSEP	.0490	.0496	.0538	.0490
PE	k	.004	.003	.02	.025
	LOORMSEP	.0609	.0610	.0618	.0620
	TESTRMSEP	.0679	.0693	.0623	.0622
EAB	k	.002	.003	.01	.003
	LOORMSEP	.0849	.0849	.0870	.0849
	TESTRMSEP	.0611	.0611	.0645	.0611
PHB	k	.006	.002	.015	.008
	LOORMSEP	.0953	.0964	.0961	.0953
	TESTRMSEP	.0662	.0694	.0678	.0661

Legend: see Table III.23.

TABLE III.41. Cross-validatory choice of ridge parameter k. Results for model 2.

solute		LOOM	HOERL	TRACE	BEST
ACP	k	.03	.003	.03	.10
	LOORMSEP	.0757	.0794	.0757	.0833
	TESTRMSEP	.0809	.0895	.0809	.0761
BAB	k	.01	.001	.01	.04
	LOORMSEP	.0817	.0835	.0817	.0873
	TESTRMSEP	.0711	.0832	.0711	.0641
EHB	k	.004	.001	.01	.006
	LOORMSEP	.0580	.0586	.0616	.0588
	TESTRMSEP	.0586	.0602	.0592	.0584
PE	k	.006	.005	.01	.00
	LOORMSEP	.0880	.0880	.0881	.0887
	TESTRMSEP	.0759	.0756	.0768	.0744
EAB	k	.02	.006	.02	.00
	LOORMSEP	.1020	.1026	.1020	.1034
	TESTRMSEP	.0903	.0867	.0903	.0849
PHB	k	.006	.001	.04	.015
	LOORMSEP	.1042	.1065	.1195	.1077
	TESTRMSEP	.0942	.0994	.0980	.0924

Legend: see Table III.24.

TABLE III.42. Cross-validatory choice of Stein parameter c

	model1			model2		
solute	LOOM	STEIN	BEST	LOOM	STEIN	BEST
ACP						
c	.995	.999	.910	.995	1.00	.940
LOORMSEP	.0841	.0842	.1021	.0813	.0814	.0894
TESTRMSEP	.0892	.0905	.0737	.0903	.0915	.0833
BAB						
c	.995	1.00	.960	.995	1.00	.970
LOORMSEP	.0919	.0923	.1003	.0840	.0841	.0902
TESTRMSEP	.0756	.0773	.0633	.0733	.0749	.0653
EHB						
c	.995	1.00	.975	.995	1.00	.960
LOORMSEP	.0574	.0576	.0601	.0595	.0596	.0673
TESTRMSEP	.0498	.0512	.0466	.0592	.0613	0508
PE						
c	.995	.999	1.00	.990	.999	1.00
LOORMSEP	.0616	.0617	.0617	.0885	.0887	.0887
TESTRMSEP	.0768	.0756	.0752	.0750	.0745	.0744
EAB						
c	.990	.999	.995	.995	.999	1.00
LOORMSEP	.0854	.0856	.0855	.1034	.1034	.1034
TESTRMSEP	.0626	.0625	.0625	.0863	.0851	.0849
PHB						
c	.990	.999	.960	.990	1.00	.930
LOORMSEP	.0982	.0988	.1042	.1076	.1080	.1292
TESTRMSEP	.0699	.0746	.0624	.0947	.1012	.0749

Legend: see Table III.25.

TABLE III.43. Cross-validatory choice of modelspecification.

solute	C_p	PRESS	PRC	R^2_{adj}
ACP				
model 1	6	.1276	.0058	.990
model 2	4.2	.1192	.0054	.990
BAB				
model 1	6	.1532	.0073	.996
model 2	2.8	.1274	.0061	.999
EHB				
model 1	6	.0597	.0025	.998
model 2	5.8	.0640	.0026	.997
PE				
model 1	6	.0685	.0033	.992
model 2	22.8	.1416	.0072	.982
EAB				
model 1	6	.1320	.0062	.990
model 2	15.4	.1926	.0102	.983
PHB				
model 1	6	.1759	.0074	.996
model 2	8.2	.2099	.0089	.994

Legend: see Table III.26.

TABLE III.44. Cross-validatory choice of estimation method.

a) model 1

solute		OLS	PLS	RIDGE	STEIN
ACP	LOOPRESS	.1276	.1212	.1165	.1274
	TESTPRESS	.1490	.1246	.1169	.1433
BAB	LOOPRESS	.1532	.1467	.1412	.1519
	TESTPRESS	.1404	.0966	.0933	.1323
EHB	LOOPRESS	.0597	.0847	.0573	.0593
	TESTPRESS	.0473	.0873	.0433	.0446
PE	LOOPRESS	.0685	.0759	.0668	.0684
	TESTPRESS	.1019	.0558	.0830	.1062
EAB	LOOPRESS	.1320	.2014	.1297	.1313
	TESTPRESS	.0705	.1360	.0672	.0705
PHB	LOOPRESS	.1759	.1753	.1634	.1736
	TESTPRESS	.1010	.1022	.0788	.0881

b) model 2

solute		OLS	RIDGE	STEIN
ACP	LOOPRESS	.1192	.1032	.1189
	TESTPRESS	.1507	.1179	.1468
BAB	LOOPRESS	.1273	.1200	.1271
	TESTPRESS	.1275	.1086	.1212
EHB	LOOPRESS	.0640	.0606	.0637
	TESTPRESS	.0676	.0618	.0630
PE	LOOPRESS	.1416	.1394	.1411
	TESTPRESS	.0997	.1037	.1013
EAB	LOOPRESS	.1926	.1873	.1924
	TESTPRESS	.1297	.1468	.1340
PHB	LOOPRESS	.2099	.1953	.2084
	TESTPRESS	.1845	.1599	.1615

Legend: all ridge and Stein estimation is done with the LOO-choice of the k and c values. PLS is done with the three significant dimensions, see text.

TABLE III.45. Diagnostics of the design matrix when the homologs are used as markers.

a) correlation matrix model 1

	A	M	MHB	EHB	PHB
A	1.000				
M	-.968	1.000			
MHB	.068	-.150	1.000		
EHB	.057	-.133	.998	1.000	
PHB	.038	-.108	.994	.998	1.000

b) variance inflation factors model 1

A	M	MHB	EHB	PHB
21.8	24.0	642.9	2430.3	809.0

c) variance decomposition proportions model 1

SV	A	M	MHB	EHB	PHB	CI
1	.0002	.0004	.0002	.0000	.0001	1
2	.0117	.0102	.0000	.0000	.0000	1.26
3	.7111	.6419	.0012	.0000	.0014	10.0
4	.2766	.3414	.1879	.0001	.1307	27.1
5	.0004	.0062	.8107	.9999	.8677	105.4

d) Loadings of the variables on the principal components model 1

PC	A	M	MHB	EHB	PHB	%explained
1	.124	-.167	.566	.566	.563	60.9
2	.699	-.685	-.107	-.116	-.133	38.4
3	.686	.685	-.156	.022	.187	0.61
4	-.158	-.185	-.708	.030	.663	0.08
5	-.002	-.006	.378	-.816	.438	0.01

Legend: see Table II.11.

TABLE III.46. OLS results model 1 with homologous markers

results of predictions in training- and test set

solute	RMSEP in training set	RMSEP C6	RMSEP C8	RMSEP PHE	RMSEP in test set
ACP	.1161 (96.8)	.0698 (98.6)	.1064 (96.8)	.0980 (96.0)	.0927 (97.3)
PE	.0976 (96.9)	.0585 (98.8)	.0669 (98.5)	.0641 (97.9)	.0633 (98.5)
EAB	.1018 (97.7)	.0568 (99.3)	.0679 (99.0)	.0647 (98.4)	.0633 (99.0)
TOL	.2643 (94.8)	.0824 (99.2)	.1703 (96.8)	.1943 (96.9)	.1566 (97.5)
BAB	.1253 (98.8)	.0591 (99.7)	.1188 (98.8)	.1074 (98.7)	.0986 (99.1)
PAR	.1123 (67.6)	.1183 (52.4)	.1043 (40.5)	.1572 (24.2)	.1286 (38.2)
mean	.1481	.0772	.1114	.1238	.1060

Legend: see Table III.12.

TABLE III.47. Ridge regression results model 1 with homologous markers

a) results of predictions in training- and test set with LOO-choice of k

solute	RMSEP in training set	RMSEP C6	RMSEP C8	RMSEP PHE	RMSEP in test set
PE (k=0.002)	.0936	.0617	.0574	.0603	.0598
mean	.1476	.0777	.1105	.1234	.1056

b) results of predictions with best choice of k

solute	RMSEP in training set	RMSEP C6	RMSEP C8	RMSEP PHE	RMSEP in test set
ACP (k=0.11)	.1333	.1033	.1102	.0492	.0917
PE (k=0.03)	.0966	.0434	.0479	.0568	.0497
TOL (k=0.002)	.3036	.1195	.1990	.0991	.1457
BAB (k=0.02)	.1824	.0778	.0875	.0412	.0717
mean	.1710	.0914	.1134	.0876	.0981

Legend: when the results of a solute are not reported in a) or b) the respective k parameter was zero, these results are presented in Table III.46.

TABLE III.48. Stein regression results model 1 with homologous markers

a) results of predictions in training- and test set with LOO-choice of c

solute	RMSEP in training set	RMSEP C6	RMSEP C8	RMSEP PHE	RMSEP in test set
ACP (c=0.995)	.1160	.0707	.1064	.0962	.0923
PE (c=0.995)	.0976	.0603	.0686	.0636	.0643
EAB (c=0.985)	.1013	.0661	.0744	.0593	.0669
PAR (c=0.88)	.1096	.1132	.0938	.1560	.1237
mean	.1477	.0776	.1107	.1228	.1054

b) results of predictions in training- and test set with best choice of c

solute	RMSEP in training set	RMSEP C6	RMSEP C8	RMSEP PHE	RMSEP in test set
ACP (c=0.97)	.1171	.0764	.1075	.0877	.0915
PAR (C=0.73)	.1136	.1119	.0856	.1559	.1213
mean	.1484	.0767	.1088	.1222	.1044

Legend: see Table III.19.

TABLE III.49. PLS results model 1 homologous markers

results of predictions in training- and test set

solute	RMSEP in training set	RMSEP C6	RMSEP C8	RMSEP PHE	RMSEP in test set
ACP	.3126	.1059	.1129	.0511	.0941
PE	.0963	.0340	.0401	.0568	.0447
EAB	.1093	.0818	.0852	.0510	.0743
TOL	.3978	.2235	.2114	.2475	.2279
BAB	.1921	.0943	.0794	.0453	.0758
PAR	.1685	.1737	.0782	.2146	.1657
mean	.2389	.1342	.1145	.1401	.1301

Legend: The solute PAR is predicted separately, only one dimension was significant. All other solutes were gathered in the Y-block and predicted with the use of a two dimensional PLS model.

TABLE III.50. Cross-validatory choice of estimation method homologous markers

solute		OLS	PLS	RIDGE	STEIN
ACP	LOOPRESS	.2425	.3126	.2425	.2421
	TESTPRESS	.1548	.1594	.1548	.1534
PE	LOOPRESS	.1715	.1670	.1576	.1713
	TESTPRESS	.0721	.0359	.0645	.0743
EAB	LOOPRESS	.1865	.2150	.1865	.1846
	TESTPRESS	.0721	.0994	.0721	.0805
TOL	LOOPRESS	1.2573	2.8486	1.2573	1.2573
	TESTPRESS	.4414	.9352	.4414	.4414
BAB	LOOPRESS	.2828	.6644	.2828	.2828
	TESTPRESS	.1749	.1035	.1749	.1749
PAR	LOOPRESS	.2268	.5112	.2268	.2161
	TESTPRESS	.2976	.4940	.2976	.2756

Legend: see Table III.27.

TABLE III.51. Diagnostics of the design matrix when bad markers are chosen

a) correlation matrix model 1

	A	M	ACP	BAB	PE
A	1.000				
M	-.968	1.000			
ACP	.075	-.149	1.000		
BAB	.072	-.144	.997	1.000	
PE	.006	-.085	.983	.990	1.000

b) variance inflation factors model 1

A	M	ACP	BAB	PE
20.5	19.2	223.7	441.4	89.2

c) variance decomposition proportions model 1

SV	A	M	ACP	BAB	PE	CI
1	.0002	.0005	.0005	.0002	.0012	1
2	.0124	.0127	.0000	.0000	.0001	1.25
3	.7829	.8693	.0009	.0009	.0063	10.1
4	.0652	.0284	.1137	.0030	.4367	14.3
5	.1393	.0893	.8850	.9959	.5556	45.8

d) loadings of the variables on the principal components model 1

PC	A	M	ACP	BAB	PE	%explained
1	.121	-.163	.568	.569	.560	60.6
2	.699	-.685	-.099	-.102	-.147	38.5
3	.688	.702	.076	.107	-.129	0.59
4	.141	.090	-.613	-.140	.759	0.30
5	.064	.050	.535	-.797	.268	0.03

Legend: see Table III.11.

TABLE III.52. OLS results model 1 bad markers

results of predictions in training- and test set:

solute	RMSEP in training set	RMSEP C6	RMSEP C8	RMSEP PHE	RMSEP in test set
EHB	.1661 (96.1)	.0312 (99.8)	.0460 (99.7)	.0406 (99.6)	.0397 (99.7)
MHB	.1501 (94.2)	.0552 (99.2)	.0347 (99.7)	.0378 (99.4)	.0435 (99.4)
EAB	.0714 (98.9)	.0932 (98.2)	.0708 (98.9)	.0236 (99.8)	.0689 (98.8)
TOL	.2888 (93.8)	.1528 (97.2)	.1481 (97.6)	.2120 (96.3)	.1734 (97.0)
PHB	.1583 (97.9)	.0389 (99.8)	.0720 (99.5)	.0357 (99.8)	.0515 (99.7)
PAR	.2021 (-5.1)	.1399 (33.5)	.0546 (83.7)	.1785 (2.2)	.1347 (32.2)
mean	.1846	.0976	.0800	.1167	.0992

Legend: see Table III.12.

TABLE III.53. Ridge regression results model 1 bad markers

a) results of predictions in training- and test set with LOO-choice of k

solute	RMSEP in training set	RMSEP C6	RMSEP C8	RMSEP PHE	RMSEP in test set
EHB (k=0.07)	.1417	.0625	.0725	.0417	.0603
MHB (k=0.07)	.1252	.0856	.0607	.0174	.0614
EAB (k=0.008)	.0661	.0857	.0657	.0255	.0641
TOL (k=0.05)	.2433	.1691	.1619	.2404	.1938
PHB (k=0.05)	.1447	.0661	.0925	.0759	.0789
PAR (k=0.04)	.1825	.1681	.0793	.1883	.1528
mean	.1600	.1153	.0951	.1302	.1145

b) results of predictions with best choice of k

solute	RMSEP in training set	RMSEP C6	RMSEP C8	RMSEP PHE	RMSEP in test set
EAB (k=0.02)	.0669	.0836	.0647	.0290	.0633
mean	.1843	.0961	.0792	.1169	.0986

Legend: when the results of a solute are not reported in a) or b) the respective k parameter was zero, these results are presented in Table III.52.

TABLE III.54. Stein regression results model 1 bad markers

a) results of predictions in training- and test set with LOO-choice of c

solute	RMSEP in training set	RMSEP C6	RMSEP C8	RMSEP PHE	RMSEP in test set
EHB (c=0.985)	.1655	.0314	.0402	.0387	.0370
MHB (c=0.97)	.1490	.0701	.0458	.0275	.0509
EAB (c=0.99)	.0713	.0981	.0757	.0191	.0724
TOL (c=0.99)	.2886	.1440	.1403	.2202	.1722
PHB (c=0.99)	.1578	.0309	.0639	.0392	.0468
PAR (c=0.55)	.1857	.1371	.0720	.1571	.1274
mean	.1814	.0966	.0800	.1135	.0977

b) results of predictions in training- and test set with best choice of c

solute	RMSEP in training set	RMSEP C6	RMSEP C8	RMSEP PHE	RMSEP in test set
EHB (c=0.98)	.1655	.0325	.0389	.0386	.0368
TOL (c=0.97)	.2898	.1273	.1260	.2367	.1714
PHB (c=0.97)	.1593	.0207	.0505	.0503	.0428
PAR (C=0.70)	.1877	.1346	.0600	.1603	.1257
mean	.1824	.0890	.0704	.1209	.0957

Legend: when the results of a solute are not reported in a) or b) the respective c parameter was one, these results are presented in Table III.52.

TABLE III.55. PLS results model 1 bad markers

a) results of predictions in training- and test set

solute	RMSEP in training set	RMSEP C6	RMSEP C8	RMSEP PHE	RMSEP in test set
EHB	.1374	.0660	.0776	.0363	.0625
MHB	.1223	.0809	.0577	.0252	.0592
EAB	.0734	.0759	.0608	.0458	.0620
TOL	.2474	.2053	.1958	.2463	.2169
PHB	.1424	.0824	.1043	.0744	.0880
PAR	.1848	.1913	.1038	.1907	.1670
mean	.1607	.1305	.1103	.1333	.1251

Legend: see Table III.21.

TABLE III.56. Cross-validatory choice of estimation method bad markers

solute		OLS	PLS	RIDGE	STEIN
EHB	LOOPRESS	.4969	.3400	.3614	.4930
	TESTPRESS	.0284	.0702	.0654	.0246
MHB	LOOPRESS	.4055	.2693	.2820	.3997
	TESTPRESS	.0341	.0631	.0679	.0467
EAB	LOOPRESS	.0918	.0969	.0787	.0915
	TESTPRESS	.0855	.0693	.0739	.0943
TOL	LOOPRESS	1.5015	1.1019	1.0658	1.4997
	TESTPRESS	.5414	.8470	.6757	.5336
PHB	LOOPRESS	.4511	.3649	.3770	.4482
	TESTPRESS	.0478	.1393	.1121	.0394
PAR	LOOPRESS	.7354	.6148	.5996	.6205
	TESTPRESS	.3265	.5023	.4202	.2919

Legend: see Table III.27.

TABLE III.57. Three-way PCA results on the whole data cube

a) results for the stationary phases

S.Phase	mean	$s^2{}_{repro}$	MS_{bef}	$MS_{res}(1)$	%expl
C1	-.0033	.0328	.0159	.0156	2.1
C6	.5548	.0046	.3408	.0019	99.4
C8	.5781	.0177	.3715	.0017	99.5
C18	.7709	.0011	.7026	.0116	98.4
CN	-1.1317	.0161	1.4057	.0100	99.3
PHE	-.7688	.0102	.6767	.0028	99.6
average	0	.0138	.5855	.0073	98.8

b) results for the solutes

solute	$s^2{}_{repro}$	MS_{bef}	$MS_{res}(1)$	%expl
ACP	.0111	.3031	.0025	99.2
BAB	.0267	1.0234	.0033	99.7
EHB	.0120	.5299	.0063	98.8
PE	.0064	.2399	.0031	98.7
TOL	.0172	1.0457	.0204	98.0
EAB	.0116	.3370	.0084	97.5
MHB	.0097	.2975	.0073	97.6
PHB	.0153	.9077	.0068	99.2

c) results for the mobile phases

M.Phase	MS_{bef}	$MS_{res}(1)$	%expl
wm1	.4671	.0063	98.7
wm2	.7493	.0081	98.9
wa1	.3413	.0070	97.9
wa2	.5477	.0070	98.7
am1	.5147	.0067	98.7
am2	.8931	.0085	99.0

d) scores of the stationary phases on the first PC

C1	C6	C8	C18	CN	PHE
-.125	4.033	4.213	5.759	-8.193	-5.688

Legend: for the meaning of MS_{bef}, $MS_{res}(1)$, $s^2{}_{repro}$ see the text. The abbreviation %expl stands for percentage of explained variation of retention values on the stationary phase.

TABLE III.58. PARAFAC results on the whole data cube

a) results for the stationary phases

S.Phase	mean	s^2_{repro}	MS_{bef}	$MS_{res}(1)$	%expl
C1	-.0033	.0328	.0159	.0156	2.1
C6	.5548	.0046	.3408	.0030	99.1
C8	.5781	.0177	.3715	.0032	99.1
C18	.7709	.0011	.7026	.0141	98.0
CN	-1.1317	.0161	1.4057	.0117	99.2
PHE	-.7688	.0102	.6767	.0042	99.4
average	0	.0138	.5855	.0086	98.5

b) results for the solutes

solute	s^2_{repro}	MS_{bef}	$MS_{res}(1)$	%expl
ACP	.0111	.3031	.0040	98.7
BAB	.0267	1.0234	.0037	99.6
EHB	.0120	.5299	.0083	98.4
PE	.0064	.2399	.0046	98.1
TOL	.0172	1.0457	.0266	97.5
EAB	.0116	.3370	.0048	98.6
MHB	.0097	.2975	.0082	97.2
PHB	.0153	.9077	.0089	99.0

c) results for the mobile phases

M.Phase	MS_{bef}	$MS_{res}(1)$	%expl
wm1	.4671	.0089	98.1
wm2	.7493	.0096	98.7
wa1	.3413	.0106	96.9
wa2	.5477	.0094	98.3
am1	.5147	.0069	98.7
am2	.8931	.0063	99.3

d) scores of the stationary phases on the first PARAFAC component

C1	C6	C8	C18	CN	PHE
.017	-.581	-.607	-.830	1.181	.760

TABLE III.59. Results of the unfold-PLS predictions

a) results for the stationary phases

S.Phase	s^2_{repro}	%X[MaMPh]	%X[NMaMPh]	MS_{mark}	%MARK	MS_{toex}	MS_{res}
C1	.0328	99.3	99.3	.0287	14.2	.0215	.0234 (---)
C6	.0046	98.8	98.7	.5500	99.1	.4773	.0058 (98.8)
C8	.0177	98.7	98.6	.7186	99.7	.4925	.0047 (99.1)
C18	.0011	99.2	99.0	1.3605	96.9	1.4814	.0275 (97.0)
CN	.0161	98.9	98.4	2.4719	97.7	1.9208	.0394 (98.0)
PHE	.0102	98.6	98.5	1.2383	99.3	.9136	.0063 (99.3)

b) root mean squared error of prediction for the solutes

	S.Phase					
	C1	C6	C8	C18	CN	PHE
solute						
ACP	.0959 (3.6)	.0543 (98.9)	.0663 (98.3)	.0789 (98.8)	.1263 (98.6)	.0646 (99.1)
BAB	.1149 (9.1)	.0657 (99.5)	.0522 (99.7)	.1062 (99.4)	.1600 (99.3)	.0662 (99.7)
EHB	.1923 (---)	.0687 (99.0)	.0360 (99.7)	.1761 (95.8)	.2180 (97.6)	.0732 (99.4)
PE	.0964 (25.2)	.0628 (98.2)	.0366 (99.4)	.1008 (97.4)	.1489 (97.3)	.0661 (98.9)
TOL	.2097 (41.6)	.0442 (99.8)	.1487 (97.3)	.3102 (96.3)	.3523 (95.9)	.1688 (98.6)
EAB	.1118 (---)	.1219 (96.0)	.0611 (98.7)	.1336 (96.5)	.1495 (98.3)	.0528 (99.5)
MHB	.1650 (---)	.0983 (96.8)	.0589 (98.8)	.2138 (86.9)	.2267 (95.6)	.0706 (98.9)
PHB	.2363 (---)	.0888 (99.0)	.0928 (98.9)	.1866 (97.6)	.2013 (98.8)	.0608 (99.8)

Legend: all values in Table III.59b are root mean squared error of predictions (RMSEP). The values between parentheses (Table III.59a, b) are percentages of explained mean sum of squares in the test set.

TABLE III.59 (continued)

c) RMSEP for the mobile phase compositions

	S.Phase					
M.Phase	C1	C6	C8	C18	CN	PHE
wm1	.0854	.0296	.0480	.1093	.1943	.0944
	(---)	(99.7)	(99.4)	(98.1)	(97.3)	(98.4)
wm2	.1661	.1156	.1157	.1681	.2043	.1133
	(---)	(98.1)	(97.7)	(97.8)	(98.3)	(99.1)
wa1	.1995	.0168	.0202	.1775	.2804	.0986
	(---)	(99.9)	(99.8)	(90.6)	(93.0)	(97.2)
wa2	.1871	.0745	.0354	.2079	.2308	.0555
	(---)	(98.8)	(99.8)	(94.5)	(97.4)	(99.6)
am1	.1371	.0620	.0758	.1562	.1277	.0370
	(44.3)	(99.1)	(98.6)	(97.8)	(99.0)	(99.8)
am2	.0715	.0893	.0348	.1334	.1134	.0515
	(50.0)	(98.8)	(99.8)	(98.5)	(99.5)	(99.8)

TABLE III.60. Results of the PARAFAC predictions

a) results for the stationary phases

S.Phase	s^2_{repro}	R^2_{train}	MS_{mark}	%MARK	MS_{toex}	MS_{res}
C1	.0328	.99	.0287	14.7	.0215	.0237 (---)
C6	.0046	.98	.5500	98.7	.4773	.0072 (98.5)
C8	.0177	.98	.7186	99.3	.4925	.0062 (98.7)
C18	.0011	.99	1.3605	96.7	1.4814	.0307 (96.7)
CN	.0161	.98	2.4719	97.1	1.9208	.0373 (98.1)
PHE	.0102	.98	1.2383	99.2	.9136	.0078 (99.1)

b) root mean squared error of prediction for the solutes

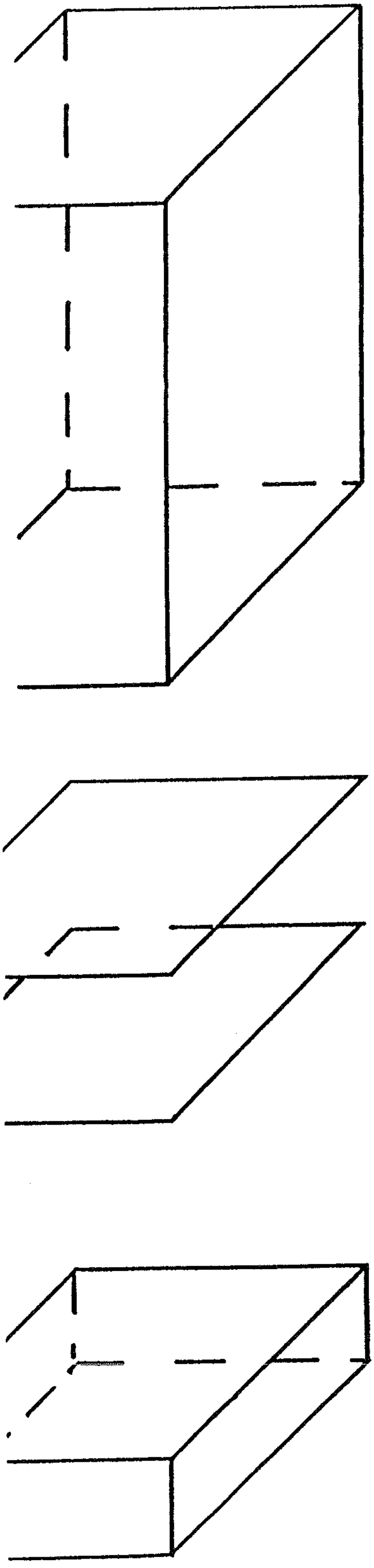
	S.Phase					
	C1	C6	C8	C18	CN	PHE
solute						
ACP	.0967 (1.9)	.0759 (97.8)	.0897 (97.0)	.0520 (99.5)	.1589 (97.7)	.0664 (99.0)
BAB	.1157 (7.9)	.0724 (99.4)	.0562 (99.7)	.1121 (99.3)	.1772 (99.1)	.0574 (99.8)
EHB	.1945 (---)	.0802 (98.6)	.0571 (99.3)	.2014 (94.6)	.1981 (98.0)	.0899 (99.1)
PE	.0984 (22.0)	.0792 (97.2)	.0383 (99.4)	.1310 (95.6)	.0731 (99.3)	.0990 (97.5)
TOL	.2071 (43.0)	.0340 (99.9)	.1665 (96.7)	.3054 (96.4)	.3891 (95.0)	.1568 (98.8)
EAB	.1117 (---)	.1199 (96.2)	.0704 (98.3)	.1323 (96.6)	.1648 (97.9)	.0359 (99.8)
MHB	.1667 (---)	.1044 (96.4)	.0591 (98.8)	.2217 (85.9)	.1989 (96.6)	.0824 (98.5)
PHB	.2386 (---)	.0925 (99.0)	.0954 (98.8)	.2111 (97.0)	.1570 (99.3)	.1018 (99.4)

TABLE III.60 (continued)

c) RMSEP for the mobile phase compositions

	S.Phase					
M.Phase	C1	C6	C8	C18	CN	PHE
wm1	.0853	.0560	.0714	.0751	.1808	.0811
	(---)	(98.8)	(98.7)	(99.1)	(97.6)	(98.8)
wm2	.1660	.1275	.1280	.1394	.2260	.1148
	(---)	(97.7)	(97.2)	(98.5)	(97.9)	(99.1)
wa1	.2027	.0352	.0412	.2189	.2267	.1200
	(---)	(99.4)	(99.3)	(85.7)	(95.4)	(95.9)
wa2	.1902	.0741	.0488	.2480	.2088	.0934
	(---)	(98.8)	(99.5)	(92.2)	(97.9)	(99.0)
am1	.1363	.0595	.0739	.1671	.1356	.0436
	(44.9)	(99.2)	(98.7)	(97.5)	(98.8)	(99.8)
am2	.0709	.1056	.0532	.1097	.1610	.0388
	(50.9)	(98.4)	(99.6)	(99.0)	(99.0)	(99.8)

Legend: see Table III.59. The abbreviation R^2_{train} is explained in the text.



PART IV
Calibration of Octadecyl Stationary Phases of Different Batches

Chapter 14 Experimental design

14.1 The choice of the stationary phases

The stationary phases, chosen to illustrate the calibration strategies, are different batches of octadecyl stationary phases of the same brand. Table IV.1a summarizes some characteristics of these stationary phases. Two batches of silica substrate material were used. Each silica substrate material was subjected to two separate silanisation procedures. This resulted in four different silica substrate/octadecyl modified materials. Each combination was duplicated, so eight stationary phases are obtained. From these eight stationary phases, six were used, as indicated in Table IV.1a. The number of six was believed to be a good compromise between investigating all four combinations in duplo and the amount of experimental effort. Measurements on these six stationary phases were performed by three different analysts, using different equipment, also indicated in Table IV.1a (details are given in Section 14.4).

Octadecyl modified material was used for two reasons. First, because its widespread use in common practice. Second, the measurements described in Part III indicated that octadecyl is a stable material for at least one month, even if used extensively.

14.2 The choice of the mobile phase compositions

Binary- and ternary mobile phase compositions were chosen for the development of calibration schemes for reversed-phase HPLC. The exact compositions are given in Table IV.1b and shown in Fig 1. These compositions were chosen in such a way that reasonable capacity factors of the test solutes (see Table IV.2 and IV.3, Section 14.3) are obtained (roughly between 1 and 40) and the compositions are regularly spread in the mixture triangle, see Chapter 4. Note that two aspects of a mobile phase are incorporated in this design. First, moving from wm1 to wm3 the eluent becomes stronger, the same holds for am1 to am3 and wa1 to wa3. Second, the kind of organic modifier is changed moving from water/methanol to water/acetonitrile mixtures.

14.3 The choice of the test solutes

Test solutes should fulfil certain conditions. First, test solutes should have reasonable retention behaviour in mobile phases as chosen in Section 14.2, hence basic solutes are not very appropriate[1,2]. In fact, in some experiments at the indicated mobile phases, the solutes N-methyl-aniline and aniline were incorporated, but they were omitted because of their very bad reproducibility and repeatability.

Second, it is convenient that the test solutes are easily detectable by a UV variable wavelength detector. Third, the test solutes should encompass a range of functional groups. Fourth, in this

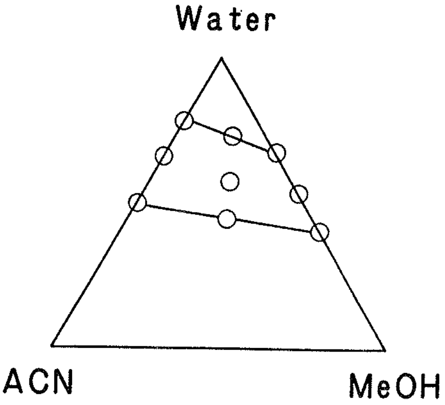


Figure IV.1. Mobile phase compositions used in the calibration of the octadecyl stationary phases. For the exact compositions, see Table IV.1.

special case the test solutes should not differ too much because the calibration strategies have been developed to work in case of a given separation problem, which is often limited to "related compounds".

In related studies (see Chapter 6 and 7) benzene derivatives were often used as test compounds. This choice was also made here, considering the above mentioned conditions. Some other solutes-steroids, phenobarbital - were incorporated to investigate their mark-power and ability to be predicted. In Table IV.2 the test solutes are shown, together with their structural formulas.

14.4 Experimental details

Methanol (MeOH) was of analytical grade and acetonitrile (ACN) was of chromatographic quality (Merck, Darmstadt, G.F.R.). Water was obtained fresh from a milli-Q water purifier (Millipore, Bedford, USA). The sixteen test solutes were obtained from various firms. The dead time was measured as the retention time of uracil and was in the range of 60 to 70 seconds. The concentrations of the injected solutes ranged from 0.02 to 0.08 mg/ml. The flow rate was 0.5 ml/min.

The six stationary phases were different batches of Chromspher Octadecyl material (Chrompack, The Netherlands), with characteristics as explained in Section 14.1. The average diameter of the silica particles was 5.0μm, the average pore volumes were 0.63±0.08 ml/g for both silica batches, the poresize was 118±10 Å and 131±10 Å for silica batches 1 and 2, respectively (Table IV.1). All columns were 100 mm long (3.0 mm i.d.).

The first series of measurements, stationary phase A, was performed with a LDC-Milton Roy Mini HPLC-pump, a Chromatronix 230 dual-wavelength detector (operated at 254 nm, except for EE (280 nm)), an injection valve (Rheodyne 7125) fitted with a 20-μl loop, and a Kipp BD40 recorder. The second series (stationary phase B) was performed with a LDC-Milton Roy Mini HPLC-pump, a Shimadzu SPD6A variable-wavelength detector (operated at 254 nm, except for EE (205 nm)), an injection valve (Rheodyne 7125) fitted with a 20-μl loop, and a Kipp BD40 recorder. The last four series (stationary phase C to F) were performed with a Waters 6000A HPLC pump, a Kratos Spectroflow 757 variable-wavelength detector (operated at 205 nm), an injection valve (Rheodyne 7010) fitted with a 10-μl loop, and an Omniscribe recorder (Houston Instruments).

The k values are reported in Table IV.3, together with the mobile phase compositions at which the measurements were made. The corresponding ln k values are reported in Table IV.6. The reproducibility and repeatability of the measurements are discussed separately in Section 14.5.

Details regarding the used software are already reported in Section 9.4.

14.5 Reproducibility and repeatability

The difference between the concepts reproducibility and repeatability is already explained in Section 9.5. On each stationary phase measurements started with the mobile phase wm2, which will be referred to as the test mobile phase composition. For stationary phase A and B this test mobile phase composition was used again in the middle and at the end of the whole series of mobile phase compositions. So three reproduced measurements, regularly spaced in the mobile phase series, of a retention time of a solute on stationary phases A and B at mobile phase composition wm2 are available. Analogous, four reproduced measurements, also regularly spaced, are available for stationary phases C to F. The standard deviation of the k (or ln k) values for each solute of these reproduced measurements is a yardstick of the reproducibility.

Contrary to the experiment described in Part III, no pattern emerged from the reproduced measurements. There was no clear drift visible in the three (or four) reproduced k values. The assumption of a constant relative error of the k values on a stationary phase can be judged with the coefficients of variation (CV). This assumption seems reasonable, except for stationary phase B. In the following calibration calculations this assumption is adopted and a logarithmic transformation of the k values is performed. The relative error in k of a solute, at constant mobile phase composition, on the six stationary phases shows variation. It is therefore questionable whether the assumption of a constant relative error in k holds also if different stationary phases are considered. For the moment, however, this assumption is adopted for the sake of convenience. Assuming a constant relative error in k, a logarithmic transformation produces ln k values with an approximately constant absolute error (see Section 9.5) which is convenient in the following model-building strategies. Summarizing, an ln k value is assumed to have a constant measurement error, not depending on the mobile phase composition, the stationary phase or the particular solute. Note that the average CV values of both stationary phase B and F are relatively high, these stationary phases have a relatively bad reproducibility.

Examining the repeatabilities (Table IV.5) a vague pattern emerges, already described in Section 9.5. Often the stronger eluent wm1 gives lower standard deviations of the repeated k values (each standard deviation is calculated from two repeated measurements) than wm3. This is in agreement with the above mentioned assumption regarding the error structure of k. Analogous reasoning holds for the water/ acetonitrile- and the ternary mixtures.

The reproducibilities and repeatabilities in ln k terms are given in Table IV.7. The s_{repeat} values are outcomes of the pooled standard deviations (actually: square root of pooled variances) over the mobile phase compositions (not only the wm2 mixture), which is allowed if the assumption of constant error variance holds. The difference between the s_{repeat} and s_{repro} is not as large as in Part III. Only for stationary phase D and F large differences can be observed. Because no clear drift was visible in the reproduced measurements, the s_{repro} values reflect day-to-day variation, which is usually higher than within-day variation, reflected by s_{repeat}. The reason for the high s_{repro} values relative to the s_{repeat} values of stationary phases D and F is not clear. Note that the bad reproducibility of stationary phase B is reflected in the s_{repeat} values of that phase, indicating one of the causes of that bad reproducibility.

14.6 Univariate description of the data set

The univariate description of the data set is divided in two parts. The first part is devoted to the variation of ln k values of a solute on a stationary phase due to changing the mobile phase composition, some representative examples are shown in Figures IV.2a to IV.2f. The second part shows the dependence of the ln k value of a solute on the stationary phases, at fixed mobile phase compositions (Figures IV.3a to IV.3d).

Examining Figures IV.2a to IV.2f, very regular patterns emerge. The dependence of the ln k of a solute on the mobile phases with the same kind modifier but of different elution strength (e.g wm1, wm2, wm3) is almost everywhere linear. These lines are the steepest for solute EE, thereby showing the strongest dependence on elution strength. Non-linear mixing behaviour is often present to some extent, e.g. comparing the ln k of PE on stationary phase B at mobile phase compositions wm2, wa2 and am2 shows that the ln k value of PE at am2 is not the mean of the ln k values at wm2 and wa2.

The differences between the stationary phases, at the constant

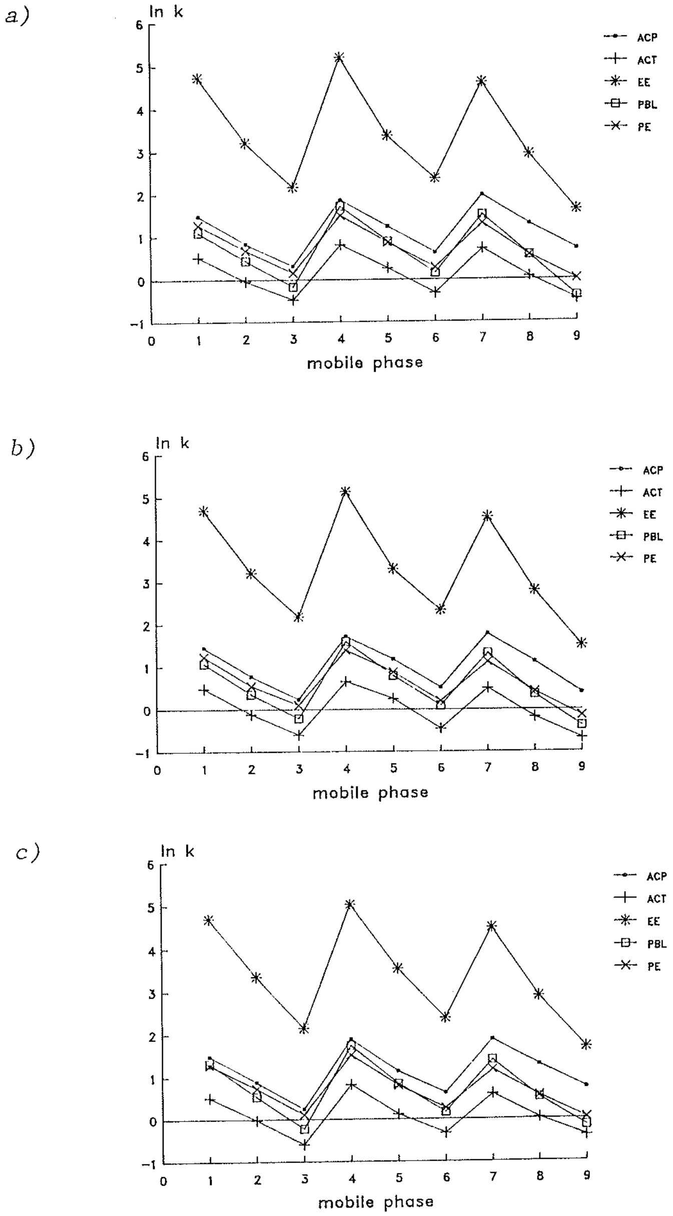


Figure IV.2. Ln k values of some of the test solutes on the different stationary phases A, B, and C. The numbers 1 to 9 correspond to the mobile phase compositions wm1, wm2, wm3, am1, am2, am3, wa1, wa2, and wa3 (see Table IV.1), respectively.

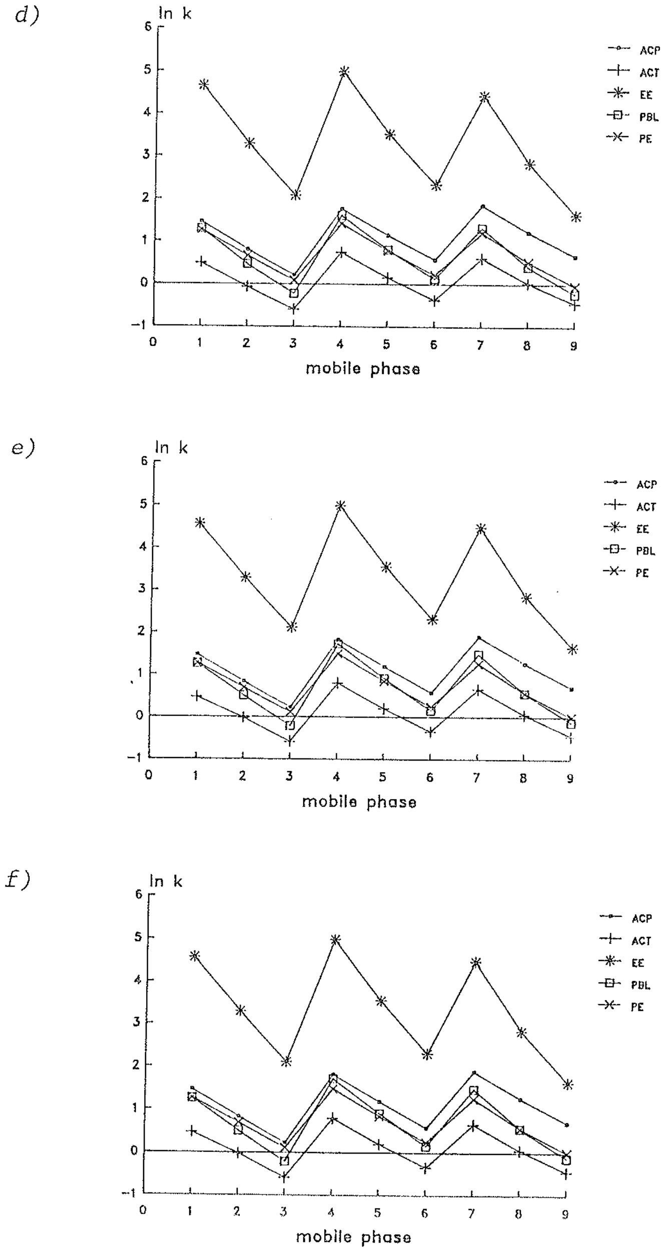


Figure IV.2 (continued). Ln k values of some of the test solutes on the different stationary phases D, E, and F. For the numbers 1 to 9 see Fig.IV.2a.

mobile phase composition wm2, are shown for each solute separately in Figures IV.3a to IV.3d. The vertical bars in the figures give the standard deviation of the associated value. It appears that stationary phases D, E, F and, to some extent, A are rather similar. Stationary phase B has generally lower ln k values than the other phases, whereas stationary phase C has higher ln k values. Although the stationary phases differ only in batch, all solutes show differences between the stationary phases.

14.7 Three-way description of the data set

Two three-way analysis methods are used to analyze the data set at hand, realizing that the data can be arranged in a data cube with three directions (modes): the solutes, the mobile phase compositions and the stationary phases. The three-way principal component analysis as described in Section 1.6 is discussed first. The data cube is unfolded in such a way that the direction of the stationary phases is left intact, thereby treating the stationary phases as objects (see Figure I.5). A datamatrix consisting of 6 objects and 9x16 (144) variables is obtained. A variable is a combination of a mobile phase composition (j=1,...,9) and a solute (k=1,...,16). Because interest focusses primarily on differences between the stationary phases, centering the data cube in the direction of the stationary phases seems appropriate. This means that, after the centering operation, the six retention values of the stationary phases for each solute/mobile phase combination, gathered in a column of the 6x144 matrix, sum up to zero. The data are not scaled because the data are measured in the same units and the influence of scaling in three-way approaches is complex. The model used to describe the data is then:

$$x_{ijk} = \text{xmean}_{mk} + \sum_{s=1}^{g} t_{is} p_{sjk} + \epsilon_{ijk} \qquad \text{(IV.1)}$$

where x_{ijk} is the ln k value of solute k at mobile phase composition j on stationary phase i, g is the number of components (latent variables) in the model, $\text{xmean}_{mk}=(\Sigma x_{ijk})/6$ (summation over i) and ϵ_{ijk} is the error with constant variance.

The results of this analysis are reported in Table IV.8. If the data cube is centered in the above described way, the mean value of all entries in one stationary phase layer (see Figure I.3), therefore averaged over all mobile phase composition/solute combinations, gives an indication of the average deviation of this stationary phase relative to a "hypothetical stationary phase" as estimated in the centering operation. The most deviating stationary phases are A and B, with the highest and lowest value for the mean, respectively. In Section 14.6 stationary phase C was found deviating, but there only the wm2 mixtures were investigated. A measure of the variation of the

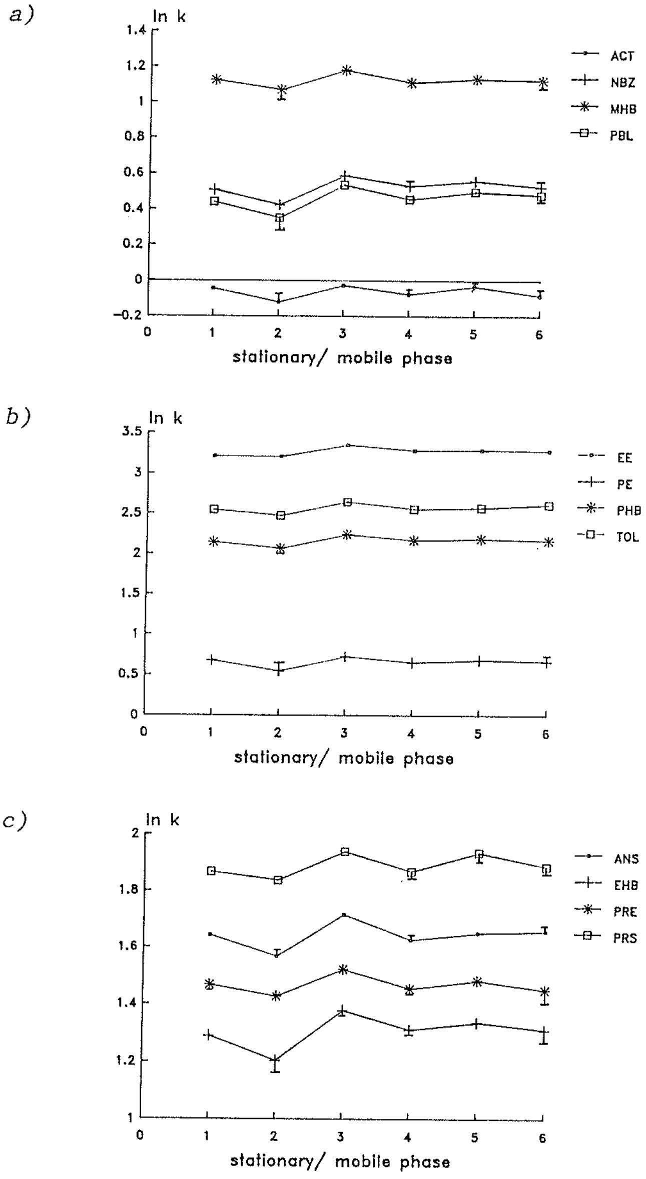


Figure IV.3. Ln k values of the solutes at mobile phase composition wm2 on the six stationary phases. The numbers 1 to 6 correspond to stationary phases A to F, respectively. The 1 sigma bars are included. If such a bar is not shown, this bar was smaller than the symbol used to indicate the measured value.

d)

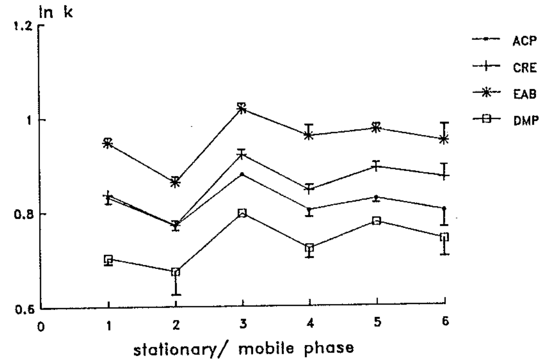


Figure IV.3 (continued).

entries in a stationary phase layer relative to the "hypothetical stationary phase" is given by MS_{bef}, the mean squared entries in the layer prior to the decomposition in components. This variation is large, of course, for the stationary phases A and B. But also stationary phase C and, to a lesser extent, stationary phase F have considerable variation. A natural yardstick to compare the MS_{bef} values with is the s^2_{repro}, as calculated for each stationary phase by pooling the appropriate individual estimates of every solute. The MS_{bef} values of stationary phases A, B and C are much higher than the corresponding s^2_{repro} values. The average signal-to-noise ratio is approximately 2 (27.87/15.46). The importance of precise measurements is illustrated: if all s^2_{repro} values are low, all stationary phases would have meaningful variation and could contribute to the decomposition. Consequently, a calibration data set with reproduced measurements at all mobile phase compositions would be advantageous and result in higher average signal-to-noise ratios than 2 because the precision of the retention values increases.

If the number of components in model (14.1) is estimated with cross-validation (see Sections 1.2 and 1.6) three significant components are found. Examining the average s^2_{repro} and the average MS_{res} after applying the first and second component (Table IV.8a), indicates that the systematic part of the data cube can be described by (at most) two components. These two components describe 86% of the variation in the data cube. The MS_{res} values for each individual stationary phase show that the variation of stationary phase B is almost completely explained by the first component. Similarly, the second component is responsible for the strong reduction of unexplained variation of stationary phase A. These phenomena are reflected in the scores on the first two components, see Table IV.8b. Stationary phase B scores high (in absolute value) on the first component and stationary phase A on the second. The scores on the first two components are depicted in Figure IV.4a. It can be observed that the stationary phases A and B are the most deviating. Note that the difference between stationary phases of the same batch but used by a different analyst/apparatus combination (A,F and B,E) is larger than the difference between batches measured by the same analyst/ apparatus combination (C,D,E,F).

The loadings of the 144 variables on the two components are shown in Figure IV.4b. High loadings (extreme values in Figure IV.4b) are observed for: ACP (wa3), ACT (am2,wa2,wa3), EAB (wa3), PHB (wa3), EE (wm1,wm2,wa1,am1,am2), PBL (wm1,wa1,wa2,wa3), PRE (wm1,wa1,am2), and PRS (wm1,wa1,wa3). A decision with regard to solute/mobile phase combinations most representative to the structure of the data cube is not easy to make. As markers the solutes ACT, EE, PBL and either PRE or PRS can be chosen. The mobile phase compositions are more difficult to select, wa3 would be part of this set in any case. The problem is that a combination of a solute and a mobile phase composition has to represent the variation in the data in order to be a suitable marker/mobile phase combination. But a high loading of a variable can be caused not only by representativeness but also by outlier characteristics. Therefore, a more quantitative method is needed to select promising combinations and validating this selection.

Tables IV.8c and IV.8d give an impression of the modelling power of the components with regard to the solutes and the mobile phase compositions, respectively. The MS_{bef} values are calculated within the appropriate layer of the data cube. Especially EE and PBL are hard to model (high MS_{res} values), despite the high loadings. This might be due to their high MS_{bef} values. The MS_{bef} values of the solutes reflect variations due to the mobile phase compositions, the stationary phases and specific combinations of those. Figures IV.3a to IV.3d do not reveal extreme dependence of EE and PBL on the stationary phases. For the solute EE this high MS_{bef} is probably due to its sensitiveness towards elution strength (see Section 14.6). Both solutes EE and PBL show relatively strong non-linear mixing behaviour (see Figures IV.2a to IV.2f), which also contributes to MS_{bef}. The ln k values of both EE and PBL are relatively low for the water/acetonitrile mobile phases, especially the wa3 mixture. Note that the influence of a specific mobile phase/stationary phase combination on the k values of EE and PBL cannot be inferred easily by Figures IV.2 and IV.3. The retention values at the mobile phases am1 and, to a lesser extent, wm1 are not explained very well by a principal component model with two components. The retention at the

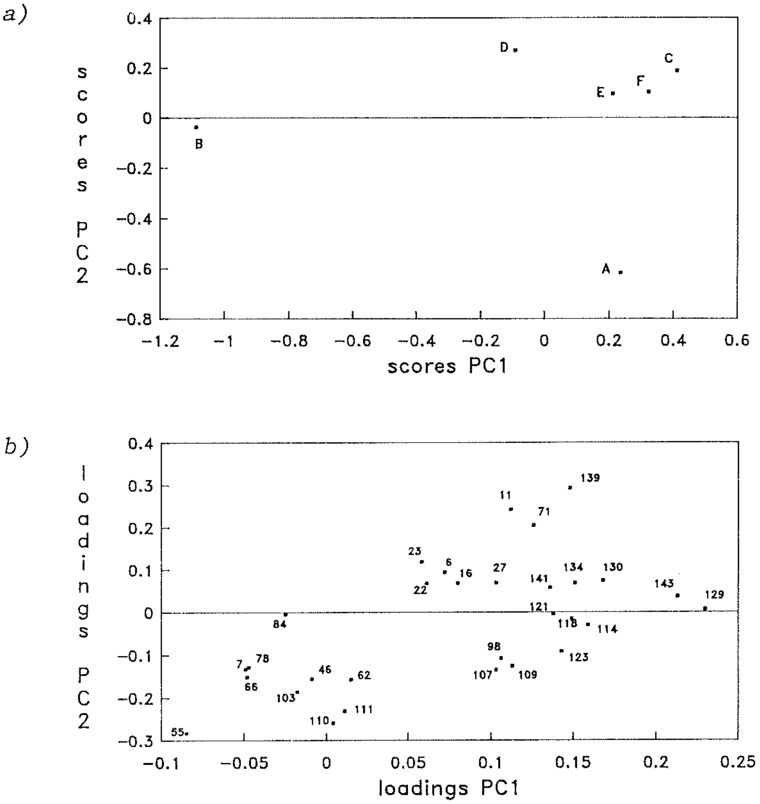


Figure IV.4. Score (a) and loading plot (b) of a three-way PCA on the data cube with the six stationary phases as objects. Legend: a) the capitals A to F refer to the stationary phases in Table IV.1; b) 6=EAB at wm1, 7=EE at wm1, 11=PBL at wm1, 16=TOL at wm1, 22=EAB at wm2, 23=EE at wm2, 27=PBL at wm2, 46=PRE at wm3, 55=EE at am1, 62=PRE at am1, 66=ACT at am2, 71=EE at am2, 78=PRE at am2, 84=CRE at am3, 98=ACT at wa1, 103=EE at wa1, 107=PBL at wa1, 109=PHB at wa1, 110=PRE at wa1, 111=PRS at wa1, 114=ACT at wa2, 118=EAB at wa2, 121=MHB at wa2, 123=PBL at wa2, 129=ACP at wa3, 130=ACT at wa3, 134=EAB at wa3, 139=PBL at wa3, 141=PHB at wa3, 143=PRS at wa3, all other variable combinations are not shown because the loadings were far less extreme.

water/acetonitrile mixtures are modelled well due to their high loadings (see above). Examining the individual residuals (not given) associated with the six stationary phases and mobile phase aml, reveals that the residuals of stationary phases A and B are small for every solute. PBL has always large residuals on the other four stationary phases and EE two times. Yet other solutes, e.g. ANS, EAB, PE and PRE have also large residuals. It is not clear whether the observed non-linear mixing behaviour, especially in the case of PBL and EE, is responsible for the high MS_{res} values of aml and wml.

The second three-way analysis is done with the PARAFAC method. The corresponding model is (adopting again the notation x_{ijk}) :

$$x_{ijk} = xmean_{jk} + \sum_{s=1}^{g} a_{is}b_{js}c_{ks} + \epsilon_{ijk} \qquad (IV.2)$$

where g, $xmean_{jk}$ and ϵ_{ijk} have the same meaning as in (IV.1). The results are reported in Table IV.9.

The reduction in unexplained variation due to the first two components (the difference between the MS_{bef} and $MS_{res}(2)$ columns in Table IV.9) is more evenly spread over the stationary phases than in case of the unfold principal components analysis. The variation associated with stationary phases A and B is not completely absorbed in the second and first PARAFAC component, respectively, as was the case in the unfold solution. The total explained variation is 72% with the PARAFAC model, slightly less than the unfold solution. But, as stated earlier (Section 1.6), the percentage of explained variation, or the size of the residuals, is not the ultimate yardstick to compare the performance of three-way decompositions with. The scores of the stationary phases on the PARAFAC components cannot be compared directly with the scores on the principal components because of scale arbitrariness of both the principal component and the PARAFAC solution (if all t_{is} values in (IV.1) are multiplied by a constant and all t_{sjk} are divided by that same constant an equivalent solution is obtained, the same holds for the PARAFAC decomposition). A score plot of the PARAFAC solution is shown in Figure IV.5. Contrary to the scores in Figure IV.4a (the unfold solution), these scores are not uncorrelated. Stationary phases B and C are the most deviating in the PARAFAC solution.

The reduction in unexplained variation of retention values of the solutes are given in Table IV.9c. For the mobile phase compositions, analogous values are given in Table IV.9d. After applying two components in the PARAFAC model, the retention of solutes EE and PBL remains largely unexplained. This effect was also observed in the unfold solution, but less dramatically. The same causes as mentioned above with respect to this phenomenon may apply for the PARAFAC solution. If the loadings of the solutes PBL and EE on both PARAFAC components are examined (see Table IV.9e), PBL appears to load high on these components contrary to the solute EE. Yet these high loadings of PBL do not model the retention of the solute PBL properly, but better than the retention of solute EE. The retention values at mobile phase compositions wml, aml and am2 still have high unex

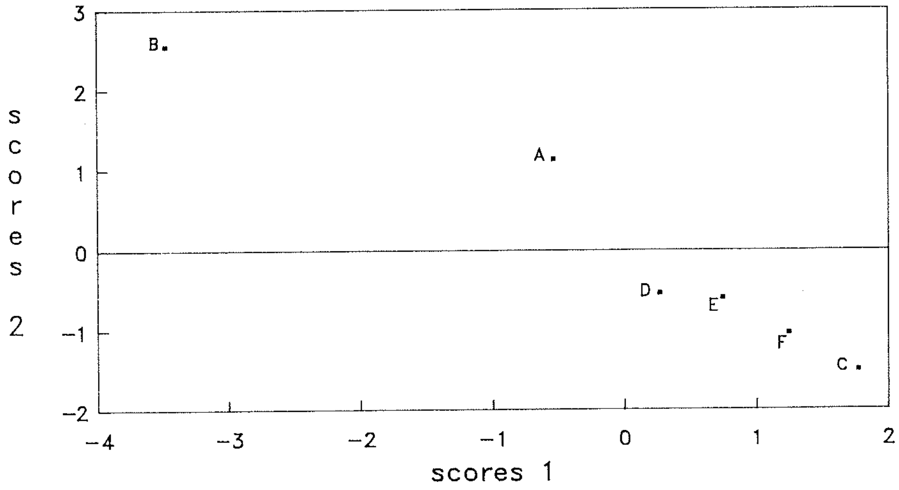


Figure IV.5. Score plot of the PARAFAC model, the capitals A to F refer to the stationary phases, see Table IV.1.

plained variation after applying the PARAFAC model (Table IV.9d). Reference is given to the discussion above regarding the modelling of retention at the mobile phase compositions with the unfold decomposition. Both mobile phase compositions aml and wal load high on both PARAFAC components (see Table IV.9f). Retention at mobile phase composition wal is modelled well, whereas retention at aml is not. Note that it is easier to select solute/mobile phase combinations which represent the systematic variation in the data cube with the use of the PARAFAC loadings than the unfolding ones. From separate loading plots - one loading plot of the solutes (Table IV.9e) and one of the mobile phases (Table IV.9f) - markers and mobile phases can be selected. Moreover, the PARAFAC decomposition is rotation invariant[3] and the rotation dependence hampers the selection of variables on the basis of loadings from a PCA (see Section 1.2). From Table IV.9e it can be inferred that ACP, ACT, PBL and PRS are suitable markers, whereas wm3, aml, wal and wa2 seem suitable mobile phase compositions. A good validation procedure is needed to evaluate this choice quantitatively.

It is hard to derive conclusions with regard to the potential of both methods (unfold and PARAFAC) to decompose the data cube. An important difference between PARAFAC and unfold-PCA is the number of degrees of freedom. In applying PARAFAC with two components 206 (16x9+2x16+2x9+2x6) parameters have to be estimated, whereas this number is 444 (16x9+16x9+16x9+2x6) in case of unfold-PCA with two components. Therefore, a considerable gain in degrees of freedom is obtained if PARAFAC is used. It is claimed[4] that the PARAFAC decomposition is more restrictive then the unfolding decomposition, which seems to be a reasonable claim considering equations (IV.1) and (IV.2). This difference in restrictiveness has two different consequences. The unfold decomposition is more flexible than PARAFAC, but leaves the principal components the freedom to explain retention on stationary phases A and B completely. This is only reasonable if these stationary phases are really extremely informative, otherwise outlying behaviour is modelled. The PARAFAC approach is more restricted, explaining less of the total variation than the unfold solution, but the stationary phases contribute more regularly to the components. Besides, in case of the PARAFAC decomposition a choice has to be made whether the components are orthogonal (which is the same as uncorrelated for the mode where centering is performed) or not. This is a model assumption which has to be made on the basis of the knowledge of the process generating the data. With the unfold-PCA method the scores are always orthogonal, so there is no choice. In the PARAFAC decomposition as applied above no orthogonality was assumed, because there was no clear reason to do this and the demand of orthogonality restricts the decomposition.

Chapter 15 Two-way approach

15.1 The choice of the training- and test set

In order to illustrate a calibration strategy as described in Subsection 8.1.2, three stationary phase are chosen. Two of these stationary phases will constitute the training set, whereas the remaining one will be the test set. Based on the scores on the first two components in both the unfold and PARAFAC solutions (Section 14.7), the stationary phases A, B and C are a reasonable choice, because these stationary phases represent the most deviating ones. Referring to Table IV.1, it is clear that these three stationary phases comprise different analyst/apparatus combinations and do not incorporate duplicate stationary phase materials.

Two analysis of variance (ANOVA) calculations are performed to illustrate the differences between the stationary phases. The first ANOVA is a one-way set up. The (only) factor is the stationary phase (varied at three levels), with three reproduced measurements for stationary phases A and B, and four reproduced measurements for stationary phase C, at mobile phase composition wm2. This ANOVA is performed for each solute separately. With the use of Scheffé confidence intervals, differences between the stationary phases can be visualized for each solute. The results are given in Table IV.10a. The level of significance is 5%. If a solute is reported in this table, the null hypothesis that the stationary phases do not differ with respect to the ln k values of that solute is rejected. Stated otherwise, the variation because of changing stationary phases is significant larger than the reproducibility of that solute. Two conclusions can be drawn from Table IV.10a. First, stationary phase C is the most deviating from the three. Second, the differences between the stationary phases depend on the solutes.

The second ANOVA is performed with two factors: the stationary phases (varied at three levels) and the amount of organic modifier in the mobile phase (varied at three levels). This two-way ANOVA is performed for each type of mobile phase: once with the binary water/methanol mixtures, once with the binary water/acetonitrile mixtures and once with the ternary mixtures as second factor. Again the three ANOVA's are performed for each solute separately. The results are reported in Table IV.10b. If a solute is shown in Table IV.10b, this means that the null hypothesis "no difference between the stationary phases" is rejected at a significance level of 5%. By comparing the results of the two-way ANOVA performed with the different types of mobile phases, it is clear that the differences between the stationary phases are not only dependent on the solutes but also on the type of the mobile phase. The differences between the stationary phases if water/methanol mixtures are considered, only show up for p-cresol (CRE). The differences between the stationary phases are much more pronounced in water/acetonitrile mixtures, which is in agreement with the remarks made in Sections 14.6 and 14.7. This stresses the notion that the measured differences between the stationary phases depend on the solutes, the kind of mobile phase and combinations of these factors.

The whole calibration procedure is illustrated with the use of stationary phase C in the test set, because retention on this stationary phase is expected to be the most difficult to predict (see the ANOVA results).

15.2 Selection of the markers.

The data set from which the markers are selected can be arranged according to Figure II.8, Subsection 8.1.2. The data are column-mean centered. In this instance the subject of scaling is not investigated. No scaling is performed because all ln k values are measured in the same units. The markers are selected with the induced-variance criterion (see Section 1.3). In Section 11.5 the determinant criterion was found to be a reasonable alternative, but the induced-variance criterion fits theoretically better the purpose of prediction.

Two versions of the calibration procedure are discussed. Version one (labelled as I) uses all mobile phase compositions in the training set, whereas version two (labelled as II) only uses the extreme mobile phase compositions i.e. wm1, wm3, am1, am3, wa1 and wa3. Version I thus makes use of 9x2 objects to chose the markers and to calculate the model with, whereas version II only uses 2x6 objects. The marker choice in case of version I is discussed first.

The four variables ANS, DMP, EE and PRE explain the highest percentage of variation in X (P=99.833%), and are therefore chosen as markers (Table IV.11a). Two of these solutes - ANS and PRE - were already expected to describe differences between the stationary phases A and B (see Table IV.10a).

The second and third best subsets are DMP, EE, PRE, TOL and ANS, DMP, EE, PRS. These subsets explain respectively 99.830% and 99.828%, illustrating the exchangeability of the three subsets of markers. The exchange of ANS and TOL between subsets one and two can be explained, as they belong to the same selectivity group (VII) as proposed by Snyder[5], and TOL is also sensitive towards differences between A and B. The solutes PRE and PRS are chemically related and can be exchanged.

Note that the markers represent moderate and slow eluting compounds, and not the fast eluting ones. This may cause problems if fast eluting compounds are to be predicted.

In case of version II, the marker choice is presented in Table IV.11b. The data matrix from which the markers are chosen is a 12x16 matrix with column mean-centered entries and is again not scaled. The resulting marker sets resemble the ones from version I, except for the presence of the solute PBL instead of the solute EE.

Whether this exchange of PBL and EE is related to the particular behaviour of these solutes as indicated in Section 14.7, is intuitively reasonable but difficult to prove.

15.3 Results of the predictions on the new stationary phase

The results of the predictions with version I are discussed firstly. The solutes ANS, DMP, EE and PRE are selected as markers. The PLS method is used to model the relationship between the behaviour of the markers and non-markers. The ln k values are ordered as indicated in Fig II.9, Subsection 8.1.2. All columns in X[M] and X[NM] are mean-centered, no scaling is performed.

With two dimensions in the PLS model, 99.5% of the variation in X[M] is used to explain 99.2% of the variation in X[NM]. These two dimensions are considered sufficient to reflect the relation between X[M] and X[NM].

The final step in the prediction procedure is the calibration of the new stationary phase C with the markers. As the measurements of the markers are available on stationary phase C at the same mobile phase compositions as in the training set, predictions on the new stationary phase are performed at these nine mobile phase compositions. It is not necessary to predict at the same mobile phase compositions as in the training set, but it is convenient with the data set at hand. The ln k values of the markers are used to predict the ln k values of the non-markers, at the nine mobile phase compositions, on the new stationary phase (see Figure II.10, Subsection 8.1.2).

The results of the calibration of the new stationary phase are shown in Table IV.12. The root mean squared error of prediction (RMSEP I) values are calculated with the use of predictions at all nine mobile phase compositions and can be compared with the s_{repro} values of the solutes on stationary phase C. The average RMSEP value is three times the average s_{repro} value, indicating some systematic prediction error. The percentage of explained variation in the test set ranged from 98.56% for TOL to 99.69 for EAB and PHB. The predictions can be considered good except for MHB, PBL, PRS and TOL. To get an impression of the performance of the calibration, the observed versus the predicted capacity factors of ACP and TOL are given (Table IV.12c, version I). The solutes represent, respectively, the best and worst calibration.

The relatively bad predictions of PBL and MHB are discussed firstly. This was expected: both compounds are fast eluting. Yet other fast eluting compounds - ACP, CRE, EAB, PE - are predicted well. Closer examination shows that PBL is worse predicted at the ternary mixtures. This could be expected because the two-way ANOVA showed differences between the measurements of PBL on the stationary phases with respect to the ternary mixtures, which is related to the non-linear mixing behaviour of PBL as discussed in Section 14.7. The conclusion is that the selective differences shown by PBL in ternary mixtures are not completely represented by the markers. The variable MHB is worse predicted at the aml and wa3 mixtures. This may be due to an analogous cause as in case of PBL, see Table IV.10b. For both mobile phase systems (water/acetonitrile and water/methanol/acetonitrile), MHB shows differences between stationary phases A,B and B,C as calculated with Scheffé intervals (with a 95% confidence level). The retention values of other solutes, e.g. EHB and PHB, show also differences between the stationary phases: for both water/acetonitrile and the ternary mixtures differences occur between stationary phases A,B and B,C. But the retention of these solutes is predicted well, so extrapolation beyond the "markers-scale" is probably one of the reasons for relatively bad predictions of MHB.

Two slow eluting markers are incorporated but nevertheless PRS and TOL are badly predicted. In a related study[6], relatively bad predictions were also observed for slowly eluting compounds. It has been suggested in that study, that this might be caused by the inherent increase in relative error associated with the measurement of long retention times. But the s_{repro} values of both PRS and TOL on stationary phase C is 0.018, so large measurement errors are not present. Because the solutes TOL and PRS are contained in the second and third best marker subsets, respectively, the suggestion is that these solutes are sensitive to the differences between the stationary/mobile phase combinations in the training set. This makes these solutes relatively difficult to predict. The solute TOL is known to be sensitive to stationary phase differences[7].

Table IV.12b gives an idea of the predictive performance at the different mobile phase compositions. Especially predictions at mobile phase aml seem troublesome. The variables MHB, PBL and PRS contribute most to the prediction errors at aml. The prediction of the retention of these solutes is already discussed.

The results of the version II calibration procedure are discussed briefly. The solutes ANS, DMP, PBL and PRE are selected as markers. The training set consists of 6x2 objects and the data are again arranged in a X[M] and a X[NM] matrix (column-mean-centered). The PLS method was used to model the relationship between X[M] and X[NM]. Two PLS components used 99.71% of the variation in X[M] to explain 99.62% of the variation in X[NM], these two components are regarded sufficient. Note that the gain of the version II procedure is in the smaller training set. In the prediction stage the markers have to be measured at the mobile phase compositions where predictions are needed, similar to the version I procedure. Predictions are made at the same mobile phase compositions as in case of version I. The root mean squared errors of prediction are given in Table IV.12a, version II. The average RMSEP value is slightly higher than the version I analogon. The predictions of the variables TOL and PRS are again troublesome as with version I. EE, a marker in case of version I, is predicted badly, especially at the ternary mobile phase compositions.

Reference is give to the discussion on the behaviour of EE in Section 14.7.

The consequences of not incorporating the wm2, am2 and wa2 mixtures in the training set can be evaluated with the use of the RMSEP values of the individual mobile phase compositions (Table IV.12b). For the wm mixtures holds that the predictions at wm2 and wm3 are slightly worse in relation to the version I case. At the aml mixture the predictions are bad, the high RMSEP value of aml is largely determined by a large residual of EE (if this residual is omitted, the RMSEP value of aml is 0.053). A slight worsening of the predictions at am2 and wa2 is observed. In order to get an impression of the quality of the version II procedure, the predicted versus observed k values of ACP and TOL are incorporated in Table IV.12c.

The question can be asked whether the above described prediction procedures really account for the variation in ln k values due to stationary phase differences. Stated otherwise, do these procedures merely model the variation due to the mobile phase influence or do they explain more? It is difficult to answer this question in depth but a good indication can be obtained by examining the following prediction procedure. Let the two stationary phases in the training set be represented by two layers in the data cube (see Figure II.7, Subsection 8.1.2). Then predictions at a third layer can be obtained straightforwardly by calculating the averages of the two ln k values of each solute/mobile phase combination. The ln k value of a specific solute/mobile phase combination on the new stationary phase can be predicted by the average of the two corresponding values of the training set stationary phases. Because no preliminary knowledge of the new stationary phase is available prior to the predictions on that stationary phase, this procedure is called the "blind" procedure. Note that the specific properties of the new stationary phase do not play a role in this blind procedure, whereas they do in the version I and II calibration procedures through the ln k values of the markers. Only the mobile phase variation is predicted with the blind procedure.

The root mean squared prediction errors (not given) of the blind procedure range from 0.061 (for NBZ) to 0.126 (for PBL) with an average of 0.086. The difference between the average RMSEP value of the blind procedure and the corresponding value of version I is roughly twice the average s_{repro} value of stationary phase C and is therefore an important difference. This means that information of the specific properties of the new stationary phase is valuable. As stated earlier in Part III, a small difference in predictive performance can have large consequences for a predicted chromatogram. Predicted versus observed values of ACP and TOL as the outcome of the blind procedure are incorporated in Table IV.12c.

Chapter 16 Three-way approaches

16.1 Selection of the markers and mobile phase compositions

The issue of the choice of the marker selection criterion is not pursued. In the two-way approaches, the criterion which is used to select the markers has to consider the multicollinearity in the resulting model. This is not the case in the three-way approaches. The influence of the markers and, therefore, of the marker selection criterion on the performance of three-way calibration has to be established still. The induced-variance criterion is adopted due to its theoretical advantage if prediction is at hand. A method is used that allows the selection of solute/mobile phase combinations representative for the structure in the data set. Each stationary phase is omitted subsequently.

Suppose the first stationary phase is completely omitted from the data cube. The remaining part of the data cube is depicted in Figure II.12, Subsection 8.1.2 (no scaling has taken place yet). First, the data cube is unfolded in such a way that the direction of the solutes is left intact and the solutes are treated as variables. This results in a (9x5)x16 data matrix. After column mean-centering of the data set thus obtained, the solutes are selected in the regular way from the centered matrix with the use of the induced-variance criterion, as explained in Section 1.3. The outcome of these calculations are given in Table IV.13a under the heading: Stationary phase A left out. The markers ANS, DMP, EE and PRS are selected and explain 99.852% of the variation in the unfolded data matrix. This outcome is in agreement with the marker selection outcome given in Section 15.2.

If this whole procedure is repeated for each subsequently omitted stationary phase, the results are as shown in Table IV.13a. The best and second best marker sets are always the same, indicating a stable choice of the markers.

An analogous procedure can be followed for the selection of the mobile phases. The first stationary phase is again omitted and the data cube is unfolded so that the direction of the mobile phases is left intact and the mobile phases are treated as variables. This results in a (5x16)x9 data matrix. Note that all solutes are kept in the data cube, not only the previously selected markers. After column mean-centering the data matrix, the mobile phases are selected that explain most of the variation at the non-selected mobile phases (induced-variance criterion). The outcome of this selection is shown in Table IV.13b, where the results are shown of the selection procedure where all stationary phases are omitted subsequently. The best choice is the mixtures wm1, wm3, wa1 and wa3, which is very appealing because those mobile phase combinations are the most extreme ones in the mixture space, see Section 14.3. The second and third best choices are always wm3, am1, wa1, wa3 and wm3, am1, wa2 and wa3, respectively. The appearance of am1 in the second and third best subsets is in agreement with the results obtained in Section 14.7 and Section 15.3. In the three-way decompositions (Section 14.7) the residuals for aml were large. The predictions at mobile phase composition aml were troublesome as indicated in Section 15.3.

The calibration procedures are performed with the solutes anisole, dimethylphtalate, ethynylestradiole and prednisolone as markers and the (binary) mobile phase compositions wml, wm3, wal and wa3 as calibration mobile phases.

16.2 Results of the calibration with unfold-PLS

In Subsection 8.2.1 a description of the unfold approach was already given (see Figures II.13 and II.14, Subsection 8.2.1). This approach will be illustrated with the data set at hand. Each stationary phase is omitted once and retention on that stationary phase is predicted thereafter with the use of PLS on the unfolded data cube. The marker/mobile phase combinations used for the calibration of the new stationary phase are ANS, DMP, EE and PRS at wml, wm3, wal and wa3 (discussed in Section 16.1). Prior to the PLS calculations, the data cubes **X[MaMPh]** and **X[NMaMPh]** are unfolded. The matrices X[MaMPh] and X[NMaMPh] are column-mean-centered, because interest focusses on the differences between the stationary phases. Note that the rank of these column-centered matrices is only four, so the number of PLS dimensions in the model is very restricted.

First the results are presented if only one component is used in the unfold-PLS model. An impression of the adequacy of the use of only one component can be obtained by comparing the average s_{repro} of the measurements in the training set with the standard deviations of the residuals in the X[MaMPh] and X[NMaMPh] block, after applying one component (not given). If each stationary phase is omitted subsequently, six different training sets are obtained. Each training set is used to build the model with and to predict retention on the omitted stationary phase. Four of these training sets contain both stationary phase A and B. Only these four training sets show that one component is not sufficient if the above mentioned comparison is made. This indicates the importance of both stationary phases in the modelling process.

The results of the unfold-PLS calibration with one component are presented in Table IV.14, in the same way as in Chapter 13. For a detailed discussion on the interpretation of Table IV.14, reference is given to Section 13.3. One of the first striking conclusions is the bad prediction of the retention values on stationary phases A and B (high MS_{res} values). Application of the model is (almost) useless. For stationary phase B this could already be inferred from the low percentage of explained variation in the X[NmaMPh] block: the model is not very good. Moreover, the percentage of variation of the markers at the selected mobile phase compositions used for calculating the score of stationary phase A and B on the PLS component is only 0.2% and 0.6%, respectively. Despite the fact that the variation of the markers, as measured with MS_{mark}, does represent systematic variation, this variation is hardly used. The bad predictions on stationary phase A and B may be due to the low number of PLS components in the model. An evaluation of these bad predictions is postponed to the discussion on results of two components unfold-PLS.

With regard to the predictions on the other stationary phases the following can be said. The predictions on stationary phase C show lack-of-fit: reduction of the unexplained variation is obtained, but not enough (the unexplained variance is still higher than s^2_{repro}). The root mean squared prediction error on stationary phase C is 0.047 (the square root of 21.98×10^{-4}), which is lower than the corresponding values in the two-way calibration procedures presented in Section 15.3, respectively 0.051 and 0.060 for the versions I and version II calibrations (see Table IV.12a). Note that for the calibration of stationary phase C in Section 15.3 the retention of four markers at nine mobile phase compositions had to be measured, whereas in the unfold-PLS case only the retention of four markers at four mobile phase compositions must be available. The pay-off is clear: at the cost of more observations in the training set less measurements are needed to calibrate a new stationary phase with. On stationary phase D, the retention values show hardly systematic variation to explain (compare 12.33 with 11.56), so conclusions cannot be drawn. Retention on stationary phase E is explained well with the one component model, the MS_{res} is only slightly higher than the s^2_{repro}. Systematic variation which has to be explained on stationary phase F is not present (19.34 is much lower than 38.44). Conclusions cannot be drawn, therefore.

The diagnostic value of %MARK is reasonable, except for stationary phase F. The reason of this failure is in the large difference between MS_{mark} and MS_{toex} for this stationary phase. The MS_{mark} value of stationary phase F is slightly higher than the s^2_{repro} of this stationary phase, whereas the MS_{toex} value is much lower than s^2_{repro}. Note that, in case of the other stationary phases, MS_{mark} is a reasonable estimator of MS_{toex}.

The root mean squared error of prediction (RMSEP) values of the individual mobile phase compositions are given in Table IV.14b. Note that the RMSEP values for the mobile phases wm1, wm3, wa1, and wa3 are averaged over the non-markers, whereas the corresponding values of the other mobile phases are averaged over all solutes because the retention of the markers are also predicted at the mobile phase compositions wm2, am1, am2, am3, and wa2. Focussing on the predictions on stationary phases A and B, the most difficult mobile phase compositions to predict at are the water/acetonitrile mixtures, this holds especially for stationary phase B.

The results of the calibration of the six stationary phases if two components in the unfold-PLS model are applied, are presented in Table IV.15. For all six training sets hold that the average s_{repro}

is always higher than the standard deviation of the residuals in the X[MaMPh] and X[NMaMPh] blocks (not given). The incorporation of a third PLS component is, therefore, not appropriate.

The percentage of explained variation in X[NMaMPh] if stationary phase B is omitted, is still low. The %MARK is again a reasonable estimator of the percentage of variation explained in the test set except for stationary phase F, which is already discussed. For stationary phase A, %MARK becomes lower than in the one component case, which is reflected in the worse predictions of the two component model on stationary phase A. On average the predictions with the two component unfold-PLS model are slightly better than with the one component model.

The predictions on stationary phase D are improved, especially at mobile phases wm3, am1, am2, wa1, and wa2 (see Table IV.14b and Table IV.15b) by using one extra component in the unfold-PLS model, but this improvement is not significant compared with the s_{repro} value of stationary phase D. Moreover, since the MS_{res} of stationary phase D becomes lower than s^2_{repro}, the danger of overfitting is present. The predictions on stationary phases E and F do not profit from the extra component: stationary phase C still shows lack-of-fit.

The predictions on stationary phase B become slightly better than in the one component model, but are still not good. High prediction errors are observed for stationary phase B at the water/acetonitrile mixtures wa2 and wa3 (see Table IV.15b). This is related to the conclusion in Section 15.1 (see Table IV.10b) where the water/acetonitrile mixtures were found to show differences between the stationary phases A, B, and C with regard to the retention of 12 solutes. Focussing on the RMSEP values of the individual solutes (see Table IV.15c) on stationary phase B, it appears that the solutes ACP, ACT, and MHB show high RMSEP values. All these solutes were present in the ANOVA tables (Table IV.10) of Section 15.1 and the three-way PCA performed on all stationary phases, discussed in Section 14.7, showed large loadings on the two principal components for the solutes ACP and ACT in combination with the water/acetonitrile mixtures. Closer examination of the prediction errors of these solutes on stationary phase B (Table IV.16), reveals that these prediction errors are systematic. Nearly all errors have a negative sign and are especially high at the water/acetonitrile mixtures.

An explanation of the bad predictions on stationary phase B may be given by the following consideration. Suppose that a systematic difference between the measurements on the stationary phase B and the other stationary phases has been introduced by the analyst/apparatus combination which was not used in the training set. This systematic error may result in a constant absolute difference between the measurements on stationary phase B and the corresponding average values of the other five stationary phases. It is questionable whether unfold-PLS, in the way applied here, can handle this situation. Incorporation of variables in the model, that describe and handle the above mentioned systematic differences, can perhaps improve the predictions. The bad predictions of retention on stationary phase A are not discussed explicitly, but notice that retention values on this stationary phase are also measured by another analyst/apparatus combination as the training set values: reference is given to the discussion on the bad predictions on stationary phase B.

16.3 Results of the calibration with the PARAFAC model

The calibration procedure with the use of the PARAFAC model was explained in Subsection 8.2.2 and an example has already been given in Section 13.4. The same set up of the six calibrations is chosen as in the unfold-PLS case, in the previous chapter, only now the PARAFAC model is applied to describe the systematic variation in the training set. The same markers and selected mobile phase compositions and the same type of centering as in the unfold-PLS case are used.

After applying one component in the PARAFAC model, the standard deviation of the residuals (not given) was always lower than the average s_{repro}. A one component model would be the most appropriate, but for the sake of comparison with the unfold-PLS calibration, a two component PARAFAC model was also calculated. The results of the calibration with the one component PARAFAC model is presented firstly.

Table IV.17a is partly a replicate of Table IV.14a; for the sake of convenience the values s^2_{repro}, MS_{mark}, and MS_{toex} are repeated. The meaning of %MARK and MS_{res} is already explained (Section 13.4) and $100xR^2_{train}$ is the percentage of variation explained by the one PARAFAC component in the training set, an analogous value to %X[MaMPh] and %X[NMaMPh] in the unfold-PLS case.

If the results of the one component PARAFAC calibration are compared with the results of the one component unfold-PLS calibration, differences are noticable. Calibration of stationary phases A and D is better with the PARAFAC model as shown by the appropriate MS_{res} values, also indicated by the %MARK value. The stationary phase E is calibrated worse with PARAFAC, also indicated by %MARK. All the other stationary phases are predicted with errors in the same order of magnitude as the unfold-PLS one component case. Note that the %MARK for stationary phase F is a more realistic value in the PARAFAC case than in the unfold-PLS case.

The gain for stationary phases A and D in calibrating with PARAFAC is obtained at the ternary mixtures (am1-am3) and the binary mixtures wa1 and wa2. The connection between this observation and the conceptual difference between the PARAFAC and unfold-PLS (or unfold-PCA), as discussed in Section 13.1, is subject of further research. The loss in calibrating with PARAFAC for stationary phase E is not attributable to a specific mobile phase composition. Note that, although the same RMSEP values are reported for the wm2 mixture on stationary phase D in Table IV.14b and Table IV.17b, the percentages of explained variation differ. This is due to rounding errors in the RMSEP value and a low mean sum of squared variation to explain for the wm2 mixture making the percentage of explained variation calculation very sensitive to small differences in unexplained variation.

Comparison of Table IV.17c with Table IV.14c shows that the loss, respectively gain, in calibrating with the PARAFAC model instead of unfold-PLS for stationary phases E, respectively stationary phase D, cannot be attributed to specific solutes.

The results of the two component PARAFAC model are presented in Table IV.18. Stationary phase B and D profit from the addition of the second component in the PARAFAC model, also indicated by %MARK. However, retention on stationary phase B is still predicted with lack-of-fit and stationary phase D has a MS_{res} considerably lower than s^2_{repro} indicating that the two component PARAFAC model overfits with respect to stationary phase D. The predictions on stationary phase A become worse by applying the extra component, but this is not indicated by %MARK. For stationary phase E holds that the inclusion of an extra PARAFAC component does not correct for the lack-of-fit with which retention on this stationary phase is predicted with the one component PARAFAC model. Calibration on stationary phases C and F is performed approximately with the same prediction error using the one- or two component PARAFAC model.

The deterioration of the predictions on stationary phase A, if an extra component is used in the PARAFAC model, is especially present at the wm1, wm2, am1, and am2 mixtures. All solutes (except EE) are predicted worse with two components on stationary phase A.

In comparing the performance of the one- and two component PARAFAC models with respect to the calibration of stationary phase B, no pattern is visible that indicates which mobile phase composition profits most (see Tables IV.18b and IV.17b) from the extra component. The variation in retention values on stationary phase B at the water/acetonitrile mixtures remains hard to explain. If an analogous comparison is made with respect to the solutes, it is striking that all solutes profit from the extra component except the steroids EE, PRS, and PRE. In Section 14.7 it was shown that these solutes load high on the unfold-PCA components. Moreover, EE could not be explained by the PARAFAC model on the data cube consisting of all stationary phases. The question rises whether the data set at hand is homogeneous enough to use the three-way models, perhaps the steroids should be discarded.

The results of the PARAFAC two component model and the unfold-PLS two components model in calibrating the stationary phases can be compared with the use of Tables IV.15 and IV.18. Significant differences can be observed in the calibration of stationary phases B and E. In case of stationary phase E the unfold-PLS model with two components predicts without lack-of-fit, contrary to the two component PARAFAC model. Whether such differences may be explained by the already discussed conceptual differences between the PARAFAC model and the unfold-PLS (or unfold-PCA) model is subject of further investigation. Diagnostic tools to check the assumption of bilinearity in mode B and mode C, as made by the PARAFAC model, are needed.

Stationary phase B is calibrated much better with the use of PARAFAC than unfold-PLS, considering the two components models, but still the predictions are not good. Examining Table IV.18c it appears that, as was the case with the unfold-PLS calibration, some of the solutes are responsible for the bad predictions, especially at the water/acetonitrile mixtures. Table IV.19 shows that the ln k values of the solutes ACP, ACT, and EAB are again overestimated (higher predicted than observed values) at the water/acetonitrile mixtures, with large error. The retention of solute EE is badly predicted at the ternary mixtures. Note that this solute is a marker, thus measurements of this solute at wm1, wm3, wa1, and wa3 are available. Despite this, the predictions at the ternary mixtures are bad, finding probably its reason in non-linear mixing behaviour of the retention values of EE on stationary phase B, see Section 14.6 and Table IV.6. Yet the PARAFAC model is capable of predicting non-linear behaviour as can be checked by comparing the average of the predictions at wm3 and wa3 (1.8464) and the prediction at am3: 2.3003. It may, however, be profitable to incorporate ternary mobile phase compositions in the set of selected mobile phases.

The idea of the difference in restrictiveness - the conceptual difference - between unfold-PLS and PARAFAC is supported by the following observations. The retention of the solute EE is worse predicted with PARAFAC than unfold-PLS, on stationary phase B, whereas the retention of the solutes ACP, ACT, and MHB is worse predicted with unfold-PLS on the same stationary phase. Assuming that EE shows deviating behaviour with regard to both three-way models PARAFAC and unfold-PCA (see Section 14.7), the unfold-PLS solution is more distorted by this deviating behaviour and predicts, therefore, the retention of ACP, ACT, and MHB worse. The PARAFAC solution is less distorted, but a consequence is that the retention of EE is not predicted well.

High RMSEP values for the individual solutes, applying the two components PARAFAC model, for the stationary phases other than stationary phase B are, if present, always for the solutes EE and PBL. This is in agreement with the conclusion in Section 14.7 where the retention of these solutes were left largely unexplained by the two components PARAFAC model applied there.

16.4 Some comments on the selected marker/mobile phase combinations

For illustrative purposes, a different set of solute/mobile phase combinations is selected to calibrate new stationary phases. This selection is based on the loadings of the solutes and mobile phases on the PARAFAC components, as indicated in Section 1.6. If, for example, stationary phase B is omitted, the resulting data cube, after the usual centering, can be decomposed with the use of two PARAFAC components. The loadings of the solutes and mobile phase compositions are given in Table IV.20b. The variables PBL, PRE, and PRS load high on both components. The fourth solute is harder to select but ACP seems a reasonable choice. In case of the mobile phases the subset wm1, wm2, am1 and wa3 is a good choice. Table IV.20a lists the chosen combinations for each omitted stationary phase. Ternary mixtures are always incorporated, contrary to the induced-variance results in Section 16.1. Some of the solutes present in the subsets chosen with the loadings and not selected with the induced-variance criterion - ACP, EAB, PBL, ACT - give trouble if their retention is predicted with the induced-variance markers (Sections 16.2.and 16.3). This sheds again some light on the bad predictions of these solutes.

If predictions with a two component PARAFAC model of the non-selected solute/mobile phase combinations on stationary phase B are calculated, the result is an RMSEP value of 0.118 (see Table IV.20a). This is clearly higher than the analogous RMSEP value of stationary phase B using the solutes/mobile phases selected with the induced-variance criterion (0.089, see Section 16.3). Each stationary phase is omitted once and all prediction results are presented in Table IV.20a. For all stationary phases the predictions are worse than the "induced-variance" predictions in Section 16.3.

It is obvious that good procedures to select solute/mobile phase combinations capable of screening and calibrating a new stationary phase are needed. Extensions of quantitative methods presented in Section 1.3 for the three-way case have to be developed.

16.5 Conclusions and suggestions for further research

1. A comparison of the two- and the three-way approaches is done with the results for stationary phase C. The average RMSEP value using the blind two-way approach (no calibration measurements on the new stationary phase) is 0.086. Note that only predictions at the mobile phases where calibration measurements were performed in the training set can be obtained. The predictions on stationary phase C, if calibration measurements are performed at the mobile phases where prediction is needed, have average RMSEP values of 0.051 and 0.060 using version I and II, respectively. The difference in the blind two-way procedure and the version I and II two-way approaches lies in the measurements needed to calibrate the new stationary phase with. The use of calibration measurements lowers the prediction errors. One step further is the prediction with the unfold-PLS and PARAFAC approaches. Only a few calibration measurements are needed and RMSEP values of 0.046 and 0.048 are obtained. The draw-back of these three-way approaches is the large training set needed to build the model with. Perhaps a number of four stationary phases in the training set is sufficient in the three-way methods provided that the measurements

are very precise, but the stationary phases in the training set have to be a representative sample. Representativeness is a problem, the retention values of the markers measured at selected mobile phase compositions are supposed to describe the properties of the new stationary phase. This is only possible if the training set is a representative sample from the stationary phases. The exact pay-off between the size of the training set, the precision of the measurements and the prediction errors on a new stationary phase is subject of further research.

2. The choice of the markers/mobile phase combinations at which calibration measurements have to be performed is not exhaustively treated and is still an open question. The results of Section 16.4 indicate that choosing with the induced-variance criterion is better than the approach using loadings. The effect of the use of either the induced-variance- or the determinant criterion on the quality of the calibration can, to some extent, be foreseen in the two-way approaches. But in the three-way approaches this is more complicated. Other marker selection criteria (see Chapter 1) are available and should be tested. A procedure that simultaneously selects markers and mobile phases is perhaps the most promising to use in the three-way calibrations. The incorporation of ternary mixtures to describe non-linear mixing behaviour is worth trying.

3. A thorough analysis of the training set, which in the three-way case consist of five stationary phases, by means of three-way analysis methods is advisable. Some of the peculiarities in the calibrations were already foreseen in the three-way analyses of the whole data cube, see the discussions on the solutes EE, PBL and the calibration of stationary phase B.

4. Outliers do influence the three-way methods. With respect to the solutes, this means that the two- and three-way calibrations are only useful for the calibration of a set of structural related compounds, which is in agreement with the conclusions drawn in Section 7.5. The steroids in the used data set are probably deviating and some doubt in this respect concerns PBL. A separation problem including steroids should be calibrated using steroid-like markers. The pay-off between the degree of heterogeneity of the calibration set of solutes, which incorporates non-related compounds, and the generality of the calibration performance has to be established.

5. Referring to conclusion 2 in Section 13.5, it can be observed that the values %MARK, MS_{mark}, and s^2_{repro} are important diagnostic tools to judge the performance of the calibration with.

6. The bad calibration of stationary phases A and B shows the consequences of non-representativeness of stationary phase samples. The

different analyst/apparatus combinations have large influence and cause, in the three-way set ups chosen here, deviating behaviour of these stationary phases. Two solutions for this problem will be sketched. The first solution is the explicit incorporation of (dummy) variables that describe the specific analyst/apparatus combination in the three-way models; this leads to hybrid models. Such hybrid models contain MANOVA aspects to account for systematic differences between the analyst/apparatus combinations, and latent structure aspects that account for pure differences between stationary phases. The second solution is less complex and uses a different kind of centering in the data cube. Suppose that centering is done across the modes containing the solutes and mobile phases. The result is that all retention values on one stationary phase are expressed as deviations from the average of the retention values of all solutes at all mobile phases on that stationary phase. This average retention value, for a given stationary phase, can be considered an estimate of the retentive power of a stationary phase. The retentive power of a new stationary phase can be estimated by averaging the retention values of the markers at the selected mobile phase compositions. Prediction of the non-selected solute/mobile phase combinations is then along the same lines as presented previously, in Sections 16.2 and 16.3. Preliminary calculations (not shown) indicate that the second solution is promising.

7. The choice between unfold-PLS and PARAFAC as calibration methods is not an easy one. Some remarks are appropriate. First, the number of degrees of freedom if the PARAFAC model is used is higher than for the unfold-PLS model, see Section 13.7 where this was shown for the comparison of PARAFAC and unfold-PCA. This may be an advantage of the PARAFAC model. Second, if it is reasonable to model the retention values of the solutes at the mobile phase compositions on one stationary phase with a bilinear model (PCA) then a natural extension to the three-way case is the PARAFAC model which is trilinear, contrary to unfold-PLS[3]. Third, referring to the remark made in conclusion 4 with respect to the heterogeneity of the training set, the robustness of both three-way methods with respect to deviating solutes and stationary phases has to be established. Fourth, a version of PARAFAC which reckons with the distinction between independent and dependent variables, like unfold-PLS, is worth developing. Finally, referring to conclusion 2 in Section 13.5, diagnostic tools besides %MARK, MS_{mark}, R^2_{train} etc have to be developed to validate the three-way calibration procedures.

8. The reasons for the difference in performance between the three-way approaches used for the calibration of the octadecyl stationary phases (Chapter 16) and used for the calibration of the different types of stationary phases (Chapter 13) are three-fold. First, the signal-to-noise ratios for the different types of stationary phases

were much higher than for the octadecyl phases, as expected. Second, all retention values on the different types of stationary phases were measured by the same analyst/apparatus combination. Finally, the set of solutes are more homogeneous in case of the calibration of the different types of stationary phases than in the octadecyl calibration, but this statement is hard to prove.

9. The two components in the PARAFAC model were not forced to be orthogonal. Some remarks with this respect were already made in Section 14.7. The calculation of the PARAFAC factor loadings is performed with an iterative procedure. The stability of the solution can be checked by performing the calculations several times, each time with other starting values. No large instabilities were observed during the calculations. Yet the stability may improve by demanding orthogonality of the PARAFAC components in some of the modes. Preliminary calculations support this suggestion. Moreover, the least squares step in which the scores of the new stationary phase are calculated might profit from orthogonality.

10. The size of the prediction errors must be evaluated keeping in mind that in optimization procedures, where the mobile phase is manipulated in order to obtain an optimal separation, relatively small prediction errors can result in non-optimal separations. Very precise predictions are, therefore, necessary.

References Part IV

1 P.C. Sadek, P.W. Carr and L.W. Bowers; *The significance of metallophilic and silanophilic interactions in reversed phase HPLC*, J.Liq.Chromatogr., 8(13) (1985) 2369-2386

2 T.L. Hafkenscheid and E. Tomlinson; *Relationship between hydrophobic (lypophilic) properties of bases and their retention in reversed-phase liquid chromatography using aqueous methanol mobile phases*, J.Chromatogr., 292 (1984) 305-317

3 H.G. Law, C.W. Snyder, J.A. Hattie and R.P. McDonald (eds); *Research Methods for Multi-mode Data Analysis*, Praeger, New York, 1984

4 S. Wold, P. Geladi, K. Esbensen and J. Öhman; *Multi-way Principal Components- and PLS-Analysis*, J.Chemometrics, 1(1) (1987) 41-56

5 L.R. Snyder; *Classification of the solvent properties of common liquids*, J.Chrom.Sci., 16 (1978) 223

6 C.H. Lochmuller, S.J. Breiner, Ch.E. Reese and M.N. Koel; *Characterization and prediction of retention behaviour in Rversed-Phase Chromatography using Factor Analytical modeling*, Anal.Chem., 61 (1989) 367-375

7 R.M. Smith, G.A. Murilla, T.G. Hurdley, R. Gill and A.C. Moffat; *Retention reproducibility of thiazide diuretics and related drugs in reversed-phase high-performance liquid chromatography*, J.Chromatogr., 384 (1987) 259-278

TABLE IV.1. Stationary- and mobile phase characteristics

a) stationary phases

S.Phase	C18 batch	Si batch	Analist/apparatus
A	1	1	1
B	2	2	2
C	3	1	3
D	4	2	3
E	2	2	3
F	1	1	3

b) mobile phase compositions

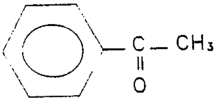
	water	acetonitrile	methanol
wm1	0.63	0.00	0.37
wm2	0.55	0.00	0.45
wm3	0.47	0.00	0.53
am1	0.71	0.11	0.18
am2	0.63	0.15	0.22
am3	0.54	0.19	0.27
wa1	0.78	0.22	0.00
wa2	0.70	0.30	0.00
wa3	0.62	0.38	0.00

Legend: the mobile phase compositions are noted down in v/v fractions. Note that the mixtures am1 to am3 are ternary mixtures.

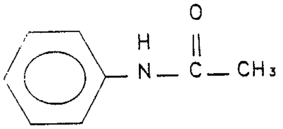
TABLE IV.2. The test solutes

nr name

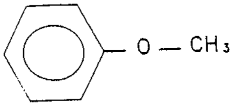
1 acetophenone (ACP)



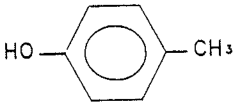
2 acetanilide (ACT)



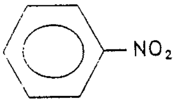
3 anisole (ANS)



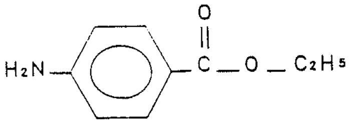
4 p-cresol (CRE)



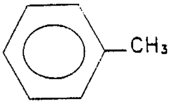
5 nitrobenzene (NBZ)



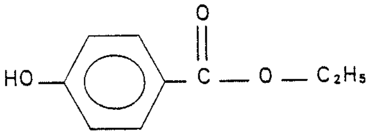
6 ethylaminobenzoate (EAB)



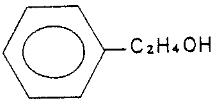
7 toluene (TOL)



8 ethylhydroxybenzoate (EHB)



9 2-phenylethanol (PE)



10 methylhydroxybenzoate (MHB)

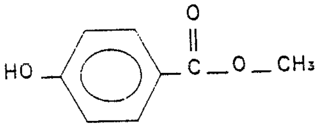
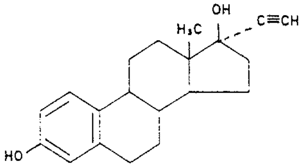


TABLE IV.2 (continued).

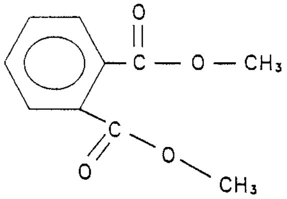
nr name

11 ethynylestradiol (EE)

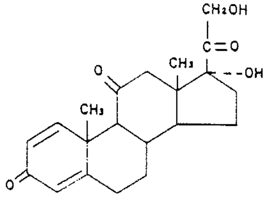


nr name

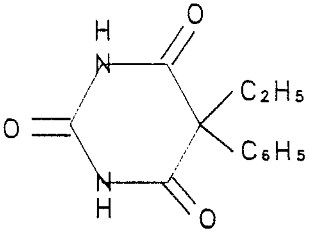
12 dimethylphtalate (DMP)



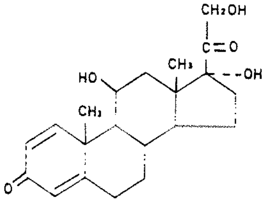
13 prednisone (PRE)



14 phenobarbital (PBL)



15 prednisolone (PRS)



16 propylhydroxybenzoate (PHB)

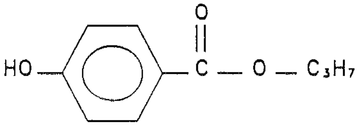


TABLE IV.3. The capacity factors of the solutes

Stationary phase A

	ACP	ACT	ANS	CRE	DMP	EAB	EE	EHB
wm1	4.44	1.68	9.25	4.36	6.09	4.46	113.78	8.17
wm2	2.30	0.96	5.16	2.31	2.58	2.02	24.76	3.63
wm3	1.37	0.62	3.08	1.34	1.31	1.09	8.76	1.78
am1	6.49	2.25	13.61	6.50	10.81	7.73	181.41	12.50
am2	3.44	1.31	7.33	3.33	4.69	3.44	28.82	5.03
am3	1.86	0.72	3.98	1.79	2.26	1.67	10.63	2.32
wa1	7.21	2.04	17.37	6.48	11.53	8.33	100.19	10.32
wa2	3.64	1.07	8.84	3.09	4.79	3.59	18.71	3.81
wa3	2.03	0.62	4.34	1.63	2.33	1.73	5.07	1.74

	MHB	NBZ	PBL	PE	PHB	PRE	PRS	TOL
wm1	3.35	5.39	2.99	3.57	21.63	14.35	20.29	23.56
wm2	1.66	3.07	1.55	1.97	8.53	4.33	6.46	12.73
wm3	0.90	1.92	0.84	1.17	3.81	1.85	2.70	7.24
am1	4.88	9.10	5.58	4.47	34.47	21.56	25.36	33.64
am2	2.25	5.18	2.43	2.36	12.22	5.99	7.05	17.40
am3	1.16	2.84	1.16	1.32	4.97	2.17	2.58	8.87
wa1	4.13	12.70	4.52	3.68	28.11	9.29	8.79	43.54
wa2	1.83	6.69	1.76	1.75	8.63	2.04	1.78	20.59
wa3	0.99	3.49	0.67	1.00	3.28	0.78	0.67	9.43

Stationary phase B

	ACP	ACT	ANS	CRE	DMP	EAB	EE	EHB
wm1	4.26	1.61	8.83	4.16	5.84	4.29	109.26	7.96
wm2	2.16	.89	4.79	2.16	2.37	1.96	24.74	3.34
wm3	1.26	.55	2.88	1.25	1.24	1.03	8.83	1.72
am1	5.61	1.92	12.09	5.75	9.64	6.88	169.95	10.76
am2	3.24	1.27	6.46	3.01	4.10	3.10	27.20	4.51
am3	1.66	.63	3.69	1.86	2.09	1.63	10.30	2.10
wa1	5.94	1.64	14.42	5.38	9.60	6.97	93.26	8.66
wa2	3.06	.83	7.77	2.61	4.21	2.89	16.37	3.29
wa3	1.47	.51	3.91	1.44	2.03	1.46	4.51	1.63

	MHB	NBZ	PBL	PE	PHB	PRE	PRS	TOL
wm1	3.24	5.21	2.91	3.44	20.97	14.31	20.04	22.17
wm2	1.53	2.91	1.42	1.74	7.88	4.16	6.27	11.85
wm3	.86	1.77	.81	1.10	3.70	1.74	2.64	6.80
am1	4.18	8.07	4.94	4.01	29.98	19.79	23.15	29.41
am2	1.96	4.92	2.19	2.38	11.00	5.89	6.62	15.73
am3	1.02	2.54	1.09	1.21	4.56	1.98	2.54	8.58
wa1	3.44	10.61	3.75	3.03	22.56	7.89	7.50	36.26
wa2	1.52	5.70	1.41	1.51	7.51	1.71	1.49	18.01
wa3	.87	3.08	.67	.86	2.84	.65	.51	8.39

TABLE IV.3 (continued).

Stationary phase C

	ACP	ACT	ANS	CRE	DMP	EAB	EE	EHB
wm1	4.46	1.66	9.66	4.47	6.43	5.03	108.79	8.57
wm2	2.41	.97	5.55	2.51	2.77	2.22	28.46	3.97
wm3	1.27	.55	2.98	1.32	1.25	1.02	8.53	1.75
am1	6.59	2.26	14.49	6.83	10.93	8.01	150.23	12.80
am2	3.09	1.12	7.09	3.13	4.42	3.26	33.97	4.81
am3	1.85	.71	4.08	1.83	2.25	1.70	10.67	2.37
wa1	6.53	1.80	16.40	6.10	10.54	7.66	87.22	9.53
wa2	3.61	1.02	8.89	3.14	4.70	3.53	17.70	3.79
wa3	2.09	.67	4.69	1.74	2.43	1.82	5.37	1.84

	MHB	NBZ	PBL	PE	PHB	PRE	PRS	TOL
wm1	3.52	5.58	3.75	3.59	22.70	14.02	20.26	26.33
wm2	1.80	3.25	1.71	2.06	9.38	4.57	6.95	14.08
wm3	.87	1.80	.79	1.11	3.72	1.68	2.55	7.08
am1	5.03	9.39	5.70	4.56	35.49	20.82	24.77	35.11
am2	2.15	4.85	2.30	2.21	11.53	5.25	6.22	17.81
am3	1.17	2.80	1.17	1.29	5.08	2.02	2.57	9.28
wa1	3.91	11.80	4.03	3.20	25.60	7.35	7.11	42.58
wa2	1.85	6.47	1.66	1.72	8.44	1.85	1.65	20.72
wa3	1.04	3.60	.84	1.02	3.51	.77	.68	10.01

Stationary phase D

	ACP	ACT	ANS	CRE	DMP	EAB	EE	EHB
wm1	4.32	1.63	9.13	4.36	6.27	4.71	105.02	8.43
wm2	2.23	.93	5.08	2.33	2.61	2.06	26.66	3.71
wm3	1.22	.55	2.84	1.27	1.22	.98	7.98	1.67
am1	5.88	2.09	12.75	6.09	10.21	7.20	145.98	11.65
am2	3.14	1.16	7.02	3.21	4.38	3.29	33.46	4.79
am3	1.77	.69	3.87	1.77	2.18	1.66	10.33	2.29
wa1	6.42	1.84	15.83	5.95	10.18	7.82	84.07	9.10
wa2	3.39	1.02	8.17	3.02	4.45	3.42	17.05	3.58
wa3	1.93	.64	4.33	1.66	2.29	1.78	5.13	1.73

	MHB	NBZ	PBL	PE	PHB	PRE	PRS	TOL
wm1	3.47	5.35	3.64	3.55	22.19	14.02	20.17	24.11
wm2	1.70	3.04	1.58	1.92	8.74	4.28	6.47	12.88
wm3	.84	1.72	.79	1.08	3.53	1.59	2.40	6.75
am1	4.62	8.39	5.02	4.12	31.88	17.87	22.34	32.54
am2	2.14	4.86	2.23	2.18	11.54	5.47	6.43	16.55
am3	1.15	2.71	1.13	1.24	4.89	2.01	2.43	8.63
wa1	3.75	11.76	3.76	3.33	24.28	7.48	7.26	39.98
wa2	1.75	6.16	1.50	1.68	7.97	1.81	1.65	19.36
wa3	.97	3.32	.79	.96	3.30	.73	.66	9.34

TABLE IV.3 (continued).

Stationary phase E

	ACP	ACT	ANS	CRE	DMP	EAB	EE	EHB
wm1	4.32	1.59	9.23	4.29	6.02	4.63	96.27	8.39
wm2	2.28	.97	5.20	2.44	2.64	2.17	26.83	3.80
wm3	1.25	.56	2.90	1.29	1.25	1.03	8.24	1.69
am1	6.19	2.21	13.25	6.41	10.44	8.19	146.95	11.76
am2	3.26	1.21	7.23	3.32	4.58	3.46	34.87	4.94
am3	1.77	.71	3.86	1.75	2.15	1.66	10.07	2.26
wa1	6.64	1.92	16.28	6.14	10.74	8.16	88.61	9.53
wa2	3.52	1.05	8.48	3.04	4.56	3.50	17.15	3.69
wa3	1.98	.64	4.40	1.65	2.32	1.80	5.14	1.76

	MHB	NBZ	PBL	PE	PHB	PRE	PRS	TOL
wm1	3.45	5.37	3.50	3.53	22.06	13.82	19.36	23.95
wm2	1.75	3.10	1.64	1.98	8.91	4.40	6.92	13.10
wm3	.86	1.77	.81	1.11	3.55	1.65	2.42	6.87
am1	4.71	8.84	5.59	4.40	31.87	19.50	23.85	33.08
am2	2.22	4.98	2.48	2.31	11.78	5.54	6.72	17.25
am3	1.13	2.72	1.16	1.27	4.80	2.05	2.48	8.58
wa1	3.96	11.92	4.40	3.49	25.34	7.71	7.43	41.86
wa2	1.83	6.42	1.74	1.74	8.10	1.90	1.69	19.75
wa3	.99	3.37	.88	1.00	3.32	.77	.67	9.42

Stationary phase F

	ACP	ACT	ANS	CRE	DMP	EAB	EE	EHB
wm1	4.39	1.58	9.53	4.33	5.80	4.70	90.43	8.39
wm2	2.23	.92	5.23	2.39	2.58	2.10	26.57	3.71
wm3	1.27	.58	2.99	1.36	1.25	1.03	8.20	1.76
am1	6.09	2.10	13.51	6.23	9.89	7.45	132.19	11.56
am2	3.26	1.18	7.32	3.33	4.47	3.56	32.73	4.93
am3	1.82	.71	3.97	1.79	2.20	1.68	10.11	2.32
wa1	6.79	1.92	16.91	6.38	10.82	8.63	88.74	9.98
wa2	3.53	1.03	8.64	3.12	4.62	3.66	17.19	3.77
wa3	2.03	.63	4.64	1.71	2.39	1.88	5.50	1.85

	MHB	NBZ	PBL	PE	PHB	PRE	PRS	TOL
wm1	3.44	5.32	3.37	3.48	22.18	13.42	19.52	24.48
wm2	1.69	3.07	1.62	1.95	8.73	4.27	6.58	13.62
wm3	.89	1.81	.83	1.14	3.71	1.68	2.54	7.13
am1	4.61	8.75	5.34	4.29	31.45	18.38	21.69	33.84
am2	2.22	5.00	2.42	2.32	11.82	5.45	6.47	17.34
am3	1.16	2.78	1.19	1.30	4.92	2.06	2.46	8.90
wa1	4.10	12.27	4.44	3.55	26.71	7.80	7.54	43.22
wa2	1.84	6.47	1.75	1.71	8.32	1.86	1.68	20.38
wa3	1.03	3.53	.91	1.00	3.57	.77	.71	9.98

TABLE IV.4. Reproducibility (k-values)

S.Phase	Solutes							
	ACP	ACT	ANS	CRE	DMP	EAB	EE	EHB
A wm2								
mean k	2.30	.96	5.16	2.31	2.58	2.02	24.76	3.63
s_{repro}	.026	.013	.009	.052	.045	.026	.585	.014
CV	1.11	1.33	.18	2.25	1.75	1.27	2.36	0.38
B wm2								
mean k	2.16	.89	4.79	2.16	2.37	1.96	24.74	3.34
s_{repro}	.019	.054	.142	.045	.163	.038	.357	.216
CV	.086	6.11	2.96	2.08	6.89	1.93	1.44	6.48
C wm2								
mean k	2.41	.97	5.55	2.51	2.77	2.22	28.46	3.97
s_{repro}	.008	.018	.044	.027	.033	.034	.908	.080
CV	.343	1.84	.787	1.09	1.18	1.55	3.19	2.01
D wm2								
mean k	2.23	.93	5.08	2.33	2.61	2.06	26.66	3.71
s_{repro}	.074	.045	.121	.081	.087	.089	1.47	.073
CV	3.34	4.83	2.38	3.50	3.35	4.33	5.50	1.96
E wm2								
mean k	2.28	.97	5.20	2.44	2.64	2.17	26.83	3.80
s_{repro}	.040	.045	.064	.082	.048	.053	1.06	.017
CV	1.75	4.69	1.23	3.35	1.82	2.44	3.97	.450
F wm2								
mean k	2.23	.92	5.23	2.39	2.58	2.10	26.57	3.71
s_{repro}	.136	.049	.195	.106	.164	.160	1.61	.233
CV	6.10	5.37	3.72	4.43	6.37	7.61	6.05	6.27

TABLE IV.4 (continued).

S.Phase	Solutes							
	MHB	NBZ	PBL	PE	PHB	PRE	PRS	TOL
A wm2								
mean k	1.66	3.07	1.55	1.97	8.53	4.33	6.46	12.73
s_{repro}	.013	.048	.016	.011	.070	.116	.049	.239
CV	.762	1.57	1.05	.573	.819	2.67	.761	1.88
B wm2								
mean k	1.53	2.91	1.42	1.74	7.88	4.16	6.27	11.85
s_{repro}	.095	.041	.120	.215	.449	.014	.096	.236
CV	6.25	1.42	8.39	12.3	5.69	.332	1.53	1.99
C wm2								
mean k	1.80	3.25	1.71	2.06	9.38	4.57	6.95	14.08
s_{repro}	.035	.031	.022	.028	.203	.043	.128	.251
CV	1.93	.972	1.32	1.37	2.17	.947	1.85	1.78
D wm2								
mean k	1.70	3.04	1.58	1.92	8.74	4.28	6.47	12.88
s_{repro}	.034	.101	.050	.073	.171	.156	.244	.197
CV	2.01	3.33	3.15	3.81	1.96	3.65	3.77	1.53
E wm2								
mean k	1.75	3.10	1.64	1.98	8.91	4.40	6.92	13.10
s_{repro}	.012	.030	.030	.005	.047	.028	.403	.164
CV	.677	.971	1.85	.232	.528	.648	5.82	1.26
F wm2								
mean k	1.69	3.07	1.62	1.95	8.73	4.27	6.58	13.62
s_{repro}	.116	.181	.105	.144	.533	.381	.264	.252
CV	6.88	5.88	6.50	7.41	6.11	8.94	4.01	1.85

Legend: the s_{repro} values are calculated with the use of three (A and B), or four (C to F) measurements, see also Table III.2 and text.

TABLE IV.5. Repeatability (k-values)

Stationary phase A

	ACP	ACT	ANS	CRE	DMP	EAB	EE	EHB
wm1	.096	.043	.032	.032	.032	.000	.319	.032
wm2	.064	.021	.074	.043	.032	.011	.383	.053
wm3	.000	.000	.043	.011	.032	.022	.011	.022
am1	.136	.052	.010	.000	.052	.021	.890	.010
am2	.000	.043	.032	.011	.043	.043	.140	.022
am3	.033	.022	.011	.000	.000	.011	.000	.000
wa1	.135	.011	.011	.022	.034	.011	.337	.045
wa2	.034	.000	.046	.011	.023	.023	.046	.023
wa3	.035	.024	.224	.024	.024	.000	.000	.012

	MHB	NBZ	PBL	PE	PHB	PRE	PRS	TOL
wm1	.011	.043	.011	.000	.064	.149	.032	.011
wm2	.011	.053	.000	.011	.085	.011	.085	.106
wm3	.011	.000	.011	.022	.000	.000	.011	.011
am1	.021	.210	.000	.000	.084	.314	.073	.063
am2	.022	.000	.054	.000	.000	.173	.032	.151
am3	.022	.033	.000	.011	.022	.011	.000	.033
wa1	.000	.112	.022	.000	.090	.045	.022	.090
wa2	.000	.046	.011	.000	.011	.023	.000	.023
wa3	.012	.035	.024	.024	.024	.000	.024	.012

Stationary phase B

	ACP	ACT	ANS	CRE	DMP	EAB	EE	EHB
wm1	.131	.101	.081	.020	.020	.000		.010
wm2	.184	.039	.058	.145	.203	.107	.523	.165
wm3	.020	.010	.020	.010	.010	.020	.000	.000
am1	.067	.029	.058	.077	.077	.048		.548
am2	.040	.010	.030	.010	.020	.020	.000	.060
am3	.020	.020	.030	.010	.000	.020	.020	.020
wa1	.038	.009	.047	.000	.076	.028		.009
wa2	.028	.000	.057	.009	.019	.000	.000	.236
wa3	.020	.071	.283	.091	.202	.010	.121	.040

	MHB	NBZ	PBL	PE	PHB	PRE	PRS	TOL
wm1	.000	.162	.000	.030	.040	.596	.000	.040
wm2	.068	.194	.223	.349	.562	.155	.048	.232
wm3	.000	.010	.010	.020	.010	.010	.000	.020
am1	.250	.087	.048	.029	1.097	.135	.058	.096
am2	.030	.010	.010	.000	.100	.060	.020	.010
am3	.010	.030	.000	.010	.051	.010	.010	.374
wa1	.000	.114	.000	.019	.057	.047	.028	.038
wa2	.151	.066	.019	.000	.415	.009	.132	.075
wa3	.061	.071	.000	.081	.020	.030	.010	.303

TABLE IV.5 (continued).

Stationary phase C

	ACP	ACT	ANS	CRE	DMP	EAB	EE	EHB
wm1	.000	.000	.012	.012	.058	.000	2.689	.023
wm2	.006	.006	.006	.000	.019	.008	.257	.008
wm3	.000	.011	.000	.000	.011	.000	.011	.000
am1	.023	.000	.046	.012	.081	.012	3.060	.151
am2	.012	.000	.012	.000	.000	.012	.071	.012
am3	.000	.000	.024	.000	.000	.012	.012	.000
wa1	.012	.000	.024	.000	.106	.012	1.508	.165
wa2	.012	.000	.061	.000	.012	.000	.012	.000
wa3	.000	.000	.012	.000	.012	.000	.000	.000

	MHB	NBZ	PBL	PE	PHB	PRE	PRS	TOL
wm1	.012	.012	.046	.000	.070	.046	.000	.046
wm2	.000	.017	.013	.014	.014	.044	.015	.064
wm3	.000	.011	.000	.000	.011	.000	.011	.000
am1	.046	.220	.046	.093	.533	.556	.023	.696
am2	.000	.000	.000	.012	.071	.024	.000	.012
am3	.000	.024	.000	.000	.000	.024	.012	.012
wa1	.082	.047	.047	.000	.401	.047	.012	.330
wa2	.012	.012	.000	.000	.012	.012	.012	.012
wa3	.000	.000	.000	.000	.000	.000	.000	.012

Stationary phase D

	ACP	ACT	ANS	CRE	DMP	EAB	EE	EHB
wm1	.011	.000	.011	.000	.055	.011	.442	.011
wm2	.011	.006	.012	.008	.029	.008	.373	.008
wm3	.000	.011	.000	.000	.000	.000	.022	.000
am1	.000	.011	.000	.011	.067	.011	.713	.134
am2	.000	.000	.000	.000	.011	.000	.430	.023
am3	.000	.000	.000	.000	.023	.000	.232	.012
wa1	.035	.000	.081	.000	.023	.000	.742	.046
wa2	.000	.000	.000	.000	.036	.012	.264	.000
wa3	.000	.000	.000	.000	.000	.000	.012	.012

	MHB	NBZ	PBL	PE	PHB	PRE	PRS	TOL
wm1	.000	.011	.044	.000	.155	.088	.000	.309
wm2	.010	.010	.020	.008	.033	.008	.017	.077
wm3	.000	.011	.000	.000	.011	.000	.000	.011
am1	.067	.000	.045	.000	.200	.011	.200	.022
am2	.000	.034	.000	.000	.079	.011	.011	.023
am3	.000	.012	.023	.012	.012	.012	.023	.058
wa1	.023	.197	.035	.000	.116	.023	.000	.023
wa2	.000	.012	.012	.000	.012	.000	.012	.168
wa3	.012	.012	.000	.012	.012	.012	.000	.012

TABLE IV.5 (continued).

Stationary phase E

	ACP	ACT	ANS	CRE	DMP	EAB	EE	EHB
wm1	.023	.011	.057	.023	.103	.023	2.874	.046
wm2	.012	.010	.019	.014	.019	.010	.198	.051
wm3	.011	.011	.011	.000	.011	.000	.274	.011
am1	.011	.000	.011	.011	.091	.034	1.277	.068
am2	.000	.000	.023	.012	.058	.000	.997	.012
am3	.011	.000	.000	.000	.000	.011	.057	.011
wa1	.012	.012	.060	.060	.300	.012	2.541	.012
wa2	.000	.000	.000	.012	.037	.000	.085	.000
wa3	.000	.000	.000	.012	.012	.012	.024	.000

	MHB	NBZ	PBL	PE	PHB	PRE	PRS	TOL
wm1	.023	.000	.068	.000	.068	.000	.103	.228
wm2	.031	.051	.013	.034	.114	.057	.092	.030
wm3	.011	.000	.000	.000	.023	.000	.000	.046
am1	.023	.023	.057	.000	.114	.091	.194	.137
am2	.012	.012	.046	.000	.035	.023	.000	.151
am3	.000	.011	.000	.011	.023	.011	.000	.023
wa1	.012	.048	.156	.000	.024	.024	.060	.527
wa2	.000	.012	.024	.000	.000	.000	.024	.037
wa3	.012	.037	.000	.024	.012	.012	.000	.037

Stationary phase F

	ACP	ACT	ANS	CRE	DMP	EAB	EE	EHB
wm1	.012	.012	.012	.023	.116	.012	.185	.023
wm2	.006	.006	.014	.012	.030	.012	.366	.015
wm3	.011	.000	.011	.000	.011	.000	.000	.011
am1	.000	.012	.012	.000	.166	.226	1.331	.083
am2	.012	.000	.024	.012	.000	.035	.118	.000
am3	.012	.012	.000	.000	.012	.012	.000	.012
wa1	.049	.012	.110	.024	.110	.305	1.561	.049
wa2	.000	.000	.000	.024	.049	.024	.171	.012
wa3	.000	.000	.012	.000	.000	.012	.000	.025

	MHB	NBZ	PBL	PE	PHB	PRE	PRS	TOL
wm1	.000	.012	.081	.000	.070	.012	.139	.255
wm2	.008	.012	.022	.006	.054	.045	.072	.104
wm3	.000	.000	.011	.034	.011	.000	.011	.023
am1	.048	.024	.095	.012	.048	.143	.024	.475
am2	.000	.000	.000	.000	.000	.000	.024	.012
am3	.012	.000	.012	.000	.012	.000	.000	.000
wa1	.024	.110	.085	.000	.146	.061	.012	.098
wa2	.000	.012	.037	.000	.012	.000	.012	.195
wa3	.012	.012	.000	.000	.049	.000	.012	.012

Legend: see text. Empty spots indicate absence of repeated measurements

TABLE IV.6. Logarithms of the capacity factors of the solutes

Stationary phase A

	ACP	ACT	ANS	CRE	DMP	EAB	EE	EHB
wm1	1.49	.52	2.22	1.47	1.81	1.49	4.73	2.10
wm2	.83	-.04	1.64	.84	.95	.70	3.21	1.29
wm3	.31	-.48	1.12	.30	.27	.09	2.17	.58
am1	1.87	.81	2.61	1.87	2.38	2.04	5.20	2.53
am2	1.24	.27	1.99	1.20	1.55	1.24	3.36	1.62
am3	.62	-.33	1.38	.58	.81	.52	2.36	.84
wa1	1.97	.71	2.85	1.87	2.45	2.12	4.61	2.33
wa2	1.29	.06	2.18	1.13	1.57	1.28	2.93	1.34
wa3	.71	-.48	1.47	.49	.85	.55	1.62	.55

	MHB	NBZ	PBL	PE	PHB	PRE	PRS	TOL
wm1	1.21	1.68	1.10	1.27	3.07	2.66	3.01	3.16
wm2	.51	1.12	.44	.68	2.14	1.47	1.87	2.54
wm3	-.10	.65	-.17	.16	1.34	.62	.99	1.98
am1	1.59	2.21	1.72	1.50	3.54	3.07	3.23	3.52
am2	.81	1.65	.89	.86	2.50	1.79	1.95	2.86
am3	.14	1.04	.14	.28	1.60	.77	.95	2.18
wa1	1.42	2.54	1.51	1.30	3.34	2.23	2.17	3.77
wa2	.60	1.90	.56	.56	2.16	.71	.58	3.02
wa3	-.01	1.25	-.41	.00	1.19	-.24	-.41	2.24

Stationary phase B

	ACP	ACT	ANS	CRE	DMP	EAB	EE	EHB
wm1	1.45	.48	2.18	1.42	1.77	1.46	4.69	2.07
wm2	.77	-.12	1.57	.77	.86	.67	3.21	1.20
wm3	.23	-.60	1.06	.22	.21	.03	2.18	.54
am1	1.72	.65	2.49	1.75	2.27	1.93	5.14	2.38
am2	1.18	.24	1.87	1.10	1.41	1.13	3.30	1.51
am3	.51	-.46	1.31	.62	.74	.49	2.33	.74
wa1	1.78	.49	2.67	1.68	2.26	1.94	4.54	2.16
wa2	1.12	-.19	2.05	.96	1.44	1.06	2.80	1.19
wa3	.39	-.68	1.36	.36	.70	.38	1.51	.49

	MHB	NBZ	PBL	PE	PHB	PRE	PRS	TOL
wm1	1.18	1.65	1.07	1.23	3.04	2.66	3.00	3.10
wm2	.42	1.07	.35	.54	2.06	1.43	1.84	2.47
wm3	-.16	.57	-.22	.10	1.31	.55	.97	1.92
am1	1.43	2.09	1.60	1.39	3.40	2.99	3.14	3.38
am2	.68	1.59	.78	.87	2.40	1.77	1.89	2.76
am3	.02	.93	.08	.19	1.52	.68	.93	2.15
wa1	1.24	2.36	1.32	1.11	3.12	2.07	2.02	3.59
wa2	.42	1.74	.35	.41	2.02	.54	.40	2.89
wa3	-.14	1.12	-.40	-.16	1.04	-.43	-.68	2.13

TABLE IV.6 (continued).

Stationary phase C

	ACP	ACT	ANS	CRE	DMP	EAB	EE	EHB
wm1	1.49	.50	2.27	1.50	1.86	1.62	4.69	2.15
wm2	.88	-.03	1.71	.92	1.02	.80	3.35	1.38
wm3	.24	-.60	1.09	.28	.22	.02	2.14	.56
am1	1.89	.82	2.67	1.92	2.39	2.08	5.01	2.55
am2	1.13	.11	1.96	1.14	1.49	1.18	3.53	1.57
am3	.61	-.34	1.41	.60	.81	.53	2.37	.86
wa1	1.88	.59	2.80	1.81	2.36	2.04	4.47	2.25
wa2	1.28	.02	2.18	1.14	1.55	1.26	2.87	1.33
wa3	.74	-.41	1.55	.55	.89	.60	1.68	.61

	MHB	NBZ	PBL	PE	PHB	PRE	PRS	TOL
wm1	1.26	1.72	1.32	1.28	3.12	2.64	3.01	3.27
wm2	.59	1.18	.54	.72	2.24	1.52	1.94	2.64
wm3	-.14	.59	-.23	.11	1.31	.52	.94	1.96
am1	1.62	2.24	1.74	1.52	3.57	3.04	3.21	3.56
am2	.77	1.58	.83	.79	2.45	1.66	1.83	2.88
am3	.16	1.03	.16	.25	1.63	.70	.94	2.23
wa1	1.36	2.47	1.39	1.16	3.24	1.99	1.96	3.75
wa2	.62	1.87	.50	.54	2.13	.62	.50	3.03
wa3	.03	1.28	-.17	.02	1.26	-.26	-.38	2.30

Stationary phase D

	ACP	ACT	ANS	CRE	DMP	EAB	EE	EHB
wm1	1.46	.49	2.21	1.47	1.84	1.55	4.65	2.13
wm2	.80	-.08	1.62	.84	.96	.72	3.28	1.31
wm3	.20	-.60	1.04	.24	.20	-.02	2.08	.51
am1	1.77	.74	2.55	1.81	2.32	1.97	4.98	2.45
am2	1.15	.15	1.95	1.17	1.48	1.19	3.51	1.57
am3	.57	-.37	1.35	.57	.78	.50	2.33	.83
wa1	1.86	.61	2.76	1.78	2.32	2.06	4.43	2.21
wa2	1.22	.02	2.10	1.10	1.49	1.23	2.84	1.27
wa3	.66	-.45	1.47	.50	.83	.57	1.63	.55

	MHB	NBZ	PBL	PE	PHB	PRE	PRS	TOL
wm1	1.24	1.68	1.29	1.27	3.10	2.64	3.00	3.18
wm2	.53	1.11	.46	.65	2.17	1.45	1.87	2.56
wm3	-.17	.54	-.23	.08	1.26	.46	.87	1.91
am1	1.53	2.13	1.61	1.42	3.46	2.88	3.11	3.48
am2	.76	1.58	.80	.78	2.45	1.70	1.86	2.81
am3	.14	1.00	.12	.21	1.59	.70	.89	2.16
wa1	1.32	2.46	1.33	1.20	3.19	2.01	1.98	3.69
wa2	.56	1.82	.41	.52	2.08	.60	.50	2.96
wa3	-.03	1.20	-.23	-.04	1.19	-.31	-.42	2.23

TABLE IV.6 (continued).

Stationary phase E

	ACP	ACT	ANS	CRE	DMP	EAB	EE	EHB
wm1	1.46	.46	2.22	1.46	1.80	1.53	4.57	2.13
wm2	.83	-.03	1.65	.89	.97	.78	3.29	1.34
wm3	.22	-.59	1.06	.25	.22	.03	2.11	.52
am1	1.82	.79	2.58	1.86	2.35	2.10	4.99	2.46
am2	1.18	.19	1.98	1.20	1.52	1.24	3.55	1.60
am3	.57	-.35	1.35	.56	.77	.51	2.31	.82
wa1	1.89	.65	2.79	1.82	2.37	2.10	4.48	2.25
wa2	1.26	.05	2.14	1.11	1.52	1.25	2.84	1.31
wa3	.68	-.45	1.48	.50	.84	.59	1.64	.56

	MHB	NBZ	PBL	PE	PHB	PRE	PRS	TOL
wm1	1.24	1.68	1.25	1.26	3.09	2.63	2.96	3.18
wm2	.56	1.13	.50	.68	2.19	1.48	1.93	2.57
wm3	-.15	.57	-.22	.11	1.27	.50	.88	1.93
am1	1.55	2.18	1.72	1.48	3.46	2.97	3.17	3.50
am2	.80	1.60	.91	.84	2.47	1.71	1.91	2.85
am3	.12	1.00	.15	.24	1.57	.72	.91	2.15
wa1	1.38	2.48	1.48	1.25	3.23	2.04	2.01	3.73
wa2	.60	1.86	.55	.55	2.09	.64	.52	2.98
wa3	-.01	1.22	-.13	.00	1.20	-.27	-.40	2.24

Stationary phase F

	ACP	ACT	ANS	CRE	DMP	EAB	EE	EHB
wm1	1.48	.46	2.25	1.47	1.76	1.55	4.50	2.13
wm2	.80	-.09	1.65	.87	.95	.74	3.28	1.31
wm3	.24	-.55	1.10	.31	.22	.03	2.10	.56
am1	1.81	.74	2.60	1.83	2.29	2.01	4.88	2.45
am2	1.18	.17	1.99	1.20	1.50	1.27	3.49	1.60
am3	.60	-.35	1.38	.58	.79	.52	2.31	.84
wa1	1.92	.65	2.83	1.85	2.38	2.15	4.49	2.30
wa2	1.26	.03	2.16	1.14	1.53	1.30	2.84	1.33
wa3	.71	-.45	1.54	.54	.87	.63	1.71	.62

	MHB	NBZ	PBL	PE	PHB	PRE	PRS	TOL
wm1	1.24	1.67	1.21	1.25	3.10	2.60	2.97	3.20
wm2	.52	1.12	.48	.66	2.16	1.45	1.88	2.61
wm3	-.12	.59	-.19	.13	1.31	.52	.93	1.96
am1	1.53	2.17	1.67	1.46	3.45	2.91	3.08	3.52
am2	.80	1.61	.88	.84	2.47	1.70	1.87	2.85
am3	.15	1.02	.18	.26	1.59	.72	.90	2.19
wa1	1.41	2.51	1.49	1.27	3.28	2.05	2.02	3.77
wa2	.61	1.87	.56	.53	2.12	.62	.52	3.01
wa3	.03	1.26	-.09	.00	1.27	-.26	-.34	2.30

TABLE IV.7. Reproducibility and repeatability (ln k values)

S.Phase	Solute							
	ACP	ACT	ANS	CRE	DMP	EAB	EE	EHB
A wm2								
mean lnk	.83	-.04	1.64	.84	.95	.70	3.21	1.29
s_{repeat}	.017	.024	.019	.009	.011	.009	.006	.007
s_{repro}	.011	.013	.002	.023	.018	.013	.024	.004
B wm2								
mean lnk	.77	-.12	1.57	.77	.86	.67	3.21	1.20
s_{repeat}	.032	.055	.025	.032	.046	.020	.055	.036
s_{repro}	.009	.061	.030	.021	.072	.020	.014	.066
C wm2								
mean lnk	.88	-.03	1.71	.92	1.02	.80	3.35	1.38
s_{repeat}	.002	.007	.003	.001	.007	.003	.012	.006
s_{repro}	.003	.018	.008	.011	.012	.016	.032	.020
D wm2								
mean lnk	.80	-.08	1.62	.84	.96	.72	3.28	1.31
s_{repeat}	.003	.007	.002	.002	.008	.003	.012	.005
s_{repro}	.033	.047	.024	.035	.033	.043	.055	.019
E wm2								
mean lnk	.83	-.03	1.65	.89	.97	.78	3.29	1.34
s_{repeat}	.005	.009	.003	.005	.012	.004	.018	.009
s_{repro}	.017	.047	.012	.033	.018	.024	.040	.005
F wm2								
mean lnk	.80	-.09	1.65	.87	.95	.74	3.28	1.31
s_{repeat}	.004	.007	.003	.004	.011	.015	.010	.006
s_{repro}	.061	.054	.037	.045	.065	.078	.061	.064

TABLE IV.7 (continued).

S.Phase	Solute							
	MHB	NBZ	PBL	PE	PHB	PRE	PRS	TOL
A wm2								
mean lnk	.51	1.12	.44	.68	2.14	1.47	1.87	2.54
s_{repeat}	.009	.012	.015	.011	.005	.012	.013	.004
s_{repro}	.008	.016	.010	.006	.008	.027	.008	.019
B wm2								
mean lnk	.42	1.07	.35	.54	2.06	1.43	1.84	2.47
s_{repeat}	.048	.027	.058	.085	.034	.025	.030	.020
s_{repro}	.064	.014	.089	.136	.059	.003	.015	.020
C wm2								
mean lnk	.59	1.18	.54	.72	2.24	1.52	1.94	2.64
s_{repeat}	.007	.008	.007	.007	.007	.011	.003	.007
s_{repro}	.019	.010	.013	.014	.022	.009	.018	.018
D wm2								
mean lnk	.53	1.11	.46	.65	2.17	1.45	1.87	2.56
s_{repeat}	.007	.006	.011	.006	.005	.005	.005	.006
s_{repro}	.020	.033	.031	.038	.020	.036	.038	.015
E wm2								
mean lnk	.56	1.13	.50	.68	2.19	1.48	1.93	2.57
s_{repeat}	.012	.010	.014	.012	.008	.009	.009	.006
s_{repro}	.007	.010	.018	.002	.005	.006	.058	.013
F wm2								
mean lnk	.52	1.12	.48	.66	2.16	1.45	1.88	2.61
s_{repeat}	.006	.004	.015	.009	.006	.007	.009	.007
s_{repro}	.070	.059	.066	.075	.062	.090	.040	.018

Legend: see text.

TABLE IV.8. Results of PCA on the unfolded data cube

a) results for the stationary phases

S.Phase	mean	s^2_{repro}	MS_{bef}	$MS_{res}(1)$	$MS_{res}(2)$
A	0.0337	2.25	30.56	26.69	0.28
B	-0.0693	31.36	82.66	0.50	0.41
C	0.0282	2.89	22.60	10.86	8.45
D	-0.0163	11.56	8.90	8.30	3.26
E	0.0093	6.25	7.35	4.24	3.62
F	0.0144	38.44	15.14	7.93	7.23
average	0	15.46	27.87	9.76	3.87

Legend: MS is the abbreviation of Mean Sum of Squares. MS_{bef} is the mean sum of squares before the PCA calculation. $MS_{res}(k)$ is the Mean Sum of Squared residuals after applying k components in the model. The cumulative percentages explained Sum of Squares of the first two components are 65.0% and 86.1%. All numbers, except the mean values, must be multiplied by 10^{-4}.

b) scores of the stationary phases on the first two Principal Components

	PC1	PC2
A	0.24	-0.62
B	-1.09	-0.04
C	0.41	0.19
D	-0.09	0.27
E	0.21	0.10
F	0.32	0.10

TABLE IV.8 (continued).

c) results for the solutes

Solute	s^2_{repro}	MS_{bef}	$MS_{res}(1)$	$MS_{res}(2)$
ACP	8.85	32.38	6.11	2.15
ACT	19.25	35.60	9.87	3.93
ANS	5.10	20.90	3.95	2.41
CRE	9.05	22.06	3.93	2.71
DMP	18.82	18.83	5.84	3.02
EAB	15.56	31.58	8.31	5.37
EE	16.90	42.83	29.51	9.88
EHB	15.42	19.64	4.81	2.37
MHB	16.45	26.01	3.34	1.81
NBZ	8.70	16.90	4.33	1.85
PBL	23.05	52.43	26.04	8.68
PE	43.00	20.41	5.46	2.99
PHB	13.83	24.25	6.93	3.23
PRE	17.09	26.37	18.60	3.90
PRS	11.70	31.83	15.25	5.34
TOL	3.01	23.87	3.89	2.30
average	15.46	27.87	9.76	3.87

d) results for the mobile phases

M.Phase	s^2_{repro}	MS_{bef}	$MS_{res}(1)$	$MS_{res}(2)$
wm1		16.97	10.68	5.30
wm2	15.46	20.97	5.77	3.19
wm3		10.17	8.80	2.09
am1		35.34	16.77	9.61
am2		20.50	10.21	3.17
am3		10.32	2.56	1.70
wa1		41.39	18.85	4.44
wa2		35.33	4.83	2.00
wa3		59.83	9.37	3.35
average	15.46	27.87	9.76	3.87

Legend: see Table IV.8a.

TABLE IV.9. PARAFAC results

a) results for the stationary phases

S.Phase	s^2_{repro}	MS_{bef}	$MS_{res}(1)$	$MS_{res}(2)$
A	2.25	30.56	22.11	8.25
B	31.36	82.66	14.76	10.78
C	2.89	22.60	13.07	10.11
D	11.56	8.90	7.09	3.96
E	6.25	7.35	5.36	4.87
F	38.44	15.14	10.39	8.80
average	15.46	27.87	12.13	7.79

Legend: see Table IV.8, the cumulative percentages explained sum of squares of the first two components are 56.5% and 72.0%.

b) scores of the stationary phases on the first two PARAFAC components

	(1)	(2)
A	-0.52	1.56
B	-3.34	3.51
C	1.70	-2.07
D	0.25	-0.74
E	0.71	-0.83
F	1.19	-1.43

TABLE IV.9 (continued).

c) results for the solutes

Solute	s^2_{repro}	MS_{bef}	$MS_{res}(1)$	$MS_{res}(2)$
ACP	8.85	32.38	8.12	5.60
ACT	19.25	35.60	13.33	9.57
ANS	5.10	20.90	4.70	3.26
CRE	9.05	22.06	6.21	4.67
DMP	18.82	18.83	5.83	3.98
EAB	15.56	31.58	10.28	7.74
EE	16.90	42.83	40.29	30.63
EHB	15.42	19.64	5.80	3.98
MHB	16.45	26.01	4.71	3.22
NBZ	8.70	16.90	4.14	3.26
PBL	23.05	52.43	29.43	20.64
PE	43.00	20.41	6.49	5.11
PHB	13.83	24.25	7.12	4.22
PRE	17.09	26.37	20.87	6.54
PRS	11.70	31.83	20.90	8.14
TOL	3.01	23.87	5.84	4.13
average	15.46	27.87	12.13	7.79

d) results for the mobile phases

M.Phase	s^2_{repro}	MS_{bef}	$MS_{res}(1)$	$MS_{res}(2)$
wm1	15.46	16.97	13.76	12.76
wm2		20.97	8.02	5.59
wm3		10.17	8.94	3.13
am1		35.34	17.94	12.12
am2		20.50	16.32	14.53
am3		10.32	4.23	3.82
wa1		41.39	17.83	5.51
wa2		35.33	4.62	3.63
wa3		59.83	17.50	9.04
average	15.46	27.87	12.13	7.79

Legend: see Table IV.9a.

TABLE IV.9 (continued).

e) loadings of the solutes on the first two PARAFAC components

	(1)	(2)
ACP	-0.48	0.22
ACT	-0.46	0.21
ANS	-0.33	0.13
CRE	-0.32	0.13
DMP	-0.35	0.16
EAB	-0.39	0.16
EE	-0.32	0.19
EHB	-0.33	0.14
MHB	-0.37	0.14
NBZ	-0.32	0.14
PBL	-0.49	0.22
PE	-0.36	0.16
PHB	-0.40	0.18
PRE	-0.41	0.24
PRS	-0.47	0.26
TOL	-0.33	0.12

Legend: all numbers must be multiplied by 10^{-1}.

f) loadings of the mobile phases on the first two PARAFAC components

	(1)	(2)
wm1	-0.81	1.01
wm2	-0.90	0.55
wm3	-1.42	2.78
am1	-2.62	4.17
am2	-1.39	2.35
am3	-1.20	1.69
wa1	-3.38	5.65
wa2	-2.58	3.58
wa3	-1.61	0.95

TABLE IV.10. Analysis of variance (ANOVA)

a) one-way ANOVA: solutes that differ between the stationary phases A,B and C

	A	B
B	ACP,ANS,CRE,NBZ PRE,TOL	
C	ACP,ANS,CRE,EAB EE,NBZ,PHB,PRE PRS,TOL	ALL SOLUTES

b) two-way ANOVA: solutes that differ between the stationary phases A,B and C

Water/methanol	Water/methanol/acetonitrile	Water/acetonitrile
CRE	ANS,DMP,EHB,MHB PBL,PHB,TOL	ACP,ACT,ANS,CRE DMP,EAB,EHB,MHB NBZ,PE,PHB,TOL

TABLE IV.11. Choice of the markers

a) markers of the version I calibration

markers	percentage explained variation
ANS,DMP,EE,PRE	99.833
DMP,EE,PRE,TOL	99.830
ANS,DMP,EE,PRS	99.828
DMP,EE,PRS,TOL	99.825
ANS,EAB,EE,PRE	99.825

b) markers of the version II calibration

markers	percentage explained variation
ANS,DMP,PBL,PRE	99.901
DMP,PBL,PRE,TOL	99.899
DMP,PBL,PHB,PRE	99.896
ANS,DMP,PBL,PHB	99.896
ANS,EAB,PBL,PRE	99.893

TABLE IV.12. Results of the version I and II calibrations

a) root mean squared errors of prediction of the solutes

solute	s_{repro}	RMSEP I	RMSEP II
ACP	.003	.036	.042
ACT	.018	.049	.044
ANS	.008		
CRE	.011	.037	.026
DMP	.012		
EAB	.016	.037	.028
EE	.032		.126
EHB	.020	.040	.035
MHB	.019	.054	.040
NBZ	.010	.040	.044
PBL	.013	.070	
PE	.014	.036	.043
PHB	.022	.045	.054
PRE	.009		
PRS	.018	.071	.077
TOL	.018	.071	.084
average	.017	.051	.060

b) root mean squared errors of prediction at the mobile phases

mobile phase	RMSEP I	RMSEP II
wm1	.045	.045
wm2	.055	.066
wm3	.034	.049
am1	.072	.091
am2	.041	.050
am3	.047	.051
wa1	.053	.063
wa2	.047	.061
wa3	.054	.051
average	.051	.060

Legend: The RMSEP values are calculated as $\surd[\Sigma(y_i - \hat{y}_i)^2/n]$, where the summation index is $i = 1,..,n$.

TABLE IV.12 (continued).

c) observed versus predicted k values for some solutes

	observed	predicted		
ACP		version I	version II	blind
wm1	1.49	1.52	1.51	1.47
wm2	0.88	0.89	0.88	0.80
wm3	0.24	0.27	0.24	0.27
am1	1.89	1.88	1.92	1.80
am2	1.13	1.17	1.19	1.21
am3	0.61	0.62	0.64	0.56
wa1	1.88	1.89	1.88	1.88
wa2	1.28	1.24	1.23	1.21
wa3	0.74	0.66	0.66	0.55
TOL				
wm1	26.33	24.46	24.31	22.86
wm2	14.08	12.72	12.54	12.28
wm3	7.08	6.64	6.44	7.01
am1	35.11	35.97	37.46	31.45
am2	17.81	17.24	17.55	16.54
am3	9.28	9.77	9.89	8.72
wa1	42.58	37.91	37.42	39.73
wa2	20.72	19.36	19.00	19.26
wa3	10.01	10.55	10.42	8.89

TABLE IV.13. Marker and mobile phase selection with induced-variance

a) selection of the markers

St.Phase left out	markers	explained variation
A	ANS,DMP,EE,PRS	99.852
	DMP,EE,PRS,TOL	99.848
	ANS,EAB,EE,PRS	99.839
	EAB,EE,PRS,TOL	99.838
	DMP,EE,PHB,PRS	99.829
B	ANS,DMP,EE,PRS	99.878
	DMP,EE,PRS,TOL	99.873
	DMP,PBL,PRS,TOL	99.870
	ANS,DMP,PBL,PRS	99.866
	ANS,EAB,EE,PRS	99.865
C	ANS,DMP,EE,PRS	99.831
	DMP,EE,PRS,TOL	99.829
	ANS,EAB,EE,PRS	99.816
	EAB,EE,PRS,TOL	99.816
	DMP,EE,PHB,PRS	99.809
D	ANS,DMP,EE,PRS	99.858
	DMP,EE,PRS,TOL	99.853
	ANS,EAB,EE,PRS	99.843
	DMP,EE,PHB,PRS	99.841
	EAB,EE,PRS,TOL	99.839
E	ANS,DMP,EE,PRS	99.832
	DMP,EE,PRS,TOL	99.827
	ANS,EAB,EE,PRS	99.816
	DMP,EE,PHB,PRS	99.815
	EAB,EE,PRS,TOL	99.814
F	ANS,DMP,EE,PRS	99.832
	DMP,EE,PRS,TOL	99.827
	ANS,EAB,EE,PRS	99.823
	EAB,EE,PRS,TOL	99.818
	DMP,EE,PHB,PRS	99.809

TABLE IV.13. (continued).

a) mobile phase selection

St.Phase left out	mobile phases	explained variation
A	wm1,wm3,wa1,wa3	99.856
	wm3,am1,wa1,wa3	99.852
	wm3,am1,wa2,wa3	99.835
	wm1,wm3,wa1,wa2	99.826
	wm3,am1,am2,wa2	99.825
B	wm1,wm3,wa1,wa3	99.877
	wm3,am1,wa1,wa3	99.858
	wm3,am1,wa2,wa3	99.845
	wm1,wm3,wa1,wa2	99.841
	wm2,am1,wa1,wa3	99.834
C	wm1,wm3,wa1,wa3	99.857
	wm3,am1,wa1,wa3	99.839
	wm3,am1,wa2,wa3	99.835
	wm1,wm3,am1,wa3	99.828
	wm2,wm3,am1,wa2	99.824
D	wm1,wm3,wa1,wa3	99.848
	wm3,am1,wa1,wa3	99.837
	wm3,am1,wa2,wa3	99.825
	wm2,wm3,am1,wa2	99.824
	wm1,wm3,wa1,wa2	99.823
E	wm1,wm3,wa1,wa3	99.843
	wm3,am1,wa1,wa3	99.821
	wm3,am1,wa2,wa3	99.807
	wm2,wm3,wa1,wa3	99.800
	wm1,wm3,wa1,wa2	99.799
F	wm1,wm3,wa1,wa3	99.840
	wm3,am1,wa1,wa3	99.825
	wm3,am1,wa2,wa3	99.808
	wm1,wm3,wa1,wa2	99.805
	wm2,wm3,am1,wa2	99.799

TABLE IV.14. Results of the unfold-PLS predictions: one component

a) results for the stationary phases

S.Phase	s^2_{repro}	%X[MaMPh] expl.	%X[NMaMPh] expl.	MS_{mark}	%MARK	MS_{toex}	MS_{res}
A	2.25	79.5	78.3	59.32	0.2	42.09	40.84 (3.0)
B	31.36	58.3	50.7	123.3	0.6	118.5	119.9 (---)
C	2.89	62.2	68.1	24.83	26.8	33.51	21.98 (34.4)
D	11.56	63.6	67.6	16.71	0.0	12.33	12.33 (0.0)
E	6.25	60.5	65.3	11.04	24.0	10.53	6.55 (37.8)
F	38.44	58.8	68.0	41.48	49.3	19.34	15.77 (18.5)
average	15.46	63.8	66.3	46.11	16.8	39.38	36.23 (8.0)

Legend: the values s^2_{repro}, MS_{mark}, MS_{toex}, and MS_{res} must be multiplied by 10^{-4}. The value in parenthesis under MS_{res} is the percentage of explained variation in the test set. When this value becomes negative, this is reported as ---. See also legend Table III.59.

b) root mean squared errors of prediction for the mobile phases

	St.Phase					
	A	B	C	D	E	F
M.Phase						
wm1	.048 (---)	.068 (10.6)	.053 (32.7)	.031 (---)	.016 (27.7)	.031 (---)
wm2	.029 (---)	.095 (4.1)	.062 (38.8)	.014 (0.1)	.024 (48.4)	.042 (---)
wm3	.067 (0.8)	.027 (---)	.020 (---)	.047 (0.0)	.024 (---)	.015 (58.0)
am1	.077 (1.7)	.104 (---)	.065 (36.8)	.052 (0.0)	.029 (31.3)	.081 (---)
am2	.064 (---)	.085 (1.9)	.048 (1.9)	.034 (---)	.032 (36.0)	.020 (62.1)
am3	.033 (10.3)	.073 (---)	.029 (42.3)	.018 (0.0)	.021 (---)	.019 (32.5)
wa1	.104 (6.4)	.136 (---)	.057 (---)	.048 (0.1)	.023 (49.9)	.020 (89.3)
wa2	.065 (12.2)	.145 (---)	.022 (69.3)	.033 (0.1)	.021 (58.0)	.030 (47.3)
wa3	.067 (---)	.180 (2.3)	.038 (73.1)	.020 (---)	.034 (47.6)	.043 (71.1)
average	.064	.110	.047	.035	.026	.040

TABLE IV.14 (continued).

c) root mean squared errors of prediction of the solutes

	S.Phase					
	A	B	C	D	E	F
solute						
ACP	.065	.134	.043	.030	.013	.038
	(9.1)	(---)	(49.4)	(0.0)	(61.2)	(---)
ACT	.071	.139	.049	.028	.025	.046
	(5.9)	(---)	(27.0)	(---)	(55.9)	(---)
ANS	.032	.099	.055	.030	.015	.022
	(16.9)	(---)	(39.8)	(0.1)	(---)	(29.8)
CRE	.035	.110	.043	.021	.020	.028
	(13.0)	(---)	(38.2)	(0.0)	(39.3)	(45.7)
DMP	.049	.090	.042	.015	.019	.053
	(8.6)	(---)	(37.9)	(0.1)	(34.9)	(---)
EAB	.042	.126	.051	.031	.033	.036
	(6.4)	(0.3)	(17.6)	(0.0)	(41.1)	(64.2)
EE	.120	.101	.046	.041	.049	.074
	(---)	(14.1)	(41.9)	(---)	(32.5)	(22.1)
EHB	.048	.098	.040	.026	.018	.036
	(8.0)	(---)	(44.0)	(0.1)	(13.2)	(---)
MHB	.040	.124	.039	.022	.016	.034
	(13.5)	(---)	(46.1)	(0.1)	(63.1)	(25.1)
NBZ	.052	.090	.039	.029	.012	.024
	(8.7)	(---)	(37.3)	(0.1)	(---)	(10.7)
PBL	.099	.124	.056	.070	.052	.050
	(---)	(4.5)	(32.3)	(0.0)	(40.4)	(55.7)
PE	.048	.102	.044	.032	.017	.033
	(9.8)	(---)	(19.9)	(0.0)	(61.6)	(---)
PHB	.058	.106	.046	.027	.024	.038
	(6.4)	(---)	(40.8)	(0.0)	(---)	(21.2)
PRE	.101	.073	.051	.054	.017	.042
	(---)	(---)	(5.0)	(0.0)	(41.7)	(---)
PRS	.070	.065	.052	.036	.027	.059
	(1.2)	(---)	(---)	(0.0)	(24.6)	(---)
TOL	.034	.109	.051	.024	.019	.016
	(9.7)	(---)	(48.0)	(0.1)	(---)	(85.5)
average	.064	.110	.047	.035	.026	.040

TABLE IV.15. Results of the unfold-PLS predictions: two components

a) results for the stationary phases

S.Phase	s^2_{repro}	%X[MaMPh] expl.	%X[NMaMPh] expl.	MS_{mark}	%MARK	MS_{toex}	MS_{res}
A	2.25	93.0	89.7	59.32	---	42.09	41.99 (0.2)
B	31.36	84.4	68.5	123.3	17.0	118.5	113.0 (4.6)
C	2.89	92.4	91.4	24.83	28.8	33.51	21.04 (37.2)
D	11.56	88.8	85.7	16.71	41.6	12.33	6.64 (46.1)
E	6.25	87.4	86.7	11.04	52.9	10.53	6.93 (34.2)
F	38.44	90.5	89.9	41.48	54.9	19.34	15.11 (21.9)
average	15.46	89.4	85.3	46.11	32.5	39.38	34.12 (13.4)

Legend: see Table IV.14a.

b) root mean squared errors of prediction for the mobile phases

	St.Phase					
	A	B	C	D	E	F
M.Phase						
wm1	.053 (---)	.089 (---)	.050 (40.9)	.019 (64.0)	.014 (42.9)	.039 (---)
wm2	.033 (---)	.110 (---)	.060 (42.2)	.020 (---)	.021 (61.1)	.048 (---)
wm3	.069 (---)	.017 (52.8)	.018 (---)	.028 (65.1)	.014 (58.2)	.023 (---)
am1	.070 (18.3)	.122 (---)	.071 (24.3)	.037 (49.4)	.033 (11.2)	.072 (---)
am2	.069 (---)	.077 (20.8)	.043 (20.3)	.021 (63.1)	.037 (18.4)	.026 (34.8)
am3	.032 (19.1)	.074 (---)	.030 (35.6)	.016 (25.4)	.019 (---)	.016 (48.0)
wa1	.109 (---)	.097 (44.5)	.050 (---)	.030 (60.0)	.033 (---)	.028 (80.1)
wa2	.064 (13.4)	.132 (13.3)	.023 (66.8)	.027 (36.2)	.028 (29.4)	.025 (64.5)
wa3	.064 (5.2)	.169 (14.7)	.034 (78.4)	.030 (---)	.024 (73.7)	.036 (79.2)
average	.065	.106	.046	.026	.026	.039

TABLE IV.15 (continued).

c) root mean squared error of prediction for the solutes

	St.Phase					
	A	B	C	D	E	F
solute						
ACP	.064 (10.9)	.135 (---)	.044 (47.7)	.017 (66.8)	.014 (50.3)	.035 (---)
ACT	.073 (---)	.140 (---)	.046 (36.6)	.023 (33.8)	.032 (26.0)	.047 (---)
ANS	.030 (25.5)	.100 (---)	.057 (35.5)	.025 (32.3)	.015 (---)	.021 (37.3)
CRE	.036 (7.3)	.106 (6.8)	.043 (38.5)	.018 (22.2)	.021 (34.0)	.031 (37.9)
DMP	.046 (16.8)	.104 (---)	.043 (33.1)	.008 (74.7)	.022 (8.6)	.051 (---)
EAB	.047 (---)	.108 (26.4)	.049 (23.5)	.027 (24.5)	.037 (27.3)	.040 (55.6)
EE	.118 (---)	.099 (16.2)	.044 (47.0)	.014 (88.0)	.032 (71.7)	.064 (41.2)
EHB	.048 (7.9)	.100 (---)	.041 (42.9)	.017 (57.5)	.016 (31.7)	.037 (---)
MHB	.041 (11.4)	.121 (3.4)	.036 (54.8)	.015 (50.8)	.018 (55.5)	.036 (16.0)
NBZ	.052 (8.7)	.087 (---)	.040 (34.8)	.020 (54.7)	.012 (---)	.023 (18.8)
PBL	.103 (---)	.103 (34.0)	.051 (45.7)	.060 (25.4)	.052 (41.1)	.052 (52.3)
PE	.051 (0.3)	.098 (3.4)	.043 (23.6)	.025 (38.3)	.024 (18.9)	.036 (---)
PHB	.057 (8.1)	.108 (---)	.047 (38.5)	.013 (75.4)	.019 (15.4)	.037 (21.3)
PRE	.102 (---)	.070 (3.3)	.049 (13.2)	.030 (68.8)	.019 (30.2)	.035 (10.7)
PRS	.070 (2.7)	.070 (---)	.053 (---)	.027 (42.9)	.036 (---)	.053 (---)
TOL	.035 (5.7)	.106 (5.0)	.051 (47.9)	.020 (34.3)	.020 (---)	.017 (82.7)
average	.065	.106	.046	.026	.026	.039

TABLE IV.16. Closer examination of stationary phase B: unfold-PLS

solute

	ACP			ACT		
	obs	pred	error	obs	pred	error
wm1	1.4500	1.4838	-.0337	.4779	.5163	-.0384
wm2	.7707	.8553	-.0846	-.1211	-.0348	-.0863
wm3	.2302	.2435	-.0133	-.6038	-.5840	-.0199
am1	1.7237	1.8539	-.1302	.6514	.8049	-.1535
am2	1.1754	1.1513	.0241	.2427	.1569	.0858
am3	.5051	.6014	-.0964	-.4646	-.3493	-.1153
wa1	1.7816	1.8835	-.1018	.4932	.6123	-.1191
wa2	1.1184	1.2635	-.1451	-.1904	.0252	-.2156
wa3	.3862	.6987	-.3126	-.6838	-.4336	-.2502

	MHB		
	obs	pred	error
wm1	1.1765	1.2437	-.0672
wm2	.4223	.5569	-.1346
wm3	-.1554	-.1502	-.0052
am1	1.4303	1.5971	-.1668
am2	.6753	.7653	-.0900
am3	.0212	.1462	-.1251
wa1	1.2363	1.3415	-.1052
wa2	.4162	.5847	-.1685
wa3	-.1388	-.0063	-.1325

TABLE IV.17. Results of the PARAFAC predictions: one component

a) results for the stationary phases

S.Phase	s^2_{repro}	R^2_{train}	MS_{mark}	%MARK	MS_{toex}	MS_{res}
A	2.25	.66	59.32	20.4	42.09	36.74 (12.7)
B	31.36	.38	123.3	0.1	118.5	116.2 (2.0)
C	2.89	.60	24.83	18.6	33.51	23.49 (29.9)
D	11.56	.59	16.71	10.2	12.33	9.93 (19.4)
E	6.25	.58	11.04	0.5	10.53	9.68 (8.1)
F	38.44	.60	41.48	23.9	19.34	15.60 (19.3)
average	15.46	.57	46.11	12.3	39.38	35.27 (10.4)

Legend: see Tables III.60 and IV.14.

b) root mean squared errors of prediction for the mobile phases

	St.Phase					
M.Phase	A	B	C	D	E	F
wm1	.060 (---)	.072 (---)	.059 (18.1)	.036 (---)	.019 (1.0)	.028 (---)
wm2	.059 (---)	.097 (---)	.063 (37.2)	.014 (9.4)	.032 (11.2)	.036 (---)
wm3	.063 (11.9)	.024 (6.7)	.021 (---)	.044 (12.9)	.022 (---)	.016 (48.6)
am1	.057 (45.9)	.101 (3.8)	.061 (44.0)	.041 (37.4)	.033 (7.2)	.085 (---)
am2	.052 (33.2)	.085 (2.0)	.058 (---)	.029 (25.2)	.040 (3.8)	.026 (34.8)
am3	.024 (53.3)	.070 (2.5)	.027 (51.0)	.015 (31.2)	.019 (---)	.018 (34.1)
wa1	.071 (55.6)	.127 (5.7)	.062 (---)	.036 (45.1)	.030 (13.2)	.027 (81.6)
wa2	.033 (76.9)	.140 (3.2)	.019 (78.3)	.026 (37.7)	.030 (14.5)	.025 (64.5)
wa3	.108 (---)	.183 (---)	.035 (77.6)	.036 (---)	.044 (11.5)	.037 (78.9)
average	.061	.108	.048	.032	.031	.039

TABLE IV.17 (continued).

c) root mean squared error of prediction for the solutes

	St.Phase					
	A	B	C	D	E	F
solute						
ACP	.041	.131	.046	.025	.018	.034
	(63.4)	(1.6)	(41.1)	(29.9)	(19.3)	(---)
ACT	.056	.136	.055	.035	.035	.043
	(40.2)	(2.0)	(9.4)	(---)	(13.1)	(---)
ANS	.023	.096	.054	.017	.014	.020
	(56.1)	(2.6)	(41.9)	(67.0)	(4.9)	(42.8)
CRE	.040	.109	.041	.022	.024	.027
	(---)	(1.7)	(43.9)	(---)	(8.6)	(50.7)
DMP	.033	.087	.039	.011	.023	.047
	(57.9)	(3.1)	(45.4)	(44.8)	(7.3)	(---)
EAB	.056	.125	.051	.031	.041	.038
	(---)	(1.6)	(19.2)	(---)	(10.1)	(58.6)
EE	.112	.108	.056	.040	.059	.093
	(7.2)	(0.4)	(14.1)	(4.6)	(---)	(---)
EHB	.045	.095	.041	.021	.019	.026
	(18.2)	(2.7)	(42.9)	(34.5)	(6.4)	(48.0)
MHB	.042	.122	.036	.023	.024	.027
	(6.4)	(1.8)	(54.8)	(---)	(16.1)	(52.9)
NBZ	.038	.086	.038	.022	.011	.020
	(51.3)	(3.2)	(38.6)	(43.9)	(12.7)	(38.5)
PBL	.121	.126	.061	.062	.065	.045
	(---)	(2.3)	(19.7)	(21.4)	(7.8)	(63.3)
PE	.032	.099	.047	.027	.025	.025
	(61.1)	(2.0)	(9.0)	(29.3)	(15.4)	(21.8)
PHB	.051	.103	.045	.024	.021	.034
	(26.8)	(2.5)	(43.4)	(22.6)	(---)	(34.7)
PRE	.088	.072	.055	.048	.023	.049
	(22.6)	(---)	(---)	(18.3)	(2.3)	(---)
PRS	.054	.062	.050	.029	.030	.067
	(42.3)	(4.2)	(6.3)	(36.4)	(8.7)	(---)
TOL	.044	.108	.053	.017	.015	.013
	(---)	(1.8)	(43.0)	(50.1)	(1.6)	(90.9)
average	.061	.108	.048	.032	.031	.039

TABLE IV.18. Results of the PARAFAC predictions: two components

a) results for the stationary phases

S.Phase	s^2_{repro}	R^2_{train}	MS_{mark}	%MARK	MS_{toex}	MS_{res}
A	2.25	.77	59.32	35.4	42.09	47.39 (---)
B	31.36	.60	123.3	27.7	118.5	79.81 (32.6)
C	2.89	.77	24.83	43.3	33.51	22.59 (32.6)
D	11.56	.73	16.71	62.4	12.33	7.18 (41.8)
E	6.25	.74	11.04	25.8	10.53	9.43 (10.4)
F	38.44	.77	41.48	31.2	19.34	15.53 (19.7)
average	15.46	.73	46.11	37.6	39.38	30.32 (23.0)

Legend: see Table IV.17a.

b) root mean squared errors of prediction for the mobile phases

	St.Phase					
	A	B	C	D	E	F
M.Phase						
wm1	.101 (---)	.057 (37.4)	.058 (20.4)	.036 (---)	.019 (1.0)	.033 (---)
wm2	.083 (---)	.036 (85.9)	.058 (47.5)	.019 (---)	.024 (50.9)	.045 (---)
wm3	.050 (43.9)	.030 (---)	.019 (---)	.028 (63.7)	.014 (57.9)	.019 (26.0)
am1	.071 (17.1)	.095 (15.0)	.072 (20.6)	.028 (70.4)	.035 (---)	.081 (---)
am2	.076 (---)	.111 (---)	.056 (---)	.029 (24.4)	.046 (---)	.024 (46.9)
am3	.026 (43.8)	.047 (56.5)	.028 (46.0)	.014 (43.2)	.019 (---)	.018 (34.1)
wa1	.074 (52.6)	.135 (---)	.053 (---)	.031 (58.1)	.037 (---)	.027 (81.6)
wa2	.028 (83.3)	.110 (40.7)	.017 (81.8)	.030 (17.7)	.031 (10.3)	.022 (71.6)
wa3	.083 (---)	.116 (59.4)	.023 (90.4)	.022 (---)	.035 (45.1)	.033 (83.7)
average	.069	.089	.048	.027	.031	.039

TABLE IV.17 (continued).

c) root mean squared error of prediction for the solutes

	St.Phase					
	A	B	C	D	E	F
solute						
ACP	.059	.103	.044	.015	.017	.035
	(23.8)	(39.6)	(46.0)	(74.0)	(32.5)	(---)
ACT	.062	.106	.052	.027	.036	.047
	(26.7)	(40.6)	(19.9)	(11.1)	(9.3)	(---)
ANS	.053	.072	.054	.025	.016	.023
	(---)	(44.4)	(41.9)	(30.7)	(---)	(22.5)
CRE	.056	.089	.040	.015	.023	.030
	(---)	(33.6)	(48.0)	(46.3)	(14.8)	(42.0)
DMP	.053	.069	.042	.008	.024	.048
	(---)	(38.0)	(38.2)	(72.2)	(---)	(---)
EAB	.080	.100	.050	.024	.040	.040
	(---)	(36.7)	(21.0)	(38.6)	(17.0)	(55.5)
EE	.084	.145	.057	.043	.061	.089
	(48.2)	(---)	(10.6)	(---)	(---)	(---)
EHB	.053	.069	.041	.019	.017	.028
	(---)	(47.9)	(42.9)	(49.2)	(17.8)	(37.5)
MHB	.059	.096	.034	.017	.021	.030
	(---)	(39.2)	(59.1)	(40.8)	(35.0)	(41.8)
NBZ	.046	.063	.038	.024	.013	.023
	(28.0)	(47.6)	(38.6)	(34.4)	(---)	(21.7)
PBL	.130	.088	.060	.059	.063	.043
	(---)	(51.7)	(24.4)	(28.3)	(14.4)	(67.6)
PE	.040	.071	.045	.024	.027	.030
	(36.8)	(50.0)	(16.3)	(43.5)	(4.2)	(---)
PHB	.064	.078	.046	.014	.018	.035
	(---)	(44.2)	(40.4)	(72.4)	(22.7)	(30.2)
PRE	.082	.094	.052	.028	.017	.044
	(33.5)	(---)	(2.7)	(72.5)	(43.7)	(---)
PRS	.066	.090	.057	.024	.036	.060
	(13.3)	(---)	(---)	(54.8)	(---)	(---)
TOL	.058	.079	.051	.021	.018	.013
	(---)	(47.0)	(48.1)	(26.6)	(---)	(90.9)
average	.069	.089	.048	.027	.031	.039

TABLE IV.18. Closer examination of stationary phase B: PARAFAC: two components

solute

	ACP			ACT		
	obs	pred	error	obs	pred	error
wm1	1.4500	1.3927	.0573	.4779	.4029	.0750
wm2	.7707	.7420	.0286	-.1211	-.1355	.0144
wm3	.2302	.2130	.0172	-.6038	-.5928	-.0110
am1	1.7237	1.6936	.0301	.6514	.6470	.0044
am2	1.1754	1.2120	-.0367	.2427	.2134	.0293
am3	.5051	.5503	-.0453	-.4646	-.3902	-.0744
wa1	1.7816	1.9087	-.1270	.4932	.6470	-.1550
wa2	1.1184	1.2285	-.1101	-.1904	.0022	-.1926
wa3	.3862	.6268	-.2406	-.6838	-.5176	-.1662

	EAB			EE		
	obs	pred	error	obs	pred	error
wm1	1.4553	1.4679	-.0126	4.6937	4.5584	
wm2	.6730	.6666	.0064	3.2084	3.2099	-.0015
wm3	.0283	.0027	.0256	2.1779	2.0962	
am1	1.9292	1.9126	.0167	5.1355	4.8986	.2369
am2	1.1309	1.2586	-.1276	3.3031	3.5184	-.2153
am3	.4877	.4737	.0139	2.3321	2.3003	.0319
wa1	1.9411	2.0981	-.1570	4.5353	4.4994	
wa2	1.0624	1.2305	-.1681	2.7957	2.8354	-.0398
wa3	.3814	.5221	-.1407	1.5071	1.5965	

Legend: all numbers are ln k values.

TABLE IV.19. PARAFAC with alternative marker/mobile phase combinations

a) results of the predictions

S.Phase	Markers	Mobile Phases	RMSEP
A	ACP,EAB,PBL,PHB	wm1,wm2,am1,am3	.076
B	ACP,PBL,PRE,PRS	wm1,wm2,am1,wa3	.118
C	ACP,PBL,PRE,PRS	wm3,am1,wa1,wa2	.054
D	ACP,ACT,PRE,PRS	am1,am2,wa1,wa2	.034
E	ACP,PBL,PRE,PRS	wm3,am1,wa1,wa2	.037
F	ACP,ACT,PRE,PRS	wm3,am1,wa1,wa2	.042
average			.067

b) loadings of the solutes and the mobile phase compositions on the two PARAFAC components when stationary phase B is left out

	(1)	(2)
ACP	1.036	0.978
ACT	0.977	0.949
ANS	0.874	0.915
CRE	0.727	0.792
DMP	0.911	0.938
EAB	0.820	0.922
EE	1.074	0.814
EHB	0.873	0.925
MHB	0.719	0.780
NBZ	0.903	0.872
PBL	1.240	1.482
PE	0.842	0.803
PHB	1.022	1.077
PRE	1.545	1.357
PRS	1.318	1.250
TOL	0.719	0.812

	(1)	(2)
wm1	1.035	-1.223
wm2	0.982	-1.224
wm3	0.744	-0.411
am1	2.097	-1.962
am2	-0.248	0.539
am3	0.699	-0.634
wa1	0.489	0.089
wa2	0.730	-0.494
wa3	0.832	-1.022

Summary

Multivariate techniques used in this thesis are described in Part I. Techniques to select markers are of special interest. Quantitative methods like the induced-variance- and determinant criterion work well in practice. Other techniques - DISNORM, Procrustus analysis and variants - are worth trying. Another important issue is the use of three-way methods. These methods provide a general framework for thinking about calibration problems.

Part II gives recent relevant developments in the field of stationary- and mobile phase optimisation of reversed-phase systems; calibration in GC, TLC and RP-HPLC and related miscellaneous topics. One of the important conclusions of Part II is the idea that correction of retention values to compensate for changing measurement conditions is performed best with a set of compounds similar to the compounds of interest. The calibration strategies presented in Chapter 8 rely on this principle: reference standards (markers) are selected which are specific for the separation problem at hand.

The calibration strategies, presented in Chapter 8, are divided in two groups: the two- and three-way approaches. If a stationary phase is conceived as an object, then a training set of retention values on, at least, five stationary phases is needed to perform the three-way approach. If the training set is smaller, the two-way approaches have to be used.

The two-way approaches have two versions. The first version tries to model the relationship between retention values of markers and of non-markers in the training set. This relationship is used subsequently to predict the retention of non-markers on a new stationary phase using the retention values of markers on that new stationary phase. This version is tested in Parts III and IV. The second version tries to model the relationship between retention values of markers on the initial stationary phase(s) and the new one(s). Predictions of retention values of non-markers on the new stationary phase(s) can be obtained using this relation and the measured retention values of the non-markers on the initial phase(s). This second version has not been tested.

Both tested three-way strategies bear the same characteristics. The training set can be represented by a data cube in which a stationary phase, the object, is characterised by the capacity factors of solutes obtained at different mobile phase compositions. This data cube is decomposed. On a new stationary phase, retention values of the markers have to be measured at a limited number of mobile phase compositions. Predictions of the retention values of the non-markers at the mobile phase compositions used in the training set, can be obtained using the previously developed decomposition. The same holds for the retention values of the markers at the non-selected mobile phase compositions.

It is important to understand clearly the differences between the two- and three-way approaches. These differences can be explained keeping in mind two aspects.

First, the two-way approaches differ from the three-way approaches with respect to the experimental effort in the training- and calibration step. The three-way methods need a large training set whereas

the two-way methods do not. In the calibration step only a few measurements are needed to calibrate a new stationary phase if a three-way approach is chosen. On the contrary, with a two-way approach more measurements are needed in the calibration step.
Second, the two-way approaches differ from the three-way approaches with respect to the way in which the mobile phase composition is handled. The calibration with two-way approaches is discussed firstly. If predictions of retention values on a new stationary phase are desired at a specific mobile phase composition, the new stationary phase has to be calibrated by measuring the retention values of the markers at that mobile phase composition. This particular mobile phase composition is not necessarily one of the mobile phases used in the training set (the initial stationary phases). For the three-way approaches the situation is different. Predictions of retention values on a new stationary phase can only be obtained at the mobile phase compositions present in the training set. However, for the calibration of a new stationary phase and contrary to the two-way case, it is not necessary to measure the retention values of the markers at each mobile phase composition on that new stationary phase. Marker retention values at a small number of selected mobile phase compositions suffice to calibrate the whole new stationary phase in the three-way case.

In Part III, the first version of the two-way approach and both three-way approaches are tested. Of this first version of the two-way approach, two different variants are used. One variant uses the mobile phase compositions explicitly in the model, contrary to the second variant. A training set of retention measurements of nine test solutes on a C1, a C18 and a CN stationary phase at six mobile phase compositions (mixtures of water, acetonitrile and methanol) is used. Retention is predicted on a C6, a C8 and a Phenyl stationary phase. For detailed discussions and conclusions, reference is made to the respective sections. The results of the two-way approaches are summarized and discussed firstly.

Four different sets of markers are evaluated: markers selected with the induced-variance criterion; selected with the determinant criterion; a homologous series and bad markers. The design matrices of these four marker-sets differed considerably with respect to the degree of multicollinearity. The design matrix of the homologous markers has a very high degree of multicollinearity, the design matrix of the bad markers and the markers chosen with the induced-variance criterion have a high degree of multicollinearity. The design matrix of the determinant markers has a moderate degree of multicollinearity.

The predictions based on the models where the induced-variance- and determinant markers are used are good: relative prediction errors of the capacity factors are between 5 and 10%. The homologous- and bad markers performed clearly worse. The predictive performance of the induced-variance- and determinant markers does not differ much. Both marker-choice criteria are sensitive to outliers; the retention of the solute paracetamol is badly predictable and therefore this solute can be regarded as an outlier. However, both the induced-variance- and determinant criterion select this solute as a marker.

The degree of multicollinearity seems to affect the performance of cross-validation. The selection of the k and c parameters in respectively ridge- and Stein regression with cross-validation leads to better results for the determinant markers (a lower degree of multicollinearity) than for the induced-variance markers. If the induced-variance markers are used, Hoerl's choice of the k parameter is better than the cross-validatory choice. The cross-validatory choice of the estimation method is also slightly better for a lower degree of multicollinearity. The selection of a model with- or without the explicit mobile phase compositions, is performed better with Amemiya's prediction criterion than cross-validation for the induced-variance markers. If the determinant markers are used, cross-validation performs slightly better in this respect. Again there seems to be a relation between multicollinearity and the performance of cross-validation.

There is no clear preference for an estimation method. If the test set (the new stationary phase) does not resemble the training set, e.g. with respect to the pattern of multicollinearity, OLS might give higher prediction errors than ridge regression, Stein regression or partial least squares. Criteria to judge the similarity between training- and test set are important. With respect to multicollinearity, such a criterion is proposed and seems to work reasonable. Yet, specific interactions between a solute, a mobile phase and a new stationary phase on which retention prediction is desired may cause high prediction errors if these specific interactions are not present in the training stage.

The average relative prediction error for a capacity factor when predicted by a three-way model, is 13%. It ranges from 3.6% (EHB on C8) to 35% (TOL on CN) for the unfold-PLS model. Although no clear difference in predictive performance was noticed between both three-way models (PARAFAC and unfold-PLS), it is worthwhile developing validation criteria with which a choice can be made in practice. Besides, more three-way models are available, but not tested in this thesis. A direct comparison between the two- and three-way methods is difficult because of the above mentioned differences between the two- and three-way approaches. On the one hand, the two-way approaches seem to predict better, but use more measurements to calibrate a new stationary phase. On the other hand, more measurements are used to build the calibration model with in the three-way case. One of the problems in the three-way calibration is the presence of non-linear behaviour of retention with respect to mixing mobile phase components. Retention of a solute measured at a ternary mixture is not the mean of the retention values of that solute at the two binary mobile phases which are mixed fifty-fifty to make the ternary mobile phase composition. Another problem which arises is drift in the measurements. During the training stage, the stationary phases changed and consequently drift in the measurements was observed. How this effects the performance of three-way (and two-way) models is not yet clear.

Three-way models reckoning with non-linear mixing behaviour should be developed. It is also worthwhile developing three-way models that explicitly account for the influence of the mobile phase constituents. If such models are available, the prediction of retention at a continuous range of mobile phase compositions is possible. This is of great importance for optimisation of separations and correction strategies.

In Part IV a data set is used consisting of retention measurements of sixteen test solutes on six octadecyl stationary phases from different batches at nine mobile phase compositions (mixtures of water, methanol and acetonitrile). Three different analyst/apparatus combinations are used to build up this training set. The differences between three of these octadecyl stationary phases, all three of them measured by a different analyst/apparatus combination, are visualized using analysis of variance. It appears that the retention values of the solutes on the three stationary phases differ significantly. The differences between the stationary phases with respect to the retention values depend on the mobile phase composition. These three different stationary phases are chosen to test a two-way approach.

Two stationary phases are chosen as the training set and the third stationary phase is used as test set. The prediction of capacity factors on this third stationary phase is performed with an average relative prediction error of 5-6%. A two-way variant in which no calibration measurements have to be performed on the new stationary phase, gives an average relative prediction error of 9%. This is worse than the value of 5-6% above, because the reproducibility is about 3%. The value of 9% can be regarded as the prediction error on a new stationary phase if *a priori* knowledge of the new stationary phase is not available. Prediction errors should be judged keeping in mind that small prediction errors may disturb completely a chromatogram. Very good predictions are needed to predict a separation correctly.

The application of three-way models for the calibration of the octadecyl (C18) stationary phases was not completely successful. All six stationary phase were used to evaluate the three-way approaches with. The first problem is the selection of a combination of solutes (markers) and mobile phase compositions which together are capable of calibrating a new stationary phase and predicting retention of all other solutes at all other mobile phase compositions. Statistical techniques for the simultaneous selection of variables from two categories, as in the three-way case, are not available. These techniques have to be developed. A "quick and dirty" approach based on the induced-variance is used in Part III and gives adequate results. In Part IV, such an approach is also used and performes better than an alternative strategy of variable selection.

Especially the aspect of the different analyst/apparatus combinations influences the performance of the three-way models. This aspect seems to hamper unfold-PLS more than PARAFAC. This may be due to the more rigid model structure of PARAFAC. Two solutions for the problem of different analyst/apparatus combinations in the training set (and perhaps in the test set) are outlined. The first idea is to use a different kind of centering and scaling in the data cube. The second idea is to make hybrid models: models with an MANOVA aspect to account for the differences between analyst/apparatus combinations and latent variable three-way models to account for the differences between the stationary phases.

Both three-way models are sensitive to outliers. The test solutes comprised benzene derivatives, some steroids and phenobarbital. Some of the test solutes - the steroids and phenobarbital - were badly predictable and showed deviating behaviour in the three-way models. The connection between the degree of heterogeneity of the set of solutes and the performance of the three-way models should be investigated. Of utmost importance are diagnostic tools to evaluate the three-way models. Some diagnostic tools are tested and seem to work well.

The cause of the differences between the PARAFAC- and unfold-PLS method with respect to their predictive performance is not clear. Unfold-PLS is perhaps more flexible, but PARAFAC uses a lower number of degrees of freedom to estimate the model parameters. Validation criteria to assess the performance of both three-way methods should be tested.

Samenvatting voor niet-ingewijden

De vloeistofchromatograaf is een modern instrument waarmee het gehalte van een stof in een medium kan worden bepaald. Denk bv. aan organische verbindingen in water, geneesmiddelresten in bloed en urine, etc.

De probleemstelling die in dit proefschrift aan bod komt, valt in twee delen uiteen.

Ten eerste: de vloeistofchromatograaf moet ingesteld worden. De kunst is om de vloeistofchromatograaf optimaal te laten functioneren. Dit is het eerste deel van de probleemstelling: in de vloeistofchromatograaf kunnen veranderingen worden aangebracht die de prestatie van het instrument beïnvloeden. Deze **bedoelde** veranderingen moeten optimaal aangewend worden.

Een probleem bij het gebruik van de vloeistofchromatograaf is dat een onderdeel van dat apparaat bij gebruik veroudert. De instelling van het apparaat is niet langer optimaal na een periode van (intensief) gebruik. Er zijn **onbedoelde** veranderingen opgetreden. Dus moet de vloeistofchromatograaf opnieuw ingesteld worden. Dit is het tweede gedeelte van de probleemstelling.

Zowel het aanwenden van de **bedoelde** veranderingen als het corrigeren voor **onbedoelde** veranderingen worden in dit proefschrift *ijken (calibratie)* genoemd.

Dit proefschrift is opgebouwd uit vier delen. De eerste twee delen zijn inleidingen: Deel I behandelt de statistiek die nodig is voor het oplossen van bovengenoemde problemen en Deel II geeft achtergrond informatie over de probleemstelling. Deel III behandelt een voorbeeld van het aanwenden van **bedoelde** verschillen om de vloeistofchromatograaf optimaal in te stellen. Deel IV, tenslotte, behandelt een voorbeeld van het opnieuw instellen van de vloeistofchromatograaf na **onbedoelde** veranderingen.

Eerst een korte schets van Deel I. Een object (bv. "een spijker") kan worden gekarakteriseerd door eraan te meten. Als er één kenmerk wordt gebruikt om het object te karakteriseren (bv. "lengte van de spijker") is één meting voldoende: het object wordt univariaat gekarakteriseerd. De statistiek die zich bezighoudt met dit soort metingen heet "*univariate* statistiek". Een object zou ook kunnen worden beschreven door meerdere kenmerken tegelijkertijd (bv. "lengte", dikte" en "gewicht" van de spijker). Er moeten nu meerdere metingen worden gedaan die in onderlinge samenhang het object beschrijven. Het moge duidelijk zijn dat deze "multivariate" aanpak vaak meer zegt over het object. De statistiek die zich bezighoudt met de analyse van dergelijke multivariate metingen heet "*multivariate* statistiek".

Deel I bevat een kort overzicht van de multivariate technieken die gebruikt zijn in het onderzoek, zoals gerapporteerd in de delen III en IV. Een belangrijk gedeelte van Deel I is de beschrijving van de methoden die in staat zijn de kenmerken (*variabelen*) te selecteren die het meest zeggen over de objecten.

In de hoge-druk vloeistofchromatografie wordt gebruik gemaakt van een vloeibare fase, die onder hoge druk door een kolom gevuld met een vaste fase wordt gepompt. De vaste fase bestaat uit kleine poreuse korreltjes waarbij een coating aan het oppervlak is gehecht. De vloeibare fase is een vloeistof: bv. een mengsel van water en een aantal organische oplosmiddelen.

Aan het begin van de kolom wordt in de vloeistofstroom een mengsel van te analyseren stoffen geïnjecteerd. Elke stof in dit mengsel heeft een bepaalde affiniteit met de vaste- en de vloeibare fase. Deze affiniteit verschilt per stof. Als een stof graag verblijft in de vaste fase, duurt het lang voor zo'n stof de kolom verlaat. De stof ondervindt veel vertraging (veel *retentie*). Een stof die weinig affiniteit vertoont met de vaste fase, maar liever verblijft in de vloeibare fase verlaat de kolom snel (weinig retentie). Door gebruikmaking van de verschillen in affiniteit (het affiniteits-patroon) kan een geïnjecteerd mengsel van stoffen zo goed mogelijk in de tijd van elkaar gescheiden met de vloeistofstroom de kolom verlaten. Door het manipuleren van de vaste- en vloeibare fase (het maken van **bedoelde** verschillen) en derhalve van het affiniteits-patroon van een groep te scheiden stoffen, kan getracht worden dit mengsel van stoffen te scheiden. Veel werk is verricht aan het systematisch manipuleren van de vloeibare fase. Aan het systematisch manipuleren van de vaste fase daarentegen nog weinig.

Een probleem bij het gebruik van hoge-druk vloeistofchromatografie is de slechte reproduceerbaarheid van de vaste fase. Als de vaste fase versleten is, moet deze vernieuwd worden, maar deze nieuwe vaste fase heeft niet precies dezelfde eigenschappen als de vorige. Dus zijn de omstandigheden van voorheen niet langer optimaal. Het vervangen van de oude vaste fase door een nieuwe leidt tot **onbedoelde** verschillen. De nieuwe vaste fase moet worden geijkt.

Deel II geeft een beschrijving van een aantal nieuwe ontwikkelingen op het gebied van de hoge-druk vloeistofchromatografie. In Hoofdstuk 8 van Deel II worden verschillende ijk-strategieën geformuleerd, die specifiek toegesneden zijn op de hierboven gesignaleerde problemen van de **bedoelde** en **onbedoelde** verschillen. Deze ijk-strategieën bevatten twee ingrediënten. Ten eerste: de keuze van speciale referentie stoffen (*markers*) gekozen uit het mengsel van te analyseren stoffen, met behulp waarvan de toestand van een systeem, bestaande uit een vaste- en vloeibare fase, zo goed mogelijk kan worden gekarakteriseerd. Ten tweede: *modellen* (wiskundige formules) worden verondersteld die een verband leggen tussen het gedrag van de markers en de niet-markers (de overige stoffen in het te analyseren mengsel) in de systemen van vaste- en vloeibare fasen. Met behulp van enkele zorgvuldig gekozen metingen aan de markers kan de status van een nieuwe vaste fase (of breder: van het nieuwe systeem) worden vastgesteld. Het gedrag van de stoffen ánders dan de markers, kan dan worden voorspeld op het nieuwe systeem. Er worden twee rivaliserende soorten modellen beschreven. Elk soort model wordt op zijn bruikbaarheid getest.

Deel III behandelt een voorbeeld van **bedoelde** variatie: het systematisch manipuleren van vaste- en vloeibare fase om daarmee een scheiding van een gegeven mengsel in zijn componenten te bewerkstelligen. Een aantal methoden om de markers te selecteren wordt getoetst. Deze methoden blijken beter te werken dan het lukraak kiezen van markers of, zoals voorheen te doen gebruikelijk, het gebruik maken van een speciale groep stoffen uit dezelfde familie (homologe reeks). Modellen worden geëvalueerd die het gedrag van de niet-markers op drie vaste fasen (de *training set*) in relatie tot het gedrag van de markers op diezelfde vaste fasen beschrijven. Het gedrag van de niet-markers op drie andere vaste fasen (de *test set*) kan dan worden voorspeld door de retentie van de markers te meten en door gebruik te maken van het model. Dit resulteert in een relatieve voorspelfout van 5-10%. Deze voorspelfout is heel redelijk te noemen daar de relatieve fout in een retentie-meting ongeveer 3% bedraagt.

Voordat een model gebruikt kan worden moeten eerst een aantal onbekenden (*parameters*) uitgerekend (*geschat*) worden. De metingen van de retentiewaarden van alle stoffen op alle combinaties van de vloeibare fasen en de drie vaste fasen in de training set kunnen worden gerangschikt in een tabel (*matrix* van getallen). Door geschikte manipulatie van deze matrix kunnen de parameters uitgerekend worden. Eén van de statistische problemen bij het schatten van de parameters is dat het gedrag van de markers op de vaste fasen een zo sterke onderlinge samenhang vertoont, dat de parameters slechts met grote onzekerheid kunnen worden uitgerekend (*multicollineariteit*). Deze multicollineariteit resulteert niet persé in slechte voorspellingen. Zodra dit dreigt, kunnen er schattingsmethoden worden gebruikt die deze dreiging het hoofd bieden. Natuurlijk moet ook de betrouwbaarheid van het model (en de schattingsmethode) worden vastgesteld: de modellen en schattingsmethoden moeten worden gevalideerd. Een in chemometrische kringen in zwang zijnde validatiemethode (*cross-validation*) wordt getoetst en blijkt niet altijd betere resultaten te geven dan andere in de statistiek bekende methoden. Speciaal het probleem van de multicollineariteit gooit roet in het eten.

Bij de modellen zoals hierboven beschreven werden de gegevens gerangschikt in een matrix en de bijbehorende modellen zijn *twee-weg* modellen. Een ander soort model kan worden gebruikt door de gegevens in een driedimensionale tabel te rangschikken. Het resultaat is nu een "kubus" van getallen: een data kubus. *Drie-weg* methoden zijn modellen die proberen zo goed mogelijk de getallen in de data kubus te benaderen. Ook voor deze modellen geldt weer dat er parameters geschat moeten worden. Het voordeel van de drie-weg modellen is dat er heel weinig metingen nodig zijn om een nieuwe vaste fase te ijken. Er zijn maar enkele metingen van de markers op de nieuwe vaste fase nodig om het gedrag van de niet-markers te voorspellen. Een aantal van deze drie-weg methoden wordt getest op hun bruikbaarheid. Het blijkt mogelijk drie-weg modellen te maken die voorspellen met een relatieve voorspelfout van 7-20%. Het nadeel van de drie-weg modellen is dat ze een grote hoeveelheid meetgegevens vergen voor het uitrekenen van de parameters: er is een grote training set nodig.

Deel IV behandelt een voorbeeld van **onbedoelde** variatie: een aantal vaste fasen die vrijwel identiek behoren te zijn, verschillen qua eiegenschappen doordat ze in verschillende productiegangen gemaakt zijn. Ook hier blijk het gedrag van de stoffen op een aantal van deze vaste fasen goed te modelleren met behulp van de twee-weg modellen. In elk geval voldoende om ook hier het gedrag op andere vaste fasen goed te voorspellen.

Het gebruik van de drie-weg methoden wordt ook hier getest. Een speciaal probleem bij deze drie-weg methoden is de selectie van de variabelen. In geval van de twee-weg modellen zijn er statistische methoden beschreven voor deze selectie. Dergelijke methoden zijn er niet of nauwelijks voor de drie-weg modellen. Twee geïmproviseerde variabele selectie methoden worden gebruikt, waarbij de ene methode iets betere resultaten oplevert dan de andere. Het voorspellen van de retentie waarden van de niet-markers verloopt niet zo gunstig als in Part III. De reden hiervoor is dat niet op elke vaste fase metingen zijn verricht door dezelfde analist op hetzelfde apparaat. Dit levert verschillen op die niet direct te corrigeren zijn. Oplossingen hiervoor worden aangegeven en suggesties voor verder onderzoek worden gedaan.

Dankwoord

Bij een dankwoord moet er een scheiding worden gemaakt tussen wie wél en wie niet bedankt wordt. Dit is een moeilijke opgave, omdat ik mij soms niet eens realiseer in hoeverre ik beïnvloed wordt door bepaalde mensen. Laat staan dat ik kan zeggen wie mij heeft geholpen bij het tot stand komen van dit proefschrift. Het is dan ook niet zo dat ik diegenen die niet met name genoemd zijn, geen dank verschuldigd ben, integendeel. Toch kan ik van een aantal mensen de bijdrage en steun duiden en dezen wil ik dan ook expliciet bedanken.

Mijn promotor, Prof. Durk Doornbos, ben ik zeer erkentelijk voor zijn moed om nieuwe wegen te bewandelen. Het vertrouwen dat hij in mij heeft, stel ik zeer op prijs.

Mijn referent, Pierre Coenegracht, heeft mij ingewijd in de geheimen van de chromatografie. Tevens is hij een kritisch lezer van mijn manuscripten. Voor beide aspecten ben ik hem veel dank verschuldigd.

De leden van de promotiecommissie, Prof. Paul Geladi, Prof. D.L. Massart en Prof. T.J. Wansbeek bedank ik voor hun bereidheid dit proefschrift te beoordelen.

A special thanks goes to Prof. Svante Wold, Umeå, Sweden. During my stay at his laboratory in december 1986, I learned a lot from him. I remember especially our discussions on biased estimation versus maximum likelihood methods which opened my eyes. On the occasions that I have met him after my first visit, he was always willing to discuss statistical problems.

Het leeuwendeel van het experimentele werk is uitgevoerd door Chris Bruins. De nauwgezetheid waarmee hij dit werk aanpakte, maakte dat ik met een gerust hart de gegevens kon gebruiken voor de vele berekeningen.

De leden (en ex-leden) van de vakgroep Analytische Chemie en Toxicologie bedank ik voor het feit dat ik mij, ondanks mijn afkomst, nooit een vreemde eend in de bijt voel.

Al zwemmend, volleyballend, voetballend en voetbalkijkend heb ik veel collega's van het Universitair Centrum voor Farmacie ontmoet. Ik ervaar deze contacten altijd als erg plezierig. Hetzelfde geldt voor de collega's die ik heb onmoet in het bestuurlijk werk dat ik in de loop van de tijd gedaan heb.

In een (te) laat stadium van mijn onderzoek heb ik Ton Steerneman erbij betrokken. Zijn kennis van de multivariate statistiek en zijn bereidheid om zich in problemen te verdiepen die niet direct op zijn pad liggen, waardeer ik zeer. Bovendien stel ik zijn hartelijkheid en vaderlijke raadgevingen zeer op prijs.

De bijvakstudenten Frank Wolbert en Jos Everts dank ik voor hun bijdrage aan het experimentele werk van Part IV. Alhoewel niet direct betrokken bij mijn onderzoek, heb ik de samenwerking met de bijvakstudenten Anne Knevelman en Paul de Wolf op prijs gesteld. Met Piet Hein van der Graaf heb ik vele gesprekken over mijn onderzoek gehad. Zijn kritisch luisterend oor en zijn snelheid van werken waren voor mij zeer stimulerend.

De wereld van de chemometrie is klein. De collega's in den lande die ik vaak tegenkom op congressen of anderszins, dank ik voor het plezierige contact dat ik met hun heb.

Jacques Duitsch verzorgde de tekeningen in Part II; de dames van het secretariaat, Jolanda en Marina, voorzagen mij altijd snel van nieuwe uitdraaien van de laserprinter die toch vaak weer kladversies bleken; Anita Sleurink nam een gedeelte van de PARAFAC berekeningen op zich; Liesbeth Jager controleerde een gedeelte van de tekst op het Engels en Wouter Daane ontwierp de voorkant van dit proefschrift. Allen bedank ik hartelijk.

Mijn ouders dank ik voor het feit dat ze me in de gelegenheid hebben gesteld een studie te voltooien. Iemand die een proefschrift schrijft, lijdt aan een ernstig syndroom: bewustzijnsvernauwing. Behalve voor een gedeelte van het monnikenwerk dat ze voor haar rekening nam, ben ik Eugenie dan ook zeer dankbaar voor het feit dat ze wil samenleven met iemand met bovengenoemd syndroom.

Age Smilde