Order Number 9315514

New approaches using chemometric methods: Categorical regression analysis and comparison of maximum-minimum distance clustering with other methods

Wu, Jinan, Ph.D.

City University of New York, 1993

U·M·I
300 N. Zeeb Rd.
Ann Arbor, MI 48106

NEW APPROACHES USING CHEMOMETRIC METHODS: CATEGORICAL REGRESSION ANALYSIS AND COMPARISON OF MAXIMUM-MINIMUM DISTANCE CLUSTERING WITH OTHER METHODS

by

JINAN WU

A dissertation submitted to the Graduate Faculty in Chemistry in partial fulfillment of the requirements for the degree of Doctor of Philosophy, The City University of New York.

1993

This manuscript has been read and accepted by the Graduate Faculty in Chemistry in satisfaction of the dissertation requirement for the degree of Doctor of Philosophy.

1/27/93
Date

Daryl G. Howery
Chairman of Examining Committee

1/27/93
Date

Executive Officer

Daryl G Howery

Gary Mennitt

Daniel C. Locke
Supervisory Committee

The City University of New York

ABSTRACT

NEW APPROACHES USING CHEMOMETRIC METHODS: CATEGORICAL REGRESSION ANALYSIS AND COMPARISON OF MAXIMUM-MINIMUM DISTANCE CLUSTERING WITH OTHER METHODS

by

Jinan Wu

Advisor: Prof. Darryl G. Howery

Two new chemometric methodologies, categorical regression analysis and maximum-minimum distance cluster analysis, are applied to a number of data sets from chemistry. Computer programs with documentation are written for both methods.

Independent variables in the models for categorical regression analysis (CRA) consist of 1's (substance has the property) and 0's (substance lacks the property) only. Comparing input data with calculated values, reasonable (and in some problems excellent) models were developed for a variety of data including retention indices, boiling points, dissociation constants and rate constants. The

simplicity of CRA models is a major advantage of the approach. Applications are limited to problems in which the substances are structurally similar.

MRA-CRA combination models for retention-index problems are developed. MRA-CRA models combine the advantage of flexibility and increased statistical significance.

Assignment of data vectors to a selected number of clusters is accomplished using maximum-minimum (max-min) distance comparisons. Clusters from the max-min method are compared to clusters from three other approaches: hierarchical cluster analysis, varimax-rotated factor analysis and target factor analysis. Several GLC retention-index matrices and TLC R_f-value matrices were studied. Nearly equivalent clusters are obtained using the various methods for the simpler GLC problems. For the more complicated GC problems and especially for the TLC problem (which involved mixed solvents), confusingly different clusters are obtained from the different methods.

Acknowledgement

My feeling of fulfillment and satisfaction upon completing this dissertation can be matched only by the gratitude I feel toward the people who helped bring it to fruition.

I especially thank Professor Darryl G. Howery, my mentor, for generously lending his time, scientific expertise and editorial know-how to this thesis.

I would like to thank Professors P. Gary Mennitt, Richard D. Pizer and David C. Locke, members of my examining and advisory committee, for their time and advice.

Finally, specially thanks my husband Q. L. Fan and my children, Yujia and Yuxing, for their support and rooting for me to the finish.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Chemometrics is the chemical discipline that uses mathematical and statistical methods to design optimal procedures and to provide maximum chemical information by analyzing chemical data [1]. Chemometrics involves the application of mathematics, statistics and computer science to classify, interpret, and predict data of chemical interest. Through chemometric analysis, data become more meaningful because additional information is extracted from measurements.

The development of chemometrics involved three stages [2]. In the first stage, data analysis methods from other fields were used. For example, most of the major methods of chemometrics were first formulated outside the physical sciences. During the 1960's, computers began to play a key role. In the second stage, covering the decade of the 1970's, methods were adapted to meet the specific needs of chemists. During the third stage, the decade of the 1980's, chemometrics matured and was introduced into the classroom. The recent publishing of two journals devoted exclusively to the field, "The Journal of Chemometrics" and "Chemometrics and Intelligent Laboratory Systems", indicates the coming of age of chemometrics.

The number of research papers and the number of books on chemometrics has increased rapidly. Some of the chemometrics books are text books, such as "Chemometrics: a text book" by Massart et al. [1], "Chemometrics" by Kowalski et al. [3] and "Practical Guide to Chemometrics" edited by Haswell [4]. Others are research monographs, such as "Applied Regression Analysis" by Draper and Smith [5], "Factor Analysis in Chemistry" by Malinowski and Howery [6], "The Interpretation of Analytical Chemical Data by the Use of Cluster Analysis" by Massart et al. [7] and "Pattern Recognition in Chemistry" by Varmuza [8]. Many applications of chemometrics in chemistry have been published in journals such as the Journal of Chromatography and Analytical Chemistry. A total of nine general reviews of chemometrics in Analytical Chemistry have been published since 1974 [9].

Most chemical problems are multivariate, and often large-scale computations are involved. Computer software has stimulated the development of chemometrics. The available software for chemometrics, both commercial and user-generated software, increases in number and quality. One of the most useful programs is the Statistical Analysis System (SAS) package [10-14], which can be carried out on IBM PC, Macintosh and mainframe computers.

Chemometrics includes more than fifteen subjects [9]. Four of

the most widely used chemometric methods are multiple regression analysis (MRA), factor analysis (FA), pattern recognition and cluster analysis. Other techniques, such as optimization, calibration and signal processing, have also proved useful in chemistry.

Applications of the above chemometric techniques in chemistry, biosciences and the environmental sciences are growing continuously in breadth and in depth. In the next four paragraphs, we give a few examples of recent applications of chemometrics to illustrate the scope of chemometrics.

Factor analysis (FA) has been applied in GC retention data on different mobile phases to explain retention characteristics [15]. FA has also been applied to study solute strength and solvent selectivity for reverse phase HPLC [16] and to study separation of components from overlapped peaks [17]. Factor analytical methods have been used to analyze three dimensional data recently, such as 2D-NMR [18] and overlapped ICP-MS responses [19]. Target FA has been used to calculate missing data and to analyze multicomponent mixtures in UV and IR spectra [20,21].

Relating, correlating and modeling measured responses involve applications of least-squares regression analysis. A tremendous number of papers have been published in the field. The simplest application, linear regression analysis, is widely used in chemistry to find

calibration curves. Multivariate regression, non-linear regression and partial regression analysis are applied in various fields, such as near IR spectroscopy [22] and the pharmaceutical industry [23]. Quantitative structure-active relationships (QSAR), which pinpoint relationships between molecular structure and chemical, physical or biological activities, use regression techniques. For example, numerous compounds have been studied for structure-retention parameter relationships in TLC [24], HPLC [25] and GC [26].

Pattern recognition, which includes cluster analysis and discriminant analysis, is a widely used chemometric technique. For example, pattern recognition techniques have been applied to classify objects, such as the sources of pollution [27], different polymer formulations [28] and chemical sensor arrays [29].

Optimization of instrumental response has been used in routine analysis. Many papers utilize optimization for automated wavelength selection in multicomponent analysis [30] and for selecting experimental conditions in chromatographic separations [31].

In our research work, two new chemometric methodologies: categorical regression analysis (CRA) and maximum-minimum distance clustering analysis (max-min), are applied to a number of data sets from chemistry. This thesis consists of two parts. Part I covers the first approach, categorical regression analysis, in Chapters 2

through 11. Part II involves the second approach, maximum-minimum distance clustering analysis, in Chapters 12 through 18.

In Chapter 2, we introduce the theory for categorical regression analysis. In Chapter 3, a simple CRA model is detailed to illustrate the application of the CRA method to a chemical problem. In Chapter 4, the statistical evaluation of CRA models using four criteria is introduced. The details of the procedures and the documentation for the computer program for CRA are given in Chapter 5. In Chapters 6 through 10, applications of the CRA method in several chemical fields are investigated. In Chapter 11, ordinary and categorical combination models are developed.

In Chapter 12, three methods for cluster analysis are introduced, and the principles of the max-min method are demonstrated in detail. The procedures of the max-min method and program documentation are given in Chapter 13. In Chapters 14 through 18, applications of the clustering methods to retention-index problems and TLC problems are investigated.

Part I

Categorical Regression Analysis

Chapter 2

Theory of Categorical Regression Analysis

Multiple regression analysis (MRA) [5,32], a well-known statistical method, plays an important role in building models for the properties of chemical substances. In MRA, the dependent variable is a physical or chemical property, and the independent variables can be decimal numbers (usually quantifying various characteristics of the substances) or ordinal numbers (usually signifying the number of a specified structural unit in the substances). For example, suppose one wants to build a model for the boiling points of several alcohols. A vector of the measured boiling points is the dependent variable in the model. The independent variables are related to the properties of the alcohols, such as the number of branched chains. In order to represent such independent variables with quantitative values, several parameters, such as the Wiener number and the connectivity index [33], are often used as independent variables in MRA models.

Categorical regression analysis (CRA) [34], involving categorized variables, is a modified form of the MRA method. Here, the independent variables are restricted to certain values. Such categorized variables have the advantage of simplicity and in some cases can be applied when the quantitative information required for standard MRA is not available.

The independent variables in a CRA model are categorized with values of either "1" or "0". "1" means that the substance belongs to that particular category; "0" means that the substance does not belong to the category. In the model, the properties are related to the structures of the substances. Each property can contain several categories. The sum of values of the category variables must equal to unity for each property of each substance, since a substance can belong to only one category in a suggested property. This restriction is unique to CRA. According to the restriction, at least two properties and two categories in each property are required for a CRA model.

Mathematically, a measured datum y_m (indexed by m = 1, 2, . . . , n for the n dependent-variable data being modeled) is taken to be a function of p properties (indexed by i = 1, 2, . . . , p). In turn, each property consists of c categories (indexed by j = 1, 2, . . . , c). Thus, each y_m is expressed as a linear combination of independent variables plus a residual e_m:

$$y_m = \sum_{i=1}^{p} \sum_{j=1}^{c} a_{ij} x_{mij} + e_m \qquad [1]$$

where each x_{mij} for the m-th substance, involving the j-th category of the i-th property, has a value of either "1" or "0". The a_{ij}'s, called category coefficients, are analogous to the standard multiple regression coefficients in MRA. According to the restriction in CRA, each property of each substance must obey the equation:

$$\sum_{j=1}^{c} x_{mij} = 1 \qquad \text{(m and i fixed)} \qquad [2]$$

Using matrix algebra, Eq. [1] can be written as:

$$Y = [X]\ A + E \qquad [3]$$

where Y is the vector of dependent-variable data being modeled, A is the vector of to-be-calculated category coefficients, [X] is the matrix of the x_{mij} independent variables consisting of 1's and 0's, and E is the vector of residuals.

According to the least-square principle, the coefficients can be found by minimizing the sum of the squared errors. Using Eq. [3] and denoting the transpose of a matrix by a prime, the square of the residuals becomes

$$\sum_{m=1}^{n} e_m^{\,2} = (Y - [X]\ A)'\ (Y - [X]\ A) \qquad [4]$$

Minimizing the sum of the squared errors by setting the derivative of the left-hand side of Eq. [4] with respect to the coefficients in A to zero, we get [1,35]:

$$A = ([X]'[X])^{-1} ([X]'Y) \quad [5]$$

where $([X]'[X])^{-1}$ is the inverse of ([X]' [X]). Eq. [5] is the standard equation for finding the multiple regression coefficients in vector A.

However, since categorized properties obey Eq. [2], the columns in [X] are not linearly independent. Thus, the determinant of [X]'[X] becomes zero and the regression-coefficient vector A then cannot be solved as it would be in MRA. The standard regression method fails whenever two or more properties are categorized. In order to obtain the category coefficients, a modification of MRA is required. A deletion method was introduced by Hayashi [36] to solve this problem. He proposed that a new matrix [X*] can be formed by removing the column for the first category of each property except for the first property. Simultaneously, a new vector A* for the category coefficients is formed by removing the coefficients for the deleted columns.

Eq. [5] then becomes

$$A^* = ([X^*]'[X^*])^{-1} ([X^*]'Y) \quad [6]$$

For this modified equation, the determinant of [X*]'[X*] does not vanish and A* can be calculated. Then the estimated values $\hat{Y}$ for the de-

pendent variables can be calculated as follows:

$$\hat{Y} = [X^*] \; A^* \qquad [7]$$

The elements in A* corresponding to the deleted columns are all equal to zero, i.e., $a_{21} = a_{31} = \ldots = a_{p1} = 0$, but a_{11} is not equal to zero. In order to obtain category coefficients having more useful physical interpretations, Hayashi showed that a normalized category coefficient for the j-th categorized property in the i-th property, a_{Nij}, can be calculated using a modified form of the a^*_{ij} from A:

$$a_{Nij} = a^*_{ij} - \sum_{j=1}^{c} a^*_{ij} \, f_{ij} \quad \text{(i fixed)} \qquad [8]$$

where f_{ij} is the fraction of 1's in the j-th column of the i-th property. Once the normalized coefficients are found, a CRA model results, since the estimated values Y for the dependent variables can be calculated as follows:

$$\hat{Y} = \bar{Y} + [X] \; A_N \qquad [9]$$

where $\bar{Y}$ is the mean value of the dependent variable in the model, A_N is a vector of the normalized category coefficients, and [X] is the matrix of the independent variables.

Eq [7] and [9] are mathematically equivalent ways to express the estimated values of the dependent variables in CRA models.

Chapter 3

An Example of a CRA Model

To demonstrate the mathematical concepts discussed in Chapter 2, we apply the CRA method to a simple chemical problem: the building of a categorical regression model for the boiling points of alcohols. We designate the problem as the 8 x 6 problem, since there are eight alcohols and a total of six categorized properties in this model. The size of the matrix of the independent variables is 8 x 6, i.e., eight rows and six columns are in matrix [X] in Eq. [3]. The vector of dependent variables is the boiling points of the eight alcohols.

The data and the CRA model are shown in Table 1. Two properties: total number of carbon atoms and type of alcohol, are selected in this model. Each property contains three categories. The first property contains three categories: 5, 6 and 7 carbon atoms. The second property contains three type categories: primary, secondary and tertiary alcohols (abbreviated as p-, s- and t-, respectively). The total number of independent variables (the categorized properties) is six in the model.

According to the restriction of CRA, categorized properties have values of either "1" or "0", and the sum of the categorized properties

Table 1. CRA model for the 8 x 6 problem involving boiling points of alcohols.

Properties (Basis of categories)

1 Total number of carbon atoms (5, 6, 7)
2 Type of alcohol (p-, s-, t-) *

		Categorized properties					
	Property	1			2		
	Category	1	2	3	1	2	3
Alcohol	B.P.						
1-pentanol	137.8	1	0	0	1	0	0
2-pentanol	119.0	1	0	0	0	1	0
2-methyl 2-butanol	102.0	1	0	0	0	0	1
1-hexanol	157.0	0	1	0	1	0	0
2-hexanol	139.9	0	1	0	0	1	0
2-methyl 2-pentanol	121.4	0	1	0	0	0	1
1-heptanol	176.2	0	0	1	1	0	0
2-methyl 2-hexanol	142.8	0	0	1	0	0	1

*) p - primary alcohol, s - secondary alcohol, t - tertiary alcohol.

must equal to 1 for each property of each substance. For example, 1-pentanol, which has 5 carbon atoms and is a primary alcohol, belongs to the first category of the first property and the first category of the second property. Thus, the first and fourth categorized properties of 1-pentanol have values of 1, while all the other categories have values of 0.

The matrix of independent variables [X] is shown in Table 2. After deleting the first column from the second property, which is the fourth categorized property, a new matrix [X*] (described in Chapter 2), also shown in Table 2, is formed. Then category coefficients A* can be calculated based on Eq. [6]:

Property	1 (No. carbon atoms)			2 (Type of alcohol)		
Category	1	2	3	1	2	3
Actual	5	6	7	p-	s-	t-
A*	137.1	156.9	177.0	0.0	-17.6	-34.9
A_N	-17.3	2.4	22.5	17.5	-0.1	-17.4

In the vector A*, the coefficient of the fourth category is 0, since that column was removed from the original matrix [X]. A coefficient having a value of zero usually is less useful chemically.

In order to obtain coefficients for a CRA model having more direct chemical interpretation, the normalized categorical coefficients can be calculated based on Eq. [8]. The computation of two

Table 2. Matrices [X] and [X*] for the 8 x 6 problem.

	[X]						[X*]				
Property	1			2			1			2	
Category	1	2	3	1	2	3	1	2	3	2	3
	1	0	0	1	0	0	1	0	0	0	0
	1	0	0	0	1	0	1	0	0	1	0
	1	0	0	0	0	1	1	0	0	0	1
	0	1	0	1	0	0	0	1	0	0	0
	0	1	0	0	1	0	0	1	0	1	0
	0	1	0	0	0	1	0	1	0	0	1
	0	0	1	1	0	0	0	0	1	0	0
	0	0	1	0	0	1	0	0	1	0	1

normalized category coefficients a_{11} and a_{21} in A_N above is illustrated as follows:

$a_{11} = 137.1 - [(137.1 \times 3/8) + (156.9 \times 3/8) + (177.0 \times 2/8)] = -17.3$

$a_{21} = 0 - [(0 \times 3/8) + (-17.6 \times 2/8) + (-34.9 \times 3/8)] = 17.5$

where the fractions f_{ij} of 1's in Eq. [8] are obtained from the j-th columns in matrix [X].

From the coefficients A* and A_N, we can see that the category coefficients increase with an increase in the number of carbon atoms. Also the coefficients decrease as the position of the alcohol group is changed from primary to secondary and tertiary. Since boiling points of alcohols are known to increase with an increase in the number of carbon atoms and to decrease as the alcohol group becomes less accessible, the model is in good agreement with chemical facts. The pattern in the normalized category coefficients is the same as that in the category coefficients.

The estimated values of the boiling points can be calculated using either Eq. [7] or Eq. [9]. The computations using both equations for 1-pentanol are as follows:

a) Using the category coefficients A* and matrix [X], the estimated boiling point (°C) for 1-pentanol is:

$$\hat{y}_{1\text{-pentanol}} = (137.1 \times 1) + (156.9 \times 0) + (177.0 \times 0) + (0.0 \times 1) + (-17.6 \times 0) + (-34.9 \times 0) = 137.1$$

b) Using the normalized category coefficients A_N and matrix [X], the estimated boiling point (°C) for 1-pentanol is:

$$\hat{y}_{1\text{-pentanol}} = 137.0 + (-17.4 \times 1) + (2.43 \times 0) + (22.46 \times 0) + (17.5 \times 1) + (-0.075 \times 0) + (-17.44 \times 0)$$
$$= 137.1$$

where the first term, 137.0, is the mean value of the boiling points of the eight alcohols in the model. The two results are exactly the same, showing the mathematical equivalence of the category coefficients and the normalized category coefficients in the CRA method.

Finally, the estimated values for the boiling points of all the alcohols (°C) in the 8 x 6 problem are:

Alcohol	Actual	Estimated	Residual
1-pentanol	137.8	137.1	0.7
2-pentanol	119.0	119.5	-0.5
2-methyl 2-butanol	102.0	102.2	-0.2
1-hexanol	157.0	156.9	0.1
2-hexanol	139.9	139.4	0.5
2-methyl 2-pentanol	121.4	122.0	-0.6
1-heptanol	176.2	176.9	-0.7
2-methyl 2-hexanol	142.8	142.0	0.8

The estimated values of the boiling points from the CRA model are very close to the actual boiling points. The residuals are slightly greater than the experimental error in the boiling point (about 0.2 - 0. 5 °C).

Chapter 4

Statistical Evaluation of CRA Models

Once a CRA model is developed, the statistical evaluation of the results becomes very important. We evaluate the precision and reliability of the models by four statistical criteria: the standard error of the residuals between actual and estimated values for the dependent variable (SE), the multiple correlation coefficient (R), the F-test value (F), and the partial correlation coefficients (R_p) [5].

The standard error, SE, in this study is expressed as:

$$SE = [\sum_{m=1}^{n} (\hat{y}_m - y_m)^2 / (n - 1)]^{1/2} \qquad [10]$$

where $\hat{y}_m$ and y_m are the estimated and the actual values of the dependent variable, respectively. The standard error is a measure of the dispersion of the residuals. The smaller the standard error, the better the model. To further evaluate a CRA model, the value of the standard error should be compared to the estimated experimental error.

The multiple correlation coefficient, R, is defined as

$$R = [\sum_{m=1}^{n} (\hat{y}_m - \bar{Y})^2 / \sum_{m=1}^{n} (y_m - \bar{Y})^2]^{1/2} \qquad [11]$$

where $\hat{y}_m$ is the estimated value, y_m is the actual value and $\bar{Y}$ is the

mean value of the dependent variable. The multiple correlation coefficient R measures how much the estimated values of the samples fit the regression plane obtained from the model. In other words, R measures the correlation between the actual values of the dependent variable and the estimated values. The value of the correlation coefficient can be between 0 and +1. A value of +1 signifies a perfect correlation between actual and estimated values. The larger the value of R, the better the model.

The F-test is a statistical criterion for measuring the significance of the regression. If the calculated F-value is larger than the value obtained from an F-distribution table at a selected confidence level (usually five percent), the model is a statistically acceptable model. The calculated F-test value is obtained from the following equation:

$$F = (S_{12} / df_1) / (S_{22} / df_2) \qquad [12]$$

where S_{12} is the variance of the regression, df_1 is the number of degrees of freedom of the model (equals the total number of categorized properties - the number of properties + 1), S_{22} is the variance of the error, and df_2 is the number of degrees of freedom for the total error in the model (equals the total number of substances in the model - df_1 - 1). F-distribution tables, calculated from an F-distribution function, can be found in most statistics textbooks [1]. The

five percent confidence level, which signifies that there is a five percent risk of reaching the wrong conclusion, is most usually used. In our studies, a two-side confidence level (total of five percent) is employed.

A partial correlation coefficient, R_{pi}, is defined as:

$$R_{pi} = [\ \sum_{m=1}^{n} (\ \hat{y}_{mi} - \bar{Y}\ )^2 / \sum_{m=1}^{n} (\ y_m - \bar{Y}\ )^2\ ]^{1/2} \qquad [13]$$

where R_{pi} is the partial correlation coefficient of the i-th property, $\hat{y}_{mi}$ is the estimated value calculated from category coefficients for the i-th property only, y_m is the actual value of the corresponding dependent variable, and $\bar{Y}$ is the mean of the values of the dependent variable in the model. Partial correlation coefficients measure the correlation between the actual values of the dependent variable and the estimated values calculated from an individual property. The partial coefficients show the relative importance of each property in the model, i.e., how significantly the property contributes to the dependent variable. Properties which have partial correlation coefficients with values near one are very important in the model. Partial correlation coefficients with values near zero indicate unimportant properties (which can be eliminated from the model).

Let's evaluate the results from the model for the boiling points of alcohols as detailed in Chapter 3. The values for the four statistical

criteria are:

SE = 0.57		R =	0.9996
F-value (calculated) =	574	R_{p1} =	0.9993
F-value (table) =	39.3	R_{p2} =	0.9993

The standard error of the residuals of the model is somewhat greater than the experimental error in boiling point (0.2 - 0.5 °C). The multiple correlation coefficient (R) is essentially equal to +1. Statistically, we can be confident that the estimated boiling points agree well with the actual boiling points for the alcohols in the model. The F-test value from our calculation (574) is much greater than the value from the F-distribution table (39.3). Thus, our categorical regression model is very significant. The partial correlation coefficients for the two properties are near unity, so that both of the properties are very important in the model.

Using each of our statistical criteria, we conclude that the CRA model for the boiling points of alcohols is a more than satisfactory model.

Chapter 5

Procedures and Computer Program

All CRA calculations are carried out on an IBM PS/2 286 personal computer. The program is written in BASIC. The series of programs in the CRA package include:

"CRA-DATA" - a program to write a data set onto a disk,

"CRA-MAIN" - a program to process all computations, and

"CRA-PRED" - a program to predict new data from the CRA model.

Once a CRA model for a problem has been selected, one can do calculations on a PC step by step, following the documentation and instructions in "CRA-DOCU" given in Appendix A. The variables involved in the three programs are also explained in "CRA-DOCU". One can build a model based on either an entire data set or a "leave-one-out" data set. In the leave-one-out method, also called the N-1 calculation, the calculations are performed after removing rows sequentially from the data set. The normalized categorical coefficients are stored on the disk automatically for predicting a set of new data. The technique is useful to assess the predictive capability of a CRA model [34]. A predicted value for the removed sample can be obtained from the reduced data matrix. An example of an application of the N-1 method is given in Chapter 8.

The computer outputs for the entire and the leave-one-out data sets for the 8 x 6 problem are given in Tables 3 and 4.

Table 3. Computer output for the 8 x 6 problem using the entire data set.

Boiling points of alcohols

Samples 8 Properties 2

Properties (categories)
1 Carbon number (5,6,7).
2 Type of alcohol (p,s,t)

				Categorized properties					
				1			2		
1	141	1-pentanol	137.8000	1	0	0	1	0	0
2	132	2-pentanol	119.0000	1	0	0	0	1	0
3	213	2-me-2-butanol	102.0000	1	0	0	0	0	1
4	141	1-hexanol	157.0000	0	1	0	1	0	0
5	115	2-hexanol	139.9000	0	1	0	0	1	0
6	122	2-me-2-pentanol	121.4000	0	1	0	0	0	1
7	118	1-heptanol	176.2000	0	0	1	1	0	0
8	220	2-me-2-hexanol	142.8000	0	0	1	0	0	1

Normalized Category Coefficients - Using Entire Set

Prop.(Category)	1 (1)	1 (2)	1 (3)	2 (1)	2 (2)	2 (3)
Entire set	-17.4042	2.4291	22.4625	17.4916	-0.0750	-17.4417

			Actual	Estimated	Residual
1	141	1-pentanol	137.8000	137.1000	0.7000
2	132	2-pentanol	119.0000	119.5334	-0.5334
3	213	2-me-2-butanol	102.0000	102.1667	-0.1667
4	141	1-hexanol	157.0000	156.9333	0.0667
5	115	2-hexanol	139.9000	139.3667	0.5333
6	122	2-me-2-pentanol	121.4000	122.0000	-0.6000
7	118	1-heptanol	176.2000	176.9667	-0.7667
8	220	2-me-2-hexanol	142.8000	142.0334	0.7666
		Mean	137.0125	137.0125	0.0000
		S.D.	21.7132	21.7056	0.5762

Multiple correlation coefficient 0.9996
F-test value 573.5861

Range of normalized category coefficients

	Prop. 1	Prop. 2
Entire set	39.8667	34.9333

Partial correlation coefficients

	Prop. 1	Prop. 2
Entire set	0.9993	0.9993

Table 4. Computer output for the 8 x 6 problem using the leave-one-out method.

Boiling points of alcohols

Samples 8 Properties 2

Properties (categories)
1 Carbon number (5,6,7)
2 Type of alcohol (p,s,t)

				Categorized properties					
				1			2		
1	141	1-pentanol	137.8000	1	0	0	1	0	0
2	132	2-pentanol	119.0000	1	0	0	0	1	0
3	213	2-me-2-butanol	102.0000	1	0	0	0	0	1
4	141	1-hexanol	157.0000	0	1	0	1	0	0
5	115	2-hexanol	139.9000	0	1	0	0	1	0
6	122	2-me-2-pentanol	121.4000	0	1	0	0	0	1
7	118	1-heptanol	176.2000	0	0	1	1	0	0
8	220	2-me-2-hexanol	142.8000	0	0	1	0	0	1

Normalized Category Coefficients - Using N-1 Method

Prop.(Category)	1 (1)	1 (2)	1 (3)	2 (1)	2 (2)	2 (3)
Entire set	-17.4042	2.4291	22.4625	17.4916	-0.0750	-17.4417
Sample removed						
1	-20.3305	0.0628	20.2362	19.5505	2.6838	-14.8229
2	-19.5857	-0.2857	20.0143	17.3667	0.6000	-17.5667
3	-19.7857	-0.0857	19.9143	14.9714	-2.6285	-19.8286
4	-17.0457	2.7343	22.8343	19.9485	2.4486	-14.9314
5	-16.8286	2.4714	22.7714	17.5952	-0.7714	-17.3381
6	-17.1600	3.1533	22.5867	14.8971	-2.7895	-19.5562
7	-14.3595	5.4738	26.6572	20.4285	2.4786	-15.2714
8	-14.0310	5.8024	24.6857	15.3286	-2.6214	-20.3714

	Actual values		Predicted values	
	Mean	S.D.	Mean	S.D.
Entire set	137.0125	21.7132	137.0125	21.7056
Sample removed				
1	136.9000	23.2101	136.9000	23.2056
2	139.5857	22.0418	139.5857	22.0360
3	142.0143	18.4029	142.0143	18.3930
4	134.1571	21.7621	134.1571	21.7535
5	136.6000	23.1830	136.6000	23.1776
6	139.2428	22.3387	139.2429	22.3330
7	131.4143	16.9734	131.4143	16.9696
8	136.1857	23.0942	136.1857	23.0915

(continued)

Table 4. (continued)

	Mult.Cor.	S.D.Res.	Deleted Datum		
Entire set	0.9996	0.5762			
Sample removed			Actual	Predicted	Error
1	0.9998	0.4550	137.8000	136.1200	1.6800
2	0.9997	0.5058	119.0000	120.6000	-1.6000
3	0.9995	0.6043	102.0000	102.4000	-0.4000
4	0.9996	0.6124	157.0000	156.8399	0.1601
5	0.9998	0.5039	139.9000	138.3000	1.6000
6	0.9998	0.4980	121.4000	122.8400	-1.4400
7	0.9998	0.3535	176.2000	178.5000	-2.3000
8	0.9999	0.3535	142.8000	140.5000	2.3000

Range of normalized category coefficients

	Prop. 1	Prop. 2
Entire set	39.8667	34.9333
Sample removed		
1	40.5667	34.3733
2	39.6000	34.9333
3	39.7000	34.8000
4	39.8800	34.8800
5	39.6000	34.9333
6	39.7467	34.4533
7	41.0167	35.7000
8	38.7167	35.7000

Partial correlation coefficients

	Prop. 1	Prop. 2
Entire set	0.9993	0.9993
Sample removed		
1	0.9995	0.9995
2	0.9994	0.9995
3	0.9991	0.9991
4	0.9993	0.9991
5	0.9995	0.9995
6	0.9995	0.9994
7	0.9997	0.9997
8	0.9996	0.9997

Chapter 6

Overview of Applications of CRA

To build a CRA model, we need 1) to select a vector of data as the dependent variable and 2) to propose a set of categorized properties as the independent variables. So, CRA models can be applied to any vector of data whenever the structures of the substances involved suggest categorized properties. Due to this advantage of the CRA method, the applications of CRA in chemistry are potentially numerous.

In order to obtain a more reliable and precise model, data for the dependent-variable vector in a model should be selected very carefully. The data being modeled preferably should be obtained under the same experimental conditions and have known experimental error.

Selection of the categorized properties is the key to modeling. The characteristics of a molecular structure are often difficult to represent as the decimal numbers required in multiple regression analysis (MRA). However, in some case, independent variables are easily categorized as "yes" (1) or "no" (0) for the suggested category. Then CRA rather than MRA can be applied. Most often, the categorized properties in a model can be related to the structures of the sub-

stances. In our studies, most of the categorized properties are directly related to molecular structures, such as the length of the carbon chain, the type of branch, the type of substituent and the position of the substituent.

To explore the potential of CRA, chemical problems in several fields have been investigated. A summary of our applications of the CRA method in chemistry is given in Table 5. In the following chapters, we will explore the applications of CRA to model the following data: a physical property (our 8 x 6 example), gas-chromatographic retention indices, dissociation constants, stability constants and rate constants. Additional applications of CRA in other fields, such as chemical shift data in nuclear magnetic resonance (NMR) and molar lattice enthalpy data, led to unsatisfactory models.

Table 5. Summary of applications of the CRA method.

Data type	Chapter	Specific problem
Physical property	3	Boiling points of alcohols
Chromatographic	7	Retention indices of hydrocarbons
	7	Retention indices of alkenes
Equilibrium constant	8	Dissociation constants of substituted benzoic acids
	8	Dissociation constants of substituted carboxylic acids
	9	Stability constants of metal-EDTA type complexes
Rate constant	10	Rate constants for solvolysis of esters

Chapter 7

CRA Models for Retention Indices of Hydrocarbons and Alkenes

The important quantity measured in a gas chromatographic experiment is the time, called the retention time, for a small sample of solute to pass through the separation column [37]. A useful systematic method for expressing such retention data is based on retention indices. Kovats [38] defined the retention index RI_x for a component x as:

$$RI_x = 100\,n + 100\,\frac{\log V_x - \log V_n}{\log V_{n+1} - \log V_n}$$

where V_x is the retention volume of the sample, and V_n and V_{n+1} are the retention volumes for the two n-alkanes with carbon numbers n and n + 1 (which eluted prior to and after the sample component).

CRA models for retention-index data of hydrocarbons and of alkenes are investigated here.

In the first CRA problem, retention indices of 21 straight- and branched-chain hydrocarbons obtained on 5% SE-30 stationary phase at 50 °C [39] are studied. The solutes are listed in Table 6.

The retention index is a property which strongly depends on

molecular structure. Properties considered as independent variables in the hydrocarbon models are: 1) the number of carbon atoms in the main chain, having specific values of 6, 7 and 8; 2) the number of methyl branches, having values of 1, 2 and 3; and 3) the number of quaternary carbon atoms, having values of 0 and 1. Various models are formed from combinations of the three properties. After investigating the various models, we find that the model with all three properties spanning eight categories as independent variables, called the 21 x 8 problem, gives the best results. Details of the model are given in Table 6.

The category coefficients and the normalized category coefficients for the 21 x 8 problem are:

Property	1			2			3	
Category	1	2	3	1	2	3	1	2
A*	574.2	674.6	768.5	0.0	-17.5	-27.05	0.0	-18.80
A_N	-126.0	-25.6	68.3	13.0	-4.45	-14.03	7.16	-11.64

The category coefficients A* of the 21 x 8 model corresponding to the removed columns of the first categories of the second and third properties, i.e., the terms 2(1) and 3(1), respectively, have values of zero. Normalized category coefficients A_N with more chemical meaning are obtained from Eq. [8].

Comparing the values of the normalized category coefficients,

Table 6. CRA model for the 21 x 8 problem involving retention indices of hydrocarbons.

Properties (Categories)

1 Total number of carbon atoms (6, 7, 8)
2 Number of methyl branches (1, 2, 3)
3 Number of quaternary carbon atoms (0, 1)

		Categorized properties							
	Property	1			2			3	
	Category	1	2	3	1	2	3	1	2
Compound	RI								
2,2-dimethyl butane	530	1	0	0	0	1	0	0	1
2,3-dimethyl butane	562	1	0	0	0	1	0	1	0
2-methyl pentane	568	1	0	0	1	0	0	1	0
3-methyl pentane	583	1	0	0	1	0	0	1	0
2,2-dimethyl pentane	622	0	1	0	0	1	0	0	1
2,4-dimethyl pentane	628	0	1	0	0	1	0	1	0
2,2,3-trimethyl butane	643	0	1	0	0	0	1	0	1
3,3-dimethyl pentane	660	0	1	0	0	1	0	0	1
2,3-dimethyl pentane	672	0	1	0	0	1	0	1	0
2-methyl hexane	667	0	1	0	1	0	0	1	0
3-methyl hexane	677	0	1	0	1	0	0	1	0
2,2,4-trimethyl pentane	692	0	0	1	0	0	1	0	1
2,2-dimethyl hexane	721	0	0	1	0	1	0	0	1
2,5-dimethyl hexane	728	0	0	1	0	1	0	1	0
2,2,3-trimethyl hexane	739	0	0	1	0	0	1	0	1
3,3-dimethyl hexane	746	0	0	1	0	1	0	0	1
2,3-dimethyl hexane	761	0	0	1	0	1	0	1	0
2-methyl heptane	764	0	0	1	1	0	0	1	0
4-methyl heptane	769	0	0	1	1	0	0	1	0
3,4-dimethyl hexane	773	0	0	1	0	1	0	1	0
3-methyl heptane	775	0	0	1	1	0	0	1	0

we find that coefficients increase about 94-100 RI units as a carbon atom is added, and the coefficients decrease with an increase in the number of side chains and the number of the quaternary carbon atoms. This is in good agreement with chemical facts.

The statistical criteria for the 21 x 8 model are

SE = 15		R	= 0.980
F-value (calculated)	= 52.8	R_{p1}	= 0.978
F-value (table)	= 3.4	R_{p2}	= 0.430
		R_{p3}	= 0.412

The statistical data show that the standard error of the residuals (15) is about three times the estimated maximum experimental error ($\pm$ 5 RI units). The multiple correlation coefficient is near unity, showing the values of the estimated RI's agree closely with the actual RI's. The calculated F-test value is about 17 times the table F-value, showing a highly significant regression. Although the partial correlation coefficients of the second and third properties are relatively small, the 21 x 8 model still is a good model.

Let's consider other combinations of the three properties. Three models are formed, called the 21 x 5 (2,3), the 21 x 5 (1,3) and the 21 x 6 (1,2) models, respectively. For example, the 21 x 5 (1,3) model involves properties 1 and 3.

The details and the results for these three models are shown in

Table 7. From the table, the standard error of the 21 x 5 (2,3) model is very high, and the multiple correlation coefficient, the calculated F-value, and the partial correlation coefficients are also quite low. All four statistical criteria indicate that the 21 x 5 (2,3) model is an unsatisfactory model and that the suggested category properties are inadequate. However, the 21 x 5 (1,3) and the 21 x 6 (1,2) models are statistically acceptable models. The standard errors are about three times the experimental error, the calculated F-values indicate the significance of the regression, the multiple correlation coefficient are close to unity, and the partial correlation coefficients are fairly large. In additional, the two models are nearly equivalent mathematically.

We next compare the three-property 21 x 8 model with the 21 x 5 (1,3) and the 21 x 6 (1,2) models. The standard error of the 21 x 8 model is slightly smaller (see Table 7) than the standard error for the other two models, the multiple correlation coefficient is slightly higher and the calculated F-test value is slightly lower than the values of other two models but still much higher than the table F-value. Therefore, the 21 x 8 model is only marginally better than the 21 x 5 (1,3) and the 21 x 6 (1,2) models.

From the discussion of the four models above, selection of the properties for a CRA model is very important. If insufficient

Table 7. Results of four models for retention indices of hydrocarbons.

Problem	No. properties (No. categories)	Details of properties (categories)	SE	R	R_p	F-value calc. (table) (df1,df2)*
21 x 5	2	Number of methyl branches (1,2,3)	73.0	0.182	0.144	0.14 (3.7)
(2,3)	(3, 2)	Number of quaternary C atoms (0,1)			0.173	(4,16)
21 x 5	2	Number of carbon atoms (6,7,8)	17.0	0.973	0.973	73 (3.7)
(1,3)	(3, 2)	Number of quaternary C atoms (0,1)			0.666	(4,16)
21 x 6	2	Number of carbon atoms (6,7,8)	16.8	0.975	0.974	56 (3.7)
(1,2)	(3, 2)	Number of methyl branches (1,2,3)			0.678	(5,15)
21 x 8	3	Number of carbon atoms (6,7,8)	15.0	0.980	0.978	53 (3.5)
	(3, 3, 2)	Number of methyl branches (1,2,3)			0.430	(6,14)
		Number of quaternary C atoms (0,1)			0.412	

*) Number of degrees of freedom

properties are selected, an unsatisfactory model is formed, such as the 21 x 5 model. Adding a property may improve the model, such as the 21 x 8 model.

Sometimes, adding additional properties may be counterproductive. Some properties do not significantly contribute to the model, so that the properties can be eliminated from the model without changing the goodness of the model. A study of the retention indices of alkenes (our second chromatographic study) illustrates this phenomenon. Retention indices of 42 alkenes were obtained on a capillary column [40]. The 29 alkenes selected are listed in Table 8.

The five properties considered in the CRA models are: 1) the number of carbon atoms in the main chain, 2) the number of methyl branches, 3) the position of the double bond, 4) steric effects and 5) the number of carbon atoms connected not only to a double bond but also to two other carbon atoms.

The complete CRA model for the retention indices of the 29 alkenes, named the 29 x 14 problem, is given in Table 8. Four models are studied. The other three models with different combinations of the above five categorized properties are the 29 x 11 (1,2,3,5), the 29 x 11 (1,2,4,5) and the 29 x 8 (1,2,5) models. Results for the four models are given in Table 9.

Table 8. CRA model for the 29 x 14 problem involving retention indices of alkenes.

Properties (Categories)

1 Number of carbon atoms in main chain (5, 6)
2 Number of methyl branches (1, 2, 3)
3 Position of double bond (1, 2, 3)
4 Steric effect (cis, trans, neither)
5 Number of carbon atoms connected to a double bond and two other carbon atoms (0, 1, 2)

Categorized properties

Property		1		2			3			4			5		
Category		1	2	1	2	3	1	2	3	1	2	3	1	2	3
Compound *	RI														
3-me cis 2-pentene	612.5	1	0	1	0	0	0	1	0	1	0	0	0	1	0
4,4-dime trans 2-pentene	614.5	1	0	0	1	0	0	1	0	0	1	0	1	0	0
2,4-dime 1-pentene	636.0	1	0	0	1	0	1	0	0	0	0	1	0	1	0
2-me 2-pentene	641.0	1	0	0	1	0	0	1	0	0	0	1	0	1	0
2-me trans 3-hexene	647.0	0	1	1	0	0	0	0	1	0	1	0	1	0	0
2,3-dime 1-pentene	648.0	1	0	0	1	0	1	0	0	0	0	1	0	1	0
4-me 1-hexene	655.5	0	1	1	0	0	1	0	0	0	0	1	1	0	0
4-me trans 2-hexene	658.0	0	1	1	0	0	0	1	0	0	1	0	1	0	0
3,4-dime trans3-pentene	677.0	1	0	0	1	0	0	1	0	0	1	0	0	1	0
3-me trans 3-hexene	684.0	0	1	1	0	0	0	0	1	0	1	0	0	1	0
2-me 2-hexene	690.5	0	1	1	0	0	0	1	0	0	0	1	0	1	0
2,2-dime trans 3-hexene	693.0	0	1	0	1	0	0	0	1	0	1	0	1	0	0
3-me cis 3-hexene	695.0	0	1	1	0	0	0	0	1	1	0	0	0	1	0
3,4,4-trime 1-pentene	698.5	1	0	0	0	1	1	0	0	0	0	1	1	0	0
2,3-dime 2-pentene	702.0	1	0	0	1	0	0	1	0	0	0	1	0	0	1
2,4,4-trime 1-pentene	702.0	1	0	0	0	1	1	0	0	0	0	1	0	1	0

(continued)

Table 8. (continued)

		Categorized properties													
	Property	1		2			3			4			5		
	Category	1	2	1	2	3	1	2	3	1	2	3	1	2	3
Compound *	RI														
5,5-dime trans 2-hexene	706.0	0	1	0	1	0	0	1	0	0	1	0	1	0	0
2,4,4-trime 2-pentene	714.5	1	0	0	0	1	0	1	0	0	0	1	0	1	0
5,5-dime cis 2-hexene	721.5	0	1	0	1	0	0	1	0	1	0	0	1	0	0
2,3,4-trime 1-pentene	724.5	1	0	0	0	1	1	0	0	0	0	1	0	1	0
2,3,3-trime 1-pentene	730.5	1	0	0	0	1	1	0	0	0	0	1	0	1	0
4,5-dime 1-hexene	734.0	0	1	0	1	0	1	0	0	0	0	1	1	0	0
2,4-dime 1-hexene	737.5	0	1	0	1	0	0	1	0	0	0	1	0	0	1
2,3-dime 2-hexene	740.0	0	1	0	1	0	0	1	0	0	0	1	0	0	1
2,5-dime 1-hexene	740.0	0	1	0	1	0	0	1	0	0	0	1	0	1	0
2,5-dime 2-hexene	749.5	0	1	0	1	0	0	1	0	0	0	1	0	1	0
3,4-dime 1-hexene	755.5	0	1	0	1	0	1	0	0	0	0	1	1	0	0
2,3,4-trime 2-pentene	765.5	1	0	0	0	1	0	1	0	0	0	1	0	0	1
2,3-dime 2-hexene	787.5	0	1	0	1	0	0	1	0	0	0	1	0	0	1

*) me - methyl.

Table 9. Results of models for retention indices of alkenes.

Problem	No. Properties (No. categories)	Details of Properties (Categories)	SE	R	R_p	F-value calc. (table) (df1, df2)
29 x 14	4 (2,3,3,3)	1. No. carbon atoms in chain (5,6)	41.0	0.376	0.542	1.8(2.35) (11,17)
		2. No. methyl branches (1,2,3)			0.364	
		3. Position of double bond (1-,2-,3-)			0.068	
		4. Steric effects (cis,trans,neither)			0.141	
		5. No. C atoms connected to C=C and two other C atoms (0,1,2)			0.473	
29 x 11 (1,2,3,5)	4 (2,3,3,3)	1. No. carbon atoms in chain (5,6)	14.6	0.945	0.923	20.9 (2.45) (8,20)
		2. No. methyl branches (1,2,3)			0.900	
		3. Position of double bond (1-,2-,3-)			0.257	
		4. No. C atoms connected to C=C and two other C atoms (0,1,2)			0.789	
29 x 11 (1,2,4,5)	4 (2,3,3,3)	1. No. carbon atoms in chain (5,6)	14.4	0.947	0.924	21.6 (2.45) (8,20)
		2. No. methyl branches (1,2,3)			0.915	
		3. Steric effects (cis, trans, neither)			0.304	
		4. No. C connected to C=C and two other C atoms (0,1,2)			0.759	
29 x 8 (1,2,5)	3 (2,3,3)	1. No. carbon atoms in chain (5,6)	15.1	0.942	0.918	28.6 (2.55) (6,22)
		2. No. methyl branches (1,2,3)			0.909	
		3. No. C atoms connected to C=C and two other C atoms (0,1,2)			0.777	

From Table 9, we find that the complete model has a high value for the standard error and a low value for the multiple correlation coefficient. The calculated F-value is significantly lower than the table value, and the partial correlation coefficients for the five properties are very low also. The four statistical criteria clearly show that the 29 x 14 model is an unsatisfactory model.

However, the four statistical values for the 29 x 11 (1,2,3,5) and 29 x 11 (1,2,4,5) models are almost the same. The multiple correlation coefficients are large (0.95). The calculated F-values are about 10 times the table values, and the standard errors are about three times the experimental error. These two models are statistically satisfactory models. However, the partial correlation coefficients for properties 1, 2, and 5 are high or relatively high in the two 29 x 11 models. The small values of the partial correlation coefficients for properties 3 and 4 (the position of the double bond and the steric effect) show that the two properties are not very important in the models. If we eliminate the two poorly correlated properties, a simpler model, the 29 x 8 (1,2,5) model, is formed. The statistical analysis of the model (see Table 9) shows that the SE, R and F-test values are very close to the corresponding values for the two 29 x 11 models. The three models are equivalent statistically, but the 29 x 8 model has the advantage of the greatest simplicity.

From these examples using retention-index data sets, the number of properties significantly contributing to a CRA model can be determined by comparing the partial correlation coefficients. Unimportant properties may be eliminated without changing the goodness of the model.

Chapter 8

CRA Models for Dissociation Constants of Substituted Acids

The extent to which an acid ionizes in an aqueous medium can be expressed by the acid dissociation constant, K_a. The value of pK_a, which is defined as $pK_a = - \log K_a$, is widely used in chemistry. The larger the value of pK_a, the weaker the acid. In this chapter, the dissociation constants of substituted benzoic acids and of substituted carboxylic acids are selected for CRA modeling.

The pK_a values for the substituted benzoic acids were obtained by measuring the conductivity of the acid solutions at 25 °C [41]. Benzoic acids involving seven substituents: chloro (-Cl), bromo (-Br), amino ($-NH_2$) , cyano (-CN), hydroxyl (-OH), methyl ($-CH_3$) and nitro ($-NO_2$), were selected for this study. The pK_a values of these acids are used as the dependent variable (see Table 10).

Several categorized properties are employed to search for suitable models. Selection of the categorized properties is focused on the type of the substituent and the position of the substituent on the parent acid. We also employed several categorized properties which are related to the individual substituents, such as the electron-donor ability of the substituent (e^- donor, e^- acceptor) and the polarity of the substituent (polar, non-polar).

Table 10. Values of pK_a for CRA models involving substituted benzoic acids.

Acid	pK_a	Acid	pK_a
2-chloro benzoic	2.877	4-cynao benzoic	3.550
3-chloro benzoic	3.830	2-hydroxyl benzoic	2.980
4-chloro benzoic	3.986	3-hydroxyl benzoic	4.076
2-bromo benzoic	2.850	4-hydroxyl benzoic	4.582
3-bromo benzoic	3.810	2-methyl benzoic	3.900
4-bromo benzoic	3.990	3-methyl benzoic	4.269
2-amino benzoic	2.090	4-methyl benzoic	4.362
3-amino benzoic	3.070	2-nitro benzoic	2.180
4-amino benzoic	2.410	3-nitro benzoic	3.460
2-cyano benzoic	3.140	4-nitro benzoic	3.440
3-cyano benzoic	3.600		

Results for three models: the 21 x 5, the 21 x 7 and the 21 x 9 models, are given in Table 11. In the three models, the categorized properties belong to two types: 1) position of substituent (o, m, p), and 2) three properties of the substituents (polarity, electron donor ability and kind of substituent).

The statistical criteria in Table 11 indicate that the 21 x 5 and the 21 x 7 models, containing the substituent properties, have large values for the standard error compared to the experimental error of 0.1 - 0.2 pK units, the correlation coefficients are much less than unity, and the calculated F-test values are very close to or even less than the table F-values. Moreover, the property "Electron e^- donor ability" does not contribute significantly to the models due to the small values of the partial correlation coefficients for this property in the 21 x 5 and 21 x 7 models.

However, the 21 x 9 model, containing two properties which are directly related to the structure of the substances, is a satisfactory model. The standard error of the model is much smaller and close to the experimental error, the correlation coefficient for the model is near unity, the calculated F-test value of the model is 3.5 times the table F-value, and both partial correlation coefficients are fairly near unity. Two properties: type of substituent and position of substituent, are sufficient for the model.

Table 11. Results of three models for dissociation constants of substituted benzoic acids.

Problem	No. properties (No. categories)	Details of properties (Categories)	SE	R	R_p	F-value calc. (table) (df_1,df_2)
21 x 5	2 (3,2)	1. Position of substituent (o,m,p)	0.54	0.614	0.610	2.42 (3.8) (4,16)
		2. e^- donor ability (e^- donor, e^- acceptor)			0.121	
21 x 7	3 (3,2,2)	1. Position of substituent (o,m,p)	0.45	0.754	0.678	3.95 (3.58) (5,15)
		2. Polarity of substituent (non-polar, polar)			0.554	
		3. e^- donor ability (e^- donor, e^- acceptor)			0.182	
21 x 9	2 (6,3)	1. Substituent (halogen, NH_2,CN,CH_3,OH,NO_2)	0.21	0.950	0.920	11.3 (3.51) (8,12)
		2. Position of substituent (o,m,p)			0.889	

Our second acid-dissociation constant problem involves the dissociation constants of substituted carboxylic acids. The pK_a values of 15 halogen-substituted carboxylic acids are used as the dependent variable [42]. The values of (pK_{HAc} - pK_a) and (pK_p - pK_a) are also studied as the dependent variable, where pK_{HAc} is the value for acetic acid, and pK_p is the pK_a value of the parent acids: acetic, propionic and butyric acid, respectively (see Table 12). Two categorized properties: type of halogen substituent (Cl, Br, I) and position of substituent (2-, 3-, 4-) are employed as independent variables.

The three models with the two same properties containing six categories but with three different kinds of dependent variable are compared. The three models are labeled as models A, B and C, having pK_a, pK_{HAc} - pK_a and pK_p - pK_a as the dependent variables, respectively (see Table 12).

Statistical analysis of the three models furnishes the values:

	Model		
	A	B	C
	pK_a	pK_{HAc} - pK_a	pK_p - pK_a
SE	0.065	0.065	0.066
F-value (calc.)	193	193	150
F-value (table)	3.5	3.5	3.5
R	0.995	0.995	0.994
R_{p1}	0.723	0.723	0.770
R_{p2}	0.995	0.995	0.994

Table 12. Values of the dependent variables for three models involving substituted carboxylic acids.

	Model		
	A	B	C
	(pK_a)	(pK_{HAc}[a] - pK_a)	(pK_p[b] - pK_a)
Acid			
Chloro acetic	2.868	1.882	1.882
Bromo acetic	2.902	1.848	1.848
Iodo acetic	3.175	1.575	1.575
2-chloro propionic	2.834	1.916	1.976
3-chloro propionic	4.076	0.675	0.734
2-bromo propionic	2.967	1.783	1.843
3-bromo propionic	4.022	0.728	0.788
2-iodo propionic	3.208	1.542	1.602
3-iodo propionic	4.045	0.705	0.765
2-chloro butyric	2.857	1.893	2.013
3-chloro butyric	4.049	0.701	0.821
4-chloro butyric	4.523	0.227	0.347
2-bromo butyric	2.975	1.775	1.895
4-bromo butyric	4.582	0.168	0.288
4-iodo butyric	4.642	0.108	0.228

a) pK_{HAc} = 4.75 for acetic acid at 25 °C.

b) pK_p = 4.81 for propionic acid, pK_p = 4.87 for butyric acid, both at 25 °C.

The three models are excellent models. The values of the four statistical criteria for models A and B are exactly the same, showing that a dependent variable plus or minus a constant does not change the nature of the model. Model C is somewhat different from models A and B, because slightly different values are used for the dependent variable. The values of (pK_p - pK_a) depend on the parent-acids. In this case, model C is statistically equivalent to models A and B because the values for pK_a of acetic, propionic and butyric acids are quite similar.

Let's compare the category coefficients and normalized category coefficients for the three models. The calculated category coefficients and normalized category coefficients are given in Table 13. The category coefficients of the first category for the second property have values of zero in the three models due to the deletion of the corresponding column. The category coefficients in models A and B are different in magnitude for the first property. However, the normalized category coefficients of models A and B have the same magnitudes but opposite sign for all six categories. Because the values of the dependent variable in model B are equal to (pK_{HAc} - pK_a), the new vector of the dependent variable has the inverse pattern to that in model A, so that the coefficients have opposite signs. The fact that the normalized category coefficients in models A and B have the

Table 13. Category coefficients and normalized category coefficients for the three models involving substituted carboxylic acids.

Property	1			2		
Category	1	2	3	1	2	3
Model	Category coefficient					
A	2.909	2.954	3.099	0.0	1.018	1.595
B	1.841	1.795	1.651	0.0	-1.018	-1.595
C	1.903	1.849	1.689	0.0	-1.059	-1.526
	Normalized category coefficient					
A	-0.066	-0.020	0.124	-0.607	0.473	0.988
B	0.066	0.020	-0.124	0.607	-0.473	-0.988
C	0.075	0.022	-0.139	0.588	-0.471	-0.938

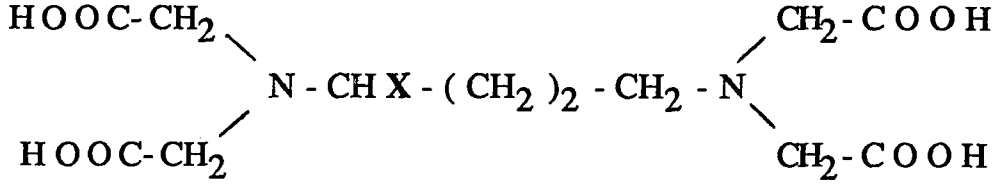
same absolute magnitude indicates the two models are identical. The normalized coefficients indicate the identical models. The normalized coefficients of model C are only slightly different from the coefficients of models A and B, confirming that the three models are statistically equivalent (as discussed above).

Chapter 9

CRA Models for Stability Constants of Chelates

The stability constants of metal-ligand complexes are dependent on the properties of the metal and the ligand. Ethylene(diamine)tetraacetate sodium salt (abbreviated EDTA) is the most widely used ligand. EDTA is a hexadentate ligand, forming a six-coordinate, 1:1 complex with many metal ions.

In this study, the stability constants of the 24 chelates of four alkaline-earth metal ions with EDTA and five of its derivatives are selected [43]. The EDTA derivatives have the general formula:



where substituent **X** may be a hydrogen atom, or methyl, ethyl, isopropyl, isobutyl or phenyl groups. To simplify the categories of the ligands, we classify the six EDTA derivatives into three categories: EDTA parent (**X** is a hydrogen atom), a-EDTA (**X** is an alkyl group), and p-EDTA (**X** is a phenyl group). The alkaline-earth metal ions are the magnesium, calcium, strontium and barium ions.

Two properties with seven categories are involved in the CRA model: 1) type of EDTA (EDTA, a-EDTA, p-EDTA) and 2) type of metal ion (Mg^{2+}, Ca^{2+}, Sr^{2+}, Ba^{2+}). A 24 x 7 model for the data is given in Table 14. The experimental error for the stability constants of the complexes are about $\pm$ 0.1 - 0.2 log K_s units [43].

The four statistical criteria calculated for the model are:

SE = 0.12	R = 0.995
F-test value (calc.) = 256	R_1 = 0.954
F-value (table) = 2.8	R_2 = 0.994

The results show that the standard error of the residuals for the model is within the range of the experimental error (0.1 - 0.2 log K_s units), the multiple correlation coefficient and the two partial correlation coefficients are close to unity, and the calculated F-test value is much greater than the table F-value. The model for the stability constants of metal-EDTA complexes is a more than satisfactory model.

The leave-one-out method, called the N-1 calculation (see Chapter 5), is applied to the 24 x 7 model. One chelate in the model is removed sequentially during the N-1 calculations. A total of 24 (N-1) calculations is performed and the normalized category coefficients for every reduced matrix (23 x 7) are calculated. The multiple correlation coefficients and the standard error of the residuals

Table 14. CRA model for the 24 x 7 problem involving stability constants of chelates.

Properties (Categories)

1 Type of ligand (EDTA, a-EDTA, p-EDTA)
2 Type of metal ion (Mg^{2+}, Ca^{2+}, Sr^{2+}, Ba^{2+})

		Categorized properties						
	Property	1			2			
	Category	1	2	3	1	2	3	4
Chelate *	log K_s							
EDTA-Mg	8.79	1	0	0	1	0	0	0
EDTA-Ca	10.69	1	0	0	0	1	0	0
EDTA-Sr	8.73	1	0	0	0	0	1	0
EDTA-Ba	7.86	1	0	0	0	0	0	1
Me-EDTA-Mg	10.02	0	1	0	1	0	0	0
Me-EDTA-Ca	11.63	0	1	0	0	1	0	0
Me-EDTA-Sr	9.60	0	1	0	0	0	1	0
Me-EDTA-Ba	8.55	0	1	0	0	0	0	1
Et-EDTA-Mg	10.15	0	1	0	1	0	0	0
Et-EDTA-Ca	11.75	0	1	0	0	1	0	0
Et-EDTA-Sr	9.66	0	1	0	0	0	1	0
Et-EDTA-Ba	8.50	0	1	0	0	0	0	1
Ip-EDTA-Mg	10.3	0	1	0	1	0	0	0
Ip-EDTA-Ca	11.77	0	1	0	0	1	0	0
Ip-EDTA-Sr	9.74	0	1	0	0	0	1	0
Ip-EDTA-Ba	8.60	0	1	0	0	0	0	1
Ib-EDTA-Mg	10.16	0	1	0	1	0	0	0
Ib-EDTA-Ca	11.84	0	1	0	0	1	0	0
Ib-EDTA-Sr	9.84	0	1	0	0	0	1	0
Ib-EDTA-Ba	8.75	0	1	0	0	0	0	1
Ph-EDTA-Mg	9.40	0	0	1	1	0	0	0
Ph-EDTA-Ca	11.25	0	0	1	0	1	0	0
Ph-EDTA-Sr	9.32	0	0	1	0	0	1	0
Ph-EDTA-Ba	8.36	0	0	1	0	0	0	1

*) Me - methyl, Et - ethyl, Ip - isopropyl, Ib - isobutyl, Ph - phenyl.

for every N-1 matrix are listed in Table 15. The values of the stability constants (log K_s) predicted from the stored (N-1) normalized coefficients for every removed chelate are given in Table 16.

From Table 15, the multiple correlation coefficients of the 24 models are all extremely high (greater than 0.99). The standard errors of the residuals are around 0.12 for the 24 models well within the experimental error (0.1 - 0.2 log K_s units). The results indicate that all 24 (N-1) models are excellent models. Removing one chelate does not change the goodness of the model at all in this case. No single chelate exhibits unique behavior.

By comparing the actual and the predicted values (Table 16), we find that the predicted log K_s for every removed chelate is very close to the actual value. Most of the errors between the actual and the predicted values are around 0.1 - 0.2 log K_s units. The standard error is 0.18, which is within the range of the experimental error. The leave-one-out technique (N-1 calculation) shows excellent ability to predict.

Table 15. Results of the 24 (N-1) models for the chelate problem using the leave-one-out method.

Chelate removed *	R	SE	Chelate removed *	R	SE
EDTA-Mg	0.996	0.109	Ip-EDTA-Mg	0.996	0.111
EDTA-Ca	0.994	0.125	Ip-EDTA-Ca	0.994	0.125
EDTA-Sr	0.994	0.125	Ip-EDTA-Sr	0.995	0.125
EDTA-Ba	0.995	0.112	Ip-EDTA-Ba	0.994	0.123
Me-EDTA-Mg	0.994	0.125	Ib-EDTA-Mg	0.995	0.123
Me-EDTA-Ca	0.994	0.122	Ib-EDTA-Ca	0.994	0.123
Me-EDTA-Sr	0.995	0.121	Ib-EDTA-Sr	0.995	0.122
Me-EDTA-Ba	0.995	0.121	Ib-EDTA-Ba	0.994	0.124
Et-EDTA-Mg	0.995	0.123	Ph-EDTA-Mg	0.995	0.115
Et-EDTA-Ca	0.994	0.125	Ph-EDTA-Ca	0.994	0.125
Et-EDTA-Sr	0.995	0.124	Ph-EDTA-Sr	0.995	0.124
Et-EDTA-Ba	0.995	0.117	Ph-EDTA-Ba	0.995	0.119

*) For designations, see Table 14.

Table 16. Predicted values of stability constants (log K_s) for 24 chelates using the leave-one-out method.

Chelate *	Actual value	Predicted value	Error
EDTA-Mg	8.79	9.15	-0.36
EDTA-Ca	10.69	10.71	-0.02
EDTA-Sr	8.73	8.68	0.05
EDTA-Ba	7.86	7.53	0.33
Me-EDTA-Mg	10.02	10.06	-0.04
Me-EDTA-Ca	11.63	11.76	-0.13
Me-EDTA-Sr	9.60	9.76	-0.16
Me-EDTA-Ba	8.55	8.72	-0.17
Et-EDTA-Mg	10.15	10.03	0.12
Et-EDTA-Ca	11.75	11.74	0.01
Et-EDTA-Sr	9.66	9.75	-0.09
Et-EDTA-Ba	8.50	8.73	-0.23
Ip-EDTA-Mg	10.30	10.00	0.30
Ip-EDTA-Ca	11.77	11.74	0.03
Ip-EDTA-Sr	9.74	9.73	0.01
Ip-EDTA-Ba	8.60	8.71	-0.11
Ib-EDTA-Mg	10.16	10.03	0.13
Ib-EDTA-Ca	11.84	11.72	0.12
Ib-EDTA-Sr	9.84	9.71	0.13
Ib-EDTA-Ba	8.75	8.67	0.08
Ph-EDTA-Mg	9.40	9.69	-0.29
Ph-EDTA-Ca	11.25	11.28	-0.03
Ph-EDTA-Sr	9.32	9.23	0.09
Ph-EDTA-Ba	8.36	8.13	0.23

*) For designations, see Table 14.

Chapter 10

CRA Models for Rate Constants of Solvolysis

As discussed in earlier chapters, the categorized properties in a CRA model are usually related to molecular structures. In the rate-constant problem described in this chapter, the categorized properties are not restricted to variations in molecular structures. The rate constant of a reaction depends not only on the properties of the reactants, but also on the reaction conditions, such as the temperature and the reaction medium.

The rate constants k for the solvolysis of 14 substituted ethyl benzoate esters in basic alcohol solution are selected from 53 rate constants [44]. Ester solvolysis of a benzoate is a homogeneous, second-order reaction:

$$ArCOOR + OR'^{-} \text{ --------> } ArCOO^{-} + ROR'$$

The values of pk (pk = - log k) for the 12 substituted benzoates are employed as the dependent variable. The properties of the reactants and the reaction conditions are selected as independent variables. The properties of the reactants are 1) the type of substituent and 2) the position of the substituent. The substituents on the ring are the halogen, nitro, methyl and methoxy groups. The position of the substituents on the ring are ortho (o), meta (m) and para (p). The

reaction conditions are temperature (25 and 50 °C) and the reaction medium (three mixed solvents). The three mixed solvents are abbreviated as E85, A40 and E40. Solvent E85 contains 85 % ethyl alcohol and 15 % water by volume; solvent A40 contains the 40 % acetone and 60 % water by volume; and solvent E40 contains 40 % ethyl alcohol and 60 % water by volume.

Three models are studied: 1) the 12 substituted benzoates in solvent E85 at 25 °C, involving two reactant properties containing seven categories (called the 12 x 7 model), 2) the 12 substituted benzoates in solvent E85 at 25 and 50 °C, involving the two reactant properties and the reaction temperatures (called the 24 x 9 model), and 3) the substituted benzoates in all three solvents (E85, A40, and E40) at 25 °C (called the 18 x 10 model). The values of log k for the 12 x 7 and the 24 x 9 models are given in Table 17. The CRA model for the 18 x 10 problem is given in Table 18. The statistical evaluation of the three models is given in Table 19.

The statistical analysis of the three models (Table 19) shows that the multiple correlation coefficients of the three models are quite high, the standard error of the residuals for the three models are about 0.2 - 0.3 pk units which is within the estimated experimental error of about 0.2 - 0.5 pk units. The calculated F-test values of the 12 x 7 model is less than the table F-value, showing an

Table 17. Values of rate constants (-log k) for ester solvolysis involving substituted ethylbenzoates in solvent E85 at 25 and 50 °C.

Ester	-log k	
	25 °C	50 °C
o-methyl benzoate	4.11	3.96
m-methyl benzoate	3.36	2.34
p-methyl benzoate	3.61	2.57
p-methoxy benzoate	3.94	2.88
o-nitro benzoate	2.27	1.41
o-fluoro benzoate	2.63	1.68
p-fluoro benzoate	2.90	1.92
o-chloro benzoate	2.86	1.96
m-chloro benzoate	2.32	1.40
p-chloro benzoate	1.63	1.67
p-bromo benzoate	2.54	1.59
p-iodo benzoate	2.56	1.61

Table 18. CRA model for the 17 x 10 problem involving solvolysis of substituted ethyl benzoates in different solvents at 25 °C.

Properties (Categories)

1 Position of substituent (o, m, p)
2 Substituent (CH_3,OCH_3,NO_2, X)
3 Solvent (E85, A40, E40) *

		Categorized properties									
	Property	1			2				3		
	Category	1	2	3	1	2	3	4	1	2	3
Ester	-log k										
o-me benzoate	4.11	1	0	0	1	0	0	1	1	0	0
m-me benzoate	3.36	0	1	0	1	0	0	0	1	0	0
m-me benzoate	2.77	0	1	0	1	0	0	0	0	1	0
p-me benzoate	2.94	0	0	1	1	0	0	0	0	1	0
p-me benzoate	3.60	0	0	1	1	0	0	0	1	0	0
p-methoxy benzoate	2.65	0	0	1	0	1	0	0	0	1	0
p-methoxy benzoate	3.94	0	0	1	0	1	0	0	1	0	0
p-methoxy benzoate	3.10	0	0	1	0	1	0	0	0	0	1
o-nitro benzoate	2.27	1	0	0	0	0	1	0	1	0	0
m-nitro benzoate	0.77	0	1	0	0	0	1	0	0	0	1
p-nitro benzoate	0.61	0	0	1	0	0	1	0	0	1	0
o-fluoro benzoate	2.64	1	0	0	0	0	0	1	1	0	0
p-fluoro benzoate	2.90	0	0	1	0	0	0	1	1	0	0
o-chloro benzoate	2.86	1	0	0	0	0	0	1	1	0	0
m-chloro benzoate	2.32	0	1	0	0	0	0	1	1	0	0
p-Bromo benzoate	2.54	0	0	1	0	0	0	1	1	0	0
p-iodo benzoate	2.56	0	0	1	0	0	0	1	1	0	0

*) For designations, see p. 56.

Table 19. Results of three models involving rate constants of ester solvolysis.

Problem	No. properties (No. categories)	Details of properties (Categories)	SE	R	R_p	F-value calc. (table) (df_1,df_2)
12 x 7	2 (3,4)	Position of subsituent (o,m,p) Substituent (CH_3,OCH_3,NO_2, X)	0.28	0.916	0.593 0.916	4.3 (4.95) (6,5)
24 x 9	3 (3,4,2)	Position of subsituent (o,m,p) Substituent (CH_3,OCH_3,NO_2, X) Temperature (25,50)	0.22	0.956	0.643 0.934 0.891	24.4 (2.71) (7,16)
18 x 10	3 (3,4,3)	Position of substituent (o,m,p) Substituent (CH_3,OCH_3 NO_2, X) Solvent (E85,A40,E40)	0.25	0.960	0.597 0.953 0.833	13.3 (3.2) (8,9)

insignificant regression for the model. The calculated F-values for the 24 x 9 and 18 x 10 models are much greater than the table F-test values. The partial correlation coefficients of the third propeties (the reaction temperature and the reaction medium) in the 24 x 9 and the 18 x 10 models are high, showing the categorized properties relating to the temperature and the solvents play very important roles in these two models. The 24 x 9 and 18 x 10 models are more satisfactory than the 12 x 7 model.

The study of the rate constant problem shows the selection of the categorized properties is not restricted to the structures of the substances. The categorized properties also may be suggested by the reaction conditions.

Chapter 11

MRA-CRA Combination Models

In categorical regression analysis (CRA), the selection of independent variables is easier than in multiple regression analysis (MRA), because categorized properties can be related in a simple way to the molecular structure as we demonstrated in previous chapters. We now develop a MRA-CRA combination model to broaden the scope of regression analysis and to relate the two approaches.

In a MRA-CRA model, the independent variables include both ordinal numbers (MRA) and categorized properties (CRA). A set of physical or chemical data can be modeled by (q + p) properties, where the q properties are standard MRA properties and the p properties are categorized properties containing a total of c categories. The model is a MRA-CRA combination model containing a total of q + c independent variables. In general, if the independent variables in a model include ordinal numbers only (indexed by k = 1, 2, . . . , q), the model is an MRA model; if the independent variables are CRA properties only (indexed as in Eq. [1] in Chapter 2), the model is a CRA model. As mentioned earlier, a CRA variable only has a value of either "1" or "0", and the sum of the independent variables must equal to unity for each CRA property for each compound.

Each CRA property must contain at least two categories. In the simplest MRA-CRA model, the independent variables include one MRA variable and one CRA property containing at least two categories.

The mathematical expressions for a MRA model and for a MRA-CRA model are similar. The calculations for a MRA-CRA model can be carried out using the statistical analysis (SAS) software. The regression coefficients A and the intercept b of a MRA-CRA model can be calculated as in MRA. The estimated values of the dependent variable for a MRA-CRA model can be obtained by using the standard equation:

$$\hat{Y} = b + [X]A \qquad [14]$$

where $\hat{Y}$ is the vector of the estimated values of the dependent variables, b is the intercept of the regression, A is the vector of the regression coefficients and [X] is the matrix of the independent variables.

In the SAS program, the regression coefficients for the last category (not the first which we chose in Chapter 2) of each CRA property in the combination model are set to zero. Thus the regression coefficients in a MRA-CRA model contain p zero's for p CRA properties. Then the CRA technique can be employed to obtain normalized regression coefficients for the categorized variables in the MRA-CRA

model as shown below.

As discussed in Chapter 2, the normalized category coefficients in a CRA model can be obtained from Eq. [8]. Then the estimated values of the dependent variables in the CRA model can be obtained from Eq. [9]. In a MRA-CRA model, the normalized category coefficients also can be obtained using Eq. [8] for the categorized properties. The estimated values of the dependent variables in the combination model then can be obtained by a modified form of Eq. [9]. Because of the two kinds of independent variables (MRA and CRA) in a combination model, Eq. [9] cannot be applied to the MRA-CRA model directly.

The required equation is derived as follows: According to Eq. [14], an estimated value $\hat{y}_m$ of the m-th substance in a MRA-CRA model can be expressed as:

$$\hat{y}_m = b + \sum_{k=1}^{q} x_{km}\, a_k + \sum_{i=1}^{p} \sum_{j=1}^{c} x_{ijm}\, a^*_{ij} \qquad [15]$$

where a_k is the regression coefficient of the k-th multiple regression variable x_{km}, and a^*_{ij} is the category coefficient of j-th category in the i-th property for the CRA variable x_{ijm}. The first term in Eq. [15] is the intercept of the standard regression, the second term gives the contribution from the multiple regression variables, and the third term gives the contribution from the categorized variables.

Rewriting Eq. [8] in Chapter 2, we have

$$a^*_{ij} = a_{Nij} + \sum_{j=1}^{c} a^*_{ij} f_{ij} \qquad [8]$$

where f_{ij}, the fraction of 1's in the j-th column of the i-th CRA property, is defined as $f_{ij} = \sum x_{ijm} / n$ for the fixed i and j.

Substituting Eq. [8] into Eq. [15] gives

$$\hat{y}_m = b + \sum_{k=1}^{q} x_{km} a_k + \sum_{i=1}^{p}\sum_{j=1}^{c} x_{ijm} (a_{Nij} + \sum_{j=1}^{c} a^*_{ij} f_{ij})$$

$$= b + \sum_{k=1}^{q} x_{km} a_k + \sum_{i=1}^{p}\sum_{j=1}^{c} x_{ijm} a_{Nij} + \sum_{i=1}^{p}\sum_{j=1}^{c} a^*_{ij} f_{ij} \sum_{j=1}^{c} x_{ijm}$$

Since $\sum x_{ijm} = 1$ for fixed i and m (see Eq. [2] in Chapter 2), Eq. [15] becomes

$$\hat{y}_m = b + \sum_{k=1}^{q} x_{km} a_k + \sum_{i=1}^{p}\sum_{j=1}^{c} x_{ijm} a_{Nij} + \sum_{i=1}^{p}\sum_{j=1}^{c} a^*_{ij} f_{ij} \qquad [16]$$

According to the definition of f_{ij}, the fourth term in Eq. [16] becomes

$$\sum_{i=1}^{p}\sum_{j=1}^{c} a^*_{ij} f_{ij} = (\sum_{m=1}^{n}\sum_{i=1}^{p}\sum_{j=1}^{c} x_{ijm} a^*_{ij}) / n = \bar{Y}_{cra}$$

This term is therefore a constant for a given data set, representing the mean value calculated from the categorized variables and catego-

ry coefficients. Thus Eq. [16] can be simplified as

$$\hat{y}_m = b + \bar{Y}_{cra} + \sum_{k=1}^{q} x_{km}\, a_k + \sum_{i=1}^{p}\sum_{j=1}^{c} x_{ijm}\, a_{Nij} \qquad [17]$$

Using matrix algebra, Eq. [17] can be rewritten as

$$\bar{Y} = (b + \bar{Y}_{cra}) + [X]\, A_{comb} \qquad [18]$$

Since from Eq. [15], $\bar{Y} = b + \bar{Y}_{mra} + \bar{Y}_{cra}$, Eq. [18] can be rewritten as

$$\hat{Y} = (\bar{Y} - \bar{Y}_{mra}) + [X]\, A_{comb} \qquad [19]$$

where Y is the vector of the estimated values for the dependent variables, $\bar{Y}$ is the mean value of the dependent variables, Y_{mra} is the mean value calculated from the k MRA variables and coefficients, A_{comb} is the vector of the normalized regression coefficients for the combination model (which contains the regression coefficients for the q MRA variables and the normalized category coefficients for the c categorized properties), and [X] is the matrix of the independent variables.

For calculational purposes, Eq. [19] is best written as

$$\hat{Y} = \sum_{m=1}^{n} (y_m - \sum_{k=1}^{q} y_{km}) / n + [X]\, A_{comb} \qquad [20]$$

Equation [20] is a general equation which summarizes the combina-

tion regression model.

Now let's set up a simple MRA-CRA combination model for the boiling points of alcohols which was chosen as our example for CRA in Chapter 3. In this 8 x 6 problem, the first property (the number of carbon atoms) can be replaced by the actual number of carbon atoms. Then the first property becomes a MRA variable. The second property (type of alcohols) remains a CRA variable, because that property is difficult to express as ordinal or decimal numbers. The boiling points of alcohols can be expressed as a MRA-CRA model, in which the first property is an MRA variable and the second property is a CRA variable containing three categories. Thus, the 8 x 6 problem in CRA becomes an 8 x 4 problem in the combination model.

The data and the combination model are shown in Table 20. Using the SAS procedure, the regression coefficient for the last category in the single CRA property is equal to zero automatically, because the SAS program checks for variables which might be linear combinations of other variables. The deletion process is carried out for even one CRA property for a MRA-CRA model, whereas at least two CRA properties are required for the deletion process in our CRA model (see Chapter 2). In order to find the normalized category coefficients, Eq. [8] is then applied to the CRA categorized property (as

Table 20. MRA-CRA combination model for the boiling points of alcohols.

Properties

1 Number of carbon atoms (MRA variable)
2 Type of alcohol (CRA variable, p-, s-, t-)

		Properties			
	Property	1	2		
	Category		1	2	3
Alcohol	B. P.				
1-pentanol	137.8	5	1	0	0
2-pentanol	119.0	5	0	1	0
2-methyl 2-butanol	102.0	5	0	0	1
1-hexanol	157.0	6	1	0	0
2-hexanol	139.9	6	0	1	0
2-methyl 2-pentanol	121.4	6	0	0	1
1-heptanol	176.2	7	1	0	0
2-methyl 2-hexanol	142.8	7	0	0	1

detailed in Chapter 3). The regression coefficient for the first property and the calculated normalized category coefficients for the second property are:

Property	1 (No. carbon atoms)	2 (Type of alcohol)		
Variable type	MRA	CRA		
Category		1	2	3
Actual	5, 6, and 7	p-	s-	t-
A	19.92	34.93	17.34	0
A_{comb}	19.92	17.5	-0.10	-17.4

The coefficient for the third category of the CRA property is equal to zero. The intercept for regression (Eq [14]) has a value of 2.53. The normalized coefficients for the second property in the 8 x 4 combination model have the same values as those in the 8 x 6 problem in Chapter 3.

The estimated values of the combination model can be calculated by using either regression coefficients (using Eq [14]) or normalized category coefficients (using Eq. [20]). The computations using both equations for 1-pentanol are as follows:

a) Using the regression coefficients A above, matrix [X] and intercept b (Eq. [14]), the estimated boiling point (°C) for 1-pentanol is

$$\hat{y}_{1\text{-PeOH}} = 2.53 + (19.92 \times 5) + [(34.93 \times 1) + (17.34 \times 0) + (0 \times 0)]$$
$$= 137.1$$

b) Using the normalized category coefficients A_{comb} and matrix [X] (Eq. [20]) , the estimated boiling point (°C) for 1-pentanol is

$$\hat{y}_{1\text{-PeOH}} = 19.98 + (19.92 \times 5) + [(17.5 \times 1) + (-0.1 \times 0) + (-17.4 \times 0)]$$

$$= 137.1$$

where the first term using Eq. [20] is:

$$[(137.8 - 19.92 \times 5) + (119.0 - 19.92 \times 5) + (102.0 - 19.92 \times 5)$$

$$+ (157.0 - 19.92 \times 6) + (139.9 - 19.92 \times 6) + (121.4 - 19.92 \times 6)$$

$$+ (176.2 - 19.92 \times 7) + (142.8-19.92 \times 7)] / 8 = 19.98$$

The estimated boiling points of the eight alcohols in the MRA-CRA model using Eq. [14] and Eq. [20] are given in Table 21. Identical estimated boiling points are obtained from the two equations.

Comparison of the estimated boiling points between the complete CRA model (Chapter 3) and the MRA-CRA model for the eight alcohols is shown also in Table 21. Except for two alcohols, exactly the same estimated boiling points are obtained, showing the complete CRA and MRA-CRA combination models are equivalent models in this problem.

Statistical comparison of the models is an important objective. The statistical evaluations for the CRA and MRA-CRA models are given in the lower portion of Table 21. The standard errors of the residuals for the two models (0.61 and 0.57, respectively) are quite similar. The multiple correlation coefficients are identical (0.9996).

Table 21. Estimated values from CRA and MRA-CRA models for boiling points of alcohols.

Compound	Actual B. P.	Estimated B. P. Eq. [9] [a]	Eq. [14] [b]	Eq. [20] [c]
1-pentanol	137.8	137.1	137.1	137.1
2-pentanol	119.0	119.5	119.5	119.5
2-methyl 2-butanol	102.0	102.2	102.1	102.2
1-hexanol	157.0	156.9	157.0	156.9
2-hexanol	139.9	139.4	139.4	139.4
2-methyl 2-pentanol	121.4	122.1	122.1	122.0
1-heptanol	176.2	176.9	176.9	177.0
2-methyl 2-hexanol	142.8	142.0	142.0	142.1
SE		0.57	0.61	0.61
Multiple correlation coefficient R		0.9996	0.9996	0.9996
F-value (calc.)		574	1900	1900
F-value (table)		39	10	10

a) CRA model in Chapter 3.

b) MRA-CRA model with category coefficients.

c) MRA-CRA model with normalized category coefficients.

The calculated F-test values of both MRA-CRA and CRA models are much greater than table F-value. The calculated F-value of the MRA-CRA model is much greater than the calculated F-value of the CRA model, in part because the number of degrees of freedom in the combination model are smaller than in the CRA model. In the CRA model, the number of degrees of freedom for the model (df_1) is five, and the number of degrees of freedom for the error (df_2) is two. However, in the MRA-CRA model, the number of degrees of freedom are three and four, respectively. The results show that the CRA and MRA-CRA models are statistically equivalent and that both are more than satisfactory models. The combination model has the advantage of fewer independent variables.

The MRA-CRA model can be applied to any CRA model if non-categorized values can be given to some of the categorized properties. To illustrate, an MRA-CRA model for retention indices is studied. In Chapter 6, we built a CRA model for the retention indices of hydrocarbons, called the 21 x 8 problem. In the CRA model, two of the three properties (total number of carbon atoms and total number of methyl branches) can be replaced by two MRA variables which are the actual numbers of carbon atoms and actual number of methyl branches. Thus, the two CRA properties including six independent variables become two MRA variables.

All the variables for the MRA-CRA models for the retention-index problems are given in Table 22. Two combination models with the following properties from Table 22: 1) the first MRA property and the second and third CRA properties, and 2) the first and second MRA properties and the third CRA property, are studied. The two combination models are called the 21 x 6 and 21 x 4 models. The normalized regression coefficients and the four statistical values for the two combination models and the complete CRA model (the 21 x 8 model from Chapter 7) are given in Tables 23 and 24, respectively.

From Table 23, we find that the normalized category coefficients in the CRA and MRA-CRA models have the same pattern and almost the same magnitude for the individual categorized properties. The behavior of the normalized coefficients implies that normalized coefficients for a certain categorized property may be nearly independent of the other MRA properties in a model.

From Table 24, the standard errors of the residuals of the three models are very close, the multiple correlation coefficients are quite high and have the same values, and the calculated F-values of the three models are much greater than the table F-values.

The statistical evaluations of the three models indicate that the 21 x 6 and 21 x 4 MRA-CRA combination models and the 21 x 8 CRA model are equivalent statistically and all are satisfactory models.

Table 22. Variables for MRA-CRA combination models for retention indices of hydrocarbons.

Properties

1 Total number of carbon atoms (MRA and CRA variables)
2 Number of methyl branches (MRA and CRA variables)
3 Number of quaternary carbon atoms (CRA variable only)

		MRA Properties		CRA properties							
Compound	R.I.	1	2	1	2			3			
2,2-dimethyl butane	530	6	2	1	0	0	0	1	0	0	1
2,3-dimethyl butane	562	6	2	1	0	0	0	1	0	1	0
2-methyl pentane	568	6	1	1	0	0	1	0	0	1	0
3-methyl pentane	583	6	1	1	0	0	1	0	0	1	0
2,2-dimethyl pentane	622	7	2	0	1	0	0	1	0	0	1
2,4-dimethyl pentane	628	7	2	0	1	0	0	1	0	1	0
2,2,3-trimethyl butane	643	7	3	0	1	0	0	0	1	0	1
3,3-dimethyl pentane	660	7	2	0	1	0	0	1	0	0	1
2,3-dimethyl pentane	672	7	2	0	1	0	0	1	0	1	0
2-methyl hexane	667	7	1	0	1	0	1	0	0	1	0
3-methyl hexane	677	7	1	0	1	0	1	0	0	1	0

(continued)

Table 22. (continued)

Compound	R. I.	MRA properties		CRA properties							
		1	2	1			2			3	
2,2,4-trimethyl pentane	692	8	3	0	0	1	0	0	1	0	1
2,2-dimethyl hexane	721	8	2	0	0	1	0	1	0	0	1
2,5-dimethyl hexane	728	8	2	0	0	1	0	1	0	1	0
2,2,3-trimethyl hexane	739	8	3	0	0	1	0	0	1	0	1
3,3-dimethyl hexane	746	8	2	0	0	1	0	1	0	0	1
2,3-dimethyl hexane	761	8	2	0	0	1	0	1	0	1	0
2-methyl heptane	764	8	1	0	0	1	1	0	0	1	0
4-methyl heptane	769	8	1	0	0	1	1	0	0	1	0
3,4-dimethyl hexane	773	8	2	0	0	1	0	1	0	1	0
3-methyl heptane	775	8	1	0	0	1	1	0	0	1	0

Table 23. Normalized category coefficients for CRA model and MRA-CRA models for retention indices.

Property	1			2			3	
Category	1	2	3	1	2	3	1	2
21 x 8 (CRA)	-126.0	-25.6	68.3	13.0	-4.45	-14.03	7.16	-11.64
21 x 6 (MRA-CRA)		96.57		12.83	-4.37	-13.91	7.11	-11.55
21 x 4 (MRA-CRA)		91.85			-17.93		7.41	-12.05

Table 24. Results of MRA-CRA models and CRA model for retention indices.

Problem	No. total properties	No. MRA variables	No. CRA variables (Categories)	SE	R	F-value calc.(table) (df_1,df_2)
21 x 8 (CRA)	3	0	3 (3,3,2)	15.0	0.980	52.8 (3.5) (6,14)
21 x 6 (MRA-CRA)	3	1	2 (3,2)	15.8	0.980	89.7 (3.8) (4,16)
21 x 4 (MRA-CRA)	3	2	1 (2)	16.8	0.980	95.7 (3.9) (3,18)

However, the two combination models have the higher F-values, showing the higher significance of the regressions due to the smaller number of degrees of freedom. So a MRA-CRA combination model has the added advantages of a higher significance of the regression.

Part II

Comparison of the Maximum-Minimum Distance Clustering Method with Other Clustering Methods

Chapter 12

Introduction to Clustering Methods

Cluster analysis is the classification of objects, characterized by their qualitative and quantitative properties, into groups [1,7]. Cluster analysis methods identify groups of similar vectors in multivariate data sets.

Each point in multivariate data sets is conveniently represented by a vector, called a pattern vector: $X = (x_1, x_2, \ldots, x_n)$. A vector composed of n measurement constitutes a set of n scalar values (coordinates). These patterns are the rows or columns of a data matrix. The n variables define an n-dimensional or n-variate pattern space.

When the points in pattern space are near each other, the associated vectors are considered to be similar. The similarity between the elements can be measured in many ways. The most common measure is the Euclidian distance between objects, defined as:

$$d_{pr} = [\sum_{i=1}^{k} (x_{ip} - x_{ir})^2]^{1/2}$$

where p and r are two samples (vectors) in a k-dimensional space. The smaller the distance, the more similar the two vectors. In cluster analysis, each vector is compared with all the other vectors. Those that are very similar will then subsequently be found in the same cluster, while dissimilar ones will be found in different clusters.

There are many methods in cluster analysis, leading to different classifications according to how the distance between the two vectors is computed. In this study, the maximum-minimum distance clustering method, a new approach for cluster analysis in chemistry, is applied to several chemical problems, and is compared to other clustering methods, such as hierarchical cluster analysis and varimax-rotated factor analysis. The target-testing technique of factor analysis is employed to check the correctness of clusters obtained from the different methods.

Maximum-Minimum Distance Cluster Analysis

The principle of the maximum-minimum distance method (abbreviated max-min here) was introduced into the pattern recognition literature in the 1960's [45]. To explain the method, the principles and an example of an application in chemistry are demonstrated in detail below.

The max-min method is a simple heuristic procedure based on

the Euclidian distance concept. To illustrate the principle, we consider a problem involving two-dimensional properties first. Multi dimensional-property problems can be solved similarly.

Suppose that there are 12 samples in a two-dimensional vector space as shown in Figure 1. We carry out the following procedures: 1) We arbitrarily let x_1 become the first cluster center, designated by C_1. 2) We determine the farthest sample (called cluster center C_2) from x_1 by comparing the distances from other points to the first cluster center C_1. Center C_2 is x_{10} in our example. 3) We compute the distances from each remaining sample to centers C_1 and C_2. For each sample, we save the minimum distance. Then, we select the maximum of these minimum distances. If this distance is an appreciable fraction of the distance between cluster centers C_1 and C_2 (say, at least one half of this distance), we are justified in creating a new cluster center. In our example x_6 is cluster center C_3. Repeating the process, we compute the distances from each of the three established cluster centers to the remaining samples and save the minimum of every group of three distances. We again select the maximum of these minimum distances, and so forth.

In the general case, the above procedure is repeated until the new maximum distance at a particular step fails to satisfy the

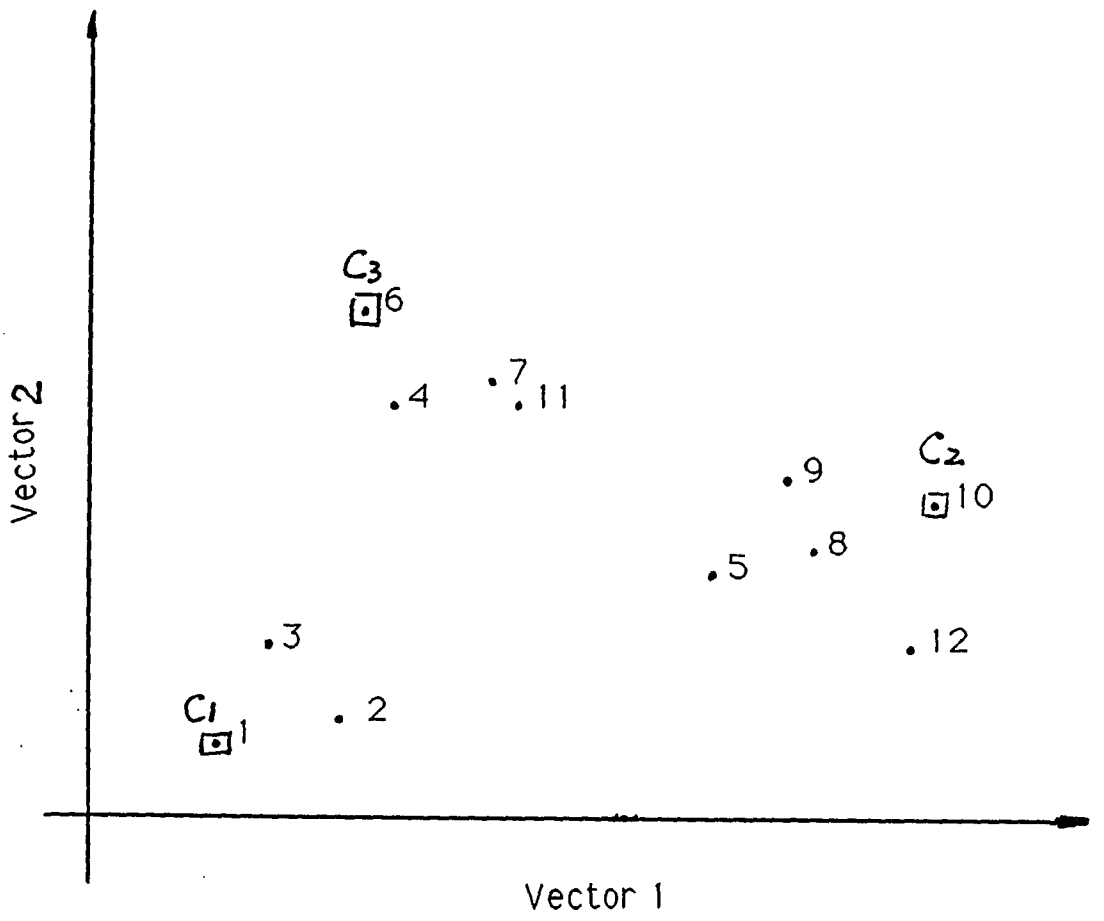


Figure 1. 12 samples in a two-dimensional space with three cluster centers C indicated.

condition for the creation of a new cluster center. In this example, we obtain three cluster centers: x_1, x_{10} and x_6. To assign the remaining samples to the domain of these centers, we assign each sample to its nearest cluster center.

Now, we apply the method to a chemical problem. Suppose seven compounds have to be clustered into groups based on their retention indices on 10 stationary phases. The retention indices of the seven compounds [46] are shown in Table 25.

First, principal-component factor analysis [6] is applied to the 7 x 10 matrix to estimate the correct number of factors. Then a reduced-size matrix with the correct number of factors is formed. Here, we omit the details of factor analysis and give the results only. Three factors are indicated to be the correct number for the data matrix, and the three factor patterns are rotated using the varimax method [47]. The varimax-rotated patterns of the three vectors are:

	Vector 1	Vector 2	Vector 3
Benzene	0.6057	0.5846	0.5395
Toluene	0.5946	0.5834	0.5530
Et butyrate	0.7245	0.5522	0.4125
Pr butyrate	0.7238	0.5556	0.4091
Pr propionate	0.7320	0.5450	0.4088
Hexanol	0.5579	0.7272	0.3998
Heptanol	0.5621	0.7244	0.3990

Table 25. Data matrix of retention indices for the 7 x 10 problem.

	Stationary phase [a]									
Compound [b]	APJ	APM	SE30	SE31	PH1	PH2	PP5	PP6	CW6K	CW20
Benzene	685	685	678	666	750	739	821	828	964	961
Toluene	797	797	784	773	860	848	928	935	1076	1061
Et butyrate	750	751	793	783	848	841	923	926	1034	1031
Pr butyrate	848	851	892	882	949	940	1022	1022	1131	1129
Pr propionate	755	755	800	790	854	846	931	935	1040	1038
Hexanol	841	841	890	862	995	984	1028	1033	1332	1331
Heptanol	942	941	991	964	1099	1087	1133	1137	1436	1436

a) For designations, see p. 95.

b) Et - ethyl, Pr- propyl.

After principal-component factor analysis, the original 10-dimensional problem (7 x 10) is reduced to a 3-dimensional rotated-pattern problem (7 x 3). Clustering the seven compounds on three factors becomes much simpler than on the original 10-factor data matrix.

Now, the max-min method is applied to the 3-dimensional matrix. Following the max-min procedures as discussed earlier, we arbitrarily select the first compound, benzene, as the first cluster center C_1. Next, we calculate the distances from each compound to the first cluster center (benzene) C_1, giving the distances:

Compound	D to C_1 (benzene)	Compound	D to C_1 (benzene)
Toluene	0.018	Pr propionate	0.186
Et butyrate	0.177	Hexanol	0.205
Pr butyrate	0.178	Heptanol	0.203

The farthest compound from benzene is hexanol which has the maximum distance 0.205. The second cluster center C_2 then is hexanol.

Then, the distances from each of the remaining compounds to benzene and hexanol are computed:

Compound	D to C_1 (benzene)	D to C_2 (hexanol)	Minimum distance
Benzene	0.0	0.205	0.0
Toluene	0.018	0.213	0.018
Et butyrate	0.177	0.242	0.177
Pr butyrate	0.178	0.239	0.178
Pr Propionate	0.186	0.252	0.186
Hexanol	0.205	0.0	0.0
Heptanol	0.203	0.005	0.005

After comparison of the two distances, the minimum distance from each compound to the two cluster centers is listed in the last column. The maximum distance between these minimum distances is 0.186, corresponding to propyl propionate which thus is identified as the third cluster center C_3.

Finally, we assign each sample to its nearest cluster center according to the three distances:

Compound	D to C_1 (benzene)	D to C_2 (hexanol)	D to C_3 (pr propionate)	Belongs to cluster
Benzene	0.0	0.205	0.186	1
Toluene	0.018	0.213	0.203	1
Et propionate	0.177	0.242	0.011	3
Pr butyrate	0.178	0.239	0.013	3
Pr propionate	0.186	0.252	0.0	3
Hexanol	0.205	0.0	0.252	2
Heptanol	0.203	0.005	0.247	2

The first two solutes, aromatic compounds benzene and toluene, belong to cluster 1. The last two solutes, hexanol and heptanol, are in cluster 2. The remaining three solutes, all esters, are in cluster 3. The max-min cluster method gives results consistent with chemical insight.

Hierarchical Cluster Analysis

Hierarchical clustering methods are widely applied in biology, geology, and the environmental sciences. Since the principles of the methods are well-known [1], we shall not discuss the method in detail.

In hierarchical clustering analysis, each vector begins in a cluster by itself. The vector distances are compared and the two closest vectors are merged to form a single new cluster. Merging of the two closest vectors is repeated stepwise until only one global cluster is left. To calculate the smallest distance between the clusters and objects, the often employed average-linkage method has been shown to give the best overall performance [7]. In the average linkage method, when a cluster has been formed, the distances between the cluster and other objects or clusters are calculated by the average distance from the cluster components to the objects. For example, if points p and q in a matrix are the most similar, they form a new cluster C.

The distance between the new object C (cluster) and another object, such as k, is obtained by averaging the distances of q and p with k: $D_{kc} = (D_{kq} + D_{kp}) / 2$. The classification is represented in a dendrogram. The dendrogram is the best way to visualize the relationship between the objects.

Unfortunately, there are no satisfactory methods for determining the total number of significant clusters. A computer simulation leads to the cubic clustering criterion (CCC), which can be used for crudely estimating the number of clusters [14]. In practice, the number of clusters may be decided arbitrarily.

Clustering with Factor Analysis

Principal factor analysis (PFA) yields an abstract solution consisting of a set of abstract eigenvalues and an associated set of abstract eigenvectors [6,48]. Each eigenvalue measures the relative importance of the associated factor. A large eigenvalue indicates a major factor, whereas a very small eigenvalue indicates an unimportant factor. PFA reduces the size of the data matrix to a smaller number of factors (vector patterns).

Two types of transformation for the abstract solutions involve 1) rotation of principal factors and 2) target testing, giving solutions with more chemical meaning. The most popular rotation is the vari-

max rotation [47], which tends to yield a few large values, many small values and relatively few intermediate values, thereby emphasizing clusters of vectors. By bringing out clustering in the data, rotated factors are valuable for classifying vectors. Two-dimensional plots of the rotated factor patterns give visual information on the clusters. In this thesis, varimax-rotated factor patterns are employed as the data matrix for the max-min method to obtain more accurate clusters.

The target-test technique [6,49] is employed to further evaluate the correctness of the clusters obtained from other methods. An input target vector, which is designed to represent a cluster, is compared with its predicted vector obtained from a least-squares calculation. If the predicted vector is in good agreement with the input target vector, the cluster is considered a valid cluster. The details are discussed in Chapter 14.

Chapter 13

Procedures and Computer Programs

In this chapter, we describe the general procedures for clustering used by us and introduce several computer programs which are employed in the calculations.

The clustering procedures involve the following steps:

1) hierarchical cluster analysis of the original data matrix, using the average-linkage method,
2) principal factor analysis of the original data matrix, followed by varimax rotation of the factor patterns,
3) maximum-minimum distance cluster analysis of the varimax-rotated factor patterns, and
4) target testing with factor analysis of the clusters obtained from steps 1 and 3.

A flow diagram of the procedures is given in Figure 2.

Several computer programs are used for the calculations. The statistical analysis system (SAS) software package includes programs for carrying out hierarchical clustering analysis and principal-component factor analysis with varimax rotation. One can use the SAS programs by following the voluminous instructions in the SAS user

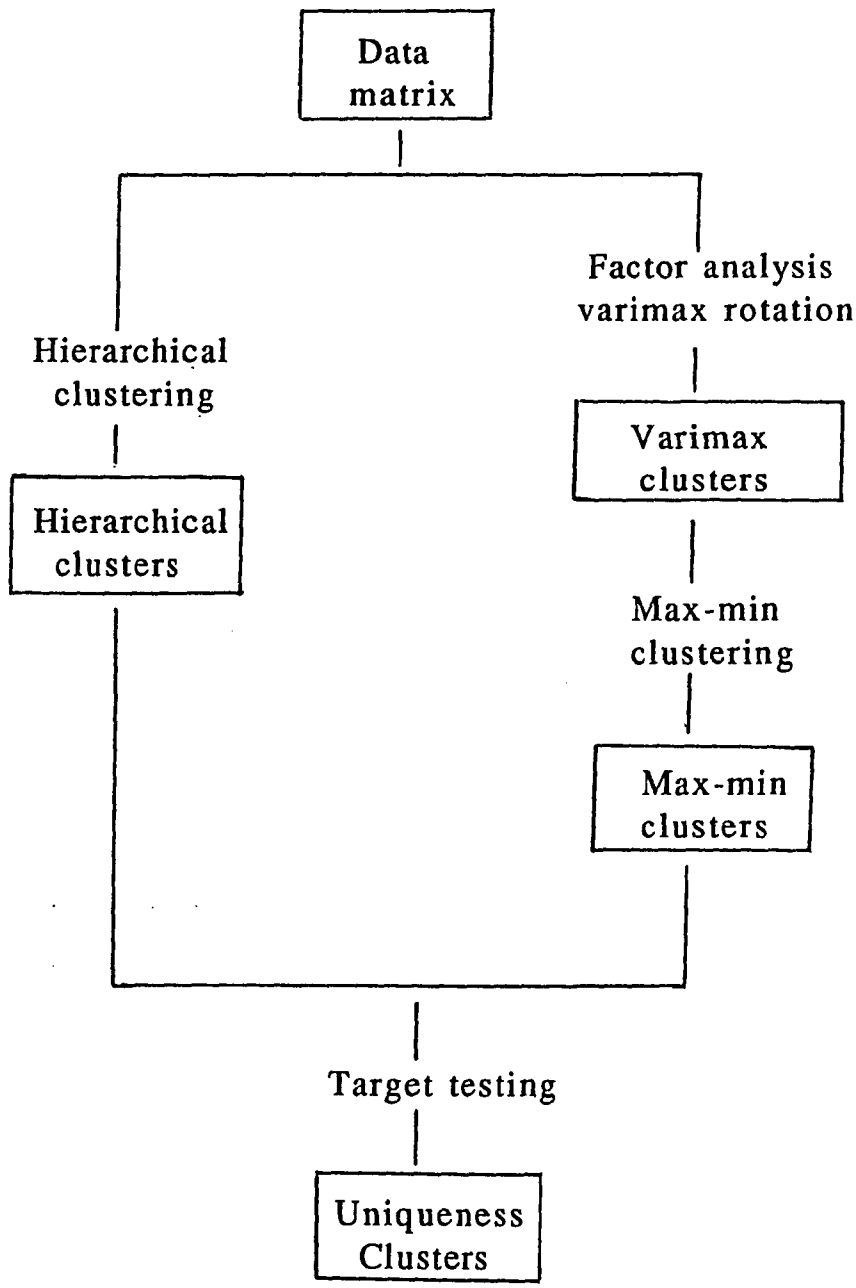


Figure 2. Flow diagram of clustering procedures.

guides [10-14]. The program for max-min clustering is written by us in BASIC, to run on the IBM PC. The instructions for using the max-min program are given in appendix C. A computer output from the max-min program for the data set for the example in Chapter 12 (the 7 x 10 problem) is given in Table 26. Computer procedures for target testing are carried out on the FORTRAN programs FACTANAL and TARGET, which were developed by Malinowski, Howery and coworkers [50,51]. Instructions for using FACTANAL and TARGET are documented in those programs.

Table 26. Computer output from the max-min method for the 7 x 10 problem.

Maximum-minimum distance cluster analysis of pattern vectors

7 samples & 3 pattern vector data matrix

SP	VECTOR 1	VECTOR 2	VECTOR 3
1	.6057	.5846	.5395
2	.5946	.5834	.553
3	.7245	.5522	.4125
4	.7238	.5556	.4091
5	.7321	.5451	.4088
6	.5579	.7272	.3998
7	.5621	.7244	.399

Distance table for 7 samples

SP	SP D (1- 1)	D (1- 2)	D (1- 3)
1	2 .0111	.0112	.0175
1	3 .1188	.1231	.1769
1	4 .1181	.1216	.1783
1	5 .1264	.1324	.1861
1	6 .0478	.1504	.2053
1	7 .0436	.1464	.2029
2	3 .1299	.1336	.1939
2	4 .1292	.1322	.1954
2	5 .1375	.1427	.2029
2	6 .0367	.1484	.2133
2	7 .0325	.1447	.2113
3	4 .0007	.0035	.0049
3	5 .0076	.0104	.011
3	6 .1666	.2416	.242
3	7 .1624	.2367	.2371
4	5 .0083	.0134	.0134
4	6 .1659	.2387	.2389
4	7 .1617	.2338	.234
5	6 .1742	.252	.2522
5	7 .17	.2471	.2473
6	7 .0042	.005	.0051

(continued)

Table 26. (continued)

Determination of cluster center 2

For 3 vectors

SP	D to SP 1	Min.
2	.0175	.0175
3	.1769	.1769
4	.1783	.1783
5	.1861	.1861
6	.2053	.2053
7	.2029	.2029

Max. dist. of min. .2053
Center of Cluster 2 SP 6

Determination of cluster center 3

For 3 vectors

SP	D to SP 1	D to SP 6	Min.
2	.0175	.2133	.0175
3	.1769	.242	.1769
4	.1783	.2389	.1783
5	.1861	.2522	.1861
6	.2053	0	0
7	.2029	.0051	.0051

Max. dist. of min. .1861
Center of Cluster 3 SP 5

Composition of clusters

For 3 clusters

For 3 vectors

SP	D to SP 1	D to SP 6	D to SP 5	To cluster
1	0	.2053	.1861	1
2	.0175	.2133	.2029	1
3	.1769	.242	.011	3
4	.1783	.2389	.0134	3
5	.1861	.2522	0	3
6	.2053	0	.2522	2
7	.2029	.0051	.2473	2

Chapter 14

Introduction to Retention-Index Problems

In gas chromatography, the retention index of a solute is an important parameter to quantify retention on a specific column. The concept and the definition of the retention index were discussed in Chapter 6. As mentioned earlier, a substance has different retention behavior on different stationary phases, due to the different interactions between solutes and stationary phases. Classifying solutes based on their retention behavior on stationary phases is selected as a clustering problem in this study.

Characteristics of Solutes and Stationary Phases

Interactions between solutes and stationary phases play an important role in gas chromatography. Classifications of stationary phases have been proposed by Rohrschneider [53] and McReynolds [54]. The McReynolds' constant system is now widely used in GC as a guide to the polarity of a column.

Retention indices of 34 solutes on the 10 stationary phases from McReynolds compilation [47] are selected for the clustering problems . The solutes include alcohols, ethers and esters with different numbers of carbon atoms and with different kinds of side

chains. A summary of the characteristics and designations of the 34 solutes is given in Table 27. In the designation of the compounds, the first letters A, T and S represent alcohols, ethers and esters, respectively. Compounds numbered from 1 to 15 (all alcohols) have the same number of carbon atoms but different positions of the alcohol functional group. The second letters P, S and T designate primary, secondary and tertiary alcohols, respectively. Compounds numbered from 16 to 34 (alcohols, ethers and esters) have different numbers of carbon atoms and different branches, with the second letter specifying their carbon chains. For example, the designation SEtP means that the compound is an ester, the substituted group is an ethyl group, and the ester is a propionate.

The 10 stationary phases selected from McReynolds' retention-index data for our studies are:

Compound	Designation	Compound	Designation
Apiezon J	APJ	Diisodecyl phthalate	PH2
Apiezon M	APM	Polyphenyl ether (5 rings)	PP5
Polymethyl silicone	SE30	Polyphenyl ether (6 rings)	PP6
Polymethylvinyl silicone	SE31	Carbowax 6000	CW6K
Dioctyl phthalate	PH1	Carbowax 20M	CW20

The retention-index data for the 34 solutes on the 10 stationary phases are given in Appendix D. From the McReynolds' constants for the 10 phases, the polarity of these stationary phases may be

Table 27. Characteristics of solutes in the retention-index problems.

Compound				Characteristics	
Family	No.	Name	Designation*	No. C atoms	No. side chains
ROH	1	Hexanol	AP1	6	0
	2	2-hexanol	AS1	6	0
	3	3-hexanol	AS2	6	0
	4	2-methyl 1-pentanol	AP2	6	1
	5	3-methyl 1-pentanol	AP3	6	1
	6	4-methyl 1-pentanol	AP4	6	1
	7	2-methyl 2-pentanol	AT1	6	1
	8	3-methyl 2-pentanol	AS3	6	1
	9	4-methyl 2-pentanol	AS4	6	1
	10	2-methyl 3-pentanol	AS5	6	1
	11	3-methyl 3-pentanol	AT2	6	1
	12	2-ethyl 1-butanol	AP5	6	1
	13	2,2-dimethyl 1-butanol	AP6	6	2
	14	2,3-dimethyl 2-butanol	AT3	6	2
	15	3,3-dimethyl 2-butanol	AS6	6	2
	16	Butanol	ABU	4	0
	17	Pentanol	APE	5	0
	18	Heptanol	AHP	7	0
	19	Octanol	AOC	8	0
ROR'	20	Dipropyl ether	TP2	6	0
	21	Dibutyl ether	TB2	8	0
	22	Dipentyl ether	TPE2	10	0
	23	Butyl ethyl ether	TBEt	6	0
	24	Isobutyl ethyl ether	TIBEt	6	1
	25	Isopropyl propyl ether	TIPP	6	1
	26	Isopropyl ether	TIP2	6	0

(continued)

Table 27. (continued)

Compound				Characteristics	
Family	No.	Name	Designation*	No. C atoms	No. side chains
RCOOR'	27	Ethyl propionate	SEtP	5	0
	28	Propyl propionate	SPP	6	0
	29	Butyl propionate	SBP	7	0
	30	Pentyl propionate	SPEP	8	0
	31	Ethyl butyrate	SEtB	6	0
	32	Propyl butyrate	SPB	7	0
	33	Butyl butyrate	SBB	8	0
	34	Pentyl butyrate	SPEB	9	0

*) 1) First letter: A - alcohol, T - ether; S - ester.

2) Second letter for compounds 1-15:

P - primary, S - secondary, T - tertiary alcohol.

3) Second and third letters for compounds 16-34:

Et - ethyl, P - propyl, B - butyl, PE - pentyl, IB - isobutyl,

IP - isopropyl, HP - heptyl, OC - octyl.

estimated. The Apiezons and the silicone are non-polar stationary phases, the Carbowaxes are polar, and the phthalates and ethers have intermediate polarity [54].

Summary of the Retention-Index Problems

Three problems are designed using the 34 solutes in our studies. Features are selected in order to determine the clustering ability of the methods. Classification of the 10 stationary phases is studied also. The features of the four problems are summarized in Table 28.

The first problem includes the retention indices of 15 alcohols numbered from 1 to 15, called the 15 x 10 problem. The essential features of this problem are: the alcohols have the same functional group and the same number of carbon atoms (6), but have different positions of the functional group and different side chains.

The second problem selected involves the retention indices of the 15 alcohols above and five ethers on the 10 stationary phases. A total of 20 compounds are in the problem, called the 20 x 10 problem. The essential features of the problem are: the solutes have the same number of carbon atoms (6) and the same empirical formula, but have different functional groups and various side chains.

The third problem involves the retention indices of alcohols, ethers and esters. A total of 16 compounds are in the problem, called

the 16 x 10 problem. The essential features of the problem are: the solutes have varying numbers of carbon atoms (from 4 to 10), and have three different functional groups (but no side chains).

The clustering of the 10 stationary phases in each problem above is also studied. In next two chapters, we will discuss the clustering results for these problems.

Table 28. Summary of the retention-index clustering problems.

Problem	Essential features	Solutes*
15 x 10	One functional group (ROH) Same number of carbon atoms (6) Different number of side chains (0-2)	1-15
20 x 10	Same number of carbon atoms (6) Two functional groups (ROH, ROR') Different number of side chains (0-2)	1-15, 20, 23-26
16 x 10	Three functional groups (ROH, ROR', RCOOR') Different number of carbon atoms (4-10) No side chains	1, 16-22, 27-34

*) For solute designations, see Table 27.

Chapter 15

Clustering of Retention Indices of Alcohols

In this chapter, we apply the clustering methods to the retention indices of the 15 alcohols on the 10 stationary phases, called the 15 x 10 problem. The essential features of the problem were given in Table 28. The properties of the 15 alcohols are very similar, each having the same functional group and the same number of carbon atoms. Differences between the alcohols result from the three possible positions of the hydroxyl group and the number of side chains.

Clusterings of the 15 alcohols are investigated by various methods. The results are detailed in this chapter. First, the hierarchical clustering method is applied to the data matrix. The calculation is carried out on the SAS program, employing the average-linkage method. The distances between compounds in a 10-dimensional space and the normalized root-mean-square (RMS) distances are calculated and the members in each cluster are indicated. A summary of the SAS printout is given in Table 29. Referring to the table, the two alcohols having the shortest distance, AS2 and AS3, are merged to form the first cluster, called cluster 14. Similarly, alcohols AT1 and AT3, and AP2 and AP4 are merged as clusters 13 and 12, respectively. Then cluster 12 merges with an alcohol AP5, forming a new clust-

Table 29. Selected output from average-linkage hierarchical cluster analysis from the SAS program for the 15 x 10 problem.

Cluster	Clusters joined *	No. alcohols in cluster	Cubic clustering criterion	Normalized RMS distance
14	AS2 AS3	2	.	0.020
13	AT1 AT3	2	.	0.050
12	AP2 AP4	2	.	0.076
11	CL12 AP5	3	.	0.081
10	AS1 CL14	3	.	0.137
9	CL11 AP3	4	.	0.166
8	AS4 AS5	2	.	0.167
7	CL13 AS6	3	.	0.186
6	CL10 AP6	4	.	0.239
5	CL7 AT2	4	.	0.249
4	AP1 CL9	5	.	0.388
3	CL5 CL8	6	0.50	0.423
2	CL6 CL3	10	-0.39	0.750
1	CL4 CL2	15	0.00	1.333

*) For designations, see Table 27. CL - cluster formed in earlier step.

ter 11. The merging is repeated until only global cluster 1 is formed. The cubic clustering criterion is calculated only for the last three clusters for this data set, because the cubic clustering criterion is not calculated when the number of clusters is greater than one-fifth the number of compounds [14].

To visually represent the clusters for the 15 alcohols, a dendrogram from the SAS program based on the RMS distances is shown in Figure 3. An easier-to-follow dendrogram is given in Figure 4. The members in each cluster are:

Cluster	Members
1	AS1, AS2, AS3, AP6
2	AP1, AP2, AP3, AP4, AP5
3	AS4, AS5, AS6, AT1, AT2, AT3

Cluster 1 contains mostly secondary alcohols, cluster 2 contains only primary alcohols, while cluster 3 contains equal numbers of secondary and tertiary alcohols. The hierarchical clustering method gives three clusters based on a combination of the position of the OH group and the number of side chains.

Next, factor analysis is applied to the 15 x 10 problem before applying the max-min clustering method. In the reproduction step of factor analysis, the eigenvalue, the standard error (SE) (also called the root-mean-square error RMS) and the indicator function for each

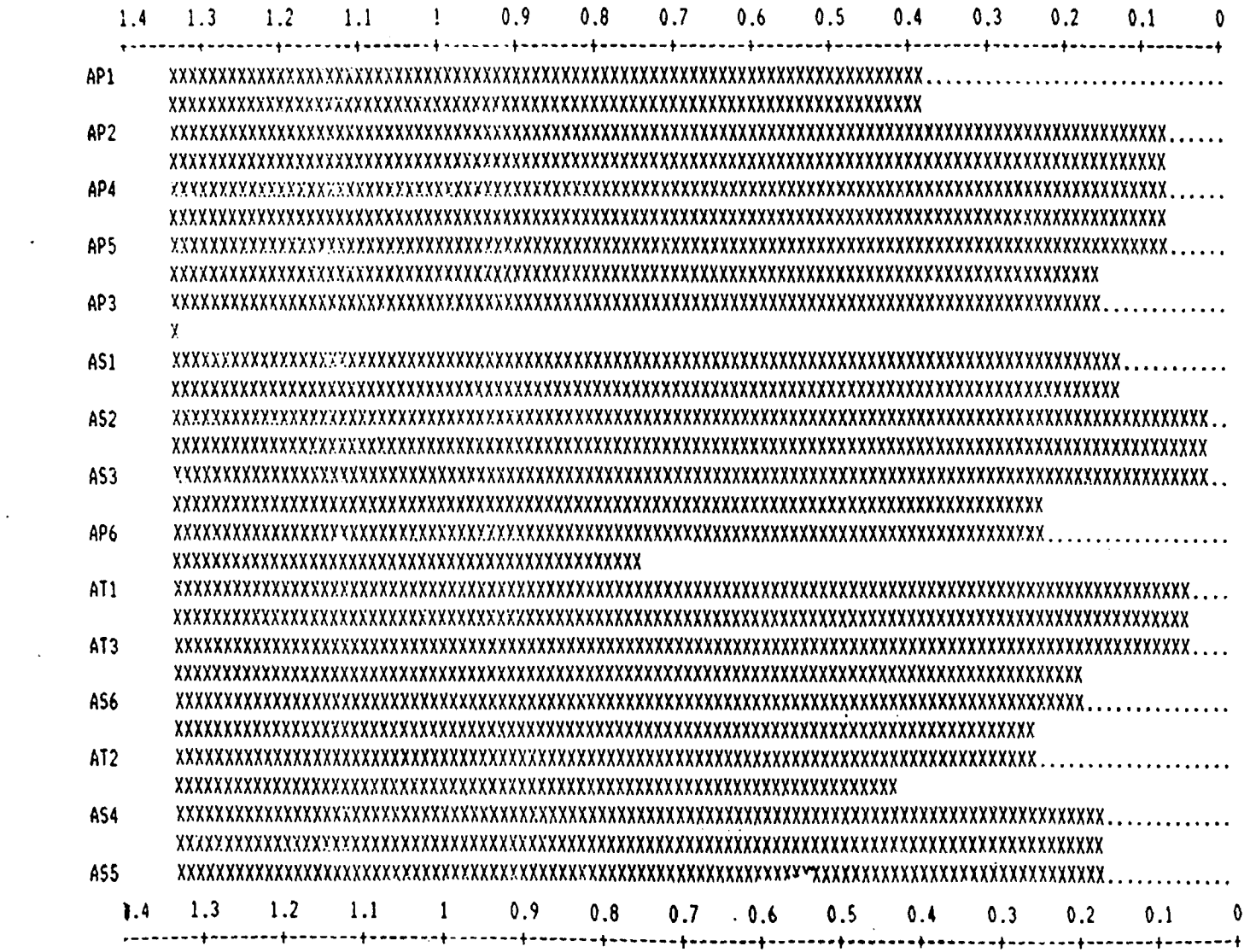


Figure 3. SAS dendrogram for the 15 x 10 problem.

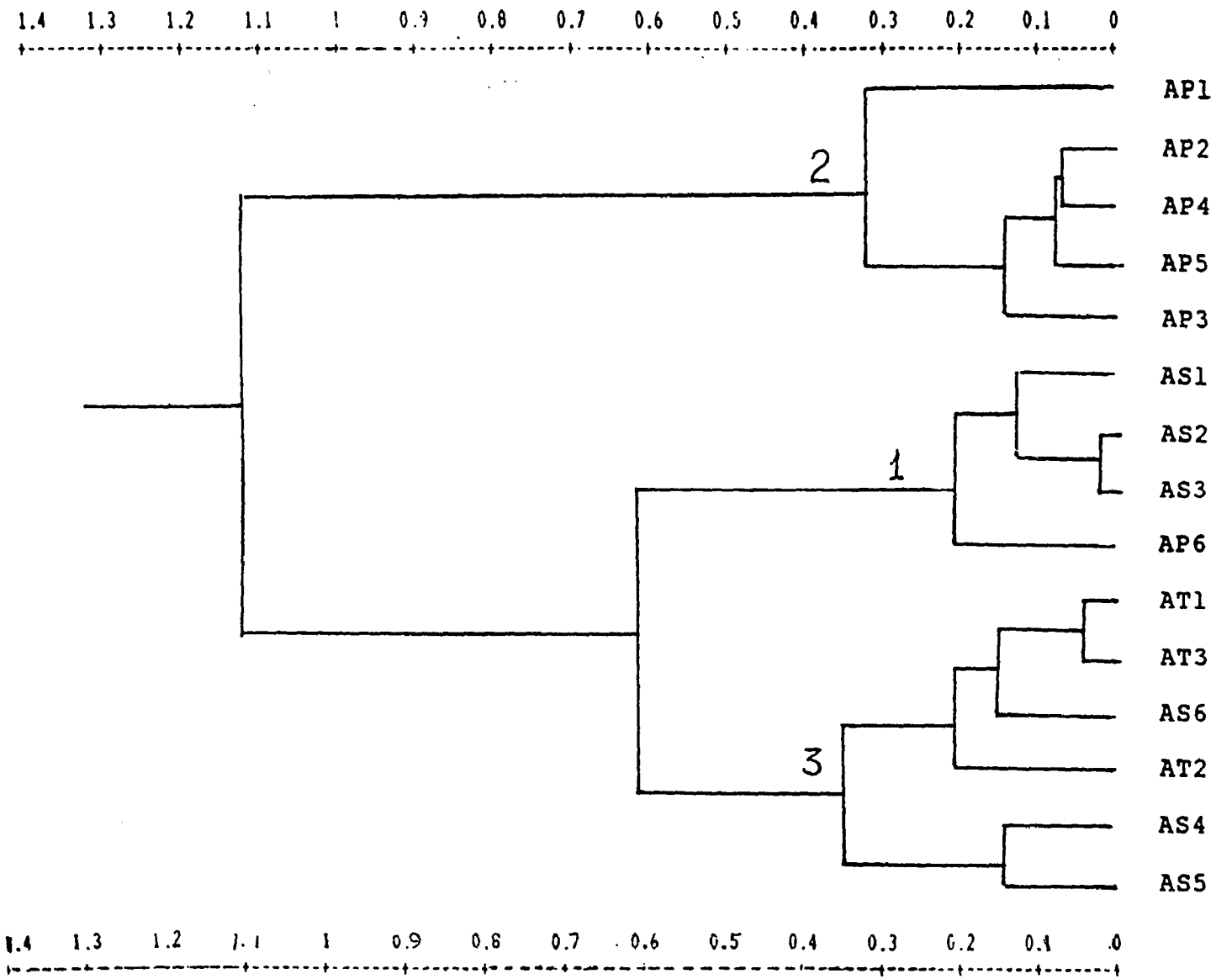


Figure 4. Redrawn dendrogram for the 15 x 10 problem with three clusters indicated.

factor are calculated. The purpose of the reproduction step in factor analysis is to determine the correct number of factors. Results for the reproduction of the 15 x 10 problem are shown in Table 30. The SE is a measure of the difference between the factor analysis-reproduced data and the experimental data. The value of the indicator function often reaches a minimum when the correct number of factors are employed [6]. Referring to Table 30, the value of the indicator function reaches a minimum value of 0.0338 at the third factor, and SE for the third factor is 1.1 which is smaller than the experimental error (about $\pm$ 5 RI units). These two values indicate that three factors are sufficient for the 15 x 10 retention-index problem.

The calculation of principal factors and rotated factors are carried out on the SAS program. The three principal factor patterns from the principal factor analysis are given in Table 31. The variances explained by the three individual factors are 99.93, 0.06 and 0.01%, respectively. The three rotated factor patterns after varimax rotation are shown in Table 31 also. The variances of the rotated factors explained by the three factors are 51.10, 48.89 and 0.01%, respectively.

The rotated factor patterns furnish classification information. A plot of the rotated factor patterns using the first two factors, shown in Figure 5, gives information in the clusters. The points A, B and C

Table 30. Reproduction table from factor analysis for the 15 x 10 problem.

Factor	Eigenvalue	SE	Indicator function
1	1.31×10^{8}	9.99	0.1400
2	1.68×10^{4}	1.86	0.0368
3	3.79×10^{2}	1.14	0.0338
4	1.33×10^{2}	0.75	0.0364
5	5.20×10^{1}	0.59	0.0468
6	3.88×10^{1}	0.42	0.0644

Table 31. Principal and varimax-rotated factor patterns for the 15 x 10 problem.

	Principal factor pattern		
Compound *	Factor 1	Factor 2	Factor 3
AP1	0.9997	-0.0199	-0.0067
AS1	0.9999	0.0081	0.0010
AS2	0.9998	0.0145	0.0008
AP2	0.9997	-0.0191	-0.0049
AP3	0.9998	-0.0153	-0.0054
AP4	0.9997	-0.0189	-0.0104
AT1	0.9987	0.0486	0.0001
AS3	0.9997	0.0193	-0.0033
AS4	0.9999	0.0016	0.0046
AS5	0.9996	0.0246	0.0029
AT2	0.9985	0.0524	-0.0072
AP5	0.9996	-0.0251	0.0043
AP6	0.9994	-0.0299	0.0156
AT3	0.9989	0.0456	0.0064
AS6	0.9999	0.0015	0.0116
% Variance	99.93	0.06	0.01

(continued)

Table 31. (continued)

	Varimax rotated factor pattern		
Compound *	Factor 1	Factor 2	Factor 3
AP1	0.7286	0.6849	-0.0072
AS1	0.7092	0.7050	0.0010
AS2	0.7046	0.7095	0.0008
AP2	0.7281	0.6855	-0.0054
AP3	0.7255	0.6882	-0.0058
AP4	0.7278	0.6856	-0.0109
AT1	0.6801	0.7331	0.0006
AS3	0.7012	0.7129	-0.0032
AS4	0.7138	0.7003	0.0044
AS5	0.6975	0.7165	0.0031
AT2	0.6772	0.7357	-0.0066
AP5	0.7322	0.6810	0.0037
AP6	0.7355	0.6773	0.0149
AT3	0.6823	0.7310	0.0069
AS6	0.7139	0.7001	0.0114
% Variance	51.10	48.89	0.01

*) For designations, see Table 27.

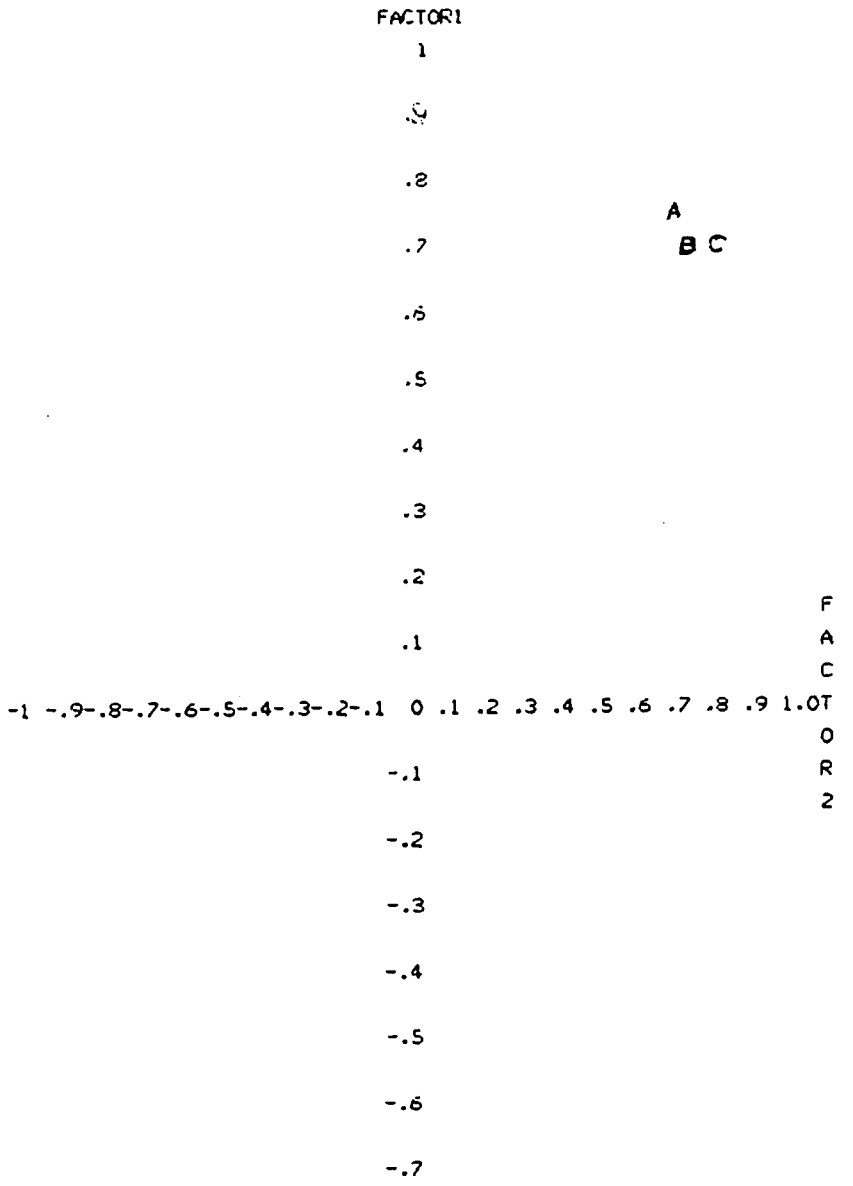


Figure 5. SAS plot of varimax factor patterns for the 15 x 10 problem.

in the factor plot represent the following compounds:

Point	Compounds
A	AP1, AP2, AP3, AP4, AP5, AP6
B	AS2, AS2, AT1, AS3, AS4, AS5, AT3, AS6
C	AT2

Point A represents mostly primary alcohols; point B represents mostly a mixture of secondary and tertiary alcohols; and point C represents a single tertiary alcohol.

Finally, the max-min distance method is applied to the rotated factor patterns above to obtain additional clusterings. After rotation, the original 15 x 10 matrix is simplified to 15 x 3 pattern-vector matrix. The max-min method then is easier to apply to the smaller matrix. Procedures for the max-min method were discussed in Chapter 12. First three cluster centers, AP1, AT2 and AS2, are determined. The distances from each sample to the three cluster centers are listed in Table 32. To assign a sample to the domain of these three cluster centers, we simply compare the distances from the sample to the three cluster centers. A sample belongs to the cluster with the shortest distance. Thus the three clusters are formed as follows:

Table 32. Distances to the first three cluster centers using the max-min distance method for the 15 x 10 problem.

Compound [a]	D to C1 [b] (AP1)	D to C2 (AT2)	D to C3 (AS2)	Belongs to cluster
AP1	0.0	0.072	0.035	1
AS1	0.029	0.045	0.006	3
AS2	0.035	0.039	0.0	3
AP2	0.002	0.072	0.034	1
AP3	0.005	0.068	0.031	1
AP4	0.004	0.071	0.035	1
AT1	0.069	0.008	0.034	2
AS3	0.039	0.033	0.003	3
AS4	0.024	0.052	0.013	3
AS5	0.046	0.030	0.010	3
AT2	0.072	0.0	0.039	2
AP5	0.012	0.078	0.040	1
AP6	0.024	0.085	0.047	1
AT3	0.067	0.015	0.032	2
AS6	0.028	0.054	0.017	3

a) For designations, see Table 27.

b) D - distance from the compound to the cluster center C.

Cluster	Members
1	AP1,AP2,AP3,AP4,AP5,AP6
2	AT1,AT2,AT3
3	AS1,AS2,AS3,AS4,AS5,AS6

In the max-min method, the 15 alcohols are clustered based solely on the position of the hydroxyl group. The members of the first cluster are all primary alcohols, the second cluster contains only tertiary alcohols, and all the secondary alcohols belong to the third cluster.

By contrast, in the hierarchical clustering method, clustering was based on a combination of the positions of the OH group and the number of side chains. From a chemical viewpoint, the clusters from the max-min method are more satisfactory.

To check the clusterings, target tests for the clusters from both methods are employed. The uniqueness test, a kind of target test, is employed for our purpose. In an uniqueness-test vector, a value of "1" is assigned to all components in a cluster, whereas a value of "0" is assigned to all other components. In other words, the target-test vector represents a cluster in the uniqueness test. If the predicted vector obtained from target testing is close to the input target-test vector, the target vector is an acceptable vector. The corresponding cluster then is considered to be a valid cluster. If the predicted vector is very different from the target-test vector in either pattern or

magnitude, the target-test vector is a non-unique vector. The corresponding cluster is an invalid cluster.

Details of clustering target tests for the 15 x 10 problem are shown in Table 33. Target vector M1 corresponds to cluster 1 from the max-min method, where all primary alcohols have values of one, while all other alcohols have values of zero. Similarly, H1 is designed to test the first cluster from the hierarchical method, and so forth. Another kind of target vector, called a combination vector, combines weighted values from more than one cluster vector. For example, a combination vector CM is designed for all three clusters from the max-min method, having values of 1, 0.5 and 0.25 based on the weighted values from the individual predicted vectors, for the first, third and second max-min clusters, respectively. The least-squares predicted vectors for each target test are given in Table 33.

Evaluation of agreement between the target and predicted vectors includes qualitative evaluation and three statistical parameters: the root-mean-square error (RMS) [6], the intraclass correlation coefficient (ICC) [55] and the spoil function [6].

Qualitative evaluation of the target test is based on a comparison of the pattern and magnitude similarity between the target vector and the predicted vector. The letter "g" means that there is good agreement between the target vector and the predicted vector;

Table 33. Details of five target tests for the 15 x 10 problem.

Vector [a]	M1		M2		CM		H1		H2	
	Target	Predicted	Target	Predicted	Target	Predicted	Target	Predicted	Target	Predicted
ROH[b]										
AP1	1.0	0.87	0.0	0.20	1.00	0.97	0.0	0.21	0.0	0.06
AS1	0.0	0.37	1.0	0.36	0.50	0.62	1.0	0.22	0.0	0.42
AS2	0.0	0.14	1.0	0.59	0.50	0.51	1.0	0.36	0.0	0.62
AP2	1.0	0.89	0.0	0.13	1.00	0.95	0.0	0.16	0.0	0.01
AP3	1.0	1.01	0.0	0.05	1.00	1.00	0.0	0.02	0.0	0.08
AP4	1.0	0.99	0.0	0.04	1.00	0.98	0.0	0.03	0.0	0.07
AT1	0.0	-0.07	0.0	0.35	0.25	0.27	0.0	0.07	1.0	0.75
AS3	0.0	0.08	1.0	0.56	0.50	0.46	1.0	0.31	0.0	0.67
AS4	0.0	0.40	1.0	0.40	0.50	0.63	0.0	0.28	1.0	0.37
AS5	0.0	-0.02	1.0	0.67	0.50	0.41	0.0	0.39	1.0	0.74
AT2	0.0	-0.30	0.0	0.62	0.25	0.18	0.0	0.24	1.0	0.97
AP5	1.0	0.80	0.0	0.44	1.00	0.97	0.0	0.44	0.0	0.09
AP6	1.0	0.82	0.0	0.52	1.00	0.98	1.0	0.54	0.0	0.03
AT3	0.0	-0.19	0.0	0.62	0.25	0.26	0.0	0.29	1.0	0.84
AS6	0.0	0.23	1.0	0.64	0.50	0.57	0.0	0.47	1.0	0.48

a) M - max-min cluster, H - hierarchical cluster, C- combination cluster.
b) For designations, see Table 27.

"f" indicates that several points are poorly predicted or that only the pattern is predicted; "p" means that there is poor agreement both in the pattern and in the magnitude between the target vector and predicted vector.

The root-mean-square error represents the difference between the individual components of the target and the predicted vectors. The smaller the RMS error, the better the prediction. The intraclass correlation coefficient compares the variance between the target vectors and the predicted vectors and the variance within the vectors. The value of the intraclass correlation coefficient varies from -1 to +1. The larger the absolute value of the intraclass correlation coefficient, the more similar the target and predicted vectors. The spoil function is based on the error theory of Malinowski [56]. The larger the value of the spoil function, the worse the prediction.

Evaluations of selected target vectors for the 15 x 10 problem are given in Table 34. Target vectors M1, M2 and CM based on the clusters obtained from the max-min method have relatively high ICC values and small RMS values and spoil values. Qualitative evaluations are in the good and fair range. The ICC value of vector CM is quite high (close to unity) and the values of RMS and spoil are the smallest among the five target vectors, showing the high validity of the combination cluster from the max-min method. The target vectors H1

Table 34. Evaluation of target tests for the 15 x 10 problem.

Target [a]	Qualitative [b] evaluation	RMS [c]	ICC [d]	Spoil [e]
M1	g	0.20	0.834	10.27
M2	f	0.26	0.538	5.24
CM	g	0.06	0.960	3.10
H1	p	0.41	-0.084	8.83
H2	p	0.35	0.448	19.08

a) For designations, see Table 33.

b) g - good agreement between target and predicted vectors,
f - several points poorly predicted or pattern only predicted,
p - agreement in neither pattern nor magnitude.

c) RMS - root-mean-square error.

d) ICC - intraclass correlation coefficient.

e) Spoil - spoil function.

and H2 for clusters obtained from the hierarchical method have smaller ICC values and larger RMS errors and spoil values. Qualitative evaluation is in the poor range.

From the statistical analysis of the clusters obtained from the various methods, clusters from the max-min method are the more reasonable clusters. The max-min clusters are in good agreement with chemical facts, since the position of the OH group is expected to be more important than the number of side chains in determing clusters.

Chapter 16

Summary of Clustering Results for Retention Indices of Solutes and Stationary Phases

In chapter 15, we detailed an application of the clustering methods to the retention-index problem for alcohols (the 15 x 10 problem). In this chapter, we summarize the clustering results for the rest of the retention-index problems described in Chapter 14, involving 34 solutes and 10 stationary phases.

In the second problem (see Table 28 in Chapter 14), 15 alcohols and five ethers are selected, called the 20 x 10 problem. The 20 isomers have the same empirical formula, but very different physical and chemical properties, due to the different functional groups. We expect to cluster the 20 isomers according to their functional groups. After applying both clustering methods on the 20 x 10 problem, the various clusters for the 20 x 10 problem are given in Table 35.

Principal factor analysis indicates that three factors are the correct number of factors for the 20 x 10 matrix. In hierarchical cluster analysis, the cubic clustering criterion indicates that four clusters are sufficient to cluster the 20 compounds. We analyze the results for two and three clusters for the 20 x 10 problem (see Table 35).

Table 35. Clusters from the max-min and hierarchical methods for the 20 x 10 problem.

Cluster	Max-min method*	Hierarchical method *
Two clusters		
1	AP1,AP2,AP3,AP4,AP5, AP6,AS1,AS2,AS3,AS4, AS5,AS6,AT1,AT2,AT3	AP1,AP2,AP3,AP4,AP5, AP6,AS1,AS2,AS3,AS4, AS5,AS6,AT1,AT2,AT3
2	TIBE,TP2,TIPP,TBEt,TIP2	TIBE,TP2,TIPP,TBEt,TIP2
Three clusters		
1	AP1,AP2,AP3,AP4,AP5, AP6,AS1,AS2,AS3,AS4 AS5,AS6,AT1,AT2,AT3	AP1,AP2,AP3,AP4,AP5
2	TIBE,TP2,TIPP,TBEt	AS1,AS2,AT1,AS3,AS4, AS5,AT2,AP6,AT3,AS6
3	TIP2	TBEt,TP2,TIBE,TIPP, TIP2

*) For designations, see Table 27.

If only two clusters are chosen, the two methods give the same classification. One cluster contains the 15 alcohols, and the other contains the five ethers. Both methods cluster the 20 isomers via the functional groups. However, if three clusters are chosen, then the hierarchical method clusters the alcohols into two clusters (see Table 35). The max-min method separates isopropyl ether with two carbon side chains as a single cluster.

Selected target tests are performed for the two-cluster classifications. Evaluation of the target tests are shown in Table 36. Qualitative evaluation is in the good range for the two clusters. The intraclass correlation coefficients are extremely high, and the root-mean-square error and the spoil function are very low. The three statistical parameters indicate that the two clusters give a correct classification for the 20 x 10 problem. Both the hierarchical and max-min methods give good (and equivalent) results for a two-cluster classification. However, if three clusters are chosen (results not shown in the table), the predicted vectors for the max-min clusters are in the good and fair range, whereas the predicted vectors for the hierarchical clusters are in the poor and fair range.

In the third problem, 16 compounds with three different functional groups are selected, called the 16 x 10 problem. The main features of the compounds are listed in Table 28. Three clusters are suf-

Table 36. Evaluation of target tests for the 20 x 10 and the 16 x 10 problems.

Target *	Qualitative evaluation *	RMS *	ICC *	Spoil *
20 x 10 problem (2 clusters and 3 factors)				
M1 (= H1)	g	0.08	0.97	2.5
M2 (= H2)	g	0.06	0.98	1.6
16 x 10 problem (3 clusters and 3 factors)				
M1	g	0.02	1.00	1.3
M2	f	0.23	0.66	21.5
M3	g	0.11	0.95	4.6
H1	p	0.39	-0.26	95.5
H2	p	0.50	-0.27	84.1
H3	p	0.40	0.13	55.9

*) For designations, see Table 34.

ficient for the 16 compounds using the hierarchical method, according to the cubic clustering criterion. Three factors are the correct number of factors for the 16 x 10 matrix from factor analysis. The three clusters from the max-min and hierarchical methods are shown in Table 37. Clusters from the max-min method are separated according to the functional groups. the first cluster contains all the alcohols, the second cluster contains all the ethers and the third cluster contains all the esters. By contrast, each cluster from the hierarchical method contains at least one alcohol, one ether and one ester. The three hierarchical clusters are mixed unsatisfyingly.

Target testing is applied to the clusters in the 16 x 10 problem. The evaluations of the target tests using the qualitative evaluation and the three statistical parameters are shown in Table 36. The designations of the target test vectors are the same as described in Table 33 of Chapter 15. Qualitative evaluations for the clusters from the max-min method are in the good range. The three statistical criteria show that the four target vectors for the clusters from the max-min method have high values of the intraclass correlation coefficients (quite close to unity), and have relatively small values for the root-mean-square error and the spoil function. By contrast, the intraclass correlation coefficients for the target vectors for the clusters from the hierarchical method are very small or even negative, the spoil-func-

Table 37. Clusters from the max-min and hierarchical methods for the 16 x 10 retention-index problem.

Cluster	Max-min method *	Hierarchical method *
1	ABU,APE,AP1,AHP,AOC	ABU,TP2,SEtP
2	TP2,TB2,TPE2	APE,AP1,TB2,SPP, SBP,SEtB,SPB
3	SEtP,SPP,SBP,SPEP,SEtB, SPB,SBB,SPEB	AHP,AOC,TPE2,SPEP, SBB,SPEB

*) For designations, see Table 27.

tion values are greater than 50, and the root-mean-square errors are relatively higher than those for the max-min vectors. Qualitative evaluations of the target-testing vectors for the hierarchical clusters are in the poor range.

The qualitative evaluation and the three statistical criteria for the target tests indicate that the clusters from the max-min method are valid clusters in the 16 x 10 problem, while the clusters from the hierarchical method are invalid. Furthermore, the clusters from the max-min method are in good agreement with the chemical facts.

The 10 stationary phases are classified into clusters for each of the three retention-index problems. The same clustering results are obtained from both the hierarchical and the max-min methods. The three clusters from both methods for the 10 stationary phases are given in Table 38. The two high-polarity phases, CW6K and CW20, are in one cluster. The second cluster contains the three intermediate polarity phases: PP5, PP6, PH1, and PH2. The third cluster consists of all the non-polar phases. The clusters are separated entirely by their polarities. In the clustering of the 10 stationary phases, the three clusters from both methods are in good agreement with chemical facts.

According to our studies on retention-index data, the clustering results of the four problems show that the max-min cluster and

Table 38. Clusters from the max-min and hierarchical methods for the 10 stationary phases.

Cluster	Max-min method *	Hierarchical method *
1	APJ, APM, SE30, SE31	APJ, APM, SE30, SE31
2	PP5, PP6, PH1, PH2	PP5, PP6, PH1, PH2
3	CW6K, CW20	CW6K, CW20

*) For designations, see p. 95.

the hierarchical cluster methods both give satisfactory clusters in general. However, clusters obtained from the max-min clustering method sometimes are in better agreement with the chemical facts.

Chapter 17
Introduction to Thin Layer Chromatographic Problems

Thin layer chromatography (TLC) is a separation technique which is widely applied in chemistry, biochemistry and medical research [54,57]. The separations may be based upon adsorption, partition or a combination of both effects. The R_f value, the widely used TLC retardation factor, is equal to the ratio of the distance from the origin to the center of a zone divided by the distance from the origin to the solvent front. The R_f value varies between 0 and 1. For convenience, the value of 100 x R_f is usually used. Identification can be effected by observation of spots obtained with an unknown and a reference sample chromatographed on the same plate. A substance has different R_f values in different developing solvents and on different plates, due to the different interactions between solutes, solvents and plates. Unfortunately, a satisfactory explanation of the very complicated interactions in TLC is lacking.

Clustering the solutes based on their R_f values obtained from different developing solvents and on different impregnated plates is selected as a problem in this study. The R_f values of 20 sugars and

13 mixed solvents developed on two silica gel plates (impregnated with sodium acetate and sodium dihydrogen phosphate) [58] are selected. The matrix of R_f (x 100) values is given in Appendix E.

Characteristics of Solutes, Solvents and Plates

The sugars selected include monosaccharides, oligosaccharides and methyl derivatives. A summary of the characteristics and designations of the 20 sugar solutes are given in Table 39. The compounds numbered from 1 to 11 are monosaccharides, designated as M1 to M11. The compounds numbered from 12 to 18 are oligosaccharides, designated as O1 to O7. The compounds numbered 19 and 20 are the derivatives of two monomers, designated as D1 and D2. The number of carbon atoms of the 20 sugars varies from five to 18. The sugars with five, six and seven carbon atoms have the same empirical formula, respectively, but different structures [42]

The developing solvent plays a very important role in TLC. The strength of the solvent strongly influences the migration of the solutes on the silica gel plate, resulting in different R_f values obtained under different developing solvents. In order to obtain a R_f value in a satisfactory range (about 0.2 to 0.7), mixed solvents are often used to adjust the polarity of the developing solvent. Thirteen mixed solvents are selected for our studies. The characteristics of the mixed

Table 39. Characteristics of the sugars in the TLC problems.

Sugar	Designation	Characteristics	
		No. carbon atoms	Type of saccharide
Ribose	M1	5	mono
Arabinose	M2	5	mono
Xylose	M3	5	mono
Lyxose	M4	5	mono
Fucose	M5	6	mono
Glucose	M6	6	mono
Galactose	M7	6	mono
Levulose	M8	6	mono
Sorbose	M9	6	mono
Mannoheptulose	M10	7	mono
Sediheptulose	M11	7	mono
Sucrose	O1	12	oligo
Turanose	O2	12	oligo
Maltose	O3	12	oligo
Lactose	O4	12	oligo
Trehalose	O5	12	oligo
Melibiose	O6	12	oligo
Raffinose	O7	18	oligo
1-methyl glucoside	D1	7	derivative
1-methyl mannoside	D2	7	derivative

solvents and their individual components are given in Table 40. The designations of the 13 solvents in the original work and in our work are given in the table. The average electric-dipole polarizabilities [42] of a mixed solvent is a measure of the polarity of the solvent. The polarizabilities of the mixed solvents are calculated from the polarizabilities of the individual components. From the values of the average polarizabilities for the mixed solvents in the table, we can see that most of solvents, except solvent B, have similar polarizabilities (about 6 to 7), indicating similar polarities. Mixed solvents B, E, H, J and K are acidic; solvent G is a basic solvent.

Two plates are used in the study: silica gel plate impregnated with sodium acetate, labeled as plate 1, and silica gel plate impregnated with sodium dihydrogen phosphate, labeled as plate 2. Sodium acetate is a salt of the weak acid, giving a basic aqueous solution. Aqueous solutions of sodium dihydrogen phosphate show weak acidity.

Summary of TLC Problems

Three problems are designed using the R_f values of the 20 sugars developed in 13 mixed solvents on plates 1 and 2. The classification of the 13 solvents on plate 1 is also studied. The features of the four problems are summarized in Table 41.

Table 40. Characteristics of the mixed solvents in the TLC problems.

Designation					
This work	Original work	Components	Volume Ratio	Mole Fraction	Polarizability* (v/v)
A	S2	n-butanol	4	0.261	6.90
		water	1	0.332	
		acetone	5	0.407	
B	S3	n-butanol	4	0.129	4.79
		acetic acid	1	0.052	
		water	5	0.819	
C	S6	isopropanol	5	0.312	6.76
		n-butanol	3	0.157	
		water	2	0.531	
D	S10	methanol	3	0.402	6.19
		n-butanol	5	0.297	
		water	1	0.302	
E	S14	isopropanol	60	0.162	7.29
		n-butanol	30	0.068	
		acetic acid	60	0.216	
		water	30	0.344	
		ethyl acetate	100	0.211	
F	S15	n-propanol	85	0.577	5.95
		water	15	0.423	
G	S18	n-propanol	5	0.311	6.51
		water	2	0.516	
		pyridine	3	0.173	

(continued)

Table 40. (continued)

Designation					
This work	Original work	Components	Volume Ratio	Mole Fraction	Polarizability* (v/v)
H	S28	isopropanol	60	0.192	7.33
		acetic acid	35	0.150	
		water	30	0.408	
		ethyl acetate	100	0.250	
I	S29	isopropanol	60	0.226	7.74
		water	30	0.480	
		ethyl acetate	100	0.294	
J	S30	methanol	3	0.403	6.19
		n-butanol	5	0.296	
		boric acid (.03M)	1	0.0001	
		water		0.301	
K	S32	methanol	15	0.206	7.22
		acetic acid	15	0.146	
		water	10	0.308	
		ethyl acetate	60	0.340	
L	S33	isopropanol	7	0.430	6.59
		water	2	0.522	
		ethyl acetate	1	0.048	
M	S35	isopropanol	4	0.485	6.38
		water	1	0.515	

*) Units of 10^{-24} cm^3

Table 41. Summary of TLC problems.

Problem	Essential features
20 x 13 - P1	20 sugars [a] 13 mixed solvents [b] Plate 1
20 x 13 - P2	20 sugars 13 mixed solvents Plate 2
20 x 26 - P1&2	20 sugars 13 mixed solvents Plates 1 & 2
13 x 20 - P1	13 mixed solvents 20 sugars Plate 1

a) For designations, see Table 39.

b) For designations, see Table 40.

First, the R_f values of the 20 sugars on the two plates are studied individually, called the 20 x 13 - P1 and the 20 x 13 - P2 problems, respectively. In the two problems, the sugars and the developing solvents are the same but the plates are different. The third problem, involving the R_f values of the 20 sugars on both plates 1 and 2 developed on the 13 mixed solvents, is called the 20 x 26 - P1&2 problem. Clustering of the 13 mixed solvents based on the R_f values on plate 1 is designated the 13 x 20 - P1 problem.

Chapter 18

Summary of Clusterings of R_f Values for Sugars and Mixed Solvents

For the TLC problems in Table 41 in Chapter 17, we give a summary of a few of the results only.

The clusters obtained from the max-min and hierarchical methods for the four TLC problems are given in Table 42. In the 20 x 13 - P1, 20 x 13 - P2 and 20 x 26 - P1&2 problems (see Table 41), the 20 sugars form four clusters in the hierarchical method according to the cubic clustering criterion. Principal-component factor analysis also indicates that four factors are the correct number for the three TLC problems. From the results in Table 42, very confusing classifications are observed. Each cluster contains mono- and oligo- saccharides. Clustering depends neither on structure (functional group position on the molecule) nor on the number of carbon atoms (mono or oligo). Comparing the clusters from the two methods, different clusters are obtained in the 20 x 13 - P1 problem. However, in the 20 x 13 - P2 and 20 x 26 - P1&2 problems, identical clusters are obtained from the hierarchical cluster method.

Evaluation of the target tests for the clusters in the 20 x 13 - P1 problem is given in Table 43. Qualitative evaluations for the problem

Table 42. Clusters from the max-min and hierarchical methods for the TLC problems.

Problem		Max-min method *	Hierarchical method *
20 x 13 - P1	1	M1,M3,M5,M8,O1,D1	O4,O6,O7
	2	O5, O6	M7,M11, O1,O2,O3,O5
	3	O3,O4,O7	M3,M4,M9
	4	M2,M4,M6,M7,M9,M10, M11, O2, D2	M1,M2,M5,M6,M8,M9, M10, D2
20 x 13 - P2	1	M1,M3,M4,M5,D1	M1,M3,M4,M5,D1
	2	O4,O6,O7	O4,O6,O7
	3	O3,O5	M2,M6,M8,M9,M10,M11,D2
	4	M2,M6,M7,M8,M9,M10, M11,O1,O2,D2	M7,O1,O2,O3,O5
20 x 26 - P1&2	1	M1,M3,M5,D1	M1,M3,M4,M5,D1
	2	O7	O4,O6,O7
	3	O4,O5,O6	M2,M6,M8,M9,M10,M11,D2
	4	M2,M4,M6,M7,M8,M9, M10,M11,O1,O2,O3,D2	M7,O1,O2,O3,O5
13 x 20 - P1	1	A,C,D,E,F,H,I,K,M	A,C,D,E,F,H,J,K,L,M
	2	B	I
	3	G,J,L	B,G

*) For designations, see Table 39 and 40.

Table 43. Evaluation of targer tests for the 20 x 13 - P1 problem.

Target *	Qualitative evaluation *	RMS *	ICC *	Spoil *
M1	f	0.29	0.583	3.16
M2	f	0.20	0.502	3.20
M3	f	0.19	0.697	2.33
M4	f	0.30	0.607	1.96
H1	f	0.24	0.511	5.07
H2	p	0.41	0.012	9.01
H3	p	2.75	0.327	6.81
H4	p	0.45	0.034	24.5

*) For designations, see Table 34.

show that the fits are in the fair to poor range, especially for the hierarchical clusters. The intraclass correlation coefficients for hierarchical clusters have very low values (0 - 0.5), while for the max-min clusters the values are also statistically low (0.5 - 0.7). The results show that the clusters from both methods are unsatisfactory.

For the 13 x 20 - P1 problem, the cubic clustering criterion for the hierarchical clustering method indicates that three clusters are the correct number for clustering 13 solvents. Principal-component factor analysis indicates also that three factors are the correct number of factors. From the results in Table 42, members in the clusters are different for both methods. Most of solvents are in the first cluster. Solvents B and G are separated from the other solvents in both the hierarchical and max-min methods. From the properties of solvents B and G in Table 40, solvent B has the smallest value of polarizability (4.7), compared to other solvents. Solvent G is the only basic solvent (pyridine).

We conclude that clustering of TLC data is quite complicated, especially when mixed solvents are utilized.

Appendix A - Documentation, List of Variables and Listing for CRA Program

Original version of program in BASIC by H. Yamazaki (Printing Bureau, Ministry of Finance, Tokyo, Japan). Modified by J. Wu and D. G. Howery (Chemistry Dept., Brooklyn College of C.U.N.Y, Brooklyn, New York).

Files in the Categorical Regression Analysis (CRA) package:

"CRA-DOCU" - documents three programs,

"CRA-DATA" - a program to write a data set onto a disk,

"CRA-MAIN" - a program to process all computations, and

"CRA-PRED" - a program to predict new data from the CRA model.

To write a data set onto disk using "CRA-DATA":

Step 1 Load "CRA-DATA"

Step 2 Input data from line 2000 as follows:

2000 DATA Sample number, Number of properties, "Name of problem"

2010 DATA "Name of 1st property", "Name of 2nd property", . . .

2020 DATA Number of categories in 1st property, Number of categories in 2nd property, . . .

2030 DATA Index number of 1st sample (arbitrary), "Name of sample", Value of independent variable, Categorized properties for 1st sample

2040 DATA Repeat 2030 for each sample, . . .

Example: Model for Boiling points of alcohols (8 x 6 problem)

2000 DATA 8,2, "Boiling points of alcohols"

2010 DATA "Carbon number (5,6,7)","Type of alcohol (p,s,t)"

2020 DATA 3,3

2030 DATA 75,"1-pentanol",137.8,1,0,0,1,0,0

2040 DATA 72,"2-pentanol",119.0,1,0,0,0,1,0

2050 DATA Repeat 2030 for each remaining sample, ...

Step 3 Save and run "CRA-DATA"

Step 4 Upon query: "Select a new filename for data set", input "filename" of choice.

To apply CRA to a data set using "CRA-MAIN":

Step 1 Load "CRA-MAIN"

Step 2 Run "CRA-MAIN". Upon query: "Give filename of desired data set", input filename of data file. Upon query: "If you desire N-1 (leave-one-out) calculations, select 1, otherwise 0", input either 1 or 0.

a) If "0" selected, model for entire data set only is calculated

b) If "1" selected, series of N-1 models are calculated.
Upon query "If you want to exclude some samples from calculation, select 1, otherwise 0", input either 1 or 0.

i) If "0" selected, N-1 model for each datum is calculated

ii) If "1" selected, some N-1 models are not calculated, Upon query "Give sample number from which you want to start the calculation", input desired sample number.

Step 4 Results are printed out and normalized category coefficients are stored in "filenameC".
Example: For data file "ROHBPT", the normalized coefficients

are stored in "ROHBPTC".

To predict new data using "CRA-PRED":

Step 1 Load "CRA-PRED"

Step 2 Run "CRA-PRED". Upon query: "Give filename of stored normalized category coefficients", input appropriate filename. Upon query "Give filename of data set to be predicted", input appropriate filename. Upon query "If entire data set is to be predicted, select 1, otherwise 0", input either 1 or 0.

a) If "1" selected, each datum is predicted.

b) If "0" selected, samples desired are selected one-by-one. Upon query: "Give the number of the sample desired", input number of 1st desired sample. Upon repeat of query, number of 2nd desired sample is input, and so forth. When all desired samples are selected, input 0.

Step 3 Results are printed out.

Variables in the CRA package of programs:

In "CRA-DATA"

N	Number of samples
NS	Number of properties
NAM2$	Name of problem
ITM$(10)	Name of properties
K(10)	Number of categories in each property
XN(100)	Researcher identification number for samples
NAM$(100)	Names of samples

XY(100)	Original independent-variable data vector
X(100,15)	Matrix of categorized properties
KC	Total number of categorized properties

In "CRA-MAIN"

Entire-set calculation

K(10)	Number of categories in each property
ITM$(10)	Name of properties
Y(100)	Original data vector (Y)
X(100,15)	Categorized properties (X)
Z(100,15)	Categorized properties after Hiyashi deletions (X*)
B(15)	Category regression coefficients (A*)
BH(15)	Normalized category coefficients (A)
YY(100)	Estimated values (Y_m)
E (100)	Residuals between original and estimated values (E)
XMNA(100)	Mean of actual values (for N-1 calculation also)
SDA(100)	Standard deviation of actual values
XMNE(100)	Mean of estimated values
SDE(100)	Standard deviation of estimated value
R3(100)	Multiple correlation coefficients
SDR(100)	Standard deviation of residuals
FE(100)	Adjusted standard deviation of residuals

N-1 calculations

XY0(100)	Original data vector
YY0(100)	Predicted values
EE0(100)	Residuals between original and predicted values
BH0(15,100)	Normalized category coefficients

RAN0(10,100)	Range of normalized category coefficents in each property
BBO(10,100)	Partial correlation coefficients

Miscellaneous variables

V(15,15)	Inverse matrix for intermediate calculations
BB(16,16)	Cofactor matrix for intermediate calculations
C(15)	For intermediate calculations
D(15)	For intermediate calculations
X1(15)	For intermediate calculation
XX(100,10)	Categorized properties in a property
RR(11,11)	Matrix of sample correlation coefficients
VV(11,11)	Inverse matrix of RR
E1(15)	Fraction of 1's in columns of categorized properties

In "CRA-PRED"

X(15)	Stored normalized category coefficients
XN	Number of samples (entire set)
NAM$(100)	Names of samples (entire set)
XY(100)	Independent-variable data vector (entire set)
XX(100,15)	Matrix of categorized properties (entire set)
XN1	Number of subset samples
NAM$(100)	Names of subset samples
XY1(100)	Subset independent-variable data vector
XX1(100,15)	Matrix of subset categorized properties
YY(100)	Predicted values
E(100)	Residuals between actual and predicted values

Listing of CRA program

```
10 REM CRAMAIN - Calculates CRA model for data set
20 LPRINT "Categorical Regression Analysis - CRAMAIN"
30 DEFINT I-N
40 DIM NAM$(100),XN(100),K(10),ITM$(10),Y(100),X(100,15),Z(100,15)
50 DIM B(15),BH(15),YY(100),XMNA(100),SDA(100),XMNE(100),SDE(100),E(100)
60 DIM XYO(100),YYO(100),BHO(15,100),EEO(100),SDR(100),RS(100),RRS(100)
70 DIM R3(100),FE(100),RANO(10,100),BBO(10,100)
80 DIM V(15,15),BB(15,15),X1(15),C(15),D(15)
90 DIM XX(100,10),XY(100),RR(11,11),VV(11,11),E1(15)
100 INPUT"Give filename of desired data set";QQ$
110 PRINT"Filename of problem: ";QQ$:PRINT
120 LPRINT:LPRINT "Filename of problem: ";QQ$;
130 LPRINT TAB(38) "Begin calculation at ";
140 LPRINT TIME$:NNO=0
150 INPUT "For N-1 method calculation, select 1, otherwise 0"; IN1
160 IF IN1=0 THEN 210
170 INPUT "To exclude some samples from calculation, select 1, otherwise 0";IHA
180 IF IHA=1 THEN 190 ELSE 210
190 INPUT"Give sample number to start the calculation"; NNO:NNNO=NNO
200 REM Read in datafile
210 OPEN "I",#1,QQ$:INPUT #1,N,NS,NAM2$
220       FOR I=1 TO NS:INPUT #1,ITM$(I):NEXT I
230       FOR I=1 TO NS:INPUT #1,K(I): NEXT I
240       KC=0:FOR I=1 TO NS:KC=KC+K(I):NEXT I
250       FOR I = 1 TO N:INPUT #1,XN(I),NAM$(I),XY(I)
260           FOR J=1 TO KC:INPUT #1,X(I,J)
270       NEXT J,I
280 CLOSE 1
290 NNN=N:IF NNO>0 THEN 480
300 REM Print out original datafile
310 LPRINT:LPRINT TAB(20) NAM2$
320 LPRINT:LPRINT "Samples ";N TAB(40) "Properties ";NS
330 LPRINT "Properties (categories)"
340 FOR I=1 TO NS :LPRINT TAB(8) I TAB(10) ITM$(I): NEXT I
350 LPRINT:LPRINT TAB(40) "Categorized properties":K(0)=0:NNT=40
360 FOR I= 1 TO NS
370     NNT=NNT+K(I-1)*3
380     LPRINT TAB(NNT)    I;
390 NEXT I
400 FOR I = 1 TO N
410     LPRINT:LPRINT I; TAB(5) XN(I);
420     LPRINT TAB(10) NAM$(I);TAB(28) USING "#####.####";XY(I);
430     FOR J=1 TO KC: LPRINT TAB(37+J*3) X(I,J);
440 NEXT J,I:LPRINT
450 REM Start calculation
460 IF IN1=0 THEN 540
470 IF NNO=0 THEN 540
480 XY1=XY(NNO)
```

```
490 FOR J =1 TO KC :X1(J)=X(NN0,J): NEXT J
500 FOR I = NN0 TO N:XY(I)=XY(I+1)
510     FOR J = 1 TO KC:X(I,J)=X(I+1,J)
520 NEXT J,I
530 NNN=N:N=N-1
540 MO=1:LL=1:NO=0:L=1:M=K(LL)
550 FOR I = 1 TO N
560     FOR J = L TO M: Z(I,J)=X(I,J+NO)
570 NEXT J,I
580 IF NS<=MO THEN 610
590 MO=MO+1:LL=LL+1:NO=NO+1
600 L=M+1:M=M+K(LL)-1:GOTO 550
610 MM=KC-NS+1
620 FOR I = 1 TO MM
630     FOR J = 1 TO MM :V(I,J)=0
640         FOR J1 = 1 TO N
650             V(I,J)=V(I,J)+Z(J1,I)*Z(J1,J)
660 NEXT J1,J,I
670 FOR I= 1 TO KC:E(I)=0
680     FOR J = 1 TO N:E(I)=E(I)+X(J,I): NEXT J,I
690 IF IN1=1 THEN 720 ELSE 700
700 FOR I=1 TO KC: PRINT TAB(37+I*3) E(I);
710 NEXT I
720 FOR I=1 TO KC:E(I)=E(I)/N:E1(I)=E(I): NEXT I
730 ES=0:YYY=0:YY2=0
740 FOR I = 1 TO N
750     Y(I)=XY(I):YYY=YYY+Y(I)
760     YY2=YY2+Y(I)*Y(I):ES=ES+Y(I)
770 NEXT I
780 ES=ES/N:YYY=YYY/N:YY2=YY2/N:XMNA(NN0)=YYY
790 ST=SQR(YY2-YYY*YYY):SDA(NN0)=ST
800 REM Transposed matrix of X * Matrix of X
810 GOSUB 3430
820 PRINT"Inverse matrix of Tr.X * X"
830 FOR I = 1 TO MM:PRINT
840     FOR J = 1 TO MM:PRINT  USING "######.###";V(I,J);
850 NEXT J,I:PRINT
860 REM Transpose of X * Y
870 FOR I = 1 TO MM
880     FOR J = 1 TO MM:VV(I,J)=V(I,J)
890 NEXT J,I
900 FOR I = 1 TO MM:YY(I)=0
910     FOR J = 1 TO N:YY(I)=YY(I)+Z(J,I)*Y(J)
920 NEXT J,I
930 PRINT"Tr.X * Y"
940 FOR I=1 TO MM:PRINT USING "######.###";YY(I):NEXT I
```

```
950 REM Category regression coefficients
960 FOR I = 1 TO MM
970     FOR J = 1 TO MM:V(I,J)=VV(I,J)
980 NEXT J,I
990 FOR I = 1 TO MM:B(I)=0
1000     FOR J = 1 TO MM : B(I)=B(I)+V(I,J)*YY(J)
1010 NEXT J,I
1020 FOR I=1 TO KC:BH(I)=0:NEXT I
1030 MO=1:LL=1:NO=0:L=1:M=K(LL)
1040 FOR I=L TO M:BH(I+NO)=B(I):NEXT I
1050 IF NS<= MO THEN 1070
1060 MO=MO+1:LL=LL+1:NO=NO+1:L=M+1:M=M+K(LL)-1:GOTO 1040
1070 PRINT:PRINT"Partial regression coefficeints"
1080 FOR I = 1 TO KC
1090         PRINT USING "#####.####";BH(I): NEXT I
1100 FOR I=1 TO KC:B(I)=BH(I):NEXT I
1110 REM Normalized category regression coefficients
1120 FOR I = 1 TO NS:NN=0
1130     FOR J = 1 TO I:NN=NN+K(J):NEXT J
1140     NP=NN-K(I)+1:SUM=0
1150     FOR JJ = NP TO NN:SUM=SUM+BH(JJ)*E1(JJ):NEXT JJ
1160     FOR JJ = NP TO NN:BH(JJ)=BH(JJ)-SUM:NEXT JJ
1170 NEXT I:K(0)=0
1180 PRINT:PRINT"Normalized Regression Coefficents"
1190 FOR I = 1 TO KC:PRINT USING "#####.####";BH(I):BH0(I,NN0)=BH(I):NEXT I
1200 REM Estimated values, mean, standard deviation, residuals
1210 KC=0:FOR I = 1 TO NS
1220         XMAX=0:XMIN=0:KCC=KC+1:KC=KC+K(I)
1230         FOR J = KCC TO KC
1240             IF XMAX<BH(J) THEN 1250 ELSE 1260
1250             XMAX=BH(J):GOTO 1280
1260             IF XMIN=>BH(J) THEN 1270 ELSE 1280
1270             XMIN=BH(J)
1280         NEXT J
1290         RAN=XMAX-XMIN:RAN0(I,NN0)=RAN
1300     NEXT I
1310 PRINT:PRINT"Tables of Residuals"
1320 FOR J = 1 TO N :YY(J)=0:E(J)=0
1330     FOR I = 1 TO KC : YY(J)=YY(J)+B(I)*X(J,I):NEXT I
1340     E(J)=Y(J)-YY(J)
1350     IF IN1=0 THEN 1360 ELSE 1370
1360     IF J=>NN0 THEN 1390
1370     PRINT J,
1380     GOTO 1400
1390     JJJ=J+1:PRINT JJJ,
1400     PRINT USING "######.###";Y(J);YY(J);E(J)
1410 NEXT J
1420 S1=0:S2=0:SS1=0
1430 FOR I = 1 TO N : S1=S1+YY(I) : SS1=SS1+YY(I)*YY(I) : S2=S2+E(I) : NEXT I
1440 S1=S1/N : S2=S2/N : XMNE(NN0)=S1
1450 SS1=SS1/N:SDE(NN0)=SQR(SS1-S1*S1)
```

```
1460 PRINT"Mean",
1470 PRINT USING "######.###";ES;S1;S2
1480 PRINT:PRINT"Mean of outside variable";YYY
1490 PRINT"Standard deviation       ";ST
1500 REM N-1 calculation:predicted values, mean, standard deviation
1510 IF IN1=0 THEN 1570 ELSE 1520
1520 IF NN0=0 THEN 1570
1530 YY1=0:FOR I = 1 TO KC:YY1=YY1+B(I)*X1(I):NEXT I
1540 E3=XY1-YY1:PRINT:PRINT"Prediction of sample #";NN0:PRINT NN0,
1550 PRINT USING "#####.####";XY1;YY1;E3
1560 XY0(NN0)=XY1:YY0(NN0)=YY1:EE0(NN0)=E3
1570 SUM1=0:FOR I = 1 TO N:SUM1=SUM1+YY(I)*Y(I):NEXT I
1580 SUM1=SUM1/N:SUM1=SUM1-S1*ES
1590 R=SUM1/(SDE(NN0)*ST):PE=ABS(1-R*R):PE=ST*SQR(PE)
1600 FPE=(N+MM)*PE*PE/(N-MM):R3(NN0)=R
1610 SDR(NN0)=PE:FE(NN0)=FPE
1620 FOR NM = 1 TO N
1630     FOR I = 1 TO NS:XX(NM,I)=0:NN=0
1640         FOR J = 1 TO I:NN=NN+K(J):NEXT J
1650         NP=NN-K(I)+1
1660         FOR JJ = NP TO NN
1670         XX(NM,I)=XX(NM,I)+B(JJ)*X(NM,JJ)
1680 NEXT JJ,I,NM
1690 FOR I = 1 TO NS:E(I)=0
1700     FOR J = 1 TO N:E(I)=E(I)+XX(J,I):NEXT J
1710     E(I)=E(I)/N
1720 NEXT I
1730 PRINT:PRINT"Tables of Criterion-Variables and Items"
1740 FOR I = 1 TO N
1750     IF IN1=0  THEN 1770
1760     IF I=>NN0 THEN 1790
1770     PRINT I,
1780     GOTO 1800
1790     III=I+1:PRINT III,
1800     PRINT USING "######.###";Y(I);
1810     FOR J = 1 TO NS
1820         PRINT USING "######.###";XX(I,J);
1830     NEXT J:PRINT
1840 NEXT I
1850 PRINT"MEAN", USING "######.###";ES;
1860 FOR I = 1 TO NS:PRINT USING "######.###";E(I);
1870 NEXT I
```

```
1880 REM Multiple correlation coefficients, partial correlation coefficients
1890 FOR I=1 TO N:XX(I,NS+1)=Y(I):NEXT I
1900 E(NS+1)=ES:N2=NS+1
1910 FOR I = 1 TO N2
1920     FOR J = 1 TO N2:S(I,J)=0
1930         FOR JJ = 1 TO N
1940             S(I,J)=S(I,J)+(XX(JJ,I)-E(I))*(XX(JJ,J)-E(J))
1950 NEXT JJ,J,I
1960 FOR I = 1 TO N2
1970     FOR J = 1 TO N2
1980         RR(I,J)=S(I,J)/SQR(S(I,I)*S(J,J))
1990 NEXT J,I
2000 PRINT:PRINT"Sample multiple correlation coefficient",R
2010 PRINT:PRINT"Sample correlation matrix"
2020 FOR I = 1 TO N2
2030     PRINT:IF I=N2 THEN 2060
2040     PRINT"I=";I,
2050     GOTO 2070
2060     PRINT"Y    ",
2070     FOR J = 1 TO N2
2080         PRINT USING "######.###";RR(I,J);
2090 NEXT J,I
2100 FOR I = 1 TO N2
2110     FOR J = 1 TO N2:V(I,J)=RR(I,J)
2120 NEXT J,I
2130 MM1=MM:MM=N2
2140 GOSUB 3430
2150 N2=MM:FOR I = 1 TO NS
2160           B(I)=-V(I,NS+1)/SQR(V(I,I)*V(NS+1,NS+1)):BB0(I,NN0)=B(I)
2170       NEXT I
2180 PRINT:PRINT"Inverse matrix of correlation matrix"
2190 FOR I = 1 TO N2:PRINT
2200     FOR J = 1 TO N2 :PRINT USING "#####.###";V(I,J);
2210 NEXT J,I
2220 PRINT:PRINT"Sample Partial Correlation Coefficients"
2230 FOR I=1 TO NS:PRINT I,B(I):NEXT I:MM=MM1
2240 IF IN1=0  THEN 2320
2250 IF NNN>NN0 THEN 2260 ELSE 2280
2260 NN0=NN0+1:GOTO 210
2270 REM Print out results
2280 IF NNN0=0 THEN N9=1 ELSE N9=NNN0
2290 LPRINT
2300 LPRINT TAB(15) "Normalized Category Coefficients - Using N-1 Method":K4=0
2310 GOTO 2340
2320 LPRINT
2330 LPRINT TAB(15) "Normalized Category Coefficients - Using Entire Set":K4=0
2340 LPRINT:LPRINT "Prop.(Category)";
2350 K4=K4+1:K9=0:JJ=0
2360 FOR I = 1 TO NS
2370     FOR J = 1 TO K(I):K9=K9+1
2380         K5=(K4-1)*6:IF K9<=K5 THEN 2420
2390         IF JJ>=6 THEN 2420
2400         LPRINT TAB(18+JJ*10) I;"(";J;")";
```

```
2410          JJ=JJ+1
2420 NEXT J,I
2430 IF NNN0>0 THEN 2550
2440 IF K4>1 THEN 2480
2450 LPRINT:OPEN "o",#1,QQ$+"C"
2460 KCC=KC+1:WRITE #1,KCC
2470 LPRINT:LPRINT "Entire set";
2480 JJ=0:K9=0
2490 FOR I = 1 TO KC:K9=K9+1:K5=(K4-1)*6
2500      IF K9<=K5 THEN 2540
2510      IF JJ>=6 THEN 2540
2520      LPRINT TAB(15+JJ*10) USING "#####.####";BH0(I,0);
2530      WRITE #1,BH0(1,0):JJ=JJ+1
2540 NEXT I
2550 IF IN1=0  THEN 2670
2560 IF K4>1 THEN 2580
2570 LPRINT:LPRINT"Sample removed":GOTO 2590:LPRINT
2580 LPRINT
2590 FOR I = N9 TO NNN:JJ=0:K9=0
2600      LPRINT TAB(5) I;
2610      FOR J = 1 TO KC:K9=K9+1:K5=(K4-1)*6
2620           IF K9<=K5 THEN 2660
2630           IF JJ>=6 THEN 2650
2640           LPRINT TAB(15+JJ*10) USING "#####.####";BH0(J,I);
2650           JJ=JJ+1
2660 NEXT J,I
2670 LPRINT:LPRINT:IF KC>K4*6 THEN 2350
2680 IF IN1=1 THEN 2710
2690 IF NNN0>0 THEN 2700
2700 IF IN1=0  THEN 2850
2710 LPRINT TAB(20) "Actual values" TAB(44) "Predicted values"
2720 LPRINT TAB(20) "Mean" TAB(30) "S.D." TAB(45) "Mean" TAB(56) "S.D."
2730 IF NNN0>0 THEN 2760
2740 LPRINT "Entire set";TAB(15) USING "#####.####";XMNA(0);SDA(0);
2750 LPRINT TAB(40) USING"#####.####";XMNE(0);SDE(0)
2760 LPRINT"Sample removed"
2770 FOR I = N9 TO NNN:LPRINT TAB(5) I;
2780           LPRINT TAB(15) USING "#####.####";XMNA(I);SDA(I);
2790           LPRINT TAB(40) USING "#####.####";XMNE(I);SDE(I)
2800 NEXT I
2810 LPRINT:LPRINT TAB(16) " Mult.Cor." TAB(28) "S.D.Res.";
2820 LPRINT TAB(50) "Deleted Datum"
2830 IF NNN0>0 THEN 2850
2840 LPRINT "Entire set"  TAB(15) USING "#####.####";R3(0) ;SDR(0)
2850 IF IN1=0 THEN 2920
2860 LPRINT"Sample removed" TAB(43) "Actual" TAB(52) "Predicted" TAB(64)"Error"
2870 FOR I = N9 TO NNN:LPRINT TAB(5) I;
2880           LPRINT TAB(15) USING "#####.####";R3(I);SDR(I);
2890           LPRINT TAB(40) USING "#####.####";XY0(I);YY0(I);EE0(I)
2900 NEXT I
```

```
2910 IF IN1=1 THEN 3130
2920 LPRINT TAB(31) "Actual";
2930 LPRINT TAB(39) "Estimated" TAB(51)"Residual"
2940 FOR J = 1 TO N: YY(J)=0: E(J)=0
2950     FOR I = 1 TO KC :YY(J) = YY(J)+BH(I)*X(J,I):NEXT I
2960     YY(J)=YY(J)+XMNE(0): E(J)=Y(J)-YY(J)
2970     LPRINT J; TAB(5) XN(J); TAB(10) NAM$(J);
2980     LPRINT TAB(28) USING "#####.####";Y(J),YY(J),E(J)
2990 NEXT J:LPRINT
3000 REM F TEST
3010 FOR I = 1 TO N : RS(I)=(YY(I)-XMNE(0))^2:RRS(I)=(Y(I)-XMNA(0))^2: NEXT I
3020 SS=0 : SSR=0
3030 FOR I = 1 TO N: SS =SS +RS(I):SSR=SSR+RRS(I):NEXT I:SSR=ABS(SS-SSR)
3040 N1=KC-NS+1:N2=N-N1-1
3050 SS=SS/N1:SSR=SSR/N2:FT=SS/SSR:FT(0)=FT
3060 PRINT:PRINT "F-TST"
3070 PRINT SSR,SS,FT
3080 LPRINT TAB(10) "Mean" TAB(28) USING "#####.####"; XMNA(0); XMNE(0);S2(0)
3090 LPRINT TAB(10) "S.D." ;TAB(28) USING "#####.####"; SDA(0);SDE(0);SDR(0)
3100 LPRINT TAB(10) "Multiple correlation coefficient";USING "#####.####";R3(0)
3110 LPRINT TAB(10) "F-test value" TAB(42) FT(0)
3120 WRITE #1,XMNE(0):CLOSE 1
3130 LPRINT:RI$="Range of normalized category coefficients"
3140 PC$="Partial correlation coefficients"
3150 FOR K = 1 TO 2:IF K=2 THEN RI$=PC$
3160                 IF K=2 THEN  3170 ELSE 3200
3170     FOR I9 = 1 TO NS
3180         FOR NN0 = NNN0 TO NNN : RAN0(I9,NN0)=BB0(I9,NN0)
3190     NEXT NN0,I9
3200     LPRINT TAB(9) RI$:J1=18
3210     FOR I = 1 TO NS:LPRINT TAB(J1) "Prop.";I;
3220         J1=J1+10
3230     NEXT I
3240     IF NNN0<>0 THEN 3300
3250     LPRINT:LPRINT "Entire set";
3260     J1=15:FOR I = 1 TO NS
3270               LPRINT TAB(J1) USING "#####.####";RAN0(I,0);
3280               J1=J1+10
3290     NEXT I
3300     IF IN1=0 THEN 3370
3310     LPRINT:LPRINT"Sample removed"
3320     FOR I = N9 TO NNN:LPRINT TAB(5) I;
3330         J1=15:FOR J = 1 TO NS
3340                   LPRINT TAB(J1) USING "#####.####";RAN0(J,I);
3350                   J1=J1+10
3360     NEXT J,I
3370 LPRINT:LPRINT: NEXT K
3380 GOTO 3400
```

```
3390 LPRINT:LPRINT"Determinant=0"
3400 LPRINT "End calculation at " ;
3410 LPRINT TIME$:LPRINT:LPRINT
3420 END
3430 REM Subroutine to take inverse of matrix
3440 T=0:FOR I=1 TO MM:T=T+V(I,I):V(I,MM+1)=V(I,I):NEXT I:D(1)=T
3450 FOR I = 1 TO MM
3460     FOR J = 1 TO MM:BB(J,I)=V(I,J):NEXT J
3470     BB(I,I)=V(I,I)-D(1)
3480 NEXT I
3490 FOR L = 2 TO MM:T=0
3500     FOR I = 1 TO MM:BB(I,MM+1)=0
3510         FOR J = 1 TO MM
3520             BB(I,MM+1)=BB(I,MM+1)+V(I,J)*BB(I,J)
3530             T=T+V(I,J)*BB(I,J)
3540     NEXT J,I
3550     AN=L:D(L)=T/AN:IF L>=MM THEN 3650
3560     FOR J = 1 TO MM
3570         FOR I = 1 TO MM:T=0
3580             FOR M = 1 TO MM:T=T+V(I,M)*BB(J,M):NEXT M
3590             IF I=J THEN 3610
3600             C(I)=T:GOTO 3620
3610             C(I)=BB(I,MM+1)-D(L)
3620         NEXT I
3630         FOR K = 1 TO MM : BB(J,K)=C(K)
3640 NEXT K,J,L
3650 T=(-1)^(MM-1)
3660 FOR I=1 TO MM:D(I)=T*D(I):FOR J=1 TO MM:BB(I,J)=T*BB(I,J):NEXT J,I
3670 T=-T:PRINT:PRINT"Determinant=";D(MM)
3680 IF D(MM)=0  THEN 3390
3690 FOR I=1 TO MM:FOR J=1 TO MM:V(I,J)=BB(I,J)/D(MM):NEXT J,I
3700 RETURN
```

Appendix B - Symbols in CRA and MRA-CRA Models

n	Number of substances
p	Number of properties
c	Number of categories
q	Number of MRA variables
x_{mij}	Categorized property
$[X]$	matrix of independent variables
$[X]'$	Transpose of matrix [X]
$([X]'[X])^{-1}$	Inverse matrix of [X]' [X]
$[X^*]$	Matrix of categorized properties after removing columns from [X]
y_m	Dependent-variable datum for m-th substance
Y	Vector of dependent-variable data
$\hat{y}_m$	Estimated value of dependent variable for m-th substance
$\hat{Y}$	Vector of estimated values of dependent variable
$\bar{Y}$	Mean value of the dependent-variable data
$\hat{y}_{km}$	Estimated value of dependent variable of m-th substance for k-th MRA variable
e_m	Residual
E	Vector of residuals

(continued)

Appendix B. (continued)

a_{ij}	Category coefficient
A	Vector of category coefficients
a^*_{ij}	Category coefficient by deletion method
A*	Vector of category coefficients by deletion method
a_{Nij}	Normalized category coefficient
A_N	Vector of normalized category coefficients
A_{comb}	Vector of normalized category coefficients for MRA-CRA combination model
f_{ij}	Fraction of 1's of the x_{mij} in j-th column
b	Intercept of MRA model
SE	Standard error
R	Multiple correlation coefficient
R_{pi}	Partial correlation coefficient of i-th property
F-value	Calculated value of F-test for regression
S_{12}	Variance of the model
S_{22}	Variance of the error
df_1	Degrees of freedom for the model
df_2	Degrees of freedom for the error

Appendix C - Documentation, List of Variables and Listing for Max-Min Program

Original version of the program in BASIC by D. Zhu. Modified by J. Wu and D. G. Howery (Chemistry Dept., Brooklyn College of C.U. N.Y, Brooklyn, New York.)

To write a data set onto disk:

Step 1 Load "MAX-MIN"

Step 2 Input data from line 1200 as follows:

1200 DATA Value of 1st factor pattern, . . . , value of n-th factor pattern for the first sample

1210 DATA repeat 1200 for each sample

Step 3 Save file as "filename" of choice.

To apply MAX-MIN to pattern vectors

Step 1 Load "filename"

Step 2 Run "filename". Upon queries "Number of samples" and "Number of pattern vectors", input the number of samples and the total number of pattern vectors.

i) Upon query "How many cluster centers are desired? To exit, input 0", input total number of cluster centers desired. Query is repeated until a zero is input.

ii) Upon query "How many clusters are desired?" input the specific number of clusters desired. Upon query "How many pattern vectors are selected for each cluster? To exit, input

0", input the number of selected vectors. Query is repeated until a zero is input.

Step 3 Results are printed out.

Variables in MAX-MIN program

XX(25,6)	Rotated vector-pattern matrix.
P(25,25,6)	Euclidian distance between samples in six-dimensional space.
B(25)	Vector of distances between samples in an individual factor space.
C(25)	Vector of squared distances between samples in an individual factor space.
D(25)	Vector of distances from the samples to a cluster center
F(25)	Vector of minimum distances to cluster centers for samples

Listing of Max-Min program

```
10 LPRINT "Maximum-minimum distance cluster analysis of pattern vectors"
20 LPRINT "  "
30 LPRINT "  "
40 PRINT "NUMBER OF SAMPLES"
50 INPUT N
60 PRINT "NUMBER OF PATTERN VECTORS"
70 INPUT M
80 LPRINT "     ";N;"samples &";M;"pattern vector data matrix"
90 LPRINT "  "
100 LPRINT " SP ";
110 DIM X(25,25),B(25),C(25),D(25),F(25),G(25),P(25,25,6),XX(25,6)
120 DEF FND(O)=INT(10000*O + .5)/10000
130      FOR I=1 TO M
140      LPRINT "     VECTOR";I;
150      NEXT I
160      LPRINT "  "
170      FOR J=1 TO N
180      LPRINT J;"     ";
```

```
190          FOR I=1 TO M
200          READ X(J,I)
210          LET XX(J,I)=FND(X(J,I))
220      LET D(K)=0
230          LPRINT XX(J,I);"       ";
240          NEXT I
250      LPRINT "  "
260      NEXT J
270 LPRINT "  "
280 LPRINT "  "
290 LPRINT "  "
300 LPRINT "  "
310 LPRINT "    Distance table for";N;"samples"
320 LPRINT "  "
330 LPRINT " SP ";"    SP";
340      FOR I=1 TO M
350      LPRINT " D (1-";I;")",
360      NEXT I
370 LPRINT "  "
380      FOR J=1 TO N-1
390          FOR K=J+1 TO N
400          LPRINT J;"    ";K;"    ";
410          LET A=0
420              FOR I=1 TO M
430              LET B(I)=X(J,I)-X(K,I)
440              LET C(I)=B(I)^2+A
450              LET A=C(I)
460              LET P(J,K,I)=SQR(C(I))
470              LET P(J,K,I)=FND(P(J,K,I))
480              LPRINT P(J,K,I),
490              LET P(K,J,I)=P(J,K,I)
500              LET P(J,J,I)=0
510              NEXT I
520          LPRINT "  "
530          NEXT K
540      NEXT J
550 LPRINT"  "
560 LPRINT"  "
570 PRINT "Determination of cluster center"
580 LPRINT "  "
590 PRINT "How many cluster centers are desired? 1-";N;"To exit,input 0."
600 INPUT Y
610 IF Y>N THEN 1740
620 LPRINT "  "
630     FOR Q=2 TO Y
640      LPRINT "Determination of cluster center ";Q
650      LPRINT"  "
660      GOSUB 800
670      NEXT Q
```

```
680 LPRINT "    Composition of clusters"
700 LPRINT "  "
710 PRINT "How many clusters are desired? 1-";Y;".To exit,input 0"
720 INPUT S
730 IF S=0 THEN 780
740 LPRINT "  "
750 LPRINT "For";S;"clusters"
760 GOSUB 1460
770 GOTO 710
780 GOTO 1740
790 STOP
800 PRINT "How many pattern vectors are selected for each cluster";Q;"? 1-";M;
    "To exit,input 0."
810 INPUT I
820 IF I=0 THEN 1190
830 IF I>M THEN 1740
840 LPRINT "For";I;" vectors"
850 LPRINT "  "
860 LPRINT "   SP ",
870 LET F(1)=1
880     FOR L=1 TO Q-1
890     LET J=F(L)
900     LPRINT "D to SP" J,
910     NEXT L
920 LPRINT "  Min."
930 LET E=0
940     FOR K=2 TO N
950 LET D(K)=P(1,K,I)
960     LPRINT"  ";K,
970         FOR L=1 TO Q-1
980         LET J=F(L)
990         LPRINT P(J,K,I);"  ",
1000         IF P(J,K,I)>=D(K) THEN 1020
1010         LET D(K)=P(J,K,I)
1020         NEXT L
1030     LPRINT D(K)
1040     IF D(K)<=E THEN 1060
1050     LET E=D(K)
1060     NEXT K
1065 LPRINT
1070 LPRINT "Max. dist. of min.";TAB(24) E
1080 PRINT "Max.of Min. ";E
1090     FOR K=2 TO N
1100     IF D(K)<E THEN 1120
1110     LET F(Q)=K
1120     NEXT K
```

```
1130 LPRINT "Center of Cluster";Q;TAB(24) " SP";F(Q)
1140 PRINT "Center of Cluster";Q;" SPL.";F(Q)
1150 LPRINT"  "
1160 LPRINT "  "
1170 LPRINT"  "
1180 GOTO 800
1190 RETURN
1200 DATA 0.6057,0.5846,0.5395
1210 DATA 0.5946,0.5834,0.5530
1220 DATA 0.7245,0.5522,0.4125
1230 DATA 0.7238,0.5556,0.4091
1240 DATA 0.7321,0.5451,0.4088
1250 DATA 0.5579,0.7272,0.3998
1260 DATA 0.5621,0.7244,0.3990
1460 PRINT "How many pattern vectors are selected for";S;" clusters? 1-";M;
     "To exit, input 0."
1470 INPUT I
1480 LPRINT"  "
1490 IF I=0 THEN 1730
1500 LPRINT "For";I;"vectors"
1510 LPRINT "  "
1520 LPRINT "  SP",
1530      FOR L=1 TO S
1540      LET J=F(L)
1550      LPRINT "D to SP" J,
1560      NEXT L
1570 LPRINT "To cluster"
1580      FOR K=1 TO N
1590      LPRINT "  ";K,
1600      LET G(K)=P(1,K,I)
1610      LET H=1
1620          FOR L=1 TO S
1630          LET J=F(L)
1640          LPRINT P(J,K,I);"  ",
1650          IF P(J,K,I)>=G(K) THEN 1680
1660          LET H=L
1670          LET G(K)=P(J,K,I)
1680          NEXT L
1690      LPRINT H
1700      NEXT K
1710 LPRINT"  "
1720 GOTO 1460
1730 RETURN
1740 END
```

Appendix D - Complete Data Matrix for Retention-Index Problems.

Compound	Designation	Stationary Phase [a]									
		APJ	APM	SE30	SE31	PH1	PH2	PP5	PP6	CW6K	CW20
Hexanol	AP1	841	841	890	862	995	984	1028	1033	1332	1331
2-hexanol	AS1	774	775	810	789	911	901	941	945	1195	1188
3-hexanol	AS2	777	776	805	786	902	893	933	938	1172	1168
2-me 1-pentanol	AP2	806	807	852	830	956	947	989	993	1285	1281
3-me 1-pentanol	AP3	817	817	864	837	971	962	1006	1011	1311	1309
4-me 1-pentanol	AP4	806	807	857	829	960	951	997	997	1294	1295
2-me 2-pentanol	AT1	714	714	739	722	834	828	865	870	1069	1066
3-me 2-pentanol	AS3	777	775	805	787	901	892	934	939	1169	1164
4-me 2-pentanol	AS4	736	736	771	750	867	859	893	896	1136	1133
2-me 3-pentanol	AS5	758	758	783	764	877	869	907	912	1130	1126
3-me 3-pentanol	AT2	743	744	763	747	854	847	890	894	1087	1085
2-et 1-butanol	AP5	818	818	860	835	963	953	990	996	1287	1283
2,2-me 1-butanol	AP6	774	775	812	790	914	905	935	939	1222	1216
2,3-me 2-butanol	AT3	721	720	741	725	835	826	863	868	1064	1059
3,3-me 2-butanol	AS6	723	721	752	734	845	837	867	871	1098	1094
Butanol	ABU	629	629	676	651	782	772	816	821	1117	1111
Pentanol	APE	736	738	784	759	890	878	925	928	1227	1223
Heptanol	AHP	942	941	991	964	1099	1087	1133	1137	1436	1435
Octanol	AOC	1043	1040	1093	1065	1202	1189	1237	1241	1541	1545

Appendix D. (continued)

Compound	Designation	Stationary Phase [a]									
		APJ	APM	SE30	SE31	PH1	PH2	PP5	PP6	CW6K	CW20
Dipropyl ether	TP2	656	660	687	677	707	702	738	740	775	773
Dibutyl ether	TB2	862	864	881	876	902	898	936	930	966	966
Dipentyl ether	TPE2	1056	1057	1076	1071	1108	1102	1136	1137	1174	1165
Butyl ethyl ether	TBEt	665	668	691	682	718	714	755	757	792	794
Isobutyl Et ether	TIBEt	583	589	618	606	642	632	662	668	700	699
Isopr pr ether	TIPP	607	613	639	931	665	660	682	687	718	712
Isopropyl ether	TIP2	561	566	593	586	617	609	621	626	663	652
Ethyl propionate	SEtP	655	654	700	688	756	749	837	840	951	950
Propyl propionate	SPP	755	755	800	790	854	846	931	935	1040	1038
Butyl propionate	SBP	859	859	900	890	960	951	1037	1038	1150	1144
Pentyl propionate	SPEP	957	954	998	988	1061	1054	1141	1140	1250	1248
Ethyl butyrate	SEtB	750	751	793	783	848	841	923	926	1034	1031
Propyl butyrate	SPB	848	851	892	882	949	940	1022	1022	1131	1129
Butyl butyrate	SBB	947	948	990	980	1050	1043	1125	1125	1230	1228
Pentyl butyrate	SPEB	1046	1045	1089	1079	1150	1141	1225	1226	1327	1325

a) For designations, see p. 92.

b) For designations, see Table 27.

Appendix E - Complete Data Matrix for R_f (x 100) Problems.

20 x 13 - P1 problem

Sugar	Designation	Solvent *												
		A	B	C	D	E	F	G	H	I	J	K	L	M
Ribose	M1	29	55	21	32	47	30	55	42	20	42	44	38	38
Arabinose	M2	25	60	20	32	38	27	51	37	13	41	43	39	35
Xylose	M3	36	58	29	43	52	39	63	45	21	52	52	50	46
Lyxose	M4	34	66	28	42	42	37	63	44	21	52	53	49	44
Fucose	M5	32	56	25	39	48	35	58	46	18	48	50	43	42
Glucose	M6	22	66	21	35	36	30	58	37	10	44	43	43	37
Galactose	M7	19	55	16	29	29	24	55	35	9	37	40	32	34
Levulose	M8	23	55	20	31	42	28	52	37	13	39	42	37	37
Sorbose	M9	26	58	23	36	43	31	55	38	12	46	43	43	41
Mannoheptulose	M10	20	60	21	33	39	29	58	36	10	42	41	41	36
Sedoheptulose	M11	21	60	18	28	35	25	56	34	10	37	48	34	31
Sucrose	O1	17	55	16	27	38	24	51	33	8	35	35	32	36
Turanose	O2	13	60	15	25	31	22	53	29	6	34	30	35	30
Maltose	O3	13	55	13	23	35	20	46	30	7	25	29	26	29
Lactose	O4	9	55	8	16	27	14	36	23	3	18	25	16	22
Trehalose	O5	12	69	15	24	25	20	48	26	17	34	26	33	30
Melibiose	O6	8	58	10	16	23	14	42	22	2	6	18	15	21
Raffinose	O7	4	55	3	10	21	9	33	17	2	6	18	15	21
Me glucoside	D1	37	55	31	45	51	41	63	44	19	51	50	52	51
Me mannoside	D2	26	55	24	34	42	34	58	37	13	47	42	43	43

20 x 13 - P2 problem

Sugar	Designation	Solvent*												
		A	B	C	D	E	F	G	H	I	J	K	L	M
Ribose	M1	45	27	42	49	34	36	69	35	26	47	36	54	49
Arabinose	M2	31	27	33	38	23	26	51	22	12	42	24	43	41
Xylose	M3	45	27	42	53	35	36	67	35	24	52	37	58	51
Lyxose	M4	47	27	42	52	34	36	64	35	23	49	36	59	47
Fucose	M5	48	27	42	53	33	37	69	31	20	51	35	58	52
Glucose	M6	24	26	27	38	19	22	50	18	10	40	20	41	35
Galactose	M7	16	27	24	30	14	19	47	13	8	31	16	36	30
Levulose	M8	29	27	32	41	23	27	58	20	13	40	24	41	41
Sorbose	M9	33	30	25	39	27	31	51	23	13	42	26	45	44
Mannoheptulose	M10	26	25	30	41	21	25	53	18	10	42	21	45	37
Sedoheptulose	M11	30	27	29	33	22	26	51	19	10	42	21	41	40
Sucrose	O1	19	27	32	34	13	22	59	10	5	35	13	45	42
Turanose	O2	19	24	24	35	11	21	55	9	5	36	10	41	36
Maltose	O3	12	27	22	29	9	17	48	6	4	30	9	35	29
Lactose	O4	6	26	14	18	5	12	33	4	3	20	9	24	16
trehalose	O5	10	25	18	25	8	15	38	6	3	38	8	31	30
Melibiose	O6	4	23	9	14	4	8	24	3	2	15	8	15	11
Raffinose	O7	2	25	8	8	2	7	22	1	2	13	8	19	9
Me glucoside	D1	54	30	51	49	44	45	74	37	22	46	36	61	60
Me mannoside	D2	30	30	36	38	25	30	70	20	14	41	20	45	46

*) For designations, see Table 40.

References

1. D. L. Massart, B. G. M. Vandeginste, S. N. Deming, Y. Michotte and L. Kaufman, "Chemometrics: a textbook", Elsevier, New York, 1988.

2. D. G. Howery and R. F. Hirsch, J. Chem. Educ., **60**, 656 (1983).

3. M. A. Sharaf, D. L. Illman and B. R. Kowalski, "Chemometrics", Wiley, New York, 1986.

4. "Practical Guide to Chemometrics", S. J. Haswell (Ed.), Marcel Dekker, New York, 1992.

5. N. R. Draper and H. Smith, "Applied Regression Analysis", 2nd ed., Wiley, New York, 1981.

6. E. R. Malinowski and D. G. Howery, "Factor Analysis in Chemistry", Wiley, New York, 1980.

7. D. L. Massart and L. Kaufman, "The Interpretation of Analytical Chemical Data by the Use of Cluster Analysis", Wiley, New York, 1983.

8. K. Varmuza "Pattern Recognition in Chemistry", Springer, New York, 1980.

9. S. D. Brown, R. S. Bear and T. B. Blank, Anal. Chem., **64**, 22R (1992).

10. "SAS User's Guide", SAS Institute, Cary, NC, 1988.

11. "SAS Introductory Guide", SAS Institute, Cary, NC, 1988.

12. "SAS Language Guide", SAS Institute, Cary, NC, 1988.

13. "SAS Procedures Guide", SAS Institute, Cary, NC, 1988.

14. "SAS/STAT User's Guide", SAS Institute, Cary, NC, 1988.

15. P. M. Coenegracht and J. Smilde, J. Chromatogr., **550**, 397 (1991).

16. S. J. Schmith and H. Zwanziger, J. Chromatogr., **544**, 381 (1991).

17. F. Walters, Anal. Lett., **22**, 635 (1989).

18. J. K. Hardy and P. L. Rinaldi, J. Magn. Reson., **88**, 320 (1990).

19. M. A. Vaughan and D. M. Templeton, Appl. Spectrosc., **44**, 1685 (1990).

20. Z. Pan, S. Siqing and M. Zhang, J. Chemom., **4**, 323 (1990).

21 R. J. Pell, J. B. Callis and B. R. Kowalski, Appl. Spectrosc., **45**, 801 (1991).

22. Z. H. Gallegos and A. Pedro, J. Med. Chem., **33**, 2813 (1990).

23. M. J. Cardone, S. A. Willaize and M. E. Lacy, Pharm. Res., **7**, 154 (1990).

24. L. Ekiert and J. Bojarski, J. Planar Chromatogr. - Med. TLC, **2**, 447 (1989).

25. P. Lu, H. Zou and Y. Zhang, J. Chromatogr., **509**, 70 (1990).

26. A. Voelkel, J. Chromatogr., **547**, 247 (1991).

27. J. Einax and K. Danzer, J. Environ. Anal. Chem., **44**, 185 (1991).

28. S. A. Stout, J. Anal. Appl. Pyrolysis, **18**, 277 (1991).

29. U. Welmar and S. Vaihinger, Chem. Sens. Technol., **3**, 51 (1991).

30. A. G. Wright, A. F. Fell and J. C. Berridge, J. Chromatogr., **458**, 335 (1988).

31. J. C. Berridge, J. Chromatogr., **485**, 27 (1989).

32. R. F. Gunst and R. L. Mason, "Regression Analysis and its Application", Marcel Dekker, New York, 1980.

33. P. G. Seybold, M. May and U. A. Bagal, J. Chem. Educ., **64**, 575 (1987).

34. H. Yamazaki and D. G. Howery, J. Chemom., submitted for publication.

35. P. E. Green, "Mathematical Tools for Applied Multivariate Analysis", Academic Press, New York, 1976.

36. C. Hayashi, Annals Inst. Math. (Tokyo), **3**, 69 (1952).

37. R. L. Grob, "Modern Practice of Gas Chromatography", Wiley, New York, 1985.

38. E. Kovats and A. I. A. Keulemans, Anal. Chem., **36**, 31A (1964).

39. G. D. Mitra and N. C. Saha, J. Chromatogr. Sci., **8**, 95 (1970).

40. D. A. Tourres, J. Gas Chromatogr., **5**, 35 (1967).

41. "Handbook of Organic Structural Analysis", Y. Yukawa (Ed.), Benjamin, New york, 1965.

42. R. Lide, "Handbook of Chemistry and Physics", 71st ed., Nat. Bureau Standards, Washington, 1990.

43. D. L. Duewer and H. Freiser, Anal. Chem., **49**, 13 (1977).

44. E. U. Condon, "Table of Chemical Kinetics, Homogeneous Reactions", Nat. Bureau Standards, Washington, 1951.

45. J. T. Tou and R. C. Gonzalez, "Pattern Recognition Principles", Addison-Wesley, Reading, MA, 1974.

46. W. O. McReynolds, "Gas Chromatographic Retention Data", Preston Technical Abstract Co., Niles, IL, 1966.

47. H. F. Kaiser, "Psychometrika", **23**, 187 (1958).

48. D. G. Howery, in "Statistics", R. F. Hirsch (Ed.), Franklin Institute Press, Philadelphia, 1978.

49. P. H. Weiner and D. G. Howery, Anal. Chem., **44**, 1189 (1972).

50. D. G. Howery, E. R. Malinowski, P. H. Weiner, J. M. Soroka, R. T. Funke, R. B. Selzer and A. Levinstone, "FACTANAL", Program 320, Quant. Chem. Program Exchange, Indiana Univ., Bloomington, IN, 1976.

51. E. R. Malinowski, "TARGET", Stevens Inst. Tech., Hoboken, NJ, 1990.

52. L. Rohrschneider, J. Chromatogr., **22**, 6 (1966).

53. W. O. McReynolds, J. Chromatogr. Sci., **8**, 685 (1970).

54. C. F. Poole and S. A. Schuette, "Contemporary Practice of Chromatography", Elseviev, New York, 1984.

55. R. Rummel, "Applied Factor Analysis", Northwestern University Press, Evanston, IL, 1970, p. 279.

56. E. R. Malinowski, Anal. Chim. Acta, **103**, 339 (1978).

57. R. Hamilton and S. Hamilton, "Thin Layer Chromatography", Wiley, New York, 1987.

58. M. Lato, B. Brunelli and G. Ciuffini, J. Chromatogr., **39**, 407 (1969).